Université du Québec
INRS-ETE

Régressions probabilistes multi-sites multivariées pour la réduction d'échelle (downscaling) des variables climatiques

Par
Mohamed Ali Ben Alaya

Thèse présentée pour l'obtention du grade de Philosophiæ Doctor (Ph.D.) en sciences de l'eau

Jury d'évaluation :

Examinateur externe	François Brissette École de Technologie Supérieure
Examinateur externe	Xuebin Zhang Environnement Canada
Examinatrice interne	André St-Hilaire INRS-ETE
Examinatrice interne	Philippe Gachon Université de Québec à Montréal
Co-directeur de recherche	Taha B.M.J. Ouarda INRS-ETE
Directeur de recherche	Fateh Chebana INRS-ETE

Février 2016

Je dédie cette thèse à ma mère Khadija, mon père Abdeljelil, mes frères Haithem et Jabeur et ma sœur Tata, pour leur soutien malgré la distance;

à ma femme Dhouha et ma fille Yasmine, pour leur patience et leur encouragement dont elles ont fait preuve pendant toute la durée de cette thèse.

À l'âme de Sinoura

REMERCIEMENTS

Ce document, tout en couronnant mes efforts, sanctionne la fin de mon cycle universitaire. Il est le résultat d'un travail de longue haleine. Cette thèse n'aurait certainement pas pu être menée à son terme si je n'avais pas bénéficié de la disponibilité, de la compréhension et de la collaboration de certaines personnes. Dans ce cadre, je tenais à remercier :

Professeur Fateh Chebana, mon directeur de thèse, pour avoir encadré ce travail. Merci d'avoir cru en moi et de m'avoir fait confiance depuis le début, et ce, jusqu'à la fin de ce doctorat. Merci aussi pour tout ce que tu as apporté à ces travaux, pour ton inestimable aide et tes conseils précieux.

Professeur Taha B.M.J. Ouarda, mon codirecteur de thèse, pour son infaillible et stimulant encadrement et pour m'avoir appris la rigueur de la progression et de l'analyse scientifique. Je ne peux ignorer ses qualités humaines et ne pas être reconnaissant au savoir qu'il m'a prodigué, la confiance qu'il m'a accordée et le grand soutien moral apporté de sa part, grâce à quoi j'ai pu accomplir mon travail.

Les membres du groupe de recherche en hydro-climatologie statistique, professeurs et étudiants, pour leur coopération et leur aide.

J'aimerais également remercier les membres de mon comité, André St-Hilaire, Philippe Gachon, François Brissette et Xuebin Zhang, je suis particulièrement reconnaissant et honoré de l'intérêt qu'ils ont porté à cette thèse en acceptant d'en être les rapporteurs.

PRÉFACE

Cette thèse présente les travaux de recherche menés au cours de mes études doctorales. La structure de la présente thèse suit la structure standard des thèses par articles de l'INRS-ETE. La première partie de la thèse comporte une synthèse générale des travaux effectués.

Cette synthèse a pour objectif de survoler la méthodologie adoptée et les principaux résultats obtenus au cours de la thèse. La deuxième partie contient six articles comme des chapitres, trois publiés, deux soumis et un finalisé et à soumettre pour publication dans une revue internationale.

ARTICLES ET CONTRIBUTION DES AUTEURS

[1] Ben Alaya, M. A., F. Chebana et T. Ouarda (2014). "Probabilistic Gaussian Copula Regression Model for Multisite and Multivariable Downscaling." Journal of Climate 27(9).

[2] Ben Alaya, M. A., F. Chebana et T. B. Ouarda (2015a). "Probabilistic Multisite Statistical Downscaling for Daily Precipitation Using a Bernoulli–Generalized Pareto Multivariate Autoregressive Model." Journal of climate 28(6): 2349-2364.

[3] Ben Alaya, M. A., F. Chebana et T. B. M. J. Ouarda (2015b). "Multisite and multivariable statistical downscaling using a Gaussian copula quantile regression model." Climate Dynamics: 1-15.

[4] Ben Alaya, M. A., T. Ouarda et F. Chebana (2016c). "Non-Gaussian spatiotemporal simulation of multisite daily precipitations: a downscaling framework." Submitted.

[5] Ben Alaya, M. A., F. Chebana et T. Ouarda (2016a). "Quantile regression multivariate autoregressive model for downscaling multisite daily precipitations." To be submitted.

[6] Ben Alaya, M. A., D. Fasbender, T. Ouarda et F. Chebana (2016b). "Application of spatial Bayesian model for downscaling daily temperatures and comparison with two probabilistic regression approaches." Submitted.

Dans le premier article, M. A. Ben Alaya a présenté un nouveau modèle de réduction d'échelle en combinant un modèle de régression probabiliste avec une copule Gaussienne. F. Chebana et T. B. M. J. Ouarda ont commenté et révisé la version finale du manuscrit.

Dans le deuxième article, M. A. Ben Alaya a proposé une nouvelle approche pour la réduction d'échelle des précipitations sur plusieurs stations. Cette approche combine un modèle de régression probabiliste en utilisant la distribution Bernoulli-Pareto Généralisée avec un champ Gaussien multivarié autorégressif. Tout au long de ce travail, F. Chebana et T. B. M. J. Ouarda ont donné de précieux conseils et suggestions et ont révisé la version finale du manuscrit.

Dans le troisième article, M. A. Ben Alaya a présenté un nouveau modèle de réduction d'échelle multisite et multivarié en combinant la régression des quantiles avec une copule Gaussienne. F. Chebana et T. B. M. J. Ouarda ont discuté l'aspect méthodologique et les résultats obtenus et ils ont révisé la version finale du manuscrit.

Dans le quatrième article, M. A. Ben Alaya a présenté un modèle de réduction d'échelle des précipitations en introduisant des structures de dépendance non Gaussiennes. F. Chebana et T. B. M. J. Ouarda ont commenté et révisé la version finale du manuscrit.

Dans le cinquième article, M. A. Ben Alaya a présenté un modèle de réduction d'échelle en combinant la régression des quantiles avec un champ Gaussien multivarié autorégressif. F. Chebana et T. B. M. J. Ouarda ont discuté la méthodologie du travail.

Dans le sixième article, M. A. Ben Alaya a présenté une comparaison de trois approches de régression probabiliste pour la réduction d'échelle des températures. D. Fasbender, F. Chebana et T. B. M. J. Ouarda ont commenté et révisé la version finale du manuscrit.

RÉSUMÉ

Les outils statistiques trouvent une large application dans la recherche climatologique, allant des méthodes simples pour déterminer l'incertitude d'une moyenne climatologique à des techniques sophistiquées qui révèlent la dynamique du système climatique. Dans le cas de réduction d'échelle (ou mise à l'échelle) climatique, le but est de prévoir les valeurs des variables météorologiques observées au niveau des stations ou à des échelles régionales à partir de la circulation atmosphérique à l'échelle synoptique, généralement pour générer des scénarios climatiques à partir de modèles climatiques globaux. Dans ce contexte, les modèles doivent non seulement être précis en termes des critères d'évaluation de performance, mais ils doivent également être en mesure de reproduire les propriétés statistiques des observations historiques selon le besoin et les exigences du domaine d'application.

L'objectif de cette étude consiste à concevoir, tester et améliorer une nouvelle structure de modélisation hybride probabiliste pour la réduction d'échelle des variables climatiques. Particulièrement, le but sera de développer de nouveaux modèles statistiques de réduction d'échelle des précipitations et des températures qui découlent de cette nouvelle structure de modélisation. Ces modèles visent à contourner les inconvénients majeurs des méthodes de régression classiques afin de fournir des informations météorologiques fiables et précises pour les applications, notamment hydrologiques. L'accent sera mis sur la reproduction de la variabilité temporelle, les extrêmes des températures et des précipitations, le caractère discret-continu des précipitations, l'intermittence spatiotemporelle multi-site et/ou multivariée et les structures de dépendance complexes. Ces nouveaux modèles sont basés sur des outils statistiques en plein essor dans la littérature hydrométéorologique au cours des dernières années, y compris les outils

multivariés tels que les copules et les approches de régression probabilistes (e.g. régression des quantiles et la forme vectorielle des modèles linéaires généralisés).

La structure de modélisation hybride proposée combine deux composantes, une composante de régression probabiliste avec une composante aléatoire. La composante de régression probabiliste permet de fournir à chaque étape de prévision toute la distribution conditionnelle univariée, tandis que la composante aléatoire permet de préserver les structures de dépendance multi-site et/ou multivariée. Pour la composante de régression probabiliste, deux outils ont été considérés, à savoir la régression des quantiles (QR) et la forme vectorielle des modèles linéaires généralisés (VGLM). Concernant la composante aléatoire, trois outils ont été considérés à savoir la copule Gaussienne, le champ Gaussien multivarié autorégressif (MAR), et l'échantillonnage non paramétrique (NB pour non-parametric bootstrapping). Dans cette thèse, cinq modèles de réduction d'échelle ont été développés en se basant sur la structure de modélisation hybride probabiliste proposée: chaque modèle correspond à un article qui sera inclus dans cette thèse sous forme d'un chapitre.

Le premier modèle est le modèle de régression probabiliste avec copule Gaussienne (PGCR). Ce modèle, présenté en détail dans le chapitre 2, utilise le VGLM comme régression probabiliste et une copule Gaussienne comme composante aléatoire. Il a été proposé comme une première évaluation de la structure hybride probabiliste. Le modèle PGCR a été appliqué pour réduire l'échelle de la température et des précipitations.

Le deuxième modèle développé est le modèle de Bernoulli-Pareto Généralisé multivarié autorégressif (BMAR). Ce modèle est présenté dans le chapitre 3. Par rapport au modèle PGCR, le modèle BMAR intègre une distribution plus appropriée afin de mieux reproduire les

précipitations extrêmes en utilisant une distribution mixte Bernoulli-Pareto Généralisée. Concernant la composante aléatoire, le modèle BMAR intègre un champ Gaussien multivarié autorégressif dans le but de préserver la corrélation spatiale et l'autocorrélation à court terme.

Le troisième modèle qui a été élaboré dans de cette thèse utilise comme composante de régression probabiliste la régression des quantiles, et comme composante aléatoire la copule Gaussienne. Ce modèle, nommé GCQR (pour Gaussian Copula Quantile Regression), est décrit dans le chapitre 4. L'avantage premier de ce modèle étant d'améliorer la composante de régression probabiliste par rapport aux modèles des chapitres précédents en relaxant l'hypothèse qui impose des distributions paramétriques. Ce modèle a été appliqué pour réduire l'échelle de la température et des précipitations.

Le quatrième modèle proposé combine le modèle VGLM en utilisant la distribution mixte Bernoulli-Pareto généralisée avec une procédure d'échantillonnage non paramétrique (NB). Le modèle est nommé VGLM-NB a été appliqué pour la réduction d'échelle des précipitations sur plusieurs stations. Le principal avantage de l'élaboration de ce modèle est de tester la structure hybride probabiliste en intégrant une composante aléatoire avec des structures de dépendance non Gaussiennes. Ce modèle est décrit dans le Chapitre 5. Ainsi ce chapitre inclut une comparaison entre les deux composantes aléatoires NB et champ Gaussien MAR pour reproduire les structures de dépendance des précipitations.

Le dernier modèle qui a été élaboré, présenté dans le chapitre 6, est un modèle intégrant la régression des quantiles avec le champ Gaussien MAR. Bien que ce modèle n'utilise pas un nouvel outil par rapport aux outils précédents, il présente une nouvelle optique pour la modélisation de la précipitation lorsqu'on veut intégrer des covariables dans le modèle. En

combinant la régression quantile avec le champ Gaussien MAR, ce modèle présente donc une extension du modèle GCQR pour intégrer l'autocorrélation des précipitations à court terme.

Dans le 7^{e} et dernier chapitre, les composantes de régression probabilistes ont été appliquées et comparées pour la réduction d'échelle de la température sur une grande région. En outre, pour des fins de comparaison, une autre formulation de la régression probabiliste a été considérée, à savoir la formulation bayésienne. Dans une première étape, un modèle spatial Bayésien (SBM) a été adapté et appliqué pour la réduction d'échelle des températures maximales et minimales ensuite ce modèle a été comparé avec la régression des quantiles et le modèle VGLM.

La structure de modélisation hybride probabiliste a été appliquée pour la réduction d'échelle des précipitations et des températures dans la partie sud de la province de Québec, Canada. Les données de réanalyse NCEP-NCAR ont été utilisées dans cette étude afin d'évaluer le potentiel de la structure de modélisation proposée. Les résultats des comparaisons avec des approches multi-sites et/ou multivariées traditionnelles montrent que la structure de modélisation proposée est plus en mesure de reproduire les caractéristiques marginales de la température et des précipitations journalières. En outre, l'approche montre une meilleure préservation des propriétés spatio-temporelle des précipitations et des températures entre les plusieurs stations. Les évaluations des modèles développés suggèrent que la structure hybride probabiliste constitue une conception de modélisation très utile pour générer l'évolution spatio-temporelle de précipitations et des températures. Cette structure de modélisation est très générale et peut être adaptée à des variables météorologiques non normalement distribuées telle que, la vitesse du vent, la couverture nuageuse et l'humidité. En plus, la facilité d'introduire des variables explicatives rend cette conception de modélisation un outil mathématique précieux non seulement en réduction d'échelle mais aussi en analyse climatique en générale.

TABLES DES MATIÈRES

LISTE DES TABLEAUX

LISTE DES FIGURES

CHAPITRE 1 : SYNTHÈSE

1. Introduction

1.1. Réduction d'échelle climatique

Les modèles climatiques globaux couplés atmosphère océan (MCGAO) sont largement utilisés pour faire des projections du climat futur. Les sorties de ces modèles sont fréquemment utilisées pour effectuer des études d'impact de changement climatiques ou des études d'adaptation dans plusieurs domaines d'application, tels que l'hydrologie, la santé et l'agriculture. Ces domaines d'application exigent des données climatiques définies à des échelles spatiales très fines, souvent de l'ordre de kilomètres ou à des sites spécifiques (Wilby et Wigley 1997). Comme les modèles climatiques globaux possèdent des résolutions horizontales de grille de l'ordre de plusieurs centaines de kilomètres, les prévisions climatiques résultantes ne peuvent pas être utilisées directement dans ces applications comme facteurs prédictifs dans les modèles définis à des échelles locales. Par ailleurs, plusieurs techniques de réduction d'échelle climatique ont été développées en vue d'affiner les sorties des modèles climatiques globaux et de fournir les informations climatiques nécessaires à des échelles plus fines. L'application la plus courante de mise à l'échelle consiste à générer des scénarios climatiques à haute résolution temporelle et spatiale, basés sur les résultats des modèles climatiques globaux.

Les techniques de réduction d'échelle peuvent être fondées physiquement, par exemple en utilisant un modèle climatique régional couvrant une aire limitée et pilotée aux frontières par les MCGAO, qui fournit des champs dynamiquement cohérents. Par contre, ces modèles dynamiques nécessitent des coûts informatiques et des ressources humaines élevés. D'autre part, les techniques de réduction d'échelles peuvent être basées sur la statistique, s'appuyant plutôt sur des données historiques observées pour établir des relations empiriques entre les prédicteurs, qui sont

des variables climatiques à grande échelle, et les prédictants qui sont des variables climatiques à petite échelle (Benestad et al. 2008).

1.2. Méthodes de réduction d'échelle statistiques

1.2.1. Théorie générale

La réduction d'échelle statistique implique l'établissement de relations empiriques entre les caractéristiques historiques de la circulation atmosphérique à grande échelle d'une part et les variables météorologiques locales d'autre part. Une fois cette relation empirique a été déterminée et validée, les conditions atmosphériques à grande échelle projetées dans le futur par les modèles MCGAO sont utilisées pour prédire les caractéristiques climatiques locales. En d'autres termes, les résultats des MCGAO à grande échelle sont utilisés comme des prédicteurs pour obtenir des variables locales ou prédictants.

Les méthodes de réduction d'échelle statistiques sont peu coûteuses en temps de calcul comparées aux méthodes dynamiques qui nécessitent une modélisation complexe des processus physiques. Ainsi, elles représentent une alternative fiable et parfois avantageuse pour les établissements qui ne disposent pas de la capacité de calcul et de l'expertise technique nécessaire à la réduction d'échelle dynamique. Contrairement aux modèles dynamiques régionaux, qui produisent des projections à échelle réduite de résolution spatiale de 20 à 50 kilomètres, les méthodes statistiques peuvent fournir des informations sur le climat à l'échelle locale des stations météorologiques (Wilby et Wigley 1997).

1.2.2. Hypothèses et mise en garde

Bien que la réduction d'échelle statistique soit efficace, peu coûteuse en temps de calcul et se compose d'un groupe varié de méthodes, elle requière les hypothèses inhérentes suivantes (Wilby et al. 2004; Benestad et al. 2008):

1. La relation statistique entre les prédicteurs et les prédictants ne change pas au cours du temps. Elle est connue comme l'hypothèse de stationnarité et postule que la relation statistique entre les prédicteurs et les prédictants demeure stable dans le futur.

2. Les prédicteurs contiennent le signal du changement climatique. Cette hypothèse suppose que la variable à grande échelle représente le système climatique et capte tout changement qui pourrait survenir dans le futur.

3. Il existe une relation forte entre les prédicteurs et les prédictants. Elle implique que l'importance de la relation doit être évaluée initialement pour déterminer sa validité.

4. Les MCGAO simulent avec précision les prédicteurs. Cette hypothèse se rapporte à la capacité d'un MCGAO pour simuler des variables climatiques observées dans le passé ainsi que leurs évolutions futures.

1.2.3. Principales catégories de réduction d'échelle statistiques

La réduction d'échelle statistique se compose d'un groupe hétérogène de méthodes qui varient dans la sophistication et l'applicabilité. Elles sont relativement simples à mettre en œuvre, mais nécessitent une quantité suffisante de données d'observation de haute qualité. Ces méthodes peuvent être regroupées en trois principales catégories: les générateurs météorologiques stochastiques, les approches par type de temps et les modèles de régression.

Les générateurs stochastiques météorologiques : Les générateurs stochastiques sont des modèles qui reproduisent les attributs statistiques des prédictants donnés. La variable de sortie de ce type de modèle est une série chronologique artificielle de données météorologiques à très haute résolution temporelle. À cause de cette haute résolution temporelle, ces modèles nécessitent de longues séquences de données quotidiennes, et sont sensibles aux données manquantes ou erronées utilisées pour la calibration (Wilks et Wilby 1999). Généralement, les générateurs stochastiques utilisent un ou plusieurs paramètres conditionnés par l'état de sortie du modèle climatique global et peuvent être appliqués aussi bien sur un seul site que sur plusieurs sites (Wilks 1998; Wilks 1999; Palutikof et al. 2002; Qian et al. 2002).

Les approches par type de temps : Dans cette méthode, la variable locale est prédite sur la base du type de la circulation atmosphérique à grande échelle. Les «types » peuvent être météorologiques, synoptiques, identifiables, ou issus des systèmes complexes cachés. Le type de l'atmosphère future, simulé par un modèle climatique global, est jumelé à l'état atmosphérique le plus semblable de la période historique. L'état de l'atmosphère historique choisi correspond alors à une valeur ou une catégorie de valeurs de la variable locale, qui sont ensuite répliquées dans le futur selon le type de la circulation atmosphérique (Zorita et Von Storch 1999). Ces méthodes sont particulièrement bien adaptées pour la réduction d'échelle des variables avec des distributions non-normales, telles que les précipitations quotidiennes. Cependant, une grande quantité de données observées quotidiennes (ex. sur 30 ans) est nécessaire afin d'évaluer toutes les conditions météorologiques possibles. L'inconvénient majeur des approches par type de temps est la mauvaise reproduction des valeurs extrêmes. En fait, ces dernières ne peuvent pas reproduire des valeurs qui dépassent les maximum observés des données de l'archive historique (Wilby et Wigley 1997).

Les modèles de régression : Les modèles de régressions permettent d'établir des relations directes entre les prédicteurs et les prédictants en utilisant des fonctions de transfert, comme par exemple la régression linéaire multiple (Wilby et al. 2002; Hammami et al. 2012; Jeong et al. 2012; Jeong et al. 2013), les réseaux de neurones artificiels (Schoof et Pryor 2001; Cannon 2008; Cannon 2011), l'analyse des fonctions orthogonales empiriques (Huth 2004), l'analyse canonique des corrélations (Palutikof et al. 2002; Huth et Pokorná 2004) et la décomposition en valeurs singulières (Widmann et al. 2003). Les modèles de régression traditionnelle sont utilisés avec succès dans la réduction d'échelle, mais leur inconvénient majeur est qu'ils reproduisent généralement la moyenne ou les prédictions centrales conditionnelles à des prédicteurs choisis. Ainsi, la variabilité de la régression est toujours inférieure à la variabilité observée (Von Storch 1999). En outre, Harpham et Wilby (2005) ont mentionné que les approches basées sur la régression montrent la difficulté de reproduire la cohérence spatiale entre les prédictants sur plusieurs stations. Les méthodes de régression sont très simples et largement utilisées.

1.3. Réduction d'échelle statistique pour les applications hydrologiques

Les modèles de réduction d'échelle doivent fournir des informations météorologiques fiables à l'échelle locale (Gachon et al. 2005). La performance d'un modèle de réduction d'échelle statistique dépend de sa capacité à reproduire les caractéristiques statistiques observées du climat local, par exemple la variabilité temporelle observée. Parce que la variabilité temporelle pourrait affecter la représentation des événements extrêmes, la bonne reproduction de la variabilité dans les applications de réduction d'échelle est un point très important. En outre, lorsque les valeurs des prédictants sont disponibles sur plusieurs stations, les modèles de réduction d'échelle doivent maintenir la cohérence des prédictions entre les sites et les variables.

Les variables météorologiques, telles que les précipitations et les températures, sont souvent d'une grande utilité pour les études d'impact hydrologiques d'un bassin versant, y compris celles liées aux changements climatiques. L'intermittence spatio-temporelle des précipitations et des températures, leurs répartitions spatiales et leurs dépendances stochastiques complexes sont quelques-unes des questions qui doivent être traitées lors d'une étude de réduction d'échelle. En hydrologie, l'écoulement dépend fortement de la distribution spatiale des précipitations dans un bassin versant, et de l'interaction entre la précipitation et la température. Par conséquent, le maintien des relations réalistes entre les précipitations et les températures sur plusieurs stations est particulièrement important dans la modélisation hydrologique (Xu 1999). Pour les ressources en eau douce, la précipitation est le plus important moteur, cependant elle est beaucoup plus difficile à modéliser que la température surtout à cause de son caractère discret-continu, sa distribution non-normale à queue lourde ainsi que sa variabilité spatiale et temporelle.

En résumé, pour les applications hydrologiques, les méthodes de réduction d'échelle sont tenues à reproduire:

- Les propriétés univariées :
  - Les valeurs extrêmes des précipitations et des températures.
  - La variabilité temporelle de la précipitation et la température.
  - L'autocorrélation à court terme principalement pour les précipitations, pour mieux produire les longues périodes sèches et les longues périodes humides.
  - Le caractère discret-continu de la précipitation.
- Les propriétés multi-sites et/ou multivariées.
  - Les dépendances spatio-temporelles des précipitations et des températures.
  - Les structures de dépendance complexes des précipitations extrêmes.

- La dépendance entre la température et la précipitation à un seul site.

2. Problématique et objectifs de recherche

2.1. Problématique

Selon la variable d'intérêt, les méthodes fondées sur la régression donnent généralement de bons résultats de réduction d'échelle, mais leur inconvénient majeur est qu'elles ne fournissent généralement que la moyenne ou la partie centrale des prédictants (Cawley et al. 2007). Par conséquent, la variance de la moyenne modélisée sera typiquement inférieure à la variance de la série observée. Un autre inconvénient est qu'elles ne peuvent pas être appliquées directement et de façon appropriée dans un cadre multi-site et/ou multivarié. En plus, l'hypothèse de normalité ne permet pas une application directe pour réduire l'échelle des précipitations. En raison de ces inconvénients, les approches basées sur la régression ne sont pas bien adaptées pour fournir les caractéristiques des précipitations et des températures requises pour les analyses hydrologiques.

Pour estimer adéquatement la variance temporelle des séries à échelle réduite en utilisant les méthodes régressives, trois principales approches ont été proposées dans la littérature, à savoir l'inflation (Huth 1999), l'ajout d'une composante aléatoire (Von Storch 1999; Clark et al. 2004) et l'expansion (Burger et Chen 2005). L'*inflation* est habituellement effectuée en multipliant les données à échelle réduite par un facteur constant, *la composante aléatoire* consiste à ajouter un bruit aléatoire et l'approche d'*expansion* consiste à forcer la correspondance entre les covariances prédites et les covariances des données observées par l'ajout d'une contrainte à la fonction de coût de régression.

Un problème qui survient avec l'approche de l'inflation réside dans le fait que les corrélations spatiales entre les sites peuvent être déformées. Cependant, la composante aléatoire peut également être appliquée dans un cadre de mise à l'échelle multi-site. Elle consiste à combiner un modèle de régression et un générateur stochastique formant ainsi un seul modèle hybride. Ainsi, le modèle stochastique hybride résultant peut surmonter les faiblesses des deux approches constituantes. Cependant, Burger et Chen (2005) ont indiqué que l'ajout d'une composante aléatoire dans une approche hybride, qui est basée sur un modèle de bruit statique, ne peut pas représenter les changements locaux dans la variabilité atmosphérique dans une simulation de changement climatique, et que cet inconvénient est bien contourné par la méthode d'expansion. En effet, dans une approche d'expansion, la variabilité observée est reproduite à l'aide d'une seule composante de régression déterministe, et donc, la variabilité reproduite n'est pas statique et peut changer dans des conditions de climat futur. Cependant, Von Storch (1999) a affirmé que les approches d'inflation et d'expansion sont des techniques inappropriées vu que l'hypothèse implicite que toute la variabilité locale peut être retracée à partir des données à grande échelle n'est pas valide et en contradiction avec la réalité. Pour cette raison, Von Storch (1999) suggère d'utiliser des composantes aléatoires qui sont plus réalistes que les approches d'inflation et d'expansion.

Même si les méthodes d'expansion permettent de reproduire la variabilité temporelle et les structures de covariances entre les prédicants, et que la variabilité totale peut changer dans des conditions climatiques futures, ces approches ne peuvent pas être interprétées physiquement, et sont contradictoires avec les connaissances physiques. Donc, même si elles donnent de bonnes réponses et de bons résultats, ces derniers sont obtenus pour une raison qui n'est pas réelle en supposant que la variabilité à l'échelle locale peut être décrite à partir d'un simple modèle de

régression déterministe. Pour cette raison les approches basées sur l'ajout d'une composante aléatoire ont été privilégiées dans la littérature. Par exemple, Jeong et al. (2012) ont utilisé des composantes aléatoires pour reproduire la corrélation spatiale de l'occurrence et la quantité des précipitations en utilisant la distribution normale multivariée. De même, Jeong et al. (2013) ont proposé un modèle multivarié multi-site pour la réduction d'échelle statistique (MMSDM) des températures minimales et maximales sur plusieurs stations simultanément. Le MMSDM emploie la régression linéaire multiple multivarié (MMLR) pour simuler des séries déterministes à partir des données de ré-analyse à grande échelle et ajoute une composante aléatoire pour compléter les variances et les structures de dépendance que le modèle MMLR n'a pas pu reproduire. Dans la même optique, Khalili et al. (2013) ont proposé une approche hybride en combinant une composante de régression linéaire avec une composante stochastique basée sur un processus d'autocorrélation spatiale. Rappelons que, même si ces approches sont couramment utilisées et peuvent être appliquées à un problème de réduction d'échelle multi-site et/ou multivariée, cette structure hybride de modélisation souffre d'un inconvénient majeur du fait que la partie de la variabilité qui est reproduite par la composante aléatoire ne dépend pas des prédicteurs et ne peut pas changer en fonction des conditions climatiques futures.

2.2. Objectifs de la thèse

L'objectif général de cette étude consiste à concevoir, tester et améliorer une nouvelle structure de modélisation hybride probabiliste qui permet de contourner le problème de la variabilité statique des méthodes hybrides traditionnelles. Particulièrement, en se basant sur cette structure de modélisation, le but sera de développer de nouveaux modèles statistiques de réduction d'échelle multi-site des précipitations et des températures. Ces nouveaux modèles sont basés sur des outils statistiques en plein essor dans la littérature hydrométéorologique au cours des

dernières années, y compris les outils multivariés tels que les copules, et les approches de régression probabilistes telles que la régression des quantiles et la forme vectorielle des modèles linéaires généralisées. Ces outils seront combinés pour la première fois en réduction d'échelle et visent à fournir des informations météorologiques fiables et nécessaires pour les applications hydrologiques. Dans le cas des précipitations, le but sera aussi de maintenir un regard à jour sur les découvertes récentes qui portent sur la nature des extrêmes et des structures de dépendance spatiotemporelles complexes.

3. Méthodologie

3.1. Structure de modélisation: la structure hybride probabiliste

Dans la structure de modélisation proposée, la première étape consiste à traiter le problème de reproduction de la variabilité temporelle. Contrairement aux approches hybrides traditionnelles, nous proposons une solution qui permet à la variabilité temporelle de changer dans le futur. Par la suite nous présentons la méthode employée pour traiter le problème de dépendance spatiotemporelle. La solution proposée consiste à reproduire la variabilité temporelle dans la composante de régression. Contrairement aux approches d'expansion, qui utilisent une composante déterministe, l'approche proposée est basée sur des outils de régression probabiliste qui permettent de reproduire toute la distribution conditionnelle. Comme le montre la Figure 1, la structure que nous allons proposer combine une composante de régression probabiliste qui fournit toute la distribution conditionnelle, avec une composante aléatoire qui contient l'information uniquement sur les structures de dépendances multi-sites et/ou multivariées.

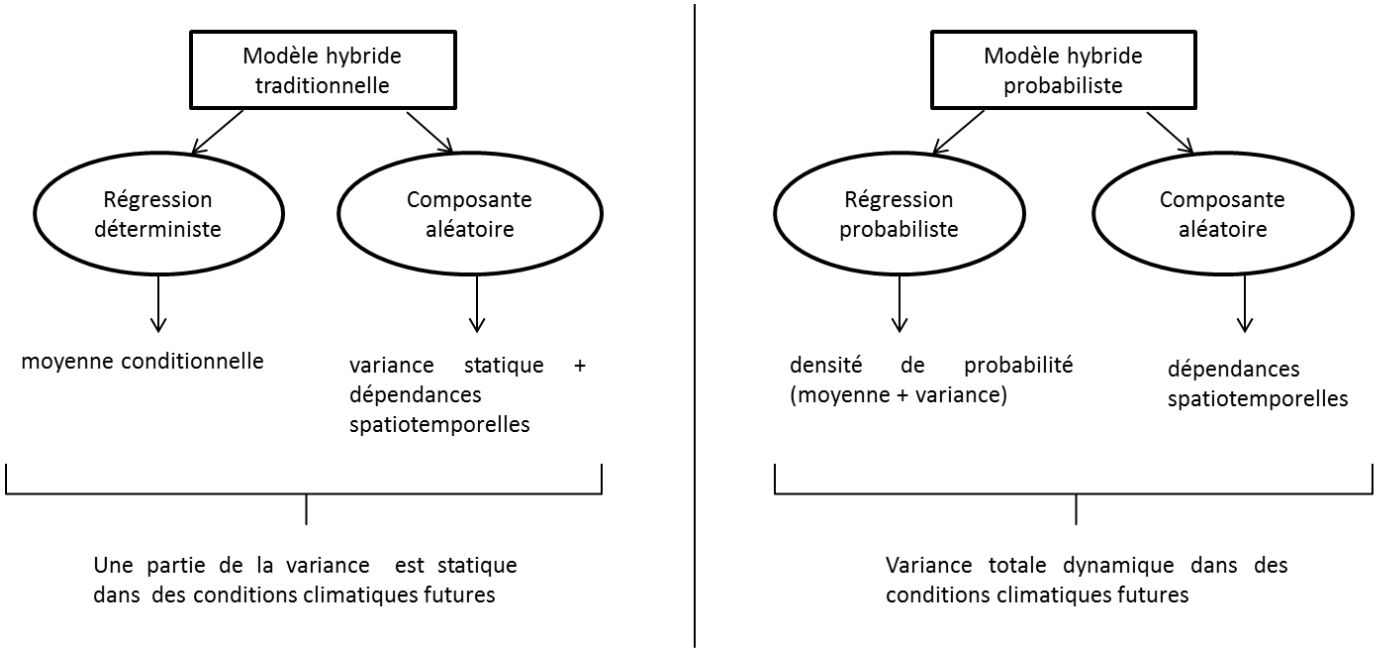

Figure 1. Différences ente la structure hybride traditionnelle et la structure hybride probabiliste

3.1.1. Régression probabiliste

Les approches de régression probabilistes ont fourni des contributions majeures dans les applications de réduction d'échelle pour mieux reproduire la variabilité temporelle observée. Les approches probabilistes comprennent: les formulations bayésiennes (Fasbender et Ouarda 2010), la régression des quantiles (Bremnes 2004a; Friederichs et Hense 2007; Cannon 2011) et les modèles de régression où les sorties sont des paramètres de la distribution conditionnelle tels que la forme vectorielle des modèles linéaires généralisés (VGLM), la forme vectorielle des modèles additifs généralisés (VGAM) (Yee et Wild 1996; Yee et Stephenson 2007) et les densités des réseaux de neurones conditionnels (CDEN pour conditional density estimation network) (Williams 1998; Li et al. 2013b). Les approches de régression probabilistes permettent la définition d'une fonction de distribution dynamique complète univariée. Dans le cas de VGLM,

VGAM et CDEN, la sortie du modèle est un vecteur de paramètres d'une distribution qui dépend des valeurs des prédicteurs. Au lieu du paramètre de position uniquement (la moyenne conditionnelle), les paramètres d'échelle et de forme peuvent varier en fonction des valeurs des prédicteurs atmosphériques mises à jour permettant ainsi un meilleur contrôle et ajustement de la dispersion, l'asymétrie et l'aplatissement. Par conséquent, la simulation d'une série temporelle à échelle réduite avec une variabilité temporelle réaliste est obtenue en simulant des valeurs aléatoires issues de la distribution conditionnelle produite à chaque étape de prévision (Williams 1998; Haylock et al. 2006). Ceci permettra de remédier au problème de la variabilité statique, du fait que la moyenne et la variance conditionnelles peuvent varier dans le futur en fonction des prédicteurs atmosphériques à grande échelle.

Un modèle de régression probabiliste univarié permet à chaque étape de prédiction de fournir la distribution cumulative $F_t(Y \mid X = x(t))$, où Y désigne un prédictant, X représente le vecteur des prédicteurs et $x(t)$ représente la valeur des prédicteurs pour un jour t allant de 1 à *n* durant la période de calibrations. Maintenant, lorsque nous disposons à la fois de plusieurs prédictants, par exemple m prédictants Y_j, avec j = 1,…, m, un modèle de régression probabiliste peut être appliqué à chaque prédicants séparément. À titre d'exemple, ces prédictants peuvent être une même variable sur plusieurs stations où des variables différentes sur une même station ou sur des stations différentes. La question qui se pose à ce niveau est: comment étendre les approches de régression probabilistes dans un contexte multi-site et/ou multivarié? Autrement dit comment générer simultanément des valeurs aléatoires à partir des m distributions conditionnelles $F_{tj}(Y_j \mid X = x(t))$ obtenues à partir des m modèles de régression probabilistes.

3.1.2. Composante aléatoire en utilisant des distributions uniformes standards

Du fait que la partie de régression probabiliste est censée reproduire toute la variabilité temporelle, la composante qu'on devrait ajouter pour compléter cette partie devrait reproduire uniquement l'information sur la structure de dépendance. En d'autres termes, nous avons besoin d'une composante aléatoire qui va expliquer les structures de dépendances sans avoir à ajouter une variance additionnelle. À ce niveau, le lecteur pose la question suivante: où peut-on trouver cette information sur les structures de dépendances non expliquée par les m modèles de régressions probabilistes, et comment modéliser ou tenir compte de cette information à l'étape de simulation.

Rappelons que, dans une structure hybride traditionnelle, la variabilité et les structures de dépendance non expliquées par la composante de régression sont contenues dans la matrice des résidus. Cette dernière est obtenue après l'évaluation de la composante de régression en utilisant les données de calibration. Par analogie à la structure hybride traditionnelle, pour reproduire les structures de dépendance spatio-temporelles dans une structure hybride probabiliste, la première étape consiste à extraire la matrice des erreurs. Cette matrice est cachée et devrait être extraite durant la période de calibration en comparant les sorties des modèles de régressions probabilistes avec les séries des prédictants observées. Extraire la matrice des erreurs représente un défi, du fait qu'à chaque étape de prédiction, la sortie du modèle de régression probabiliste est toute une densité de probabilité conditionnelle alors qu'une seule valeur ponctuelle observée du prédictant est disponible pour la comparaison. Pour remédier à ce point, une matrice de variables cachées $\mathbf{U}$ de dimension $n \times m$ est extraite. Premièrement, l'évaluation des m modèles probabilistes durant la

période de calibration, permettra d'obtenir les fonctions de densités conditionnelles à chaque étape de prédiction t, et ainsi les distributions conditionnelles cumulatives F_{tj} pour chaque prédictant j. Les éléments u_{tj} de la matrice $\mathbf{U}$ sont obtenus à partir de l'équation suivante :

$$u_{tj} = F_{tj}(y_{tj}) \tag{1}$$

où y_{tj} représente la valeur observée du prédictant Y_j au jour t. Par analogie aux modèles hybrides traditionnels, la quantité $F_{tj}(y_{tj})$ est équivalente à l'erreur d'un modèle de régression déterministe qui produit comme sortie une seule valeur prédite du prédictant. La Figure 2 montre les étapes nécessaires à l'obtention des variables cachées pendant la période de calibration.

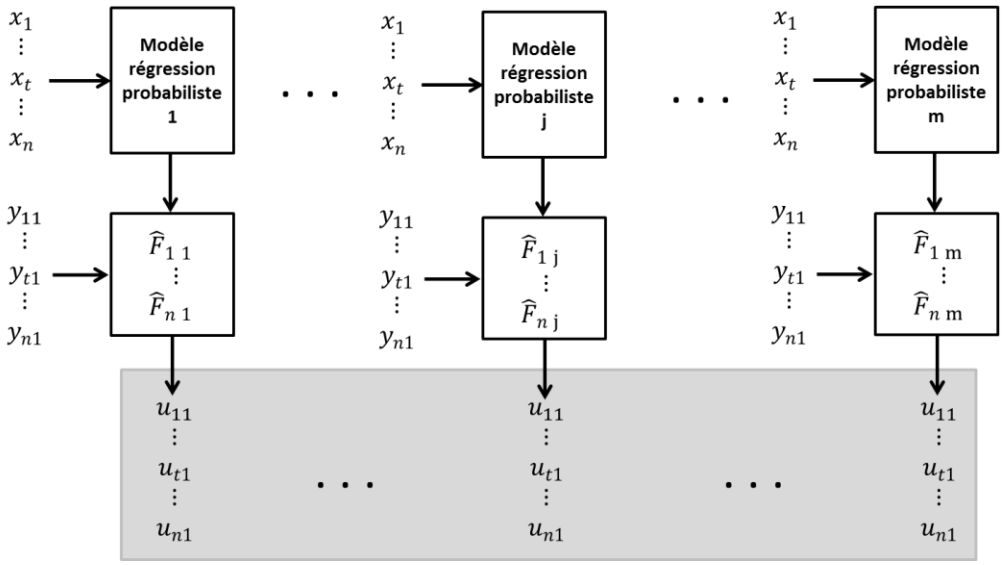

Figure 2. Méthode pour l'extraction de la matrice des variables cachées

La matrice des variables cachées $\mathbf{U}$ résultante représente des valeurs entre 0 et 1 qui contiennent les informations inexpliquées par les modèles des régressions probabilistes. Elle contient notamment l'information sur les structures de dépendances spatiotemporelles multi-sites et/ou

multi-variables incluant l'autocorrélation à court et à long termes. Si le modèle de régression probabiliste reproduit bien les caractéristiques marginales d'un prédictant j, la matrice variable U_j correspondant au prédictant Y_j sera uniformément distribuée entre 0 et 1, si ce n'est pas le cas, ceci résulte du fait que le modèle de régression n'est pas approprié pour reproduire les caractéristiques marginales du prédicant en question.

À ce stade, il nous reste à montrer comment faire des simulations multi-sites et/ou multivariée à partir de la structure de modélisation hybride probabiliste que nous avons proposée. En règle générale, la simulation à partir d'un modèle de régression probabiliste peut être réalisée par l'échantillonnage d'une valeur aléatoire issue d'une distribution uniforme sur [0, 1], ensuite d'appliquer à cette valeur l'inverse de la fonction de distribution cumulative obtenue à partir du modèle de régression probabiliste. Nous devons toujours garder à l'esprit que, les paramètres de la distribution conditionnelle, sorties du modèle de régression probabiliste, varient à chaque étape de prévision en fonction des valeurs mises à jour des prédicteurs atmosphériques à grande échelle. Maintenant, pour obtenir des simulations spatialement corrélées, nous avons besoin de simuler des variables aléatoires issues des distributions uniformes sur [0, 1] (distributions uniformes standards) et qui sont corrélées.

Ainsi les séries synthétiques des prédictants au cours de la période de validation peuvent être obtenues en utilisant les deux étapes suivantes :

(i) Générer des séries aléatoires issues d'une distribution uniforme standard correspondant à l'ensemble des prédictants et qui ont les mêmes structures de dépendance que la matrice des variables cachées $\mathbf{U}$.

(ii) Appliquer sur les séries générées à l'étape (i) l'inverse de la distribution cumulative obtenue à partir du modèle de régression probabiliste à chaque étape de prédiction t' et pour chaque prédictant j correspondant. Où t' désigne un jour durant la période de validation.

3.2. Outils statistiques

Dans cette section nous présentons les différents outils statistiques qui ont été utilisés dans la structure probabiliste hybride proposée.

3.2.1. Composante 1 : régression probabiliste

Les efforts déployés dans le domaine des statistiques ont été consacrés à l'élaboration du modèle de régression linéaire et de méthodes d'estimation associées en minimisant une somme des carrés des résidus appelée méthode des moindres carrés. Les modèles de régression élaborés en utilisant l'estimateur des moindres carrés produisent la moyenne conditionnelle de prédictant connaissant la valeur des prédicteurs choisis. Cette démarche est plus appropriée si les prévisions sont générées à partir d'une fonction déterministe qui est altérée par un processus de bruit normalement distribué avec une variance constante (Cannon 2008). Lorsque le processus de bruit a une variance non constante ou est non-normal, il est plus approprié d'utiliser un modèle qui décrit la densité conditionnelle du prédictant dans un cadre probabiliste.

3.2.1.1. Forme vectorielle des modèles linéaires généralisés (VGLM)

La distribution de la quantité des précipitations, à une échelle de temps journalière, tend à être fortement asymétrique et est communément supposée suivre une distribution Gamma (Stephenson et al. 1999; Giorgi et al. 2001; Yang et al. 2005). Dans une perspective de régression, le modèle linéaire généralisé (GLM) étend la régression classique pour gérer

l'hypothèse de normalité de la sortie du modèle. Ainsi, la sortie peut suivre une gamme de distributions qui permettent à la variance de dépendre de la moyenne comme, par exemple, la famille de distribution exponentielle et en particulier la distribution Gamma (Coe et Stern 1982; Stern et Coe 1984; Chandler et Wheater 2002). Néanmoins, les résultats récents suggèrent que la distribution Gamma peut ne pas convenir pour la modélisation des précipitations extrêmes car elle est très restrictive et ne peut pas tenir compte de certaines caractéristiques des précipitations comme par exemple les queues lourdes. Pour traiter ce problème, d'autres options ont été proposées dans la littérature en particulier les distributions Pareto généralisée (GP) et Weibull (WEI) (Ashkar et Ouarda 1996; Serinaldi et Kilsby 2014). Toutefois, du fait que la variance ne dépend pas de la moyenne, ces deux distributions ne peuvent pas être utilisées dans un GLM. La forme vectorielle des modèles linéaires généralisés (VGLMs) a été développée pour gérer cette insuffisance (Yee et Stephenson 2007). Au lieu de la moyenne conditionnelle uniquement, le modèle VGLM fournit toute la distribution conditionnelle de la réponse en utilisant un modèle de régression linéaire dont les sorties sont des vecteurs de paramètres de la distribution conditionnelle sélectionnée (Kleiber et al. 2012). En outre, dans les applications de réduction d'échelle, le VGLM a un avantage particulier lui permettant de reproduire la variabilité temporelle.

3.2.3.1. Régression des quantiles

Du fait que la moyenne d'un échantillon peut être définie comme la solution au problème de minimisation d'une somme des carrés des résidus, l'estimateur du moindre carrés couramment utilisé en analyse de régression fournit la moyenne conditionnelle de la réponse. Cependant, la médiane d'un échantillon peut être définie comme la solution au problème de minimisation d'une somme des résidus absolus. À cet égard, la régression médiane, également connu comme

régression des moindres écarts absolus (LAD pour least absolute deviation), minimise la somme des résidus absolus. La régression médiane est plus robuste aux valeurs aberrantes que la régression des moindres carrés et évite l'imposition d'une distribution paramétrique du processus des erreurs. La question qui se pose est la suivante: puisque la médiane correspond au quantile d'ordre 0.5, pourquoi ne pas utiliser d'autres quantiles également? Autrement dit, si la moyenne et la médiane d'un échantillon peuvent être définies comme des solutions à des problèmes de minimisation appropriés, quel est le problème d'optimisation qui peut avoir comme solution un quantile quelconque de l'échantillon? En cherchant la réponse à cette question, Koenker et Bassett (1978) ont introduit une nouvelle technique de régression appelée régression quantile qui fournit le quantile conditionnelle de la variable réponse.

Même si les approches de régression probabilistes, qui intègrent l'influence de prédicteurs atmosphériques à grande échelle sur le vecteur des paramètres de la distribution conditionnelle, (tel que le modèle VGLM) sont avantageuses, ces approches souffrent des problèmes inhérents à savoir l'imposition d'une forme paramétrique et la supposition que cette forme paramétrique restent la même à chaque étape de prédiction. Dans ce contexte, la régression des quantiles représente une solution alternative avantageuse qui permet de décrire toute la distribution conditionnelle sans imposer aucune forme paramétrique. En effet, une description complète de la distribution conditionnelle peut être obtenue en fournissant directement ses quantiles, par exemple les quantiles de non-dépassement d'un ordre de 0.01 à 0.99 avec un pas de 0.01. Ceci revient à appliquer un modèle de régression quantile pour chacun de ces ordres de quantile.

Au cours de la dernière décennie, l'application de la régression quantile dans la modélisation environnementale et l'évaluation de l'impact des changements climatiques a augmenté considérablement. Les modèles basés sur la régression quantile ont été introduits pour décrire les

effets des variables météorologiques sur la concentration d'ozone (Baur et al. 2004), étudier l'écoulement fluvial annuel (Luce et Holden 2009), prédire l'énergie éolienne (Bremnes 2004b), estimer l'incertitude hydrologique (Weerts et al. 2011), prédire les quantile des crues dans un climat en évolution (Sankarasubramanian et Lall 2003) et modéliser l'intensité et la tendance des cyclones tropicaux (Elsner et al. 2008; Jagger et Elsner 2009). En outre, l'application de la régression quantile a apporté d'importantes contributions pour la réduction d'échelle des précipitations (Bremnes 2004a; Friederichs et Hense 2007; Cannon 2011; Tareghian et Rasmussen 2013).

3.2.2. Composante 2 : générateur à valeurs entre 0 et 1

Dans cette section nous présentons les différents outils qui ont été utilisés pour modéliser les variables cachées de la matrice $\mathbf{U}$ de la section 3.1.2.

3.2.2.1. Copule Gaussienne

La composante aléatoire, dans la structure de modélisation hybride probabiliste proposée, consiste à générer simultanément des valeurs uniformément distribuées entre 0 et 1 et qui ont les mêmes caractéristiques que les variables cachées de la matrice $\mathbf{U}$. De ce fait, la matrice $\mathbf{U}$ peut être modélisée en utilisant une distribution multivariée dont les marges sont uniformément distribuées entre 0 et 1. Une telle distribution est appelée copule. Récemment, les copules sont devenues très populaires, en particulier dans certains domaines comme l'économétrie, les finances, la gestion des risques et de l'assurance. Au cours des dernières années, l'application des copules a également apporté d'importantes contributions dans le domaine de l'hydrométéorologie. Une introduction à la théorie de la copule est fournie dans Joe (1997), Nelsen (2013), Genest et Chebana (2015) et Salvadori et De Michele (2007). Schölzel et Friederichs (2008) donnent un

bref aperçu des copules pour des applications dans la météorologie et le climat. Les modèles basés sur les copules ont été introduits également dans l'analyse fréquentielle hydrologique multivariée (Chebana et Ouarda 2007; El Adlouni et Ouarda 2008; Chebana et Ouarda 2011), l'évaluation des risques, l'interpolation géostatistique et l'analyse des valeurs extrêmes multivariées (De Michele et Salvadori 2003; Bárdossy 2006; Renard et Lang 2007; Kazianka et Pilz 2010). En plus, les copules ont été largement utilisés pour décrire la structure de dépendance des variables climatiques extrêmes (AghaKouchak 2014; Guerfi et al. 2015; Hobæk Haff et al. 2015; Mao et al. 2015; Vernieuwe et al. 2015). Les copules permettent de décrire la structure de dépendance indépendamment des distributions marginales, et donc, en utilisant différentes distributions marginales en même temps sans transformation (Sklar 1959; Dupuis 2007). Dans cette thèse, nous avons utilisé une copule Gaussienne, comme un choix initial pour tester la structure proposée et aussi puisqu'elle ne pose pas de difficultés pour les dimensions élevées comme le cas ici. Une copule Gaussienne est définie par:

$$\mathbb{C}(w;C) = \Phi_m\left[\Phi^{-1}(w_1),\ldots,\Phi^{-1}(w_m);C\right] \tag{2}$$

où Φ est la fonction de distribution cumulative normale standard univariée, et $\Phi_m(w;C)$ désigne la distribution cumulative pour un vecteur normale multivariée w de dimension *m*, de moyenne 0 et de matrice de covariance C.

3.2.2.2. Champ Gaussien multivarié autorégressif

Reproduire l'autocorrélation des séries météorologiques est un point très important pour bien reproduire les périodes sèches et les périodes humides. Toutefois, la copule Gaussienne ne permet pas de tenir compte de l'autocorrélation d'une série. Pour remédier à ce point, la copule Gaussienne peut être utilisée dans un cadre autorégressif multivarié. Les variables cachées de la

matrice $\mathbf{U}$ peuvent être transformées en des variables Gaussiennes de matrice $\mathbf{Z}$. Un élément z_{tj} de la matrice $\mathbf{Z}$ est obtenu en utilisant la transformation suivante:

$$z_{tj} = \Phi^{-1}[u_{tj}] \tag{3}$$

Utiliser une copule Gaussienne est équivalent à modéliser la matrice $\mathbf{Z}$ en utilisant une distribution Gaussienne multivariée. Pour tenir compte de l'autocorrélation des séries dans un seul site où la corrélation croisée entre deux prédictants décalés d'un jour, la matrice $\mathbf{Z}$ peut être modélisée en utilisant une distribution Gaussienne multivariée autorégressive (MAR).

3.2.2.3. Échantillonnage non-paramétrique

Des résultats récents démontrent qu'un examen attentif de la structure de dépendance dans les processus hydrométéorologiques révèle que le cadre méta-Gaussien est très restrictif et ne peut pas tenir compte de certaines caractéristiques telles que l'asymétrie et les queues lourdes. Par conséquent, il ne sera plus possible de simuler de façon réaliste la structure de dépendance multi-site des précipitations quotidiennes (El Adlouni et al. 2008; Bárdossy et Pegram 2009; Lee et al. 2013).

Pour exploiter cette connaissance pour la simulation des précipitations, Li et al. (2013a) et Serinaldi (2009) ont considéré les copules pour introduire des structures temporelles non Gaussiennes dans un site unique. Bargaoui et Bárdossy (2015) ont utilisé des copules bivariées non Gaussiennes pour modéliser les courtes durées des précipitations extrêmes. Pour la simulation multi-site des précipitations, Bárdossy et Pegram (2009) et AghaKouchak et al. (2010) ont introduit des structures spatiales de dépendance de queue non Gaussiennes en simulant à partir d'une copule normale v-transformée proposée par Bárdossy (2006). D'autres modèles

théoriques de copules peuvent également être utilisés pour reproduire ces propriétés de dépendances spatiales telles que les copules Meta-elliptiques (Fang et al. 2002) ou les copules vignes (Gräler 2014).

À ce niveau, il est presque évident que nous devons sélectionner une copule flexible qui fournit à la fois la dépendance temporelle et spatiale. Cependant, la question clé dans notre problème n'est pas de simuler les structures de dépendance à partir d'une copule qui donne le meilleur ajustement aux données. La principale question est plutôt : "comment extraire l'information sur la structure de dépendance des données, et la façon de préserver cette information dans l'étape de simulation?" Avant de répondre à cette question, nous devons d'abord savoir où nous pouvons trouver cette information. La matrice de données de rang peut être considérée comme le support d'information de la copule empirique. Rappelons que les rangs de données sont les statistiques qui contiennent la plus grande quantité d'informations sur la structure de dépendance de données (Oakes 1982; Genest et Plante 2003; Song et Singh 2010). Dans ce contexte, l'information sur la structure de dépendance des données peut être reproduite à l'étape de simulation en utilisant un échantillonnage basé sur les rangs des données (Vinod et López-de-Lacalle 2009; Vaz de Melo Mendes et Leal 2010; Srivastav et Simonovic 2014). Ainsi, au lieu d'utiliser une copule flexible (donc complexe et non parcimonieuse), une technique d'échantillonnage non paramétrique simple peut être adoptée comme une alternative aux copules. La procédure consiste à générer des séries aléatoires uniformes entre 0 et 1, puis à les ordonner en fonction des rangs observés de la matrice des variables cachées $\mathbf{U}$. Par conséquent, les rangs observés seront conservés ce qui permet de préserver une grande quantité de la structure de dépendance spatio-temporelle sans avoir supposé des dépendances Gaussiennes ou une forme particulière de copules.

4. Applications et résultats

4.1. Zone d'étude et base de données

Des données observées quotidiennes de températures maximales et minimales et des précipitations des stations situées dans la province de Québec (Canada) ont été utilisées dans la présente étude. Les emplacements géographiques de ces stations sont représentés sur la Figure 3.

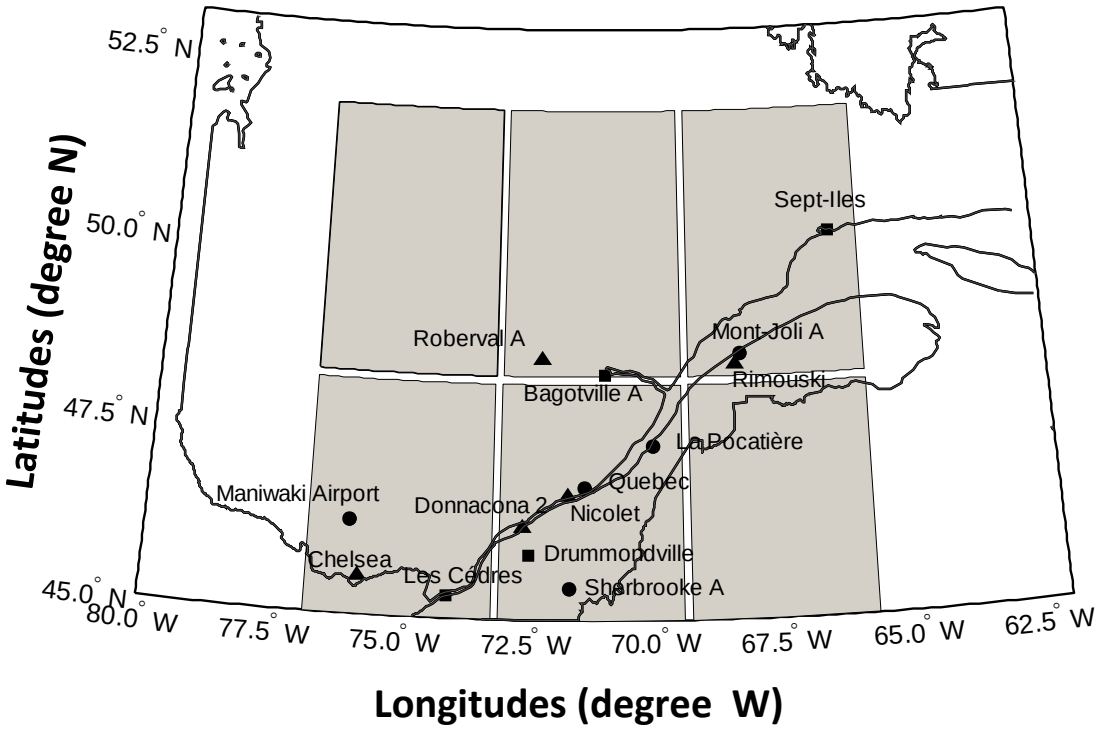

Figure 3. Position des grilles CGCM3 et des stations d'observations pour les précipitations et les températures. Les stations Les Cèdres, Drummondville, Sept-Îles et Bagotville contiennent à la fois des données de précipitations et de températures. Ces quatre stations sont représentées par des carrés. Le reste des stations sont représentées par des cercles pour les températures et par des triangles pour les précipitations.

La liste des stations est présentée dans le Tableau 1 pour les températures et dans le Tableau 2 pour les précipitations. Neuf stations sont considérées pour les températures maximales et minimales ainsi que neuf stations sont disponibles pour les précipitations. Notons que les stations Cedars, Drummondville, Sept-Îles et Bagotville, contiennent à la fois des données de précipitations et de températures. Toutes les séries des prédictants sont obtenues à partir des stations météorologiques d'Environnement Canada.

Tableau 1. Liste des neuf stations de températures.

No.	Nom de la station	Latitude (°N)	Longitude (°W)
1	Les Cèdres	45.30	74.05
2	Drummondville	45.88	72.48
3	Sept-Îles	50.22	66.27
4	Bagotville A	48.33	71.00
5	Québec	46.79	71.38
6	Sherbrooke A	45.43	71.68
7	Maniwaki Airport	46.27	75.99
8	La Pocatière	47.36	70.03
9	Mont-Joli A	48.60	68.22

Tableau 2. Liste des neuf stations des précipitations.

No.	Nom de la station	Latitude (°N)	Longitude (°W)
1	Chelsea	45.52	75.78
2	Les Cèdres	45.30	74.05
3	Nicolet	46.25	72.60
4	Drummondville	45.88	72.48
5	Donnacona	46.69	71.73
6	Roberval A	48.52	72.27
7	Bagitville A	48.33	71.00
8	Rimouski	48.45	68.53
9	Sept-Îles	50.22	66.27

Les prédicteurs sont obtenus à partir du produit de ré-analyse NCEP / NCAR interpolés sur la grille Gaussienne MCCG3 (3,75 ° de latitude et longitude). Six grilles couvrant la zone des stations des prédictants ont été sélectionnées (voir la Figure 2) dans lesquelles 25 prédicteurs NCEP sont disponibles (voir Tableau 3). Au total, 150 prédicteurs sont disponibles pour le processus de réduction d'échelle. Pour réduire le nombre des prédicteurs, une analyse en composantes principales (ACP) a été effectuée. Les premières composantes principales qui conservent plus de 97% de la variance totale ont été sélectionnées. L'ensemble des données couvrent la période entre le 1er Janvier 1961 et le 31 Décembre 2000. Cette période d'enregistrement est divisée en deux sous-périodes désignées pour la calibration (1961-1990) et la validation (1991-2000).

Tableau 3. Liste des prédicteurs NCEP dans une grille CGCM3.

No	Prédicteurs	No	Prédicteurs
1	Pression au niveau moyen de la mer	14	Divergence à 500 hPa
2	Vitesse du vent 1000 hPa	15	Vitesse du vent 850 hPa
3	Composante U à 1000 hPa	16	Composante U à 850 hPa
4	Composante V à 1000 hPa	17	Composante V à 850 hPa
5	Tourbillon à 1000 hPa	18	Tourbillon à 850 hPa
6	Direction du vent à 1000 hPa	19	Géopotentiel à 850 hPa
7	Divergence à 1000 hPa	20	Direction du vent à 850 hPa
8	Vitesse du vent à 500 hPa	21	Divergence à 1000 hPa
9	Composante U à 500 hPa	22	Humidité spécifique à 500 hPa
10	Composante V à 500 hPa	23	Humidité spécifique à 850 hPa
11	Tourbillon à 500 hPa	24	Humidité spécifique à 1000 hPa
12	Géopotentiel à 500 hPa	25	Température à 2m
13	Direction du vent à 500 hPa		

4.2. Les différents modèles élaborés

Dans cette thèse, cinq modèles de réduction d'échelle ont été développés en se basant sur la structure de modélisation hybride probabiliste proposée:

Le premier modèle est le modèle de régression probabiliste avec copule Gaussienne (PGCR). Ce modèle utilise le VGLM comme régression probabiliste et une copule Gaussienne comme composante aléatoire. Il est présenté dans Ben Alaya et al. (2014). Le modèle PGCR a été proposé comme une première évaluation de la structure hybride probabiliste. Le modèle a été appliqué pour réduire l'échelle de la température et des précipitations sur quatre stations. Pour la comparaison avec la structure hybride traditionnelle, PGCR a été comparé avec le modèle multi-site multivariée de réduction d'échelle statistique (MMSDM pour multisite multivariate statistical downscaling model) et le modèle classique de régression linéaire multiple multivarié (MMLR pour multiple multivariate lineaire regression). Le modèle MMSDM ajoute une composante aléatoire au modèle MMLR pour compléter la variabilité et les structures de dépendances non expliquées par la composante MMLR.

Le deuxième modèle qui a été développé est le modèle de Bernoulli-Pareto Généralisée multivariée autorégressif (BMAR), et est décrit dans Ben Alaya et al. (2015a). Le modèle BMAR combine un modèle de régression VGLM de distribution Pareto généralisée avec une copule Gaussienne autorégressive. Par rapport au modèle PGCR, le modèle BMAR intègre une distribution plus appropriée afin de mieux reproduire les précipitations extrêmes en utilisant une distribution mixte Bernoulli-Pareto Généralisée. Concernant la composante aléatoire, le modèle BMAR intègre un champ Gaussien multivarié autorégressif (MAR) dans le but de préserver la corrélation spatiale et l'autocorrélation temporelle à court terme. Le modèle BMAR a été

comparé avec un modèle de précipitations issu de la structure hybride traditionnelle. Il s'agit du modèle hybride proposé par Jeong et al. (2012) pour la réduction d'échelle multi-site des précipitations quotidiennes. Comme fonction de transfert, Jeong et al. (2012) ont utilisé une régression linéaire multiple multivariée (MMLR). Ensuite, pour la reproduction de la variabilité temporelle et la dépendance spatiale des observations multi-sites, une composante aléatoire a été intégrée en utilisant (i) une distribution normale multivariée, (ii) un modèle de chaîne de Markov de premier ordre, et (iii) une technique de correction du biais.

Le troisième modèle qui a été élaboré utilise comme composante de régression probabiliste la régression des quantiles, et comme composante aléatoire la copule Gaussienne. Ce modèle, nommé GCQR pour Gaussian Copula Quantile Regression, est décrit dans Ben Alaya et al. (2015b). Ce modèle a été appliqué pour réduire l'échelle de la température et des précipitations sur toute la base de données antérieurement présentée, et a été comparé avec les modèles MMLR et MMSDM.

Le quatrième modèle qui a été développé combine le modèle VGLM en utilisant la distribution mixte Bernoulli-Pareto généralisée avec la procédure d'échantillonnage non paramétrique. Ce modèle est nommé VGLM-NB (pour VGLM non-parametric bootstrapping). Le premier avantage de l'élaboration de ce modèle est de tester la structure hybride probabiliste en intégrant une composante aléatoire avec des structures de dépendances non Gaussiennes. Le modèle a été appliqué pour la réduction d'échelle multi-site des précipitations et comparé avec le modèle VGLM-MAR (VGLM avec un champ Gaussien MAR). Le modèle est décrit dans Ben Alaya et al. (2016c).

Le dernier modèle qui a été élaboré est un modèle intégrant la régression des quantiles avec le champ Gaussien MAR. Ce modèle est présenté dans Ben Alaya et al. (2016a). Bien que ce modèle n'utilise pas un nouvel outil par rapport aux outils précédents de la structure hybride probabiliste, il présente une nouvelle optique de la modélisation des précipitations lorsqu'on veut intégrer des covariables dans le modèle. En combinant la régression des quantiles avec la copule Gaussienne autorégressive, ce modèle présente donc une extension du 3^{e} modèle GCQR pour intégrer l'autocorrélation des précipitations à court terme.

Le Tableau 4 présente un récapitulatif des différents modèles élaborés au cours de cette thèse, en se basant sur la structure hybride probabiliste proposée.

Tableau 4. Les différents modèles élaborés en se basant sur la structure hybride probabiliste

Modèle	**Composantes**		**Distribution conditionnelle des prédictants**		
	Régression	**Aléatoire**	**Température**	**Occurrence des précipitations**	**Quantité des précipitations**
PGCR	VGLM	Copule Gaussienne	Gaussienne	Bernoulli	Gamma
BMAR	VGLM	MAR		Bernoulli	Pareto Généralisée
GCQR	QR	Copule Gaussienne	Non-paramétrique	Bernoulli	Non-paramétrique
VGLM-NB	VGLM	NB		Bernoulli	Pareto Généralisée
QRMAR	QR	MAR		Bernoulli	Gamma

Dans Ben Alaya et al. (2016b), les composantes de régression probabilistes ont été appliquées et comparées, dans un cadre de réduction d'échelle, sur une base de données qui couvre la totalité de la province du Québec. En outre, pour des fins de comparaison, une autre formulation de la

régression probabiliste a été considérée, à savoir la formulation bayésienne. Dans une première étape, le modèle spatial Bayésien (SBM) de Fasbender et Ouarda (2010) a été adapté et appliqué pour la réduction d'échelle de la température sur une zone d'étude qui couvre la totalité de la province du Québec. Le modèle Bayésien est proposé en vue de contourner l'incapacité des méthodes de régression classiques à produire des estimations spatiales à des sites non jaugés. En utilisant cette méthode, la distribution de la moyenne à priori est estimée à l'aide des caractéristiques locales dans un modèle de régression géographique (GRM). L'approche employée repose sur un cadre Bayésien afin de combiner un modèle spatial mensuel commun pour les températures minimales et maximales avec les fluctuations quotidiennes induites par les prédicteurs atmosphériques. Dans une deuxième étape, ce modèle spatiale Bayésien a été comparé avec la régression des quantiles et le modèle VGLM, pour la réduction d'échelle des températures maximales et minimales.

4.2. Principaux résultats et discussions

Dans cette section nous présentons les principaux résultats d'application des approches proposées et présentées dans la section précédente. Cette section contient deux sous-sections. Dans la première, on s'intéresse à présenter les principaux résultats associés à un seul site. Dans la deuxième sous-section, l'accent est mis sur les résultats liés aux cas multi-sites et multivairiés.

4.2.3. Résultats univariés (dans un seul site)

Le Tableau 5 résume les résultats de l'évaluation des modèles GCQR, PGCR, MMLR et MMSDM pour la réduction d'échelle des précipitations et des températures associés aux stations Les Cèdres, Drummondville, Sept-Îles, Bagotville. Pour évaluer et comparer la performance des modèles, trois critères statistiques ont été considérés: la racine carrée de l'erreur quadratique

moyenne (*RMSE*), l'erreur moyenne (*ME*) et la différence entre la variance observée et la variance modélisée (*D*). Les résultats obtenus montrent que, en termes de *RMSE* et *ME*, les performances des quatre modèles sont similaires pour les températures maximales et minimales. D'autre part, on constate que les modèles PGCR et GCQR performent mieux que les modèles MMSDM et MMLR pour réduire l'échelle des précipitations en termes de *RMSE* et *ME*. En plus, dans Ben Alaya et al. (2014) et Ben Alaya et al. (2015b), la comparaison basée sur des indices des températures et des précipitations montre que les modèles PGCR et GCQR représentent mieux les extrêmes et la variabilité observée des températures et des précipitations sur une base saisonnière et interannuelle. De même, dans Ben Alaya et al. (2015a) et Ben Alaya et al. (2016a) les résultats de comparaison montrent que les modèles BMAR et QRMAR conduisent à des estimations plus efficaces que le modèle hybride traditionnel de Jeong et al. (2012) en termes de *ME*, *RMSE* et *D*.

Tableau 5. Évaluation de la qualité de la série estimée pour PGCR, MMLR, MMSDM et GCQR au cours de la période de validation (1991-2000) pour les quatre stations météorologiques. Les critères sont *ME*, *RMSE*, et les différences entre la variance observé et modélisé *D*. Pour les critères des modèles de PGCR et MMSDM ont été calculées à partir de la moyenne conditionnelle.

		ME			RMSE			D		
		Tmax (°C)	Tmin (°C)	Prec (mm)	Tmax (°C)	Tmin (°C)	Prec (mm)	Tmax (°C)	Tmin (°C)	Prec (mm)
Les Cèdres	PGCR	0.55	0.22	**-**0.14	3.28	3.70	**6.35**	-1.34	-0.85	**-1.16**
	MMLR	0.55	0.22	2.59	3.28	3.70	7.02	8.81	9.21	45.49
	MMSDM	0.55	0.22	1.97	3.30	3.71	6.48	-1.84	-2.83	17.92
	GCQR	0.55	**0.21**	**0.07**	3.28	3.70	6.38	**-0.50**	**-0.50**	9.03
Drummondville	PGCR	**0.46**	0.49	0.48	3.31	4.08	5.42	-3.83	6.31	10.69
	MMLR	0.47	0.49	2.44	3.31	4.08	6.14	7.22	18.15	34.53
	MMSDM	0.47	0.49	0.97	3.32	4.10	5.57	-4.10	**3.47**	11.27
	GCQR	0.47	0.49	**0.41**	3.31	4.08	**5.35**	**-3.40**	6.13	**10.48**
Sept-Îles	PGCR	0.20	**-0.52**	-0.67	3.18	3.59	5.57	**5.51**	**2.12**	-10.02
	MMLR	0.19	-0.53	2.11	3.17	**3.58**	5.99	16.36	13.42	33.31
	MMSDM	0.19	-0.53	0.72	3.17	3.60	5.59	6.5	2.34	13.43
	GCQR	0.19	-0.53	**-0.35**	3.17	3.59	**5.33**	6.74	2.33	**0.89**
Bagotville	PGCR	-0.05	0.14	0.87	3.53	3.85	**6.13**	**1.12**	-0.28	24.40
	MMLR	-0.05	0.14	2.37	3.53	3.84	6.72	15.42	12.76	41.51
	MMSDM	-0.05	0.14	1.11	3.53	3.85	6.24	3.48	-1.77	28.16
	GCQR	-0.05	**0.13**	**0.72**	3.55	3.84	6.35	1.95	**-0.15**	**22.88**

Le caractère gras indique le meilleur résultat.

Dans Ben Alaya et al. (2016b) la comparaison entre les trois modèles probabilistes, VGLM, le modèle de la régression des quantiles et le modèle spatiale Bayésien (SBM), a été effectuée en se basant sur les RMSEs d'un ensemble de plusieurs indices climatiques mensuels et saisonniers. Les résultats de cette comparaison montrent qu'en général, le modèle QR fournit une meilleure performance par rapport au modèle VGLM et au modèle SBM. On peut conclure également que les résultats des modèles VGLM et QR sont légèrement supérieurs à ceux du modèle Bayésien en termes de *RMSE* des indices climatiques. Cette constatation n'est pas surprenante puisque le

modèle VGLM et le modèle QR déterminent une relation directe entre les prédicteurs et les 22 prédictants au niveau des stations, contrairement au modèle SBM qui introduit des informations de prédicteurs par leurs liens à travers des prédictants à grandes échelles. Cependant, ces deux modèles (VGLM et QR) sont incapables de fournir des estimations à des endroits non-jaugés. Même si ces derniers sont plus précis que le SBM, ce gain de précision peut être négligé devant l'avantage de fournir des estimations dans des sites non-jaugés.

En conclusion, les résultats obtenus exhibent le rôle de la composante de régression probabiliste aussi bien au niveau du QR qu'au niveau du VGLM. En effet, la composante de régression probabiliste permet de prédire non seulement la moyenne conditionnelle mais aussi toute la répartition conditionnelle. De ce fait, la variabilité temporelle peut changer dans des conditions des climats futurs, ce qui n'est pas le cas pour les modèles hybrides traditionnels, tels que les modèles MMSDM et le modèle hybride de Jeong et al. (2012).

4.2.3. Résultats sur les propriétés multi-sites et multivariées

Pour évaluer la capacité de la structure de modélisation hybride probabiliste, les résultats ont été comparés avec des approches hybrides traditionnelles en se basant sur les nuages des points observés et simulés des corrélations croisées entre les paires des stations. La Figure 4 présente les nuages des points des corrélations croisées entre les stations pour les modèles GCQR, MMLR et MMSDM pour la température maximale, la température minimale, la quantité des précipitations et l'occurrence des précipitations. Comme le montre la Figure 4, le modèle MMLR généralement surestime la corrélation croisée pour tous les prédictants et donne la plus mauvaise performance comparativement aux GCQR et MMSDM. Ce résultat est attendu, du fait que le modèle MMLR n'est pas un modèle multi-site. Nous pouvons apercevoir que les modèles GCQR et MMSDM

préservent bien la corrélation entre les stations de la température maximale, la température minimale ainsi que la quantité de précipitations. Pour l'occurrence des précipitations, les deux modéles GCQR et MMSDM sous-estiment les corrélations entre les stations. Toutefois, en se basant sur les valeurs de *RMSE* associées à chaque modèle, on peut noter que le modèle GCQR performe légèrement mieux (*RMSE* = 0.0746) que le modèle MMSDM (*RMSE* = 0.1511). Il convient également de mentionner que des résultats similaires au modèle GCQR ont été obtenus dans Ben Alaya et al. (2014) en utilisant le modèle PGCR qui intègre la copule Gaussienne avec le modèle VGLM. Par conséquent, le gain réel en utilisant les modèles PGCR et GCQR ne consiste pas seulement dans la modélisation de la dépendance en utilisant une copule Gaussienne mais notamment en incluant l'avantage de la régression probabiliste dans un cadre multivarié et/ou multi-site.

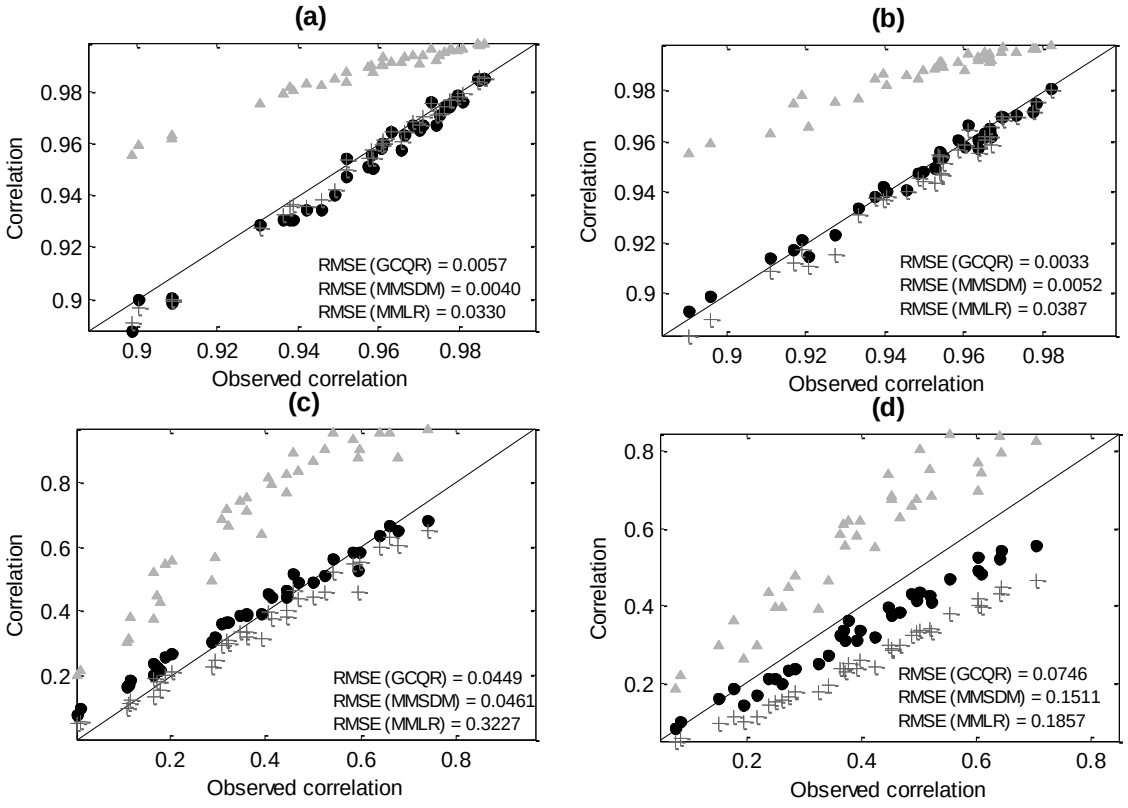

Figure 4. Nuage de points des corrélations observés et modélisés pour chaque paire de stations obtenue par le modèle GCQR (points noirs), le modèle MMLR (triangle gris) et le modèle MMSDM (de plus de gris) pour la température maximale (a), la température minimale (b), la quantité de précipitations (c) et l'occurrence des précipitations (d). Les valeurs des corrélations des modèles GCQR et MMSDM sont obtenues en utilisant la moyenne des valeurs de corrélation calculés à partir de 100 simulations.

Il est à noter également que le modèle MMSDM présente la difficulté de reproduire les corrélations entre les prédictants lorsque la quantité des précipitations est considérée. Ce résultat peut être expliqué par l'utilisation du processus de « probability mapping » qui a été employé pour corriger la distribution des précipitations. Cependant, il n'est pas nécessaire d'avoir recours à des mesures de transformation ou des procédures de correction de distribution lors de

l'évaluation du modèle de GCQR. En effet, la reproduction de la distribution conditionnelle est automatique à l'aide de sa composante de régression quantile.

Les modèles BMAR et QRMAR utilisent le champ Gaussien MAR comme composante aléatoire, dans le but de préserver les corrélations entre des séries décalées de 1 jour. La Figure 5.a présente les nuages de points pour des corrélations croisées des séries des précipitations non décalées pour le modèle BMAR, le modèle hybride traditionnel et le modèle MMLR au cours de la période de validation. La Figure 5.b illustre les nuages de points des corrélations des séries décalées de 1 jour. La Figure 5.a montre que le modèle BMAR et le modèle hybride traditionnel préservent convenablement les corrélations croisées des séries de précipitations non décalées. Cependant, en termes de *RMSE*, le modèle BMAR performe mieux que le modèle hybride traditionnel. À partir de la Figure 5.b on peut voir qu'en comparant avec le modèle hybride traditionnel, le modèle BMAR reproduit de manière plus adéquate les corrélations croisées des séries décalées de 1 jour. Ce résultat est bien confirmé par la valeur obtenue de la *RMSE*. En fait, par construction, le modèle hybride traditionnel est uniquement capable de tenir compte de l'autocorrélation dans une même série, contrairement au BMAR qui est censé préserver les corrélations croisées mêmes pour des séries de différentes stations.

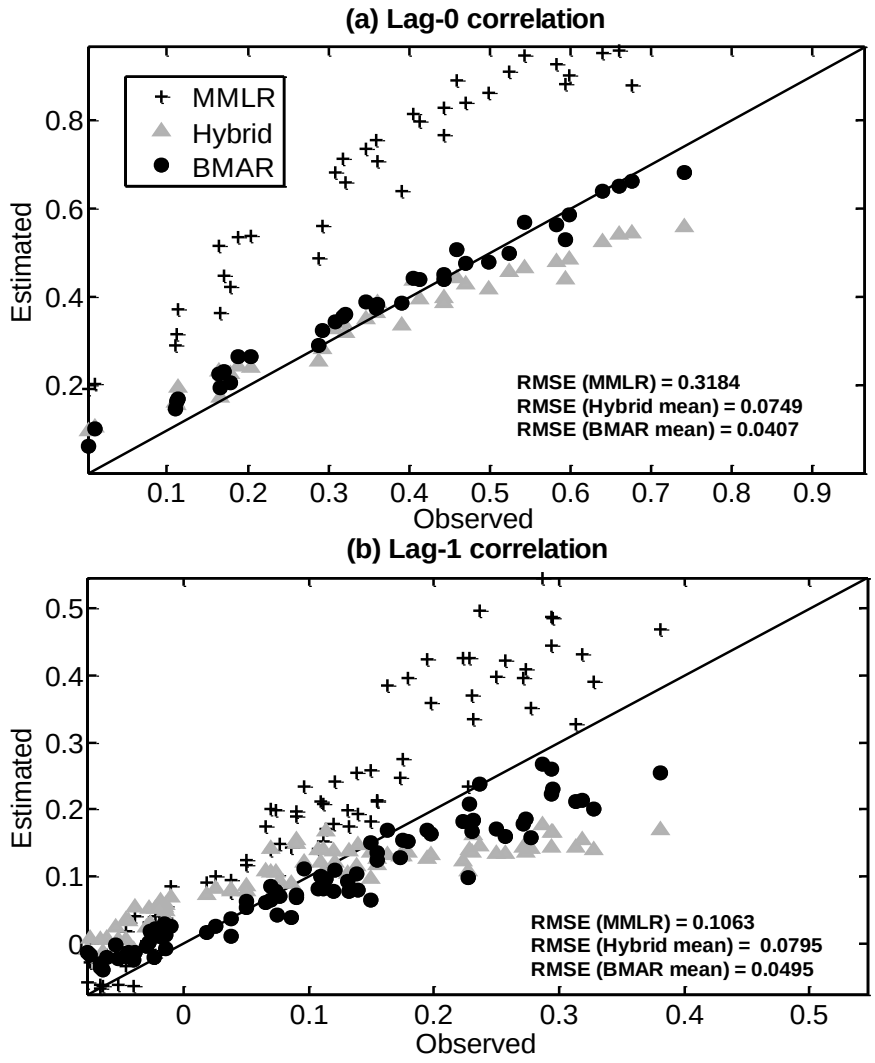

Figure 5. Nuage des points des corrélation observées et modélisées pour chaque paire de stations (a) et chaque paire de stations décalées de 1 jour (b) pour le modèle BMAR, le modèle hybride traditionnelle et le modèle MMLR au cours de la période de validation. Les valeurs de corrélation du modèle BMAR et du modèle hybride sont obtenues en utilisant la moyenne des valeurs de corrélation calculées à partir de 100 simulations.

La corrélation entre des paires de stations est souvent utilisée pour la spécification de modèles multi-sites de précipitation (c'est bien le cas de la copule Gaussienne). Cependant, la dynamique des événements de crue est fortement liée à l'apparition simultanée des précipitations extrêmes

sur plusieurs stations (Serinaldi et al. 2014). À cet égard, une vérification des propriétés multi-sites des précipitations extrêmes basée sur un ordre de corrélation supérieur à deux est nécessaire mais souvent ignoré. Dans ce contexte, Bárdossy and Pegram (2009) ont présenté l'entropie binaire en tant que mesure de dépendance des extrêmes de précipitation à un triplé de stations. Cette mesure permet de surmonter une validation seulement par paires de stations afin de rechercher les propriétés de dépendance d'ordre élevé. La théorie de l'entropie a été formulée par Shannon (1948) pour fournir une mesure de l'information contenue dans un ensemble de données. Pour calculer l'entropie binaire à un triplé de stations, un seuil de quantile est premièrement fixé pour définir des séries binaires de valeur égale à 0 si la valeur de précipitation est inférieure à ce seuil, et 1 dans le cas contraire. L'entropie binaire d'un triplé de stations est l'entropie de trois séries binaires obtenues après la fixation d'un seuil de quantile. Par conséquent, si l'association entre les variables à un seuil donné est forte, la quantité de l'information qui est contenue dans ces séries binaires sera plus faible et donc l'entropie H sera plus faible. Plus de détails sur le calcul de l'entropie binaire H est présenté dans Ben Alaya et al. (2016c).

Dans Ben Alaya et al. (2016c), l'échantillonnage non paramétrique basé sur les rangs a été utilisé comme une alternative aux copules. Pour évaluer le gain réel de cette procédure par rapport à la copule Gaussienne MAR, la Figure 6 montre les nuages des points entre l'entropie binaire observée et modélisée pour l'occurrence des précipitations et à trois seuils de quantile de non dépassement: 0.75, 0.90 et 0.95. Les points correspondent à toutes les combinaisons de triplés de stations.

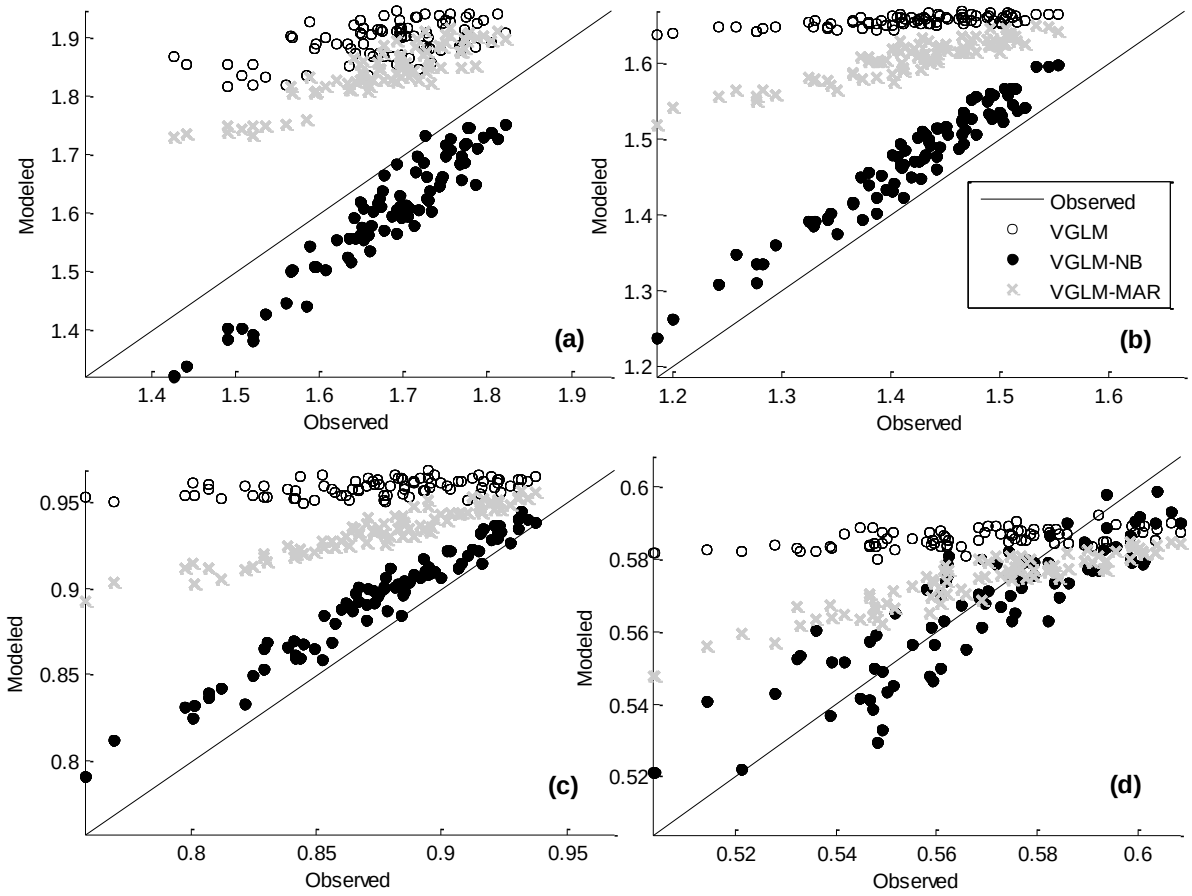

Figure 6. Nuage de points de l'entropie binaire observée et modélisée pour les occurrences de précipitation (a), et à trois seuils de quantiles: 0.75 (b), 0,90 (c) et 0.95 (d). Les points correspondent à toutes les combinaisons de triplés de stations.

On constate d'après la Figure 6a que l'occurrence de précipitations simulées en utilisant les deux modèles VGLM et VGLM-MAR (VGLM avec un champ Gaussien multivarié autorégressif) montrent des valeurs d'entropie binaire plus élevées que les valeurs observées. Des résultats similaires ont été obtenus pour l'entropie binaire correspondant aux seuils des quantiles 0.75, 0.90 et 0.95. Ce résultat indique que la structure de dépendance Gaussienne ne suffit pas pour capturer la forte association des précipitations extrêmes. Comparé au VGLM-MAR, le modèle VGLM-NB donne des valeurs d'entropie plus proches aux valeurs observées, indiquant que la simulation

en utilisant l'échantillonnage non-paramétrique est une amélioration par rapport au cadre Gaussien multivarié autorégressif. En réalité, ce résultat est prévu, du fait que le VGLM-MAR capte la structure spatiale en modélisant une combinaison de relations bivariées utilisant la copule Gaussienne.

Contrairement au modèle VGLM-MAR, une caractéristique intéressante du modèle VGLM-NB proposé est que les corrélations entre les paires de stations ne sont pas utilisées pour la définition du modèle. En effet, la méthode d'échantillonnage non paramétrique employée ne modélise pas les structures de dépendance, mais imite les rangs des données observées pour préserver les propriétés multi-sites inexpliquées par le VGLM. Comme c'est le cas pour la plupart des méthodes de ré-échantillonnage (Ouarda et al. 1997; Buishand et Brandsma 1999; Buishand et Brandsma 2001; Mehrotra et Sharma 2009; Lee et al. 2012), cette approche est guidée par les données de façon non-paramétrique et permet ainsi d'éviter toute erreur de spécification du modèle lors de la conservation des propriétés multi-sites. Cependant, tandis que les modèles de ré-échantillonnage traditionnels souffrent de l'incapacité à produire des valeurs qui sont plus extrêmes que celles observées, la composante de régression probabiliste de la structure hybride probabiliste proposée dans cette thèse permet de surmonter cet inconvénient.

5. Conclusions, contributions et perspectives

5.1. Conclusions

Cette étude a permis de développer et d'évaluer une nouvelle structure de modélisation hybride probabiliste pour la réduction d'échelle statistique des précipitations et des températures sur plusieurs stations. Le principal objectif de cette structure est de remédier aux inconvénients des

méthodes de régression classiques pour fournir la qualité de l'information requise principalement pour les études d'impact hydrologiques. Cette structure combine deux composantes: une composante de régression probabiliste et une composante aléatoire. La première permet de fournir à chaque étape de prédiction toute la distribution conditionnelle et permet ainsi de contourner les problèmes inhérents des méthodes de régression déterministe classique, à savoir : les sous-estimations de la variabilité, le problème de variance constante et la non normalité des prédictants. La composante aléatoire consiste à générer des valeurs corrélées et uniformément distribuées entre 0 et 1. Cette composante est adoptée pour compléter l'information sur les structures de dépendances spatio-temporelles des précipitations et des températures.

Pour les régressions probabilistes deux outils ont été considérés à savoir la régression des quantiles (QR) et la forme vectorielle des modèles linéaires généralisés (VGLM). Pour la composante aléatoire, trois outils ont été adoptés à savoir la copule Gaussienne, le champ Gaussien multivarié autorégressif (MAR), et l'échantillonnage non paramétrique (NB). En se basant sur cette structure de modélisation hybride probabiliste, cinq modèles de réduction d'échelle ont été développés qui sont: PGCR, BMAR, GCQR, VGLM-NB et QRMAR. La Figure 7 résume les différents outils de régression probabiliste et aléatoire qui ont été utilisées dans chaque modèle.

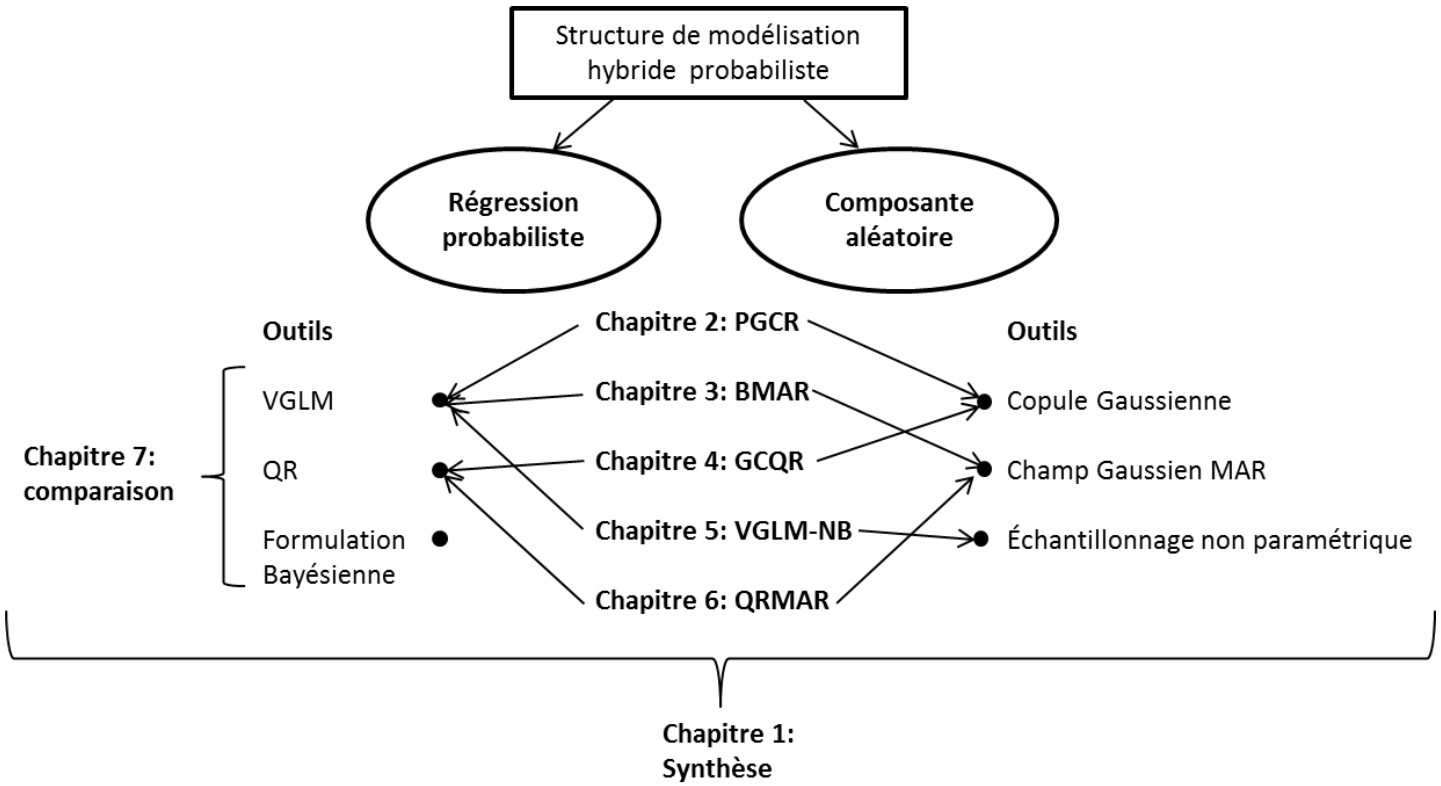

Figure 7. Résumé des différents chapitres de la thèse, et des différents modèles élaborés en utilisant la structure hybride probabiliste.

La structure de modélisation hybride probabiliste a été ensuite appliquée pour la réduction d'échelle des précipitations et des températures dans la partie sud de la province du Québec, Canada. Les données de ré-analyse NCEP-NCAR ont été utilisées dans cette thèse afin d'évaluer le potentiel de la structure de modélisation proposée, bien que l'objectif final est d'utiliser les prédicteurs AOGCM. Les modèles PGCR et GCQR ont été comparés avec les modèles MMLR et MMSDM qui sont des approches traditionnelles pour la réduction d'échelle des précipitations et des températures sur plusieurs stations. Les modèles BMAR et QRMAR ont été comparés avec le modèle hybride traditionnel de Jeong et al. (2012) pour la réduction d'échelle multi-site des précipitations. Les principaux résultats obtenus montrent que la structure de modélisation hybride

probabiliste est plus en mesure de reproduire les caractéristiques marginales des températures et des précipitations journalières. En outre, l'approche montre une meilleure préservation des propriétés spatio-temporelle des précipitations et des températures sur plusieurs stations.

5.2. Perspectives de recherches

Non linéarité: La dynamique du climat n'est pas linéaire (Von Storch et Zwiers 2001). Les composantes non linéaires de la partie hydrodynamique comprennent des termes d'advection importants. La partie thermodynamique contient divers autres processus non linéaires comme par exemple le processus de condensation. Pour en tenir compte, la composante de régression probabiliste peut être améliorée par l'utilisation d'autres outils de régression probabiliste non linéaires, notamment en utilisant le VGAM qui est une extension non linéaire du modèles VGLM, ainsi que les densités des réseaux de neurones conditionnelles qui sont une généralisation des modèles de régression des réseaux de neurones dans un cadre probabiliste.

Régionalisation: Dans un cadre de régionalisation où la température et les précipitations sur des sites non jaugés sont nécessaires, étendre la simulation à des endroits non jaugés peut être fait en deux étapes. Dans la première, les paramètres des modèles de régression probabiliste sont régionalisés, par exemple, en utilisant l'approche proposée par Reich (2012) pour la régression des quantiles ou en interpolant les paramètres des distributions conditionnelles dans le cas du VGLM. Le modèle résultant fournira la CDF à des endroits non jaugés pour chaque jour. Dans la seconde étape, à travers l'étape de simulation, des valeurs uniformément distribuées entre 0 et 1 et spatialement corrélées sont générées à des emplacements non jaugés. Ceci peut être réalisé en utilisant un modèle de krigeage basé sur les copules appliqué aux séries de variables cachées **U**.

Études de la sensibilité aux prédicteurs NCEP: Les données NCEP / NCAR sont utilisées pour la calibration et la validation de la structure hybride probabiliste. Même si les données NCEP sont complètes et physiquement compatibles, car elles sont essentiellement des interpolations de données d'observation basées sur un modèle dynamique, elles sont soumises à des biais de modélisation (Hofer et al. 2012). Les variables NCEP qui ne sont pas assimilables, mais générées par les paramétrages basés sur un modèle dynamique, peuvent différer sensiblement des conditions météorologiques réelles. L'utilisation de ces variables pour la calibration et la validation des techniques de réduction d'échelle empiriques peut induire un écart significatif des relations prédicteurs / prédictants modélisées par rapport à la réalité. Ainsi, ceci rend l'évaluation des techniques de réduction d'échelle plus difficile. De ce fait, la sélection des prédicteurs NCEP pour la calibration du modèle de réduction d'échelle exige un examen complet. De cette façon, l'étude de la sensibilité de la structure de modélisation hybride probabiliste aux prédicteurs NCEP est importante, non seulement pour une meilleure sélection des prédicteurs mais aussi pour une élaboration plus réaliste des scénarios climatiques futurs.

Combinaison avec les modèles dynamiques: La réduction d'échelle statistique a reçu une attention considérable de la part des statisticiens. Leurs contributions n'ont, cependant, pas été assez utilisées par la communauté climatique, bien qu'ils tentent de répondre aux besoins de l'utilisateur final. Cependant, l'étude du système climatique est, dans une large mesure, l'étude des statistiques de la météorologie; ainsi, il n'est pas surprenant que l'analyse, la modélisation et le raisonnement statistique sont omniprésents dans les sciences climatologiques. De ce point de vue, l'analyse statistique aide à quantifier les effets de l'incertitude, à la fois en termes de mesures d'observations qu'en fonction de notre compréhension des processus qui gouvernent la variabilité du climat. L'analyse statistique nous aide également à identifier lesquels des nombreux éléments

d'information provenant des observations du système climatique sont dignes de la synthèse et de l'interprétation (Zwiers et Von Storch 2004).

Les méthodes de réduction d'échelle dynamiques et les méthodes statistiques sont basées sur deux philosophies différentes. Considérant que les méthodes dynamiques résolvent les équations pour le vent local, la température et l'humidité à travers des formulations directes de tous les processus pertinents connus (dynamique et thermodynamique), les méthodes statistiques permettent l'utilisation d'informations à partir de données empiriques qui peuvent également intégrer des processus inconnus. Partant de ce point, combiner la structure de modélisation hybride probabiliste proposée avec des méthodes dynamiques peut maximiser l'utilité des informations sur le climat local et donc mener plus efficacement les études d'impacts, et en particulier les impacts hydrologiques.

Des scénarios futurs en utilisant les statistiques des sorties des modèles: dans cette thèse les différentes approches qui ont été développées utilisent des données NCEP qui représentent les conditions réelles du système climatique à grande échelle. L'établissement des relations statistiques à partir des données NCEP dans un problème de réduction d'échelle conduit à une modélisation du lien réel entre les conditions climatiques à grande échelle et les conditions à l'échelle locale. Ce cadre de réduction d'échelle est souvent désigné par la terminologie des prévisions parfaites (PP pour perfect prognosis). Notons que l'objectif de la thèse n'était pas d'élaborer des scénarios futurs mais plutôt de proposer de nouvelles approches de réduction d'échelle statistiques. Ainsi, les données NCEP ont été utilisées dans le but d'évaluer les approches proposées à modéliser le lien réel entre le climat à grande échelle et le climat à l'échelle locale. Cependant, pour simuler des scénarios futurs, les modèles de réduction d'échelle qui ont été calibrés en utilisant des données NCEP ne peuvent être appliqués que si la simulation

des modèles climatiques futures est parfaite. Ceci implique la nécessité de faire des corrections de biais des modèles climatiques globaux avant d'appliquer la réduction d'échelle. Pour contourner ce problème, la littérature sur l'étude d'impact s'est graduellement orientée vers l'utilisation des statistiques des sorties des modèles climatiques (MOS pour model output statistics). Dans ce cadre, le lien entre le climat à grande échelle et le climat à l'échelle locale est établi en utilisant comme prédicteurs les sorties des modèles climatiques. Ainsi, dans le but d'élaborer des simulations futures du climat, la structure hybride probabiliste proposée dans cette thèse peut notamment être utilisée dans un cadre de réduction d'échelle MOS. Ceci permet de mieux exploiter cette structure hybride pour faire des simulations futures sans avoir à imposer que les sorties des modèles climatiques globaux sont parfaites.

5.3. Originalités et contributions de l'étude

L'originalité de cette thèse de doctorat se manifeste à plusieurs niveaux. Si les méthodes de régression probabilistes et les approches par copules ont été séparément étudiées par plusieurs auteurs, la combinaison des deux outils dans une seule structure n'a pas encore été amplement explorée en hydrométéorologie et en climatologie, ce qui constitue la pierre angulaire de ce projet de recherche. Le but derrière le développement d'une structure hybride probabiliste réside dans l'intérêt à combiner ces deux outils pour profiter de leurs avantages dans un problème de réduction d'échelle de variables climatiques sur plusieurs stations. Dans Ben Alaya et al. (2014), le modèle PGCR qui découle de cette structure de modélisation en combinant le VGLM et la copule Gaussienne constitue donc une première application pour tester et évaluer cette structure de modélisation. Par la suite, quatre autres modèles qui découlent de cette structure ont été

développés en utilisant des outils statistiques différents en vue de montrer sa flexibilité et son efficacité.

Dans le cas des précipitations, son intermittence spatio-temporelle, son caractère discret-continu et sa distribution fortement asymétrique rendent sa modélisation beaucoup plus difficile que la modélisation de la température. À cet égard, une option généralement considérée consiste à l'intégration de deux modèles spatiaux, l'un pour le processus de l'occurrence (la transition entre les jours humides et secs) et l'autre pour le processus de quantité de précipitations (valeurs positives de précipitations dans les jours de pluie). Dans Ben Alaya et al. (2015a), le modèle BMAR constitue une tentative pour éviter la division entre les processus de l'occurrence et de la quantité. En effet, la distribution mixte discrète-continue Bernoulli-Pareto Généralisée a été considérée dans le modèle VGLM pour modéliser à la fois l'occurrence et la quantité des précipitations, tandis que dans la composante aléatoire le champ méta-Gaussien multivarié a été considéré pour reproduire l'intermittence spatio-temporelle. De cette manière, le nombre des variables dans la composante aléatoire à utiliser dans l'étape de simulation est réduit de deux (un pour le processus de l'occurrence et l'autre pour le processus de quantité) à un, ce qui rend la modélisation plus parcimonieuse.

Bien que l'application de la régression des quantiles dans la modélisation environnementale et l'évaluation de l'impact du changement climatique ait augmenté considérablement, cette technique n'a jamais été appliquée dans un contexte multi-site de réduction d'échelle. C'est dans cette optique que le modèle GCQR a été développé dans Ben Alaya et al. (2015b) et c'est ainsi que les deux outils, (i) régression des quantiles et (ii) copule Gaussienne, ont été combinés pour la première fois dans un contexte de réduction d'échelle multi-site multivarié.

Vue les exigences de plus en plus strictes afin d'assurer une meilleure modélisation des structures de dépendances complexes des précipitations extrêmes, plusieurs auteurs ont souligné la nécessité d'utiliser des copules non Gaussiennes et flexibles pour remédier à ce point. Du fait que plusieurs modèles théoriques de copules existent dans la littérature statistique et qui n'ont pas encore été explorés dans la littérature hydrométéorologique, la tendance dans la littérature s'oriente à chercher une réponse à la question: quelle copule devrait-on choisir que ce soit pour simuler ou bien pour modéliser? Bien que des copules flexibles puissent résoudre ce problème, leur utilisation n'est pas toujours nécessaire. De ce point de vue, l'utilité des copules a été révisée dans Ben Alaya et al. (2016c) pour revenir à la question clé: comment générer des valeurs uniformément distribuées entre 0 et 1 pour reproduire des structures de dépendance complexes. C'est ainsi que la méthode d'échantillonnage basée sur les rangs a été proposée dans Ben Alaya et al. (2016c). Cette méthode facile à implémenter et ne modélise pas la dépendance, mais imitent les caractéristiques historiques observées et permet ainsi d'éviter toute erreur de spécification du modèle. Cependant, il convient de mentionner que cette approche ne peut pas remplacer les copules, lorsque la modélisation des structures de dépendance est nécessaire.

La structure de modélisation hybride probabiliste est mathématiquement riche, car elle offre un moyen simple pour simuler l'évolution spatio-temporelle de variables discrètes-continues. En outre, elle peut être adaptée à différents domaines ainsi que l'introduction de forçage par des covariables exogènes. Cette structure de modélisation représente donc un outil mathématique précieux dans la recherche de l'hydrométéorologie et les analyses climatiques où des variables aléatoires non normalement distribuées, comme les précipitations, la vitesse du vent, la couverture nuageuse et l'humidité, sont souvent impliquées.

6. Références bibliographiques

AghaKouchak, A. (2014). "Entropy–copula in hydrology and climatology." Journal of Hydrometeorology **15**(6): 2176-2189.

AghaKouchak, A., A. Bárdossy and E. Habib (2010). "Conditional simulation of remotely sensed rainfall data using a non-Gaussian v-transformed copula." Advances in Water Resources **33**(6): 624-634.

Ashkar, F. and T. B. Ouarda (1996). "On some methods of fitting the generalized Pareto distribution." Journal of Hydrology **177**(1): 117-141.

Bárdossy, A. (2006). "Copula-based geostatistical models for groundwater quality parameters." Water Resour. Res. **42**(11): W11416.

Bárdossy, A. and G. G. S. Pegram (2009). "Copula based multisite model for daily precipitation simulation." Hydrology and Earth System Sciences **13**(12): 2299-2314.

Bargaoui, Z. K. and A. Bárdossy (2015). "Modeling short duration extreme precipitation patterns using copula and generalized maximum pseudo-likelihood estimation with censoring." Advances in Water Resources **84**: 1-13.

Baur, D., M. Saisana and N. Schulze (2004). "Modelling the effects of meteorological variables on ozone concentration—a quantile regression approach." Atmospheric Environment **38**(28): 4689-4699.

Ben Alaya, M. A., F. Chebana and T. Ouarda (2014). "Probabilistic Gaussian Copula Regression Model for Multisite and Multivariable Downscaling." Journal of Climate **27**(9).

Ben Alaya, M. A., F. Chebana and T. Ouarda (2016a). "Quantile regression multivariate autoregressive model for downscaling multisite daily precipitations." To be prepared and submitted.

Ben Alaya, M. A., F. Chebana and T. B. Ouarda (2015a). "Probabilistic Multisite Statistical Downscaling for Daily Precipitation Using a Bernoulli–Generalized Pareto Multivariate Autoregressive Model." Journal of climate **28**(6): 2349-2364.

Ben Alaya, M. A., F. Chebana and T. B. M. J. Ouarda (2015b). "Multisite and multivariable statistical downscaling using a Gaussian copula quantile regression model." Climate Dynamics: 1-15.

Ben Alaya, M. A., D. Fasbender, T. Ouarda and F. Chebana (2016b). "Application of spatial Bayesian model for downscaling daily temperatures and comparison with two probabilistic regression approaches." Submitted.

Ben Alaya, M. A., T. Ouarda and F. Chebana (2016c). "Non-Gaussian spatiotemporal simulation of multisite daily precipitations: a downscaling framework." Submitted.

Benestad, R. E., I. Hanssen-Bauer and D. Chen (2008). Empirical-statistical downscaling, World Scientific.

Bremnes, J. B. (2004a). "Probabilistic forecasts of precipitation in terms of quantiles using NWP model output." Monthly Weather Review **132**(1).

Bremnes, J. B. (2004b). "Probabilistic wind power forecasts using local quantile regression." Wind Energy **7**(1): 47-54.

Buishand, T. A. and T. Brandsma (1999). "Dependence of precipitation on temperature at Florence and Livorno (Italy)." Climate Research **12**(1): 53-63.

Buishand, T. A. and T. Brandsma (2001). "Multisite simulation of daily precipitation and temperature in the Rhine basin by nearest-neighbor resampling." Water Resources Research **37**(11): 2761-2776.

Burger, G. and Y. Chen (2005). "A Regression-based downscaling of spatial variability for hydrologic applications." Journal of Hydrology **311**: 299-317.

Cannon, A. J. (2008). "Probabilistic multisite precipitation downscaling by an expanded Bernoulli-gamma density network." Journal of Hydrometeorology **9**(6): 1284-1300.

Cannon, A. J. (2011). "Quantile regression neural networks: Implementation in R and application to precipitation downscaling." Computers & Geosciences **37**(9): 1277-1284.

Cawley, G. C., G. J. Janacek, M. R. Haylock and S. R. Dorling (2007). "Predictive uncertainty in environmental modelling." Neural Networks **20**(4): 537-549.

Chandler, R. E. and H. S. Wheater (2002). "Analysis of rainfall variability using generalized linear models: a case study from the west of Ireland." Water Resources Research **38**(10): 10-11-10-11.

Chebana, F. and T. B. Ouarda (2011). "Multivariate quantiles in hydrological frequency analysis." Environmetrics **22**(1): 63-78.

Chebana, F. and T. B. M. J. Ouarda (2007). "Multivariate L-moment homogeneity test." Water Resources Research **43**(8).

Clark, M., S. Gangopadhyay, L. Hay, B. Rajagopalan and R. Wilby (2004). "The Schaake Shuffle: A Method for Reconstructing Space–Time Variability in Forecasted Precipitation and Temperature Fields." Journal of Hydrometeorology **5**(1): 243-262.

Coe, R. and R. Stern (1982). "Fitting models to daily rainfall data." Journal of Applied Meteorology **21**(7): 1024-1031.

De Michele, C. and G. Salvadori (2003). "A Generalized Pareto intensity-duration model of storm rainfall exploiting 2-Copulas." Journal of Geophysical Research D: Atmospheres **108**(2): ACL 15-11 ACL 15-11.

Dupuis, D. J. (2007). "Using copulas in hydrology: Benefits, cautions, and issues." Journal of Hydrologic Engineering **12**(4): 381-393.

El Adlouni, S., B. Bobée and T. Ouarda (2008). "On the tails of extreme event distributions in hydrology." Journal of Hydrology **355**(1): 16-33.

El Adlouni, S. and T. Ouarda (2008). "Study of the joint distribution flow-level by copulas: Case of the Chateauguay river." Canadian Journal of Civil Engineering **35**(10): 1128-1137.

Elsner, J. B., J. P. Kossin and T. H. Jagger (2008). "The increasing intensity of the strongest tropical cyclones." Nature **455**(7209): 92-95.

Fang, H.-B., K.-T. Fang and S. Kotz (2002). "The meta-elliptical distributions with given marginals." Journal of Multivariate Analysis **82**(1): 1-16.

Fasbender, D. and T. B. M. J. Ouarda (2010). "Spatial Bayesian Model for Statistical Downscaling of AOGCM to Minimum and Maximum Daily Temperatures." Journal of Climate **23**(19): 5222-5242.

Friederichs, P. and A. Hense (2007). "Statistical Downscaling of Extreme Precipitation Events Using Censored Quantile Regression." Monthly Weather Review **135**(6): 2365-2378.

Gachon, P., A. St-Hilaire, T. B. M. J. Ouarda, V. Nguyen, C. Lin, J. Milton, D. Chaumont, J. Goldstein, M. Hessami, T. D. Nguyen, F. Selva, M. Nadeau, P. Roy, D. Parishkura, N. Major, M. Choux and A. Bourque (2005). "A first evaluation of the strength and weaknesses of statistical downscaling methods for simulating extremes over various regions of eastern Canada " Sub-component, Climate Change Action Fund (CCAF), Environment Canada **Final report**(Montréal, Québec, Canada): 209.

Genest, C. and F. Chebana (2015). "Copula modeling in hydrologic frequency analysis." In Handbook of Applied Hydrology (V.P. Singh, Editor) **McGraw-Hill, New York,** (in press).

Genest, C. and J. F. Plante (2003). "On Blest's measure of rank correlation." Canadian Journal of Statistics **31**(1): 35-52.

Giorgi, F., J. Christensen, M. Hulme, H. Von Storch, P. Whetton, R. Jones, L. Mearns, C. Fu, R. Arritt and B. Bates (2001). "Regional climate information-evaluation and projections." Climate Change 2001: The Scientific Basis. Contribution of Working Group to the Third Assessment Report of the Intergouvernmental Panel on Climate Change [Houghton, JT et al.(eds)]. Cambridge University Press, Cambridge, United Kongdom and New York, US.

Gräler, B. (2014). "Modelling skewed spatial random fields through the spatial vine copula." Spatial Statistics **10**: 87-102.

Guerfi, N., A. A. Assani, M. Mesfioui and C. Kinnard (2015). "Comparison of the temporal variability of winter daily extreme temperatures and precipitations in southern Quebec (Canada) using the Lombard and copula methods." International Journal of Climatology.

Hammami, D., T. S. Lee, T. B. M. J. Ouarda and J. Le (2012). "Predictor selection for downscaling GCM data with LASSO." Journal of Geophysical Research D: Atmospheres **117**(17).

Harpham, C. and R. L. Wilby (2005). "Multi-site downscaling of heavy daily precipitation occurrence and amounts." Journal of Hydrology **312**(1): 235-255.

Haylock, M. R., G. C. Cawley, C. Harpham, R. L. Wilby and C. M. Goodess (2006). "Downscaling heavy precipitation over the United Kingdom: A comparison of dynamical and statistical methods and their future scenarios." International Journal of Climatology **26**(10): 1397-1415.

Hobæk Haff, I., A. Frigessi and D. Maraun (2015). "How well do regional climate models simulate the spatial dependence of precipitation? An application of pair-copula constructions." Journal of Geophysical Research: Atmospheres **120**(7): 2624-2646.

Hofer, M., B. Marzeion and T. Mölg (2012). "Comparing the skill of different reanalyses and their ensembles as predictors for daily air temperature on a glaciated mountain (Peru)." Climate Dynamics **39**(7-8): 1969-1980.

Huth, R. (1999). "Statistical downscaling in central Europe: Evaluation of methods and potential predictors." Climate Research **13**(2): 91-101.

Huth, R. (2004). Sensitivity of local daily temperature change estimates to the selection of downscaling models and predictors. Boston, MA, ETATS-UNIS, American Meteorological Society.

Huth, R. and L. Pokorná (2004). "Parametric versus non-parametric estimates of climatic trends." Theoretical and Applied Climatology **77**(1): 107-112.

Jagger, T. H. and J. B. Elsner (2009). "Modeling tropical cyclone intensity with quantile regression." International Journal of Climatology **29**(10): 1351.

Jeong, D., A. St-Hilaire, T. Ouarda and P. Gachon (2013). "A multivariate multi-site statistical downscaling model for daily maximum and minimum temperatures." Climate Research **54**(2): 129-148.

Jeong, D. I., A. St-Hilaire, T. B. M. J. Ouarda and P. Gachon (2012). "Multisite statistical downscaling model for daily precipitation combined by multivariate multiple linear regression and stochastic weather generator." Climatic Change **114**(3-4): 567-591.

Joe, H. (1997). Multivariate models and multivariate dependence concepts, CRC Press.

Kazianka, H. and J. Pilz (2010). "Copula-based geostatistical modeling of continuous and discrete data including covariates." Stochastic Environmental Research and Risk Assessment **24**(5): 661-673.

Khalili, M., V. T. Van Nguyen and P. Gachon (2013). "A statistical approach to multi-site multivariate downscaling of daily extreme temperature series." International Journal of Climatology **33**(1): 15-32.

Kleiber, W., R. W. Katz and B. Rajagopalan (2012). "Daily spatiotemporal precipitation simulation using latent and transformed Gaussian processes." Water Resources Research **48**(1).

Koenker, R. and G. Bassett (1978). "Regression quantiles." Econometrica: journal of the Econometric Society: 33-50.

Lee, T., R. Modarres and T. Ouarda (2013). "Data-based analysis of bivariate copula tail dependence for drought duration and severity." Hydrological Processes **27**(10): 1454-1463.

Lee, T., T. B. Ouarda and C. Jeong (2012). "Nonparametric multivariate weather generator and an extreme value theory for bandwidth selection." Journal of Hydrology **452**: 161-171.

Li, C., V. P. Singh and A. K. Mishra (2013a). "A bivariate mixed distribution with a heavy-tailed component and its application to single-site daily rainfall simulation." Water Resources Research **49**(2): 767-789.

Li, C., V. P. Singh and A. K. Mishra (2013b). "Monthly river flow simulation with a joint conditional density estimation network." Water Resources Research **49**(6): 3229-3242.

Luce, C. H. and Z. A. Holden (2009). "Declining annual streamflow distributions in the Pacific Northwest United States, 1948–2006." Geophysical Research Letters **36**(16).

Mao, G., S. Vogl, P. Laux, S. Wagner and H. Kunstmann (2015). "Stochastic bias correction of dynamically downscaled precipitation fields for Germany through Copula-based integration of gridded observation data." Hydrology and Earth System Sciences **19**(4): 1787-1806.

Mehrotra, R. and A. Sharma (2009). "Evaluating spatio-temporal representations in daily rainfall sequences from three stochastic multi-site weather generation approaches." Advances in Water Resources **32**(6): 948-962.

Nelsen, R. B. (2013). An introduction to copulas, Springer Science & Business Media.

Oakes, D. (1982). "A model for association in bivariate survival data." Journal of the Royal Statistical Society. Series B (Methodological): 414-422.

Ouarda, T. B. M. J., J. W. Labadie and D. G. Fontaine (1997). "Indexed sequential hydrologic modeling for hydropower capacity estimation." Journal of the American Water Resources Association **33**(6): 1337-1349.

Palutikof, J. P., C. M. Goodess, S. J. Watkins and T. Holt (2002). "Generating rainfall and temperature scenarios at multiple sites: Examples from the Mediterranean." Journal of Climate **15**(24): 3529-3548.

Qian, B., J. Corte-Real and H. Xu (2002). "Multisite stochastic weather models for impact studies." International Journal of Climatology **22**(11): 1377-1397.

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.

Salvadori, G. and C. De Michele (2007). "On the use of copulas in hydrology: theory and practice." Journal of Hydrologic Engineering **12**(4): 369-380.

Sankarasubramanian, A. and U. Lall (2003). "Flood quantiles in a changing climate: Seasonal forecasts and causal relations." Water Resources Research **39**(5).

Schölzel, C. and P. Friederichs (2008). "Multivariate non-normally distributed random variables in climate research - Introduction to the copula approach." Nonlinear Processes in Geophysics **15**(5): 761-772.

Schoof, J. T. and S. C. Pryor (2001). "Downscaling temperature and precipitation: A comparison of regression-based methods and artificial neural networks." International Journal of Climatology **21**(7): 773-790.

Serinaldi, F., A. Bárdossy and C. G. Kilsby (2014). "Upper tail dependence in rainfall extremes: would we know it if we saw it?" Stochastic Environmental Research and Risk Assessment **29**(4): 1211-1233.

Serinaldi, F. and C. G. Kilsby (2014). "Simulating daily rainfall fields over large areas for collective risk estimation." Journal of Hydrology **512**: 285-302.

Shannon, C. (1948). "A mathematical theory of communication." Bell Syst Tech J **27**(3): 379–423.

Sklar, A. (1959). Fonctions de répartition à n dimensions et leurs marges.

Song, S. and V. P. Singh (2010). "Meta-elliptical copulas for drought frequency analysis of periodic hydrologic data." Stochastic Environmental Research and Risk Assessment **24**(3): 425-444.

Srivastav, R. K. and S. P. Simonovic (2014). "Multi-site, multivariate weather generator using maximum entropy bootstrap." Climate Dynamics **44**(11-12): 3431-3448.

Stephenson, D. B., K. Rupa Kumar, F. J. Doblas-Reyes, J. F. Royer, F. Chauvin and S. Pezzulli (1999). "Extreme daily rainfall events and their impact on ensemble forecasts of the Indian monsoon." Monthly Weather Review **127**(9): 1954-1966.

Stern, R. and R. Coe (1984). "A model fitting analysis of daily rainfall data." Journal of the Royal Statistical Society. Series A (General): 1-34.

Tareghian, R. and P. F. Rasmussen (2013). "Statistical downscaling of precipitation using quantile regression." Journal of Hydrology **487**: 122-135.

Vaz de Melo Mendes, B. and R. P. C. Leal (2010). "Portfolio management with semi-parametric bootstrapping." Journal of Risk Management in Financial Institutions **3**(2): 174-183.

Vernieuwe, H., S. Vandenberghe, B. De Baets and N. E. Verhoest (2015). "A continuous rainfall model based on vine copulas." Hydrology and Earth System Sciences Discussions **12**(1): 489-524.

Vinod, H. D. and J. López-de-Lacalle (2009). "Maximum entropy bootstrap for time series: the meboot R package." Journal of Statistical Software **29**(5): 1-19.

Von Storch, H. (1999). "On the Use of "Inflation" in Statistical Downscaling." Journal of Climate **12**(12): 3505-3506.

Von Storch, H. and F. W. Zwiers (2001). Statistical analysis in climate research, Cambridge university press.

Weerts, A., H. Winsemius and J. Verkade (2011). "Estimation of predictive hydrological uncertainty using quantile regression: examples from the National Flood Forecasting System (England and Wales)." Hydrology and Earth System Sciences, 15,(1).

Widmann, M., C. S. Bretherton and E. P. Salathé Jr (2003). "Statistical precipitation downscaling over the northwestern united states using numerically simulated precipitation as a predictor." Journal of Climate **16**(5): 799-816.

Wilby, R. L., S. P. Charles, E. Zorita, B. Timbal, P. Whetton and L. O. Mearns (2004). Guidelines for use of climate scenarios developed from statistical downscaling methods. Supporting material of the Intergovernmental Panel on Climate Change (IPCC), prepared on behalf of Task Group on Data and Scenario Support for Impacts and Climate Analysis (TGICA).

Wilby, R. L., C. W. Dawson and E. M. Barrow (2002). "sdsm - a decision support tool for the assessment of regional climate change impacts." Environmental Modelling and Software **17**(2): 145-157.

Wilby, R. L. and T. M. L. Wigley (1997). "Downscaling general circulation model output: a review of methods and limitations." Progress in Physical Geography **21**(4): 530-548.

Wilks, D. (1998). "Multisite generalization of a daily stochastic precipitation generation model." Journal of Hydrology **210**(1): 178-191.

Wilks, D. S. (1999). "Multisite downscaling of daily precipitation with a stochastic weather generator." Climate Research **11**: 125-136.

Wilks, D. S. and R. L. Wilby (1999). "The weather generation game: a review of stochastic weather models." Progress in Physical Geography **23**(3): 329-357.

Williams, P. M. (1998). "Modelling seasonality and trends in daily rainfall data." Advances in neural information processing systems: 985-991.

Xu, C.-y. (1999). "From GCMs to river flow: a review of downscaling methods and hydrologic modelling approaches." Progress in Physical Geography **23**(2): 229-249.

Yang, C., R. E. Chandler, V. S. Isham and H. S. Wheater (2005). "Spatial-temporal rainfall simulation using generalized linear models." Water Resources Research **41**(11): 1-13.

Yee, T. W. and A. G. Stephenson (2007). "Vector generalized linear and additive extreme value models." Extremes **10**(1-2): 1-19.

Yee, T. W. and C. Wild (1996). "Vector generalized additive models." Journal of the Royal Statistical Society. Series B (Methodological): 481-493.

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

Zwiers, F. W. and H. Von Storch (2004). "On the role of statistics in climate research." International Journal of Climatology **24**(6): 665-680.

CHAPITRE 2: PROBABILISTIC GAUSSIAN COPULA REGRESSION MODEL FOR MULTISITE AND MULTIVARIABLE DOWNSCALING

Probabilistic Gaussian Copula Regression model for multisite and multivariable downscaling

M. A. Ben Alaya[1], F. Chebana[1] and T.B.M.J. Ouarda[2, 1]

[1]INRS-ETE, 490 rue de la Couronne, Québec (QC), Canada G1K 9A9
[2] Institute Center for Water and Environment (iWater), Masdar Institute of science and technology
P.O. Box 54224, Abu Dhabi, UAE

[*]**Corresponding author:** Tel: +1 (418) 654 2530#4468
Email: mohammed_ali.ben_alaya@ete.inrs.ca

Accepted December 9, 2013

(Journal of climate)

Abstract

Atmosphere–ocean general circulation models (AOGCMs) are useful to simulate large-scale climate evolutions. However, AOGCM data resolution is too coarse for regional and local climate studies. Downscaling techniques have been developed to refine AOGCM data and provide information at more relevant scales. Among a wide range of available approaches, regression-based methods are commonly used for downscaling AOGCM data. When several variables are considered at multiple sites, regression models are employed to reproduce the observed climate characteristics at small scale, such as the variability and the relationship between sites and variables.

This study introduces a probabilistic Gaussian copula regression (PGCR) model for simultaneously downscaling multiple variables at several sites. The proposed PGCR model relies on a probabilistic framework to specify the marginal distribution for each downscaled variable at a given day through AOGCM predictors, and handles multivariate dependence between sites and variables using a Gaussian copula. The proposed model is applied for the downscaling of AOGCM data to daily precipitation and minimum and maximum temperatures in the southern part of Quebec, Canada. Reanalysis products are used in this study to assess the potential of the proposed method. Results of the study indicate the superiority of the proposed model over classical regression-based methods and a multivariate multisite statistical downscaling model.

Keywords: Downscaling, Gaussian copula, Probabilistic regression, Temperature, Precipitation, Multisite, Multivariable.

1. Introduction

Atmosphere-ocean general circulation models (AOGCM) (Atmosphere-Ocean General Circulation Models) are commonly used to simulate large-scale climate evolution. Information provided by these models is widely used to produce future climate projections. AOGCM data are generally produced on regular grids with a low horizontal resolution around 2.5° longitude and latitude (approximately 250 to 300 km). However, this resolution is coarse for regional and local climate studies. Downscaling techniques have been developed to refine AOGCM data and provide information at more relevant scales. These techniques can be classified into two main categories: dynamic methods and statistical methods (Wilby and Wigley 1997; Herrera et al. 2006). Dynamic methods use regional climate models (RCM), which have the same basic principles as AOGCM, with a high resolution between 25 and 50 km. RCMs only cover a limited portion of the globe. Dynamic methods require large computational capabilities and substantial human resources. Statistical methods, on the other hand, consider statistical relationships between large-scale variables (predictors) and small-scale variables (predictands). The main advantages of these methods are their simplicity and their low computational costs. Thus they represent a good alternative to dynamic methods in the case of limited resources.

The performance of a statistical downscaling model depends on its ability to reproduce the observed statistical characteristics of local climate (Wilby 1998; Wilby et al. 2002; Gachon et al. 2005; Hessami et al. 2008). Statistical downscaling methods are required to take into account the climatic characteristics of predictands, such as their observed variability, in order to provide reliable meteorological information at the local scale (Wilby and Wigley 1997). The proper reproduction of the variability in downscaling applications is a very important issue, since a poor

representation of the variability could lead to a poor representation of extreme events. In addition, if data are required at multiple stations, then a spatial or multisite model is employed to better represent the observed correlation between sites and predictands.

Once a downscaling model has been developed, downscaled data can be entered as input into an environmental model, for example, a hydrological model for streamflow in a watershed that requires climate information at a finer scale (Cannon 2008). Precipitation and temperature are commonly considered as predictands in a downscaling problem. In hydrology, streamflows depend strongly on the spatial distribution of precipitation in a watershed, and on the interactions between temperature and precipitation which determines whether precipitation falls as rain or snow (Lindström et al. 1997). Therefore, maintaining realistic relationships between sites and variables in downscaled results is particularly important for a number of applications such as hydrological modelling.

Statistical downscaling techniques can be grouped into three main approaches: stochastic weather generators (Wilks and Wilby 1999), weather typing (Conway et al. 1996) and regression methods (Wilby et al. 2002). Regression based methods are commonly used for downscaling AOGCM data. They can find a direct relationship between large-scale predictors and local predictands. A variety of regression methods have been used in the literature for downscaling purposes, such as: multiple linear regression (MLR) (Wilby 1998; Hellström et al. 2001; Huth 2004; Hessami et al. 2008; Jeong et al. 2012), principal component analysis (PCA) with MLR (Huth 2004), canonical correlation analysis (CCA) combined to MLR (Huth 2002; Palutikof et al. 2002), artificial neural networks (ANN) (Schoof and Pryor 2001; Cannon 2008; Cannon 2011) and singular value decomposition (SVD) (Widmann et al. 2003). Regression-based methods usually perform well

for downscaling purposes, but their major drawback is that they generally provide only the mean or the central part of the predictands (Cawley et al. 2007). Therefore the variance of the modeled mean will typically be smaller than the variance of the observed series. The reason is in part that regression models cannot represent the influence of small scale phenomena. On the other hand, stochastic weather generators use a random number conditioned upon large-scale model output state and can be applied at a single site as well as multisite (Wilks 1998, 1999; Qian et al. 2002; Palutikof et al. 2002). However, the main limitation of weather generators is the difficulty in adjusting the parameters in a physically realistic and consistent manner under future climate states (Wilby et al. 1998).

To correctly estimate the temporal variance of downscaled data series, three main approaches have been proposed in the literature: inflation (Huth 1999), randomization (Von Storch 1999; Clark et al. 2004) and expansion (Burger and Chen 2005). Inflation is usually carried out by multiplying the downscaled data by a constant factor, randomization consists in adding a random noise and in the expansion approach, the predicted variances are constrained to match the variance of the observed data by adding a constraint to the regression cost function. A problem with the inflation approach is that the spatial correlations between sites can be misrepresented. However randomization can also be applied in a multisite downscaling framework. In a randomization procedure, regression and stochastic weather generator approaches can be combined in a single hybrid model. Thus, the resulting stochastic hybrid model can overcome weaknesses of both approaches. For instance, Jeong et al. (2012) employed a randomization procedure to reproduce the cross-site correlation of precipitation occurrence and amount among the observation sites using the multivariate normal distribution. As well, Jeong et al. (2013) proposed a multivariate multi-site statistical downscaling model (MMSDM) for simultaneous

downscaling of climate variables including daily maximum and minimum temperatures for multiple observation sites. The MMSDM employs multivariate multiple linear regression (MMLR) to simulate deterministic series from large-scale reanalysis data and ads spatially correlated random series to the deterministic series of the MMLR to complement the underestimated variance and to reproduce the spatial correlation of variables from multiple sites and an at-site temporal correlation between variables. In the same way, Khalili et al. (2013) proposed a hybrid approach by combining a linear regression component with a stochastic component based on a spatial moving average process to reproduce the observed spatial dependence between extremes temperatures at different sites. However, Burger and Chen (2005) indicated that randomization in a hybrid approach is based on a static noise model and thus it failed to represent local changes in atmospheric variability in a climate change simulation, which is well explained by expended downscaling. The expanded downscaling can be applied to multiple predictands by constraining the covariance matrix of the predicted series to be equal to the observed covariance matrix. For example, Cannon (2009) introduced a multivariate ridge regression with negative ridge parameters for improving the covariance structure of multivariate linear downscaling models. This procedure is conceptually similar to expanded downscaling since both force the covariance structure of the predictions to match that of the observations. Thereby, the observed variability is reproduced using a single deterministic regression component, and thus, the reproduced variability is not static and may change in a climate change simulation. On the other hand, Von Storch (1999) suggested that the inflation and expansion approaches are inappropriate techniques, because the implicit assumption that all local variability can be traced back to the large-scale using a deterministic regression model is improper and is not

the case in reality. For this reason, Von Storch (1999) suggests using randomisation approaches that are more realistic than expanded approaches.

As an alternative to the three existing techniques for specifying predictand variance, the variability of predictand can also be captured by modeling the whole distribution. In this regard, probabilistic approaches have provided significant contributions in downscaling applications (e.g. Bates et al. 1998; Hughes et al. 1999; Bellone et al. 2000; Vrac and Naveau 2007; Cannon 2008; Fasbender and Ouarda 2010). They allow reproducing the whole distribution by modeling the effect of the AOGCM predictors on the parameters of the predictand distributions. For example, Williams (1998) employs an artificial neuronal network (ANN) to model parameters of a mixed Bernoulli-gamma distribution for precipitation at a single site. For this purpose, the problem that arises is how to extend probabilistic approaches in multisite downscaling tasks. In this context, Cannon (2008) applied the principles of expanded downscaling in a probabilistic modeling framework to allow for a realistic representation of spatial relationships between precipitation at multiple sites. On the other hand, as indicated previously, expanded approaches are not appropriate in a climatic downscaling problem.

Given the disadvantages of these three existing techniques for specifying predictand variances and covariance structure, we propose in this paper a copula based approach as an alternative solution to extend the probabilistic modeling framework in multisite downscaling tasks. The methodology proposed here presents another advantage in that multiple climatic variables are downscaled at multiple sites simultaneously and consistently to produce realistic relationships between both sites and variables.

Multivariate dependence structures can be modeled using classical distributions such as the multivariate normal distribution. However, the multivariate normal approach cannot adequately reproduce the dependence structure of hydro-meteorological data when existing asymmetries are significant such as for precipitation variables. To address this limitation, the use of copula-based approaches can be beneficial since copula functions are more flexible and may be better suited to the data (e.g. Schölzel and Friederichs 2008). Copulas have recently become very popular, especially in fields like econometrics, finance, risk management, and insurance. In recent years, the application of copulas has also made significant contributions in the field of hydrometeorology. Schölzel and Friederichs (2008) provide a brief overview of copulas for applications in meteorology and climate research. Models based on copulas have been introduced for multivariate hydrological frequency analysis (Chebana and Ouarda 2007; El Adlouni and Ouarda 2008), risk assessment, geostatistical interpolation and multivariate extreme values (e.g. De Michele and Salvadori 2003; Bárdossy 2006; Renard and Lang 2007; Kazianka and Pilz 2010) . In addition, copulas allow describing the dependence structure independently from the marginal distributions, and thus, using different marginal distributions at the same time without any transformations (Sklar 1959; Dupuis 2007).

The goal of the present study is to develop and test a probabilistic Gaussian copula regression (PGCR) model for multisite and multivariable downscaling. The model can reproduce observed spatial relationships between sites and variables and specify at each site and for each day, the conditional distributions of each variables. To this end, PGCR uses a probabilistic framework to address the limitations of regression-based approaches, namely (i) the poor representation of extreme events, (ii) the poor representation of observed variability and (iii) the assumption of normality of data. The PGCR model specifies a marginal distribution for each predictand through

AOGCM predictors as well a Gaussian copula to handle multivariate dependence between margins. The proposed model can be considered as a hybrid approach combining a probabilistic regression based downscaling model with a stochastic weather generator component. The main advantage of this model compared to conventional hybrid approaches is that the temporal variability may change in future climate simulations.

The present paper is structured as follows: After a presentation of the multiple multivariate linear regressions (MMLR) model and the multivariate multisite statistical downscaling model (MMSDM) as a classical regression based method in statistical downscaling, the proposed PGCR model is presented. The PGCR model is then applied to the case of daily precipitations and maximum and minimum temperatures in the southern part of the province of Quebec, Canada. Reanalysis data are used in order to assess the potential of the proposed method. After the calibration of PGCR model, an independent dataset is used to assess the downscaling quality. Results are compared with those obtained with MMLR and MMSDM using statistical criteria and climatic indices that describe the frequency, intensity and duration of the variables of interest. Finally, discussion and conclusions are given.

2. Methodology

The MMLR model, the MMSDM and the PGCR model are presented respectively in sections 2.1,2.2 and 2.3. The probabilistic framework for the PGCR model is presented with a description of the different selected marginal distributions for each predictand. Then, a simulation procedure is presented using Gaussian copula to produce the dependence structure between several predictands at multiple sites.

2.1. Multivariate multiple linear regression

Statistical downscaling from multiple AOGCM predictors to numerous meteorological observation sites is one of the situations where predictions of several dependent variables are required from a set of independent variables. Multivariate regression approaches have been used in many scientific areas in order to analyze relationships between multiple independent variables and multiple dependent variables (Jeong et al. 2012).

We consider s meteorological stations, where for each site $j = 1 \ldots s$, we consider three predictands, maximum daily temperature $Tmax_j$, minimum daily temperature $Tmin_j$ and daily precipitation $Prec_j$. The precipitation data $Prec_j$ is of particular interest since the responses are the product of an occurrence process, which decides whether or not there is any rainfall on a particular day, and an amount process, which governs the amount of rainfall, given that some rainfall is observed. Therefore, $Prec_j$ can be decomposed into two separate variables: one for precipitation occurrence Poc_j and one for wet-day precipitation amount Pam_j . To define wet days, the occurrence was limited to events with a precipitation amount larger than or equal to 1 mm/day to avoid problems associated to trace measurements and low daily values. A dry day is defined as a day having less than 1mm of precipitation (Jeong et al. 2012).

The precipitation amount vector Pam_j for a site j is not normally distributed. Indeed, the gamma distribution has often been found to provide a good fit to rainfall amounts in many studies (e.g. Stephenson et al. 1999; Yang et al. 2005). Thus, appropriate transformation should be performed before developing a regression-based precipitation amount model. Yang et al. (2005) proposed the Anscombe transformation for transforming the precipitation amount to a normal

distribution. If the vector Pam_j is Gamma distributed, the distribution of $R_{ij} = Pam_{ij}^{1/3}$ on a day i at a site j, where the Rs is the Anscombe residuals, is normal (e.g. Yang et al. 2005; Jeong et al. 2012).

We denote $\mathbf{A}$ the matrix grouping all the predictand vectors $\mathbf{Tmax}$, $\mathbf{Tmin}$, $\mathbf{Poc}$ and $\mathbf{R}$, of dimension $n \times m$ where $m = 4s$:

$$\mathbf{A} = \left(\mathbf{Tmax}_1, \mathbf{Tmin}_1, \mathbf{Poc}_1, \mathbf{R}_1, \ldots, \mathbf{Tmax}_s, \mathbf{Tmin}_s, \mathbf{Poc}_s, \mathbf{R}_s\right). \quad (1)$$

We define $\mathbf{X}$ the $n \times l$ dimensional matrix that contains the multiple predictor variables. One can estimate the parameter matrix $\mathbf{B}$ of dimensions $l \times m$, which can define the linear relationship between the two matrices $\mathbf{X}$ and $\mathbf{A}$. Therefore, the MMLR can be expressed as:

$$\mathbf{A} = \mathbf{X} \times \mathbf{B} + \mathbf{E} \quad (2)$$

where $\mathbf{E}$ is the residual matrix of dimensions $n \times m$. The parameter matrix $\mathbf{B}$ can be estimated using the Ordinary Least Squares (OLS) method which is given by:

$$\hat{\mathbf{B}} = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{A} \quad (3)$$

Under the assumption that the errors are normally distributed, $\hat{\mathbf{B}}$ by the OLS is the maximum likelihood estimator. Then the deterministic series of predictands can be obtained using the following MMLR equations and atmospheric predictors:

$$\hat{\mathbf{A}} = \mathbf{X} \times \hat{\mathbf{B}} \quad (4)$$

Note that the wet day was determined when deterministic series of daily probability of precipitation occurrence by the MMLR occurrence model was larger than 0.5. The parameter matrix $\hat{\mathbf{B}}$ is affected by multi-collinearity which produces large standard errors of estimated parameters (Jeong et al. 2012). A number of methods have been employed in order to limit the influence of multi-collinearity, such as ridge regression, principal component analysis (Fasbender and Ouarda 2010), canonical correlation regression (Huth 2004), stepwise regression and lasso regression (Hammami et al. 2012). In this study, principal component analysis was employed to deal with multi-collinearity problem.

2.2. Multivariate multiste statistical downscaling model (MMSDM)

The MMLR predicts only deterministic components explainable by linear regression and the independent atmospheric variables X for the different multiple sites. As mentioned previously, this deterministic component underestimates the temporal variability of each predictand and cannot adequately reproduce the correlation between sites and variables. To this end, a classical stochastic randomisation procedure is commonly employed by adding correlated random series to the deterministic component. The resulting model is a multivariate multisite statistical downscaling model (MMSDM) (Jeong et al. 2013). The MMSDM employs MMLR to simulate deterministic series from large-scale reanalysis data and ads spatially correlated random series to the deterministic series of the MMLR to complement the underestimated variance and to reproduce the correlation between sites and variables.

The residual (or error) matrix E [$n \times m$] of the MMLR model is described as:

$$\mathbf{E} = \mathbf{A} - \hat{\mathbf{A}} \tag{5}$$

For the MMSDM, correlated random noise among the predictands at multiple sites was generated from multivariate normal distribution and added to the deterministic series of the MMLR.

The cross-correlated error matrix H[$n \times m$] is generated from a multivariate normal distribution having zero error mean and an error covariance matrix $[\boldsymbol{\Sigma} = \mathbf{SCS}]$ equal to that of the residual matrix $\mathbf{E}(\mathbf{H} \sim \mathbf{N}_m(0, \boldsymbol{\Sigma}))$, where S is a diagonal matrix of standard deviations and C is a correlation matrix of the residual matrix E. Generated residuals were then added to the downscaled predictand as:

$$\tilde{\mathbf{A}} = \hat{\mathbf{A}} + \mathbf{H} \tag{6}$$

The residual vectors for precipitation amount at each site may be not normally distributed and may be skewed. To overcome this problem, a probability distribution mapping technique was adapted and the generated precipitation amount was adjusted using the gamma distribution.

2.3. Probabilistic Gaussian copula regression

2.3.1. Probabilistic regression

In most applications, regression models are performed to describe a mapping that approximates the conditional mean of the prediction and data. This mapping is appropriate if the data are generated from a deterministic function that is corrupted by a normally distributed noise process with constant variance (Cannon 2008). When the noise process has non-constant variance or is non-normal, it is more appropriate to use a model that fully describes the conditional density of the predictand in a probabilistic framework. Thus, the distribution of each predictand at the observed sites must be represented by an appropriate probability density function (PDF), and then we employ a regression model with outputs for each parameter in the assumed PDF of noise

process. In this paper, the normal distribution is chosen for the temperature variables. According to Dorling et al. (2003), for a normally distributed noise process with non-constant variance, the conditional density regression would have two outputs: one for the conditional mean and one for the conditional variance. For $Tmax$ and $Tmin$, the model is described by:

$$\mu_{\max j}(t) = \mathrm{E}\left[Tmax_j(t) \mid x(t)\right] = a^T_{\max j} x(t) \tag{7}$$

$$\mu_{\min j}(t) = \mathrm{E}\left[Tmin_j(t) \mid x(t)\right] = a^T_{\min j} x(t) \tag{8}$$

$$\sigma_{\max j}(t) = \sqrt{Var\left[Tmax_j(t) \mid x(t)\right]} = \exp\left[b^T_{\max j} x(t)\right] \tag{9}$$

$$\sigma_{\min j}(t) = \sqrt{Var\left[Tmin_j(t) \mid x(t)\right]} = \exp\left[b^T_{\min j} x(t)\right] \tag{10}$$

Where $x(t)$ is the value of predictors at the day t, and coefficients $a_{\max j}$, $a_{\min j}$, $b_{\max j}$ and $b_{\min j}$ are estimated separately. Then the conditional normal PDF of $Tmax_j$ for a day t is given by:

$$f_{t\ \max j}\left[Tmax_j(t) \mid x(t)\right]) = \frac{1}{\sqrt{2\pi\sigma_{\max j}{}^2(t)}} \exp\left[-\frac{(Tmax_j(t) - \mu_{\max j}(t))^2}{2\sigma_{\max j}{}^2(t)}\right] \tag{11}$$

And the conditional normal PDF of $Tmin_j$ for a day t is given as:

$$f_{t\ \min j}\left[Tmin_j(t) \mid x(t)\right] = \frac{1}{\sqrt{2\pi\sigma_{\min j}{}^2(t)}} \exp\left[-\frac{(Tmin_j(t) - \mu_{\min j}(t))^2}{2\sigma_{\min j}{}^2(t)}\right] \tag{12}$$

Note that for Poc_j, a standard problem is that the dry-wet leads to a Bernoulli process. In this case we use a logistic regression given by:

$$p_j(t) = \frac{1}{1+\exp\left[-c_j^T x(t)\right]} \tag{13}$$

where $p_j(t)$ is the probability of precipitation occurrence at a site j on a day t and c_j is the coefficient of the logistic model. Thus the conditional distribution of Poc_j is given by:

$$f_{t\ Pocj}\left[Poc_j(t) \mid x(t)\right] = \begin{cases} p_j(t) & if \quad Poc_j(t) = 1 \\ 1 - p_j(t) & if \quad Poc_j(t) = 0 \end{cases} \tag{14}$$

The Pam_j is modeled through a conditional gamma distribution with shape parameters $\alpha_j(t)$ and scale parameter $\beta_j(t)$ given by (Cannon 2008):

$$\alpha_j(t) = \exp\left[d_j^T x(t)\right] \text{ and } \beta_j(t) = \exp\left[e_j^T x(t)\right] \tag{15}$$

where d_j and e_j are the coefficients of the model. Thus, the conditional gamma PDF for Pam_j on a day t is given by:

$$f_{t\ Pamj}\left[Pam_j(t) \mid x(t)\right] = \frac{\left[Pam_j(t)/\beta_j(t)\right]^{\alpha_j(t)-1} \exp\left[-Pam_j(t)/\beta_j(t)\right]}{\beta_j(t)\Gamma(\alpha_j(t))} \tag{16}$$

where $\Gamma(\cdot)$ is the gamma function.

Finally, let us define the random vector $\mathbf{Y}$ of dimension $n \times m$ grouping each predictand at each site with:

$$\mathbf{Y} = (Tmax_1, Tmin_1, Poc_1, Pam_1, \ldots, Tmax_s, Tmin_s, Poc_s, Pam_s) \tag{17}$$

It is important to mention that the vector of predictand $\mathbf{Y}$ is different from the vector $\mathbf{A}$, since the latter contains the transformed variables for the precipitation amount. The conditional PDF $f_{tk}\left[y_k(t) \mid x(t)\right]$ at the time t of the k^{th} element of $\mathbf{Y}$, where $k = 1,\ldots,m$ is then given by Eq. (9) if y_k is a *Tmax* , Eq. (10) if it is a *Tmin*, Eq. (12) if it is a *Poc* and Eq. (14) if it is a *Pam*. Then all coefficients $a_{\max j}$, $a_{\min j}$, $b_{\max j}$, $b_{\min j}$, c_j, d_j and e_j for all sites are set following the method of maximum likelihood by minimizing the negative log predictive density (NLPD) cost function (Haylock et al. 2006; Cawley et al. 2007; Cannon 2008).

$$\mathscr{L}_k = \sum_{t=1}^{n} \log\left\{ f_{tk}\left[y_k(t) \mid x(t)\right]\right\} \text{ for } k = 1,...,m \tag{18}$$

This is carried out via the simplex search method (Lagarias et al. 1999). This is a direct search method that does not use numerical or analytical gradients.

2.3.2. Conditional simulation using Gaussian copula

Once the proposed probabilistic regression model has been trained, it can be used to estimate the PDF of each predictand for a given day when we have the AOGCM predictors. Then, it is possible to create synthetic predicted series of each predictand by sampling in the obtained PDF on each day. In these steps, it is important to maintain realistic relationships between sites and variables. Indeed, consistency of predictions between sites and variables is very important particularly in hydrological modelling.

Reproducing the relationships of multiple random variables when each variable is normally distributed is possible by using the multivariate normal distribution. However, this is not the case in a number of studies such as the present one (*Pam* and *Poc*). Indeed, *Pam* variables are not

usually normally distributed. In this regard, (Pitt et al. 2006) proposed a Gaussian copula framework that can describe the dependence part of the model but allows the margins to be normal or not, discontinuous or continuous.

A copula is a multivariate distribution whose marginals are uniformly distributed on the interval $[0,1]$. The multivariate function $\mathbb{C}$ is called a copula if it is a continuous distribution function and each marginal is a uniform distribution function on [0, 1]; that is $\mathbb{C}[0,1]^q \rightarrow [0,1]$ with

$$\mathbb{C}(u) = \Pr(U_1 \leq u_1, \ldots, U_q \leq u_q) \tag{19}$$

In which each $U_i \sim \mathrm{Un}(0,1)$ and $u = (u_1, \ldots, u_q)$. If $\mathbb{C}$ is a Gaussian copula, then:

$$\mathbb{C}(w; C) = \Phi_q \left\{ \Phi^{-1}(w_1), \ldots, \Phi^{-1}(w_q); C \right\} \tag{20}$$

where Φ is the standard normal cumulative distribution function and $\Phi_q(w; C)$ is the cumulative distribution function for a multivariate normal vector w having zero mean and covariance matrix C. Following Pitt et al. (2006), we use latent variables to transform the marginal distributions of each predictand to a standard normal distribution. The dependence structure between predictands is reproduced by assuming a multivariate Gaussian distribution for the latent variables $z(t) \approx \mathrm{N}_m(0, C)$, using the following equations:

$$z_k(t) = h_{tk}(y_k(t)) \text{ where } h_{tk}(y_k(t)) = \Phi^{-1}[F_{tk}(y_k(t))] \tag{21}$$

Note that Poc is a discrete variable for which the cumulative distribution function F_{tk} is discontinuous. Thus, in order to map $z_k(t)$ onto the full range of the normal distribution, the

cumulative probabilities $F_{tk}(y_k(t))$ for Poc are randomly drawn from a uniform distribution on [0, $1-p(t)$] for dry days and [$1-p(t)$, 1] for wet days. Finally, the goal in this step of calibration is the estimation of the copula parameter, which is the correlation matrix C of the latent variables $z(t)$. Then, for a new day t' it is possible to generate $v=[v_1,\ldots,v_m]$ that can be randomly generated by the Gaussian copula $\mathbb{C}(z;C)$. Figure 1 illustrates how the probabilistic regression model and the Gaussian copula are combined to produce one simulation at a day t'. The value of the synthetic predictand time series $\hat{y}_k(t')$ is given by $F_{tk}^{-1}[v_k]$, except for precipitation occurrences. For precipitation occurrences, $\hat{y}_k(t')=1$ if v_k is greater than $1-p_k(t')$, and $\hat{y}_k(t')=0$ if v_k is less than $1-p_k(t')$.

For the application of a copula model,Vogl et al. (2012) and (Laux et al. 2011) mentioned that it is an indispensable prerequisite that the marginals are "iid" (independent and identically distributed). If this is not the case, an appropriate transformation has to be applied to the data to generate "iid" variates. Thereby, for the PGCR model, all marginal distributions obtained from the probabilistic regression model for a given day are assumed to be "iid".

3. Data and study area

The study area is located in Quebec (Canada), in latitudes between 45° N and 60° N and longitudes between 65° W and 75° W (see Figure 2). Observed daily *Tmax*, *Tmin* and *Prec* are selected as predictands at four stations: Cedars, Drummondville, Seven Islands and Bagot-ville. These stations are located around the St-Lawrence River and the St-Lawrence Gulf (see Figure 2). The corresponding series are provided by Environment Canada weather stations and cover the period between 1 January 1961 and 31 December 2000.

The reanalysis product NCEP / NCAR are used to evaluate the potential of the downscaling method. All the NCEP / NCAR data are averaged on a daily basis from 6-hours data for on the original regular grid of 2.5 ° lat. x 2.5 ° long. The obtained predictors are linearly interpolated using spline functions on the CGCM3 Gaussian grid 3.75 ° x 3.75 ° corresponding to the third version of the coupled globe climate model. These predictors (except the wind direction) are then normalized to the reference period 1961-1990. Six grid points covering the study area are selected (see Figure 2), and for each grid point, 25 NCEP predictors are provided (see Table 1). For each day, 150 variables are available for the downscaling process. The total data set is divided into two independent sets: a calibration period between 1961 and 1990 and a validation period between 1991 and 2000. A principal component analysis (PCA) is first performed in order to reduce the number of AOGCM predictors and to deal with the multi-collinearity problem. The l first components that preserve more than 97% of the variance of the original NCEP predictors are then used in the proposed study as predictor variables.

4. Results

The PGCR model was trained for the calibration period, using $Tmax$, $Tmin$ and $Prec$ data from the 4 stations and the 40 predictors obtained by the PCA**. A**ll the coefficients $a_{\max j}$, $a_{\min j}$, $b_{\max j}$, $b_{\min j}$, c_j, d_j and e_j for each site were set following maximum likelihood estimator by minimising the negative log predictive density (NLPD) cost function for each predictand. Then, once the parameters of the PDF have been estimated for each day t and for each predictand, all obtained conditional marginal distributions are used to transform the observed data to have a new data set for the calibration period in the open interval (0,1). These new data sets are then used to obtain the latent variables $z(t)$ and to fit the parameter of the Gaussian copula. Thereafter, 100

realizations are generated of the precipitation and maximum and minimum temperature series for the validation period (1991-2000), as shown in Figure 1.

Figure 3 illustrates an example of the obtained result using the PGCR model at cedars station during 1991 for both $Tmax$ and $Tmin$. Results indicate that the estimated series are close to the true observed series for both $Tmax$ and $Tmin$. Moreover, the majority of the observations are within to the 95% confidence intervals based on the estimated standard deviations. This indicates that the proposed model adequately depicts the natural process and its fluctuations. Figure 4 illustrates PGCR results for precipitations at cedars station during 1991. Figure 4.a and Figure 4.b show respectively the estimated series of the shape and the scale for conditional gamma distribution. The estimated series of the probability of precipitation occurrences is shown in Figure 4.c, and the synthetic precipitation series and the observed series are shown in Figure 4.d. We can see that the PGCR model provides interesting results for both Poc and Pam.

4.1 Univariate results

For assessing the downscaling quality, data between 1991 and 2000 are used. Two approaches are considered to validate the PGCR model. The first approach is based on a direct comparison between the estimated and observed values using statistical criteria, while the second approach is based on calculating climate indices. In the two validation approaches, the PGCR model results are compared to those obtained using the MMLR and the MMSDM models.

In the first validation approach, three statistical criteria are used for model validation. These criteria are given by:

$$ME = \frac{1}{n}\sum_{t=1}^{n}\left(y_{obs_t} - y_{est_t}\right) \tag{22}$$

$$RMSE = \sqrt{\frac{1}{n}\sum_{t=1}^{n}\left(y_{obs_t} - y_{est_t}\right)^2} \tag{23}$$

$$D = \sigma^2(y_{obs}) - \sigma^2(y_{est}) \tag{24}$$

where n denotes the number of observations, y_{obs_t} refers to the observed value, y_{est_t} is the estimated value, t denotes the day and σ is the standard deviation. The first criterion is the mean error (ME) which is a measure of accuracy. The second criterion is the root mean square error (RMSE) which is given by an inverse measure of the accuracy and must be minimized, and the last criterion D measures the difference between observed and modeled variances, this criterion evaluates the performance of the model in reproducing the observed variability.

The PGCR and MMSDM models give probabilistic predictions. Point forecasts can be made by estimating the conditional mean for each day. For each predictand, values of the RMSE, ME and D for all the three models PGCR, MMLR and MMSDM are given in Table 2. The RMSE and the ME for PGCR and MMSDM were calculated using the conditional mean for each day. However the differences between observed and modeled variances are obtained using the mean D values of 100 realisations. From Table 2 it can be seen that all the three models give similar results in terms of RMSE and ME for maximum and minimum temperatures. However, for downscaled precipitations, PGCR shows the best performance, since it has lower RMSE and close to zero ME compared to both the MMSDM and MMLR. Table 3 indicates also that MMSDM performs better than MMLR in terms of both RMSE and ME for downscaled precipitations. This result is

due to the fact that the MMLR model is in reality biased for precipitation variables. The main reason for this bias is that zero precipitation amounts were included to calibrate MMLR amount model. Furthermore, the Anscombe residuals R from the observed precipitation amount may not be normally distributed. For this reason, the MMSDM model employs a probability mapping technique to correct this bias. Moreover, from Table 2 It can also be noted that, in terms of D, the PGCR model reproduces better the temporal variability compared to MMLR and MMSDM. In the second validation approach, we consider a set of several climate indices that have been proposed for northern climates for maximum and minimum temperatures. The definitions of these climate indices are presented in Table 3. These indices are chosen to evaluate the performance of downscaling models and reflect temperature characteristics including the frequency, intensity, and duration of temperature extremes (Wilby 1998; Wilby et al. 2002; Gachon et al. 2005; Hessami et al. 2008). Likewise, for downscaled precipitation several climate indices defined in Table 4 are considered to assess the downscaling quality of precipitation. PGCR and MMLR are then compared by computing the RMSE for each of the climatic indices for both precipitation and temperature. For the PGCR model the RMSE of this indices are calculated using the mean RMSE values of 100 realisations.

The results for temperature indices are presented inTable 5. These results indicate that PGCR performs better than both MMLR and MMSDM. MMLR gives better results only for the FSL at Bagot-ville, the GSL at Drummondville and the 90^{th} percentile for minimum Temperature at Seven-Island. Similarly, Table 6 summarizes the RMSE of downscaled precipitations climatic indices for both PGCR and MMLR models and for the 4 weather stations during the validation period (1991–2000). For precipitation amounts five indices are considered: the mean

precipitation of wet days (MPWD), the 90th percentile of daily precipitation (P90), the maximum 1-day precipitation (PX1D), the maximum 3-day precipitation (PX3D), and the maximum 5 day precipitations (PX5D). Of more, three other indices are considered for precipitation occurrences: the maximum number of consecutive wet days (WRUN), the maximum number of consecutive dry days (DRUN) and the number of wet days (NWD). In terms of RMSE, results indicate that PGCR gives better results than MMLR and MMSDM for all precipitation indices except for WRUN at cedars, Pmax90 at Seven-Island and DRUN at Bagot-ville for which the MMSDM gives the best results. This result shows the role of the gamma distribution in the PGCR model to replicate adequately the characteristic of precipitation amounts. Furthermore, the use of the logistic regression allows PGCR model to better reproduce observed monthly characteristics of precipitation occurrences based on the indices WRUN, NWD and DRUN.

4.2. Inter-station and inter-variable results

To evaluate the ability of the conditional Gaussian copula in the PGCR model to replicate the observed cross-site correlations for each predictand, the scatter plots of observed and modeled cross-site correlations of each predictand for PGCR, MMLR and MMSDM are plotted (Figure 5). The correlation values of the PGCR and MMSDM model were obtained using the mean of the correlation values calculated from a 100 realisations. For all predictands, MMLR overestimates the cross-site correlations as shown in Figure 5, and PGCR and MMSDM reproduce well these cross-site correlations. On the other hand, PGCR outperformed the MMSDM for precipitation occurrences. Similarly, Figure 6 shows the scatter plots of observed and modeled cross-predictand correlations for PGCR, MMLR and MMSDM during the validation period. This figure shows that both PGCR and MMSDM are able to reproduce more adequately theses cross-

predictand correlations. This is a great achievement of the two multivariable models, in comparison with the univariate MMLR model. Basically, both PGCR and MMSDM precisely simulate these cross-predictand correlations and there is no clear difference between them except when precipitation amount is present. Indeed, MMSDM has difficulty reproducing the cross-correlation predictand. Indeed theses cross-correlation values can be affected by the probability mapping step that is used to correct the bias and to reproduce the adequate distribution of precipitation. However, when evaluating the PGCR model there is no need to rely on transformation steps or on bias correction procedures and the mapping in the conditional distribution is automatic using its probabilistic regression component.

For precipitation, joint probabilities of the events that two sites are both dry or both wet on a given day are displayed in Figure 7. PGCR and MMSDM adequately simulate these joint probabilities and there is no clear difference between PGCR and MMSDM simulations, both having almost better results compared to MMLR model. This is a great finding of PGCR and MMSDM, in comparison with the joint probabilities from the single-site MMLR model.

Finally, to evaluate the consistency of local weather variables, Figure 8 compares differences of mean temperatures on wet and dry days (mean temperature on wet days minus the corresponding value on dry days) in the synthetic data sets with the observed ones. The values in the plots are for each site and each month (48 data points in total). It appears that the self-consistency between local precipitation and temperatures is reproduced very well in synthetic data sets, from either PGCR or MMLR.

5. Conclusions and discussions

A PGCR model is proposed in this paper for the downscaling of AOGCM predictors to multiple predictands at multi-sites simultaneously and to preserve relationships between sites and variables. This model relies on a probabilistic framework in order to describe the conditional density of each predictand for a given day. The PGCR model uses a Gaussian distribution for maximum and minimum temperature, a Bernoulli distribution for precipitation occurrences and a gamma distribution for precipitation amounts. In the probabilistic framework, PGCR adopts a regression model with outputs for each parameter in the specified probability density function. To maintain realistic relationships between sites and variables, the PGCR model uses a Gaussian copula that describes dependences between all predictands.

The developed model was then applied to generate daily maximum and minimum temperatures and precipitations of four observation sites located in the southern part of the province of Quebec (Canada). NCEP reanalysis data were used as predictors in order to assess the potential of the method, although the final objective is to use AOGCM predictors. Application results of the PGCR model were compared with those of the regression-based MMLR model and a MMSDM as a classical multisite and multivariable model.

Results show that all the three models give similar results in terms of RMSE and ME for maximum and minimum temperatures. On the other hand, PGCR performs better for downscaled precipitations in terms of ME and RMSE. In addition the comparison based on temperature and precipitation indices shows that the PGCR model is more able to reproduce extremes and observed variability on a seasonal and interannual basis for both temperature and precipitation. In

terms of reproducing spatial and inter-variable properties, both PGCR and MMSDM models provide interesting results without significant differences.

Reproduction of the temporal variability in both precipitation and temperature fields are among the most important achievements for the proposed PGCR model. The MMLR model showed difficulty in reproducing the observed variability. Indeed, regression models generally reproduce the mean of the process conditionally to the selected independent variables. As a consequence, the variability of the regression is always smaller than the initial variability. In this regard, Von Storch (1999) mentioned that predictors generated from synoptic-scale fields cannot represent all variability at the sub-grid scale. Hence, in this way, the PGCR model presents the advantage of modeling the entire conditional distribution, and thereby gives the mean of the downscaled predictands as well as their variability. The reproduction of the variability is not a typical characteristic of PGCR, indeed stochastic and hybrid models, such as the MMSDM model, have this characteristic. Nevertheless, these hybrid approaches are based on static noises that are not dependent on the predictors. Thus, hybrid approaches may not represent local change in atmospheric variability in a climate change simulation. Thereby, one advantage of the proposed PGCR model compared to hybrid approaches is that the temporal variability could change in the future, which may explain the fact that PGCR gives better results compared to the MMSDM model as shown in Table 2. This finding shows the role of the probabilistic regression component model of PGCR that allows predicting not only the conditional mean but also the whole conditional distribution including variability. Therefore, the real gain when using the proposed PGCR model consists not only in the dependence modeling trough Gaussian copula but also in including the advantages of probabilistic regression in a multi-site multivariable framework. This is due to the characteristic of the marginal effect elimination through the copula. This attractive

characteristic helps to model and understand the dependence structure effectively, as it has no relationship with the marginal behavior. However, it should be mentioned that it is more difficult to reproduce correlations when the precipitation amount is considered, since precipitation amount results are affected by the precipitation occurrence results. For this reason it would be more judicious to model the occurrence and amount of precipitation simultaneously by employing only one PDF that describes the discrete-continuous behavior of precipitation, for example by using a Bernoulli-gamma or Poisson-gamma PDF.

An important fact that has not been considered in this work is that the time structure of the downscaled precipitations must be poorly simulated, such us lag-1 correlation for downscaled precipitations. The number of AOGCM grid points in this study can also be increased, which would improve the precision of both MMLR and PGCR. In that context, the use of regional-scale predictors from regional climate models (RCM) instead of coarse-scale AOGCM will be strongly beneficial for improving downscaled results. Finally, the results of PGCR model can be improved by developing the PGCR model for each month. This would allow **taking** into account **the** seasonal variability of each predictand.

6. References

Bárdossy, A. (2006). "Copula-based geostatistical models for groundwater quality parameters." Water Resources Research **42**(11).

Bates, B. C., S. P. Charles and J. P. Hughes (1998). "Stochastic downscaling of numerical climate model simulations." Environmental Modelling and Software **13**(3-4): 325-331.

Bellone, E., J. P. Hughes and P. Guttorp (2000). "A hidden Markov model for downscalling synoptic atmospheric patterns to precipitation amounts." Climate Research **15**(1): 1-12.

Burger, G. and Y. Chen (2005). "A Regression-based downscaling of spatial variability for hydrologic applications." Journal of Hydrology **311**: 299-317.

Cannon, A. J. (2008). "Probabilistic multisite precipitation downscaling by an expanded Bernoulli-gamma density network." Journal of Hydrometeorology **9**(6): 1284-1300.

Cannon, A. J. (2009). "Negative ridge regression parameters for improving the covariance structure of multivariate linear downscaling models." International Journal of Climatology **29**(5): 761-769.

Cannon, A. J. (2011). "Quantile regression neural networks: Implementation in R and application to precipitation downscaling." Computers and Geosciences **37**(9): 1277-1284.

Cawley, G. C., G. J. Janacek, M. R. Haylock and S. R. Dorling (2007). "Predictive uncertainty in environmental modelling." Neural Networks **20**(4): 537-549.

Chebana, F. and T. B. M. J. Ouarda (2007). "Multivariate L-moment homogeneity test." Water Resources Research **43**(8).

Clark, M., S. Gangopadhyay, L. Hay, B. Rajagopalan and R. Wilby (2004). "The Schaake Shuffle: A Method for Reconstructing Space–Time Variability in Forecasted Precipitation and Temperature Fields." Journal of Hydrometeorology **5**(1): 243-262.

Conway, D., R. Wilby and P. Jones (1996). "Precipitation and air flow indices over the British Isles." Climate Research **7**: 169-183.

De Michele, C. and G. Salvadori (2003). "A Generalized Pareto intensity-duration model of storm rainfall exploiting 2-Copulas." Journal of Geophysical Research D: Atmospheres **108**(2): ACL 15-11 ACL 15-11.

Dorling, S. R., R. J. Foxall, D. P. Mandic and G. C. Cawley (2003). "Maximum likelihood cost functions for neural network models of air quality data." Atmospheric Environment **37**(24): 3435-3443.

Dupuis, D. J. (2007). "Using copulas in hydrology: Benefits, cautions, and issues." Journal of Hydrologic Engineering **12**(4): 381-393.

El Adlouni, S. and T. B. M. J. Ouarda (2008). "Study of the joint law flow-level by copulas: Case of the Chateauguay River." Étude de la loi conjointe débit-niveau par les copules: Cas de la rivière Châteauguay **35**(10): 1128-1137.

Fasbender, D. and T. B. M. J. Ouarda (2010). "Spatial Bayesian Model for Statistical Downscaling of AOGCM to Minimum and Maximum Daily Temperatures." Journal of Climate **23**(19): 5222-5242.

Gachon, P., A. St-Hilaire, T. B. M. J. Ouarda, V. Nguyen, C. Lin, J. Milton, D. Chaumont, J. Goldstein, M. Hessami, T. D. Nguyen, F. Selva, M. Nadeau, P. Roy, D. Parishkura, N. Major, M. Choux and A. Bourque (2005). "A first evaluation of the strength and weaknesses of statistical downscaling methods for simulating extremes over various regions of eastern Canada " Sub-component, Climate Change Action Fund (CCAF), Environment Canada **Final report**(Montréal, Québec, Canada): 209.

Hammami, D., T. S. Lee, T. B. M. J. Ouarda and J. Le (2012). "Predictor selection for downscaling GCM data with LASSO." Journal of Geophysical Research D: Atmospheres **117**(17).

Haylock, M. R., G. C. Cawley, C. Harpham, R. L. Wilby and C. M. Goodess (2006). "Downscaling heavy precipitation over the United Kingdom: A comparison of dynamical and statistical methods and their future scenarios." International Journal of Climatology **26**(10): 1397-1415.

Hellström, C., D. Chen, C. Achberger and J. Räisänen (2001). "Comparison of climate change scenarios for Sweden based on statistical and dynamical downscaling of monthly precipitation." Climate Research **19**(1): 45-55.

Herrera, E., T. B. M. J. Ouarda and B. Bobée (2006). "Downscaling methods applied to Atmosphere-Ocean General Circulation Models (AOGCM)." Méthodes de désagrégation Appliquées aux Modèles du Climat Global Atmosphère-Océan (MCGAO) **19**(4): 297-312.

Hessami, M., P. Gachon, T. B. M. J. Ouarda and A. St-Hilaire (2008). "Automated regression-based statistical downscaling tool." Environmental Modelling & Software **23**(6): 813-834.

Hughes, J. P., P. Guttorp and S. P. Charles (1999). "A non-homogeneous hidden Markov model for precipitation occurrence." Journal of the Royal Statistical Society. Series C: Applied Statistics **48**(1): 15-30.

Huth, R. (1999). "Statistical downscaling in central Europe: Evaluation of methods and potential predictors." Climate Research **13**(2): 91-101.

Huth, R. (2002). "Statistical downscaling of daily temperature in central Europe." Journal of Climate **15**(13): 1731-1742.

Huth, R. (2004). Sensitivity of local daily temperature change estimates to the selection of downscaling models and predictors. Boston, MA, ETATS-UNIS, American Meteorological Society.

Jeong, D., A. St-Hilaire, T. Ouarda and P. Gachon (2013). "A multivariate multi-site statistical downscaling model for daily maximum and minimum temperatures." Climate Research **54**(2): 129-148.

Jeong, D. I., A. St-Hilaire, T. B. M. J. Ouarda and P. Gachon (2012). "CGCM3 predictors used for daily temperature and precipitation downscaling in Southern Québec, Canada." Theoretical and Applied Climatology **107**(3-4): 389-406.

Jeong, D. I., A. St-Hilaire, T. B. M. J. Ouarda and P. Gachon (2012). "Multisite statistical downscaling model for daily precipitation combined by multivariate multiple linear regression and stochastic weather generator." Climatic Change **114**(3-4): 567-591.

Kazianka, H. and J. Pilz (2010). "Copula-based geostatistical modeling of continuous and discrete data including covariates." Stochastic Environmental Research and Risk Assessment **24**(5): 661-673.

Khalili, M., V. T. Van Nguyen and P. Gachon (2013). "A statistical approach to multi-site multivariate downscaling of daily extreme temperature series." International Journal of Climatology **33**(1): 15-32.

Lagarias, J. C., J. A. Reeds, M. H. Wright and P. E. Wright (1999). "Convergence properties of the Nelder-Mead simplex method in low dimensions." SIAM Journal on Optimization **9**(1): 112-147.

Laux, P., S. Vogl, W. Qiu, H. Knoche and H. Kunstmann (2011). "Copula-based statistical refinement of precipitation in RCM simulations over complex terrain." Hydrology and Earth System Sciences **15**(7): 2401-2419.

Lindström, G., B. Johansson, M. Persson, M. Gardelin and S. Bergström (1997). "Development and test of the distributed HBV-96 hydrological model." Journal of Hydrology **201**(1-4): 272-288.

Palutikof, J. P., C. M. Goodess, S. J. Watkins and T. Holt (2002). "Generating rainfall and temperature scenarios at multiple sites: Examples from the Mediterranean." Journal of Climate **15**(24): 3529-3548.

Pitt, M., D. Chan and R. Kohn (2006). "Efficient Bayesian inference for Gaussian copula regression models." Biometrika **93**(3): 537-554.

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.

Schölzel, C. and P. Friederichs (2008). "Multivariate non-normally distributed random variables in climate research - Introduction to the copula approach." Nonlinear Processes in Geophysics **15**(5): 761-772.

Schoof, J. T. and S. C. Pryor (2001). "Downscaling temperature and precipitation: A comparison of regression-based methods and artificial neural networks." International Journal of Climatology **21**(7): 773-790.

Sklar, A. (1959). Fonctions de répartition à n dimensions et leurs marges.

Stephenson, D. B., K. Rupa Kumar, F. J. Doblas-Reyes, J. F. Royer, F. Chauvin and S. Pezzulli (1999). "Extreme daily rainfall events and their impact on ensemble forecasts of the Indian monsoon." Monthly Weather Review **127**(9): 1954-1966.

Vogl, S., P. Laux, W. Qiu, G. Mao and H. Kunstmann (2012). "Copula-based assimilation of radar and gauge information to derive bias-corrected precipitation fields." Hydrology and Earth System Sciences **16**(7): 2311-2328.

Von Storch, H. (1999). "On the Use of "Inflation" in Statistical Downscaling." Journal of Climate **12**(12): 3505-3506.

Vrac, M. and P. Naveau (2007). "Stochastic downscaling of precipitation: From dry events to heavy rainfalls." Water Resources Research **43**(7).

Widmann, M., C. S. Bretherton and E. P. Salathé Jr (2003). "Statistical precipitation downscaling over the northwestern united states using numerically simulated precipitation as a predictor." Journal of Climate **16**(5): 799-816.

Wilby, R. (1998). "Statistical downscaling of daily precipitation using daily airflow and seasonal teleconnection indices." Climate Research **10(3)**.

Wilby, R. L., C. W. Dawson and E. M. Barrow (2002). "sdsm - a decision support tool for the assessment of regional climate change impacts." Environmental Modelling and Software **17**(2): 145-157.

Wilby, R. L. and T. M. L. Wigley (1997). "Downscaling general circulation model output: a review of methods and limitations." Progress in Physical Geography **21**(4): 530-548.

Wilks, D. S. and R. L. Wilby (1999). "The weather generation game: a review of stochastic weather models." Progress in Physical Geography **23**(3): 329-357.

Williams, P. M. (1998). "Modelling seasonality and trends in daily rainfall data." Advances in neural information processing systems: 985-991.

Yang, C., R. E. Chandler, V. S. Isham and H. S. Wheater (2005). "Spatial-temporal rainfall simulation using generalized linear models." Water Resources Research **41**(11): 1-13.

Table 1. NCEP predictors on the CGCM3 grid.

No	Predictors	No	Predictors
1	mean pressure at the sea level	14	Divergence at 500 hPa
2	Wind speed at 1000 hPa	15	Wind speed at 850 hPa
3	Component U at 1000 hPa	16	Component U at 850 hPa
4	Component V at 1000 hPa	17	Component V at 850 hPa
5	Vorticity at 1000 hPa	18	Vorticity at 850 hPa
6	Wind direction at 1000 hPa	19	Geopotential at 850 hPa
7	Divergence at 1000 hPa	20	Wind direction at 850 hPa
8	Wind speed at 500 hPa	21	Divergence at 1000 hPa
9	Component U at 500 hPa	22	Specific humidity at 500 hPa
10	Component V at 500 hPa	23	Specific humidity at 850 hPa
11	Vorticity at 500 hPa	24	Specific humidity at 1000 hPa
12	Geopotential at 500 hPa	25	Temperature at 2m
13	Wind direction at 500 hPa		

Table 2. Quality assessment of the estimated series for PGCR, MMLR and MMSDM during the validation period (1991–2000) for the four weather stations. Criteria are ME, RMSE, and differences between observed and modeled variance D. For PGCR and MMSDM models criteria were calculated from the conditional mean.

		ME			RMSE			D		
		Tmax (°C)	Tmin (°C)	Prec (mm)	Tmax (°C)	Tmin (°C)	Prec (mm)	Tmax (°C)	Tmin (°C)	Prec (mm)
Cedars	PGCR	0.55	0.22	**-0.14**	3.28	3.70	**6.35**	**-1.34**	**-0.85**	**-1.16**
	MMLR	0.55	0.22	2.59	3.28	3.70	7.02	8.81	9.21	45.49
	MMSDM	0.55	0.22	1.97	3.30	3.71	6.48	-1.84	-2.83	17.92
Drummondville	PGCR	**0.46**	0.49	**0.48**	3.31	4.08	**5.42**	**-3.83**	6.31	**10.69**
	MMLR	0.47	0.49	2.44	3.31	4.08	6.14	7.22	18.15	34.53
	MMSDM	0.47	0.49	0.97	3.32	4.10	5.57	-4.10	**3.47**	11.27
Seven-Island	PGCR	0.20	**-0.52**	**-0.67**	3.18	3.59	**5.57**	**5.51**	**2.12**	**-10.02**
	MMLR	0.19	-0.53	2.11	3.17	**3.58**	5.99	16.36	13.42	33.31
	MMSDM	0.19	-0.53	0.72	3.17	3.60	5.59	6.5	2.34	13.43
Bagot-ville	PGCR	-0.05	0.14	**0.87**	3.53	3.85	**6.13**	**1.12**	**-0.28**	**24.40**
	MMLR	-0.05	0.14	2.37	3.53	3.84	6.72	15.42	12.76	41.51
	MMSDM	-0.05	0.14	1.11	3.53	3.85	6.24	3.48	-1.77	28.16

Table 3. Definition of the climatic indices used for the performance assessment of downscaled temperatures.

Indices	Definition	Unite	Scale Time
DTR	Mean of diurnal temperature range	°C	Season
FSL	Frost season length: Days between 5 consecutive T_{mean}<0 °C and 5 consecutive T_{mean}> 0°C	Days	Years
GSL	Growing Season length: Days between 5 consecutive T_{mean}<5 °C and 5 consecutive T_{mean}> 5°C	Days	Years
FR-Th	Days with freeze and thaw (T_{max}> 0°C,T_{min}<0°C)	Days	Months
Tmax90	90th percentile of daily maximum temperature	°C	seasons
Tmin90	90^{th} percentile of daily minimum temperature	°C	seasons

$$T_{mean} = \frac{T_{max} + T_{min}}{2}$$

Table 4. Definition of the climatic indices used for the performance assessment of downscaled precipitations.

Indices	Definition	Unite	Scale Time
MWD	Mean precipitation of wet day	Mm	Months
Pmax90	90^{th} percentile of daily precipitation	Mm	Seasons
PX1D	Maximum 1-days precipitation	Mm	Months
PX3D	Maximum 3-days precipitation	Mm	Months
PX5D	Maximum 5-days precipitation	Mm	Months
WRUN	Maximum number of consecutive wet days	Days	Months
DRUN	Maximum number of consecutive dry days	Days	Months
NWD	Number of wet day	Days	Months

Table 5. RMSE of climatic indices of downscaled temperatures for PGCR, MMLR and MMSDM on the 4 weather stations during the validation period (1991–2000). RMSE of PGCR and MMSDM models were calculated using the mean of 100 realisations.

		DTR	FSL	GSL	FR-Th	Tmax90	Tmin90
Cedars	PGCR	**0.93 (0.05)**	**3.01 (0.97)**	**21.84 (2.59)**	**2.56 (0.54)**	**1.32 (0.12)**	**1.32 (0.10)**
	MMLR	0.93	3.91	22.81	2.94	1.73	1.52
	MMSDM	0.94 (0.05)	3.48 (0.84)	24.29 (2.26)	2.87 (0.55)	1.5 (0.10)	1.46 (0.10)
Drummondville	PGCR	0.75 (0.05)	**2.82 (0.79)**	25.29 (2.28)	**2.29 (0.49)**	**1.19 (0.12)**	**1.17 (0.12)**
	MMLR	0.75	3.84	**23.74**	2.68	1.59	1.98
	MMSDM	**0.74 (0.06)**	3.66 (0.64)	25.63 (2.38)	2.94 (0.50)	1.69 (0.12)	1.55 (0.12)
Seven-Island	PGCR	1.06 (0.05)	**1.18 (0.23)**	**11.93 (2.77)**	**2.30 (0.50)**	**0.95 (0.11)**	1.42 (0.12)
	MMLR	1.06	1.89	12.71	2.67	1.62	**1.39**
	MMSDM	1.06 (0.05)	1.64 (0.16)	12.10 (2.90)	2.74 (0.66)	1.09 (0.11)	1.60 (0.12)
Bagot-ville	PGCR	**0.74 (0.06)**	1.27 (0.27)	**17.87 (2.84)**	**2.27 (0.53)**	**1.12 (0.15)**	**1.03 (0.13)**
	MMLR	0.75	**0.01**	20.93	2.44	1.44	1.65
	MMSDM	0.75 (0.06)	1.48 (0.20)	19.19 (2.88)	2.48 (0.61)	1.21 (0.13)	1.38 (0.12)

Bold means better result, and values between brackets mean standard deviation of the 100 RMSE.

Table 6. RMSE of climatic indices of downscaled precipitations for PGCR, MMLR and MMSDM models on the 4 weather stations during the validation period (1991–2000). RMSE of PGCR and MMSDM models were calculated using the mean of 100 realisations.

		Precipitation amount					**Precipitation occurences**		
		MPWD (mm)	Pmax90 (mm)	PX1D (mm)	PX3D (mm)	PX5D (Days)	WRUN (Days)	DRUN (Days)	NWD (Days)
Cedars	PGCR	**1.85 (0.28)**	**5.81 (0.29)**	**24.74 (3.38)**	**26.14 (4.54)**	**26.06 (5.13)**	1.12 (0.32**)**	**1.78 (0.63)**	**2.81 (0.55)**
	MMLR	3.30	9.01	35.52	42.95	47.42	1.62	5.02	5.88
	MMSDM	2.18 (0.25)	4.08 (0.20)	26.92 (3.61)	30.01 (4.04)	30.85 (4.14)	**1.09 (0.27)**	2.03 (0.72)	2.94 (0.55)
Drummondville	PGCR	**0.93 (0.19)**	**4.24 (0.18)**	**8.23 (2.70)**	**15.89 (3.51)**	**15.77 (3.33)**	**1.06 (0.37)**	1.65 (0.78)	**2.55 (0.58)**
	MMLR	2.41	7.88	17.25	29.48	34.42	1.82	3.68	5.64
	MMSDM	1.29 (0.19)	3.41 (0.16)	8.94 (3.03)	18.37 (3.18)	19.47 (3.55)	1.16 (0.28)	1.65 (0.72)	3.16 (0.56)
Seven-Island	PGCR	**1.10 (0.27)**	3.47 (0.31**)**	**10.14 (3.59)**	**13.22 (4.01)**	**16.25 (4.85)**	**0.83 (0.28)**	**2.25 (0.62)**	**3.16 (0.58)**
	MMLR	2.55	7.10	23.41	31.32	37.87	1.35	6.04	6.04
	MMSDM	1.46 (0.18)	**3.06 (0.18)**	13.11 (2.09)	17.34 (2.82)	21.54 (3.00)	0.87 (0.27)	2.32 (0.84)	3.36 (0.65)
Bagot-ville	PGCR	**1.10 (0.14)**	**4.71 (0.20)**	**7.66 (1.24)**	**12.26 (1.82)**	**13.42 (2.34)**	**1.63 (0.24)**	2.45 (0.79)	**3.74 (0.53)**
	MMLR	2.44	8.10	18.86	28.47	32.67	2.54	5.47	7.22
	MMSDM	1.35 (0.15)	4.21 (0.18)	8.23 (3.01)	14.06 (2.88)	15.85 (3.15)	1.73 (0.23)	**2.42 (0.83)**	3.90 (0.54)

Bold means better result, and values between brackets mean standard deviation of the 100 RMSE.

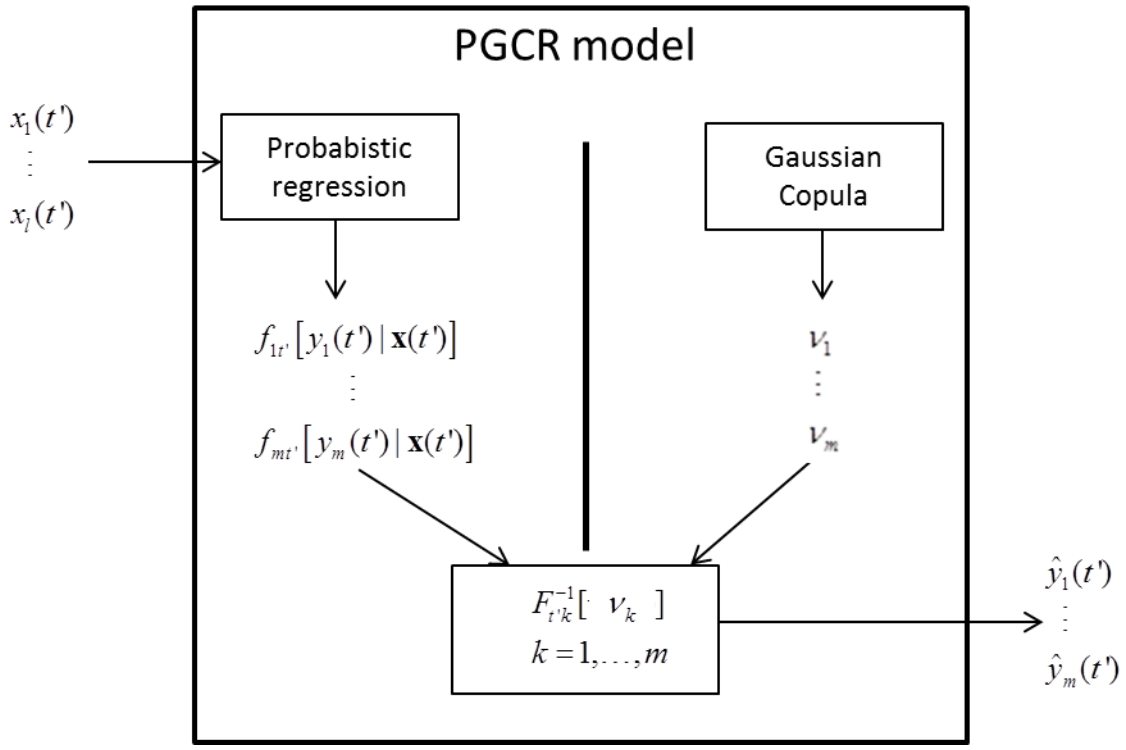

Figure 1. Simulation procedure of the PGCR model, using the probabilistic regression model and the Gaussian copula.

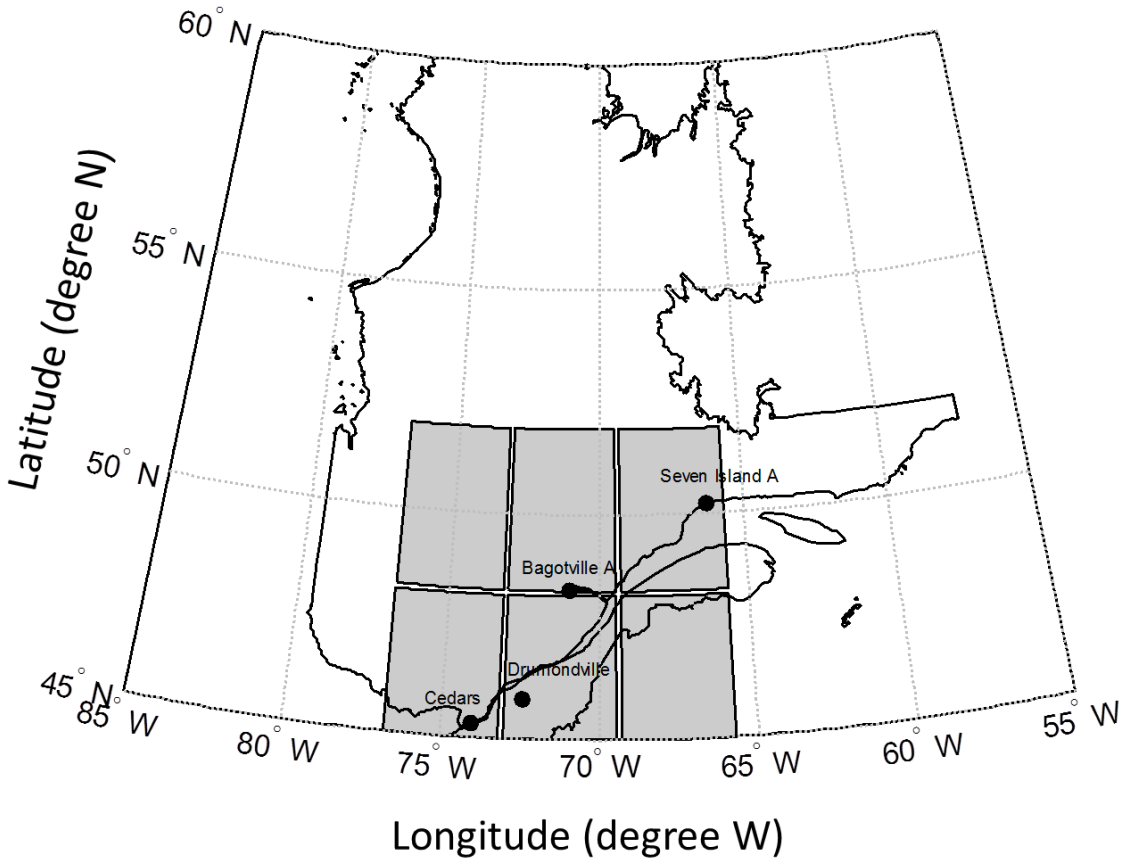

Figure 2. Location of CGCM3 grid and observation stations of daily precipitation and daily maximum and minimum temperatures.

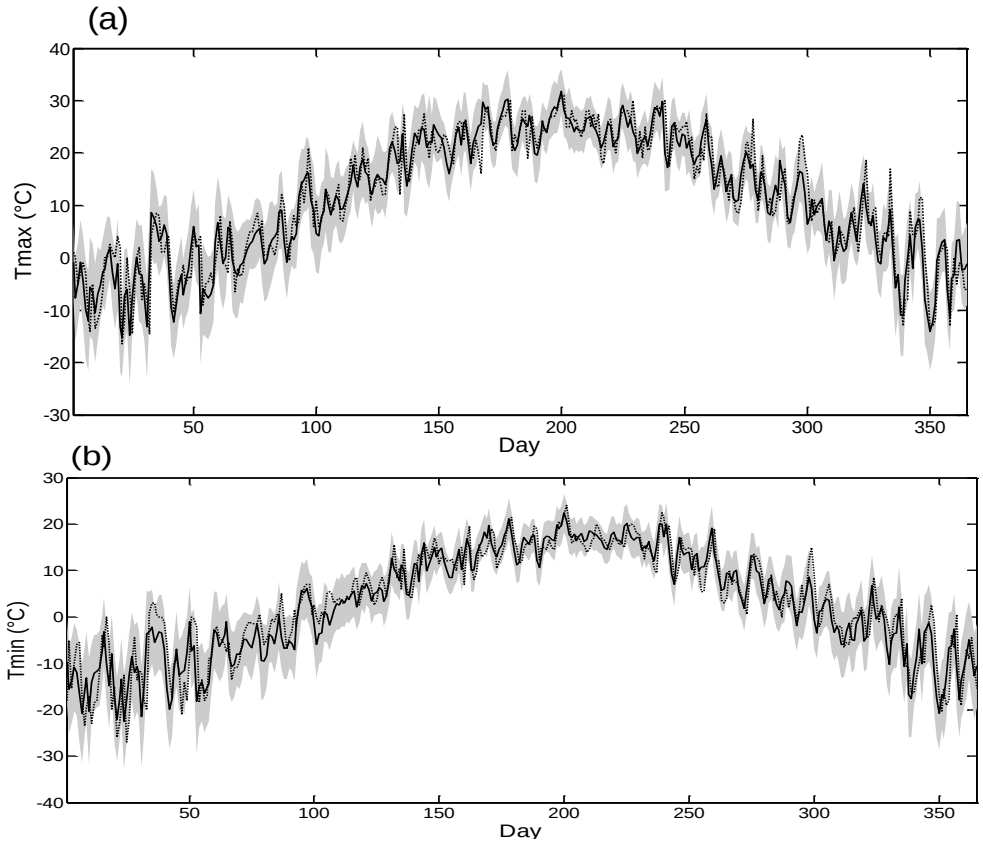

Figure 3. PGCR result for cedars station during 1991. Time series of (a) maximum temperature and (b) minimum temperature. The conditional mean is shown by the solid line, the observed series is shown by the dashed line, and the 95% confidence interval obtained from the estimated standard deviations is illustrated with the gray shaded area.

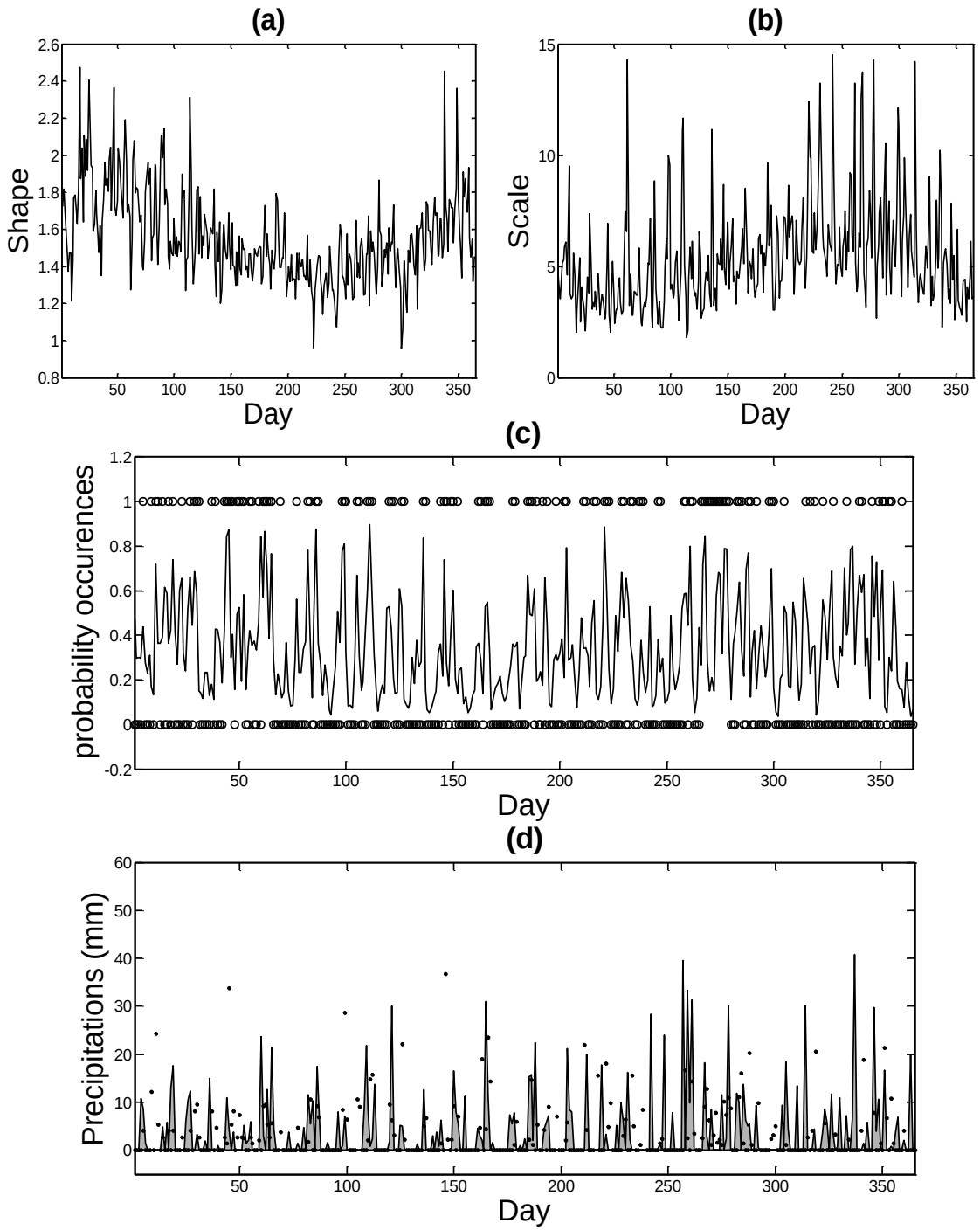

Figure 4. PGCR results for precipitations at cedars station during 1991. The conditional gamma parameters are shown in (a) for the shape and (b) for the scale. The probability of precipitation occurrences is shown by the solid line in (c) where circles indicate the observed precipitation occurrences. (d) indicates the observed precipitation values (dots) and a synthetic precipitation obtained by combining (a) and (b) and (c).

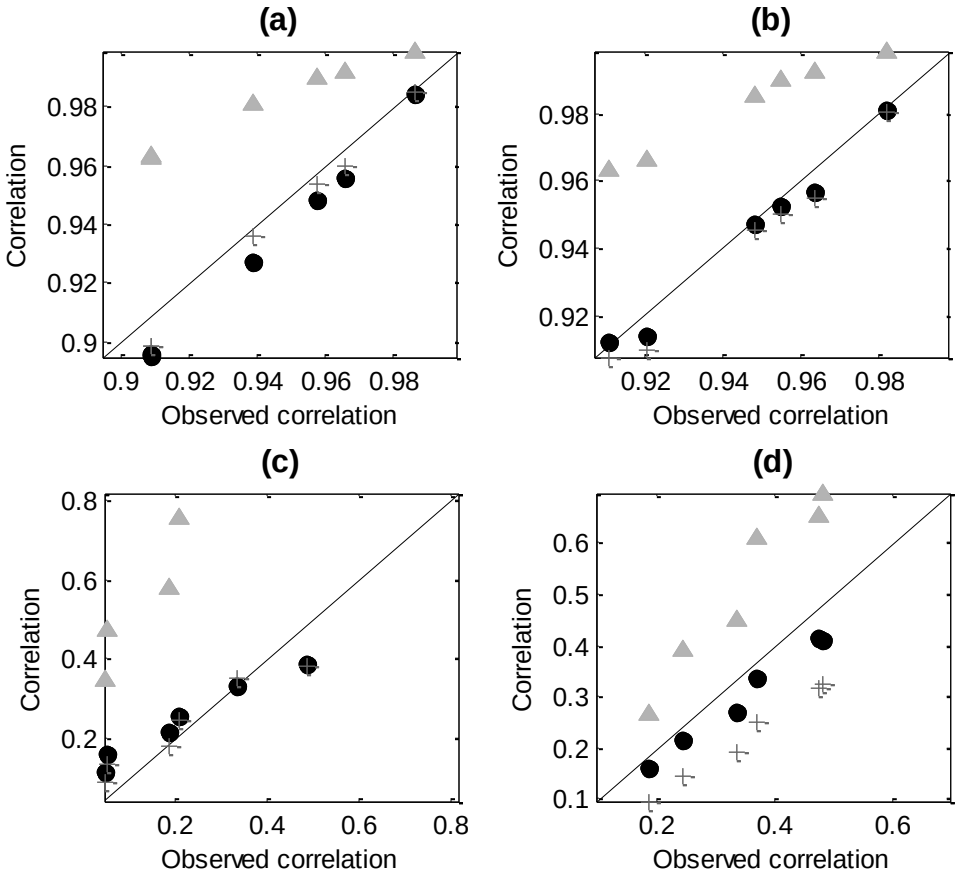

Figure 5. Scatter plot of observed and modeled cross-site correlations by PGCR (black dots), MMLR (gray triangle) and MMSDM (gray plus) for maximum temperature (a), minimum temperature (b), precipitation amount (c) and precipitation occurrences (d). Correlation values of PGCR and MMSDM models are obtained using the mean of the correlation values calculated from 100 simulations.

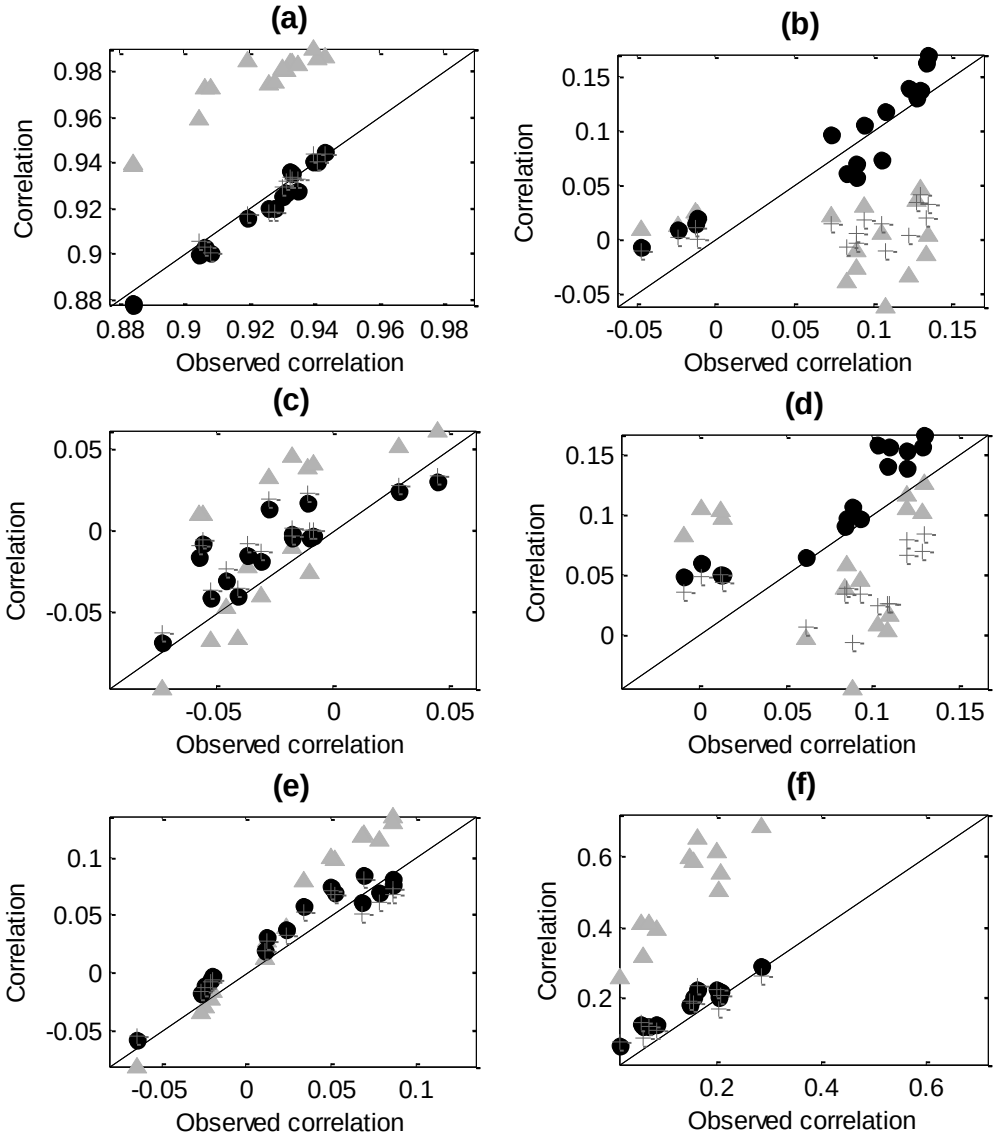

Figure 6. Scatter plot of observed and modeled correlations by PGCR (black dots), MMLR (gray triangle) and MMSDM (gary plus) for *Tmax-Tmin* (a), *Tmax-Pam* (b), *Tmax-Poc* (c), *Tmin-Pam* (d), *Tmin-Poc* (e) and *Pam-Poc* (f). Correlation values of PGCR and MMSDM models are obtained using the mean of the correlation values calculated from 100 simulations.

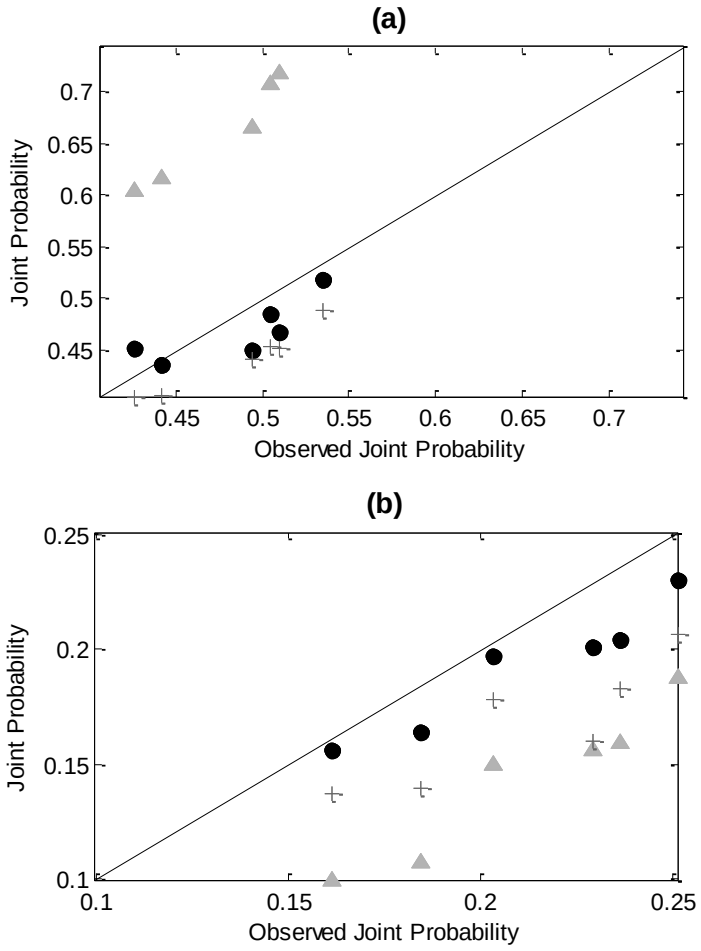

Figure 7. Joint probabilities that station pairs are (a) both wet, or (b) both dry, on a given day, for the observed and modeled joint probability by PGCR (black dots), MMLR (gray triangles) and MMSDM (gray plus) during the validation period. Values of PGCR and MMSDM models are obtained using the mean of the joint probability values calculated from 100 simulations.

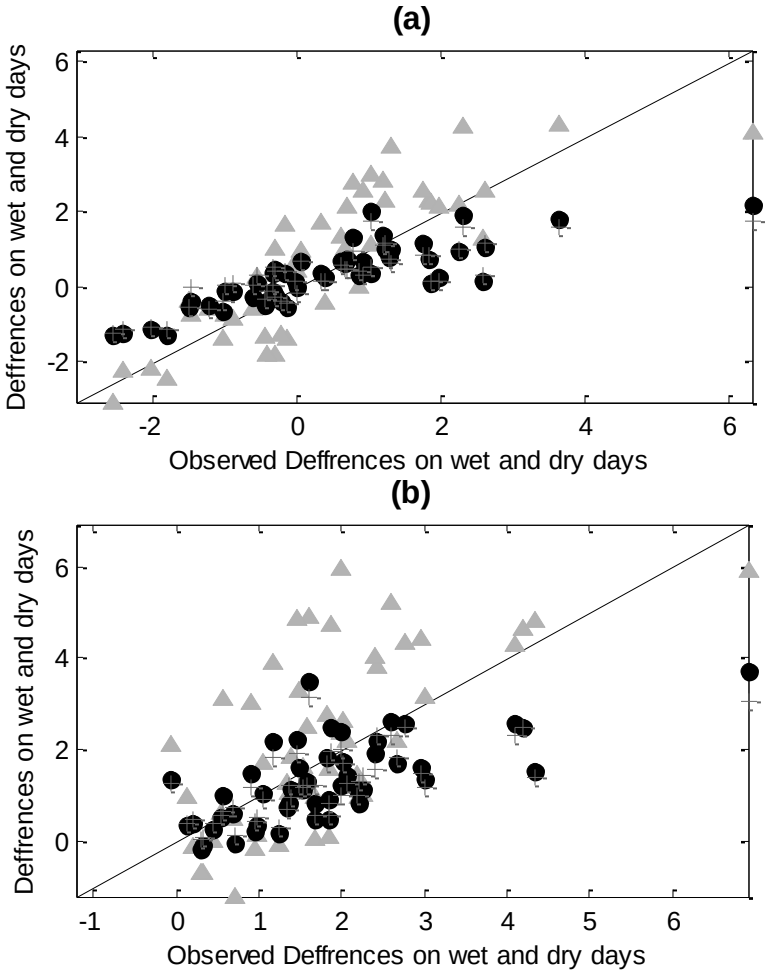

Figure 8. Observed versus modeled differences of daily maximum temperatures on wet days and dry days in (a), as well as differences of daily minimum temperatures in (b) for PGCR (black dots), MMLR (gray triangles) and MMSDM (gray plus) for all stations and all months during the validation period.

CHAPITRE 3: PROBABILISTIC MULTISITE STATISTICAL DOWNSCALING FOR DAILY PRECIPITATION USING A BERNOULLI–GENERALIZED PARETO MULTIVARIATE AUTOREGRESSIVE MODEL

Probabilistic multisite statistical downscaling for daily precipitation using a Bernoulli-Generalized Pareto multivariate autoregressive model

M. A. Ben Alaya[1], F. Chebana[1] and T.B.M.J. Ouarda[2, 1]

[1]*INRS-ETE, 490 rue de la Couronne, Québec (QC),*

Canada G1K 9A9

[2] Institute Center for Water and Environment (iWATER), *Masdar Institute of science and technology, P.O. Box 54224, Abu Dhabi, UAE*

[*]**Corresponding author:** Tel: +1 (418) 654 2530#4468

Email: moahammed_ali.ben_alaya@ete.inrs.ca

Accepted November 13, 2014

(Journal of climate)

Abstract

A Bernoulli-Generalized Pareto multivariate autoregressive (BMAR) model is proposed in this paper for multisite statistical downscaling of daily precipitations. The proposed model relies on a probabilistic framework in order to describe the conditional probability density function of precipitation at each station for a given day and handles multivariate dependence in both time and space using a multivariate autoregressive model. In a probabilistic framework, BMAR employs a regression model whose outputs are parameters of the mixed Bernoulli-Generalized Pareto distribution. As a stochastic component, the BMAR employs a latent multivariate autoregressive Gaussian field to preserve lag-0 and lag-1 cross-correlations of precipitation at multiple sites. The proposed model is applied for the downscaling of AOGCM data to daily precipitation in the southern part of Quebec, Canada. Reanalysis products are used in this study to assess the potential of the proposed method. Based on the mean errors (ME), the root mean square errors (RMSE), precipitations indices, and the ability to preserve lag-0 and lag-1 cross-correlation, results of the study indicate the superiority of the proposed model over a multivariate multiple linear regression (MMLR) model and a multisite hybrid statistical downscaling procedure that combines MMLR and a stochastic generator schemes.

Keywords: Statistical downscaling, Bernoulli-Generalized Pareto distribution, Vector generalized linear model, Multisite daily precipitation, Multivariate autoregressive Gaussian field, Spatio-temporal dependence.

1. Introduction

Stochastic weather generators are statistical models designed to provide realistic random sequences of atmospheric variables such as precipitation, temperature and wind speeds (see, e.g., Wilks and Wilby 1999). In particular, precipitation poses a number of challenges including, for instance, its spatio-temporal intermittence, its highly skewed distribution and its complex stochastic dependencies. For example, maintaining realistic relationships between precipitations at several sites is particularly important in hydrology. Indeed, streamflow depends strongly on the spatial distribution of precipitation in a watershed, and generated precipitations can be entered directly into a hydrological model to estimate streamflow in a given watershed (Xu 1999). A large number of stochastic precipitation models have been proposed in the literature, including resampling based approaches (e.g. Buishand and Brandsma 2001), hidden Markov models for occurrence (e.g. Robertson et al. 2004) and for intensity (e.g. Charles et al. 1999), power transformation to normality (e.g. Yang et al. 2005), copula-based approaches (e.g. Bárdossy and Pegram 2009) or artificial neural networks (e.g. Cannon 2008). Wilks and Wilby (1999) and Baigorria and Jones (2010) provided an overview of precipitation models.

At a single site, a commonly used approach for modelling precipitation involves a two-stage model that simulates the occurrence of wet and dry days before simulating precipitation amounts. To preserve the local properties of precipitation (i.e., marginal distributions and temporal correlation), a number of variations of this general method have been proposed in the literature. From a two-state representing wet and dry days, first or higher order Markov chains are commonly used to generate the occurrence process. However this approach may underestimate the observed occurrence of prolonged droughts (Katz and Parlange 1998). Alternatively, wet and dry spell lengths could be simulated alternately from distributions fitted to corresponding

observed records (Racsko et al. 1991). Once the days with occurrence of precipitation have been determined, the precipitation amount on wet days can be generated from a statistical distribution fitted to observed precipitation amounts. The short term autocorrelation of precipitation amounts has been modeled by a parametric autocorrelation function (e.g. Katz and Parlange 1998), by an autoregressive process (e.g. Hutchinson 1995), or more recently using a copula framework (e.g. Serinaldi 2009; Li et al. 2013a).

To properly account for the stochastic dependence between sites, a number of techniques employed two spatial models, one for precipitation occurrence process and one for the precipitation amount (e.g. Jeong et al. 2012). To avoid splitting occurrence and amount processes, Bardossy and Plate (1992) employed a censored power-transformed Gaussian distribution. In the same context, Ailliot et al. (2009) combined the latter with hidden Markov model for daily precipitation. An alternative solution to avoid the split between occurrence and amount process when reproducing the stochastic dependence structure is to use uniform marginal distributions of a meta-Gaussian random field. The latter can also be employed in an autoregressive form in order to reproduce both spatial dependence and short term autocorrelation. This last procedure avoids a sequential simulation conditioned on the simulation of the rainfall random fields at the previous time steps. Serinaldi and Kilsby (2014) opted for combining this autoregressive random field with a generalized additive model whose outputs are parameters of the at-site mixed discrete-continued marginal distribution of the precipitation process. This modular structure is mathematically rich because it offers a simple way to generate space-time evolution of discrete-continues variables. Furthermore, it can be adapted to different areas as well as introducing exogenous forcing covariates. Thus, making it a valuable tool in hydrometeorology and climate research analyses where often non-normally distributed random

variables, like precipitation, wind speed, cloud cover, humidity, are involved. After recent successful applications, in simulating daily rainfall fields over large areas (Serinaldi and Kilsby 2014), and modeling radar rainfall uncertainties (Villarini et al. 2014), it is very likely that this modular structure will have growing impact in climate downscaling applications where stochastic weather generators are routinely adapted for these purposes. In this context, the aim of the present paper is to propose a multisite probabilistic regression-based model that adapts this approach for daily precipitations downscaling.

Downscaling techniques have been developed to refine Atmosphere-Ocean Global Climate Models (AOGCMs) data and to provide information at more relevant scales. These techniques include dynamic downscaling, which uses regional climate models (RCM) over a limited area, and statistical downscaling which considers statistical relationships between large-scale variables (predictors) and small-scale variables (predictands) (Wilby et al. 1998) and provide climate information at the equivalent of point climate observations (Wilby et al. 2002). Statistical downscaling techniques represent a good alternative to dynamic methods in the case of limited resources, because of their ease of implementation and their low computational requirements (Benestad et al. 2008; Maraun et al. 2010). Stochastic weather generators can be used for climate-change downscaling through appropriate adjustments to their parameters. These adjustments can be accomplished in two ways: (i) through imposed changes in the corresponding monthly statistics, (ii) or by controlling the generator parameters by daily variations in simulated atmospheric circulation (Wilks 2010). The considered approach in the present study focuses on the second way, since the modular structure proposed by Serinaldi and Kilsby (2014) allows the introduction of exogenous forcing covariates.

Precipitation is one of the most important predictand in a downscaling perspective. Maraun et al. (2010) provided an overview of downscaling precipitation techniques. An alternative to stochastic weather generators in statistical downscaling is to find a direct relationship between large scale predictors and local predictands using a transfer function in a regression framework. For example, a transfer function can include: multiple linear regression (Wilby et al. 2002; Hammami et al. 2012; Jeong et al. 2012; Jeong et al. 2013), empirical orthogonal functions analysis (Huth 2004), canonical correlation analysis (Palutikof et al. 2002; Huth and Pokorná 2004), artificial neural networks (Schoof and Pryor 2001), singular value decomposition (Widmann et al. 2003), generalized linear model (GLM) (Beecham et al. 2014) and generalized additive model (GAM) (Levavasseur et al. 2011). Regression models are successfully used in downscaling, but their major drawback is that they generally reproduce the mean or the central predictions conditional to the selected predictors. Therefore, regression variability is always lower than the observed variability (Von Storch 1999). In addition, Wilby et al. (2003) mentioned that regression-based approaches show difficulty to preserve spatial dependence among multisite precipitations.

To correctly estimate the temporal variability in a regression model, three main approaches have been proposed in the literature: inflation (Huth 1999), randomization (Von Storch 1999; Clark et al. 2004) and expansion (Burger and Chen 2005). Inflation is usually performed by multiplying the downscaled data by a constant factor, but in this case the spatial correlations between sites can be misrepresented. Randomization consists in adding a random noise. In this way, regression and unconditional resampling techniques can be combined in a single hybrid model which can overcome weaknesses of both approaches (Jeong et al. 2012). However, Burger and Chen (2005) indicated that hybrid approach based on a static noise failed to represent local changes in

atmospheric variability in a climate change simulation, which is well explained using expended downscaling (Bürger 1996). Expanded downscaling is applied to multisite predictands by constraining the covariance matrix of the predicted series to be equal to the observed covariance matrix (Cannon 2009). On the other hand, Von Storch (1999) suggested that the inflation and expansion approaches are inappropriate techniques, because the implicit assumption that all local variability can be traced back to the large-scale is improper and is not the case in reality.

Given the drawbacks of these three existing techniques to reproduce the observed temporal variability, it is relevant to build the whole conditional distribution in order to capture the variability of the process. In this regard, probabilistic regression approaches have provided useful contributions in downscaling applications. Probabilistic approaches include: Bayesian formulation (Fasbender and Ouarda 2010), qunatile regression (Bremnes 2004; Friederichs and Hense 2007; Cannon 2011) and regression models where outputs are parameters of the conditional distribution. The last regression approach includes the vector form of generalized linear model (VGLM), the vector form of the generalized additive model (VGAM) (Yee and Wild 1996; Yee and Stephenson 2007) and conditional density estimation network (Williams 1998; Li et al. 2013b). Probabilistic regression approaches have been extended to multisite downscaling by Cannon (2008), following the methodology used in expanded downscaling. But this method is based on the assumption that all spatial dependence structures could be reproduced using synoptic scale atmospheric predictors. Alternatively, Ben Alaya et al. (2014) proposed a Probabilistic Gaussian Copula Regression (PGCR) model for multisite and multivariable downscaling. However, the PGCR model does not take into account cross-correlations lagged in time.

The aim of the present paper is to propose a multisite probabilistic regression-based downscaling model for daily precipitations, namely, Bernoulli-Generalized Pareto multivariate autoregressive (BMAR) model. BMAR specifies the conditional marginal distribution of precipitation for each site through AOGCM predictors, by using a VGLM whose outputs are parameters of the Bernoulli-Generalized Pareto distribution. Thus, with this component, the BMAR is able to model the occurrence and the amount of precipitation simultaneously and reproduce the observed temporal variability. In addition, a latent meta-Gaussian autoregressive random field is employed by the BMAR as a stochastic component to extend the probabilistic modeling framework in multisite downscaling tasks. This component allows the BMAR model to reproduce the observed spatial relationships between sites (such as the observed lag-0 and lag-1 cross-correlations), and to randomly generate realistic synthetic precipitation series.

The present paper is structured as follows: After a brief presentation of the multisite hybrid statistical downscaling model of Jeong et al. (2012) as a classical model to compare, the proposed BMAR model is presented. The BMAR model is then applied to the case of daily precipitations in the southern part of the province of Quebec, Canada. Reanalysis data are used in order to assess the potential of the proposed method. After the calibration of the BMAR model, an independent dataset is used to assess the downscaling quality. Based on statistical criteria and climatic indices that describe the frequency, intensity and duration of precipitation, results are compared with those obtained using a multisite hybrid model of Jeong et al. (2012) and the multivariate multiple linear regression model (MMLR). Finally a discussion and conclusions are given.

2. Data and study area

The study area is located in Quebec, in the latitudes between 45 ° N and 60 ° N and the longitudes between 60 ° W and 80 ° W. Nine series of observed daily precipitations (see Figure 1) are selected as predictands. These series, are provided by Environment Canada's hydro-meteorological network, have been rehabilitated by Mekis and Hogg (1999) and cover the period from 1 January 1961 to 31 December 2000. Table 1 reports the names and latitude-longitude locations of the nine selected meteorological stations. These stations are mapped in Figure 1 with respect to their numbers as in Table 1.

The reanalysis data from the National Center for Environmental Prediction (NCEP)/ National Center for Atmospheric Research (NCAR) over the period 1961-2000 (Kalnay et al. 1996; Kistler et al. 2001) are used to evaluate the potential of the downscaling method. NCEP / NCAR data are averaged on a daily basis from 6-hour data on the original regular grid of 2.5 ° latitude and longitude. Obtained predictors are then linearly interpolated on the CGCM3 Gaussian grid (3.75 ° latitude and longitude) and normalized to the reference period 1961-1990. The study area is covered by six grid points (see Figure 1), and for each grid point, 25 NCEP predictors are provided (see Table 2). For each day, 150 predictors are thus available. In order to reduce the number of predictors, a principal component analysis (PCA) is employed and the first 40 components that preserve more than 97% of the variance of the original NCEP predictors are then preserved as predictor variables. Finally, data from 1961 to 1990 are used for the calibration, whereas data from 1991 to 2000 are used for the validation.

3. Methodology

The multisite hybrid downscaling model of Jeong et al. (2012) and the BMAR model are presented in section 3.1 and section 3.2 respectively. The probabilistic framework for the BMAR model is presented with a description of the conditional Bernoulli-Generalized Pareto distribution. Then, a simulation procedure is presented using a latent multivariate autoregressive Gaussian field to reproduce the dependence structure of precipitations at multiple sites.

3.1. Multisite hybrid statistical downscaling of Jeong et al. (2012)

Let $\mathbf{X}$ denote a multiple atmospheric predictor variables matrix of dimension $n \times l$ and $\mathbf{Y}$ a multivariate predictand variables matrix of dimension $n \times m$. The linear relationship between the two matrices $\mathbf{X}$ and $\mathbf{Y}$ can be defined using the following MMLR equation:

$$\mathbf{Y} = \mathbf{X} \times \mathbf{W} + \mathbf{E} \tag{1}$$

where $\mathbf{W}$ is parameter matrix of dimension $l \times m$, and $\mathbf{E}$ is the residual matrix of dimension $n \times m$. The parameter matrix $\mathbf{W}$ can be estimated using the Ordinary Least Squares (OLS) method which is given by:

$$\hat{\mathbf{W}} = (\mathbf{X}^{\mathbf{T}}\mathbf{X})^{-1}\mathbf{X}^{\mathbf{T}}\mathbf{Y} \tag{2}$$

3.1.1. Precipitation occurrences

Let $\mathbf{O}[n \times m]$ be the observed binary (0 or 1) matrix of precipitation occurrence. For a given day $i = 1, 2, \ldots, n$, and a given site $j = 1, 2, \ldots, m$, an element O_{ij} of the matrix $\mathbf{O}$, is equal to 0 for a

dry day and 1 for a wet day. The matrix of the downscaled deterministic series of daily precipitation occurrence probabilities $\hat{\mathbf{O}}$ can be modeled using the following MMLR equation:

$$\hat{\mathbf{O}} = \hat{a}_0 + \mathbf{X}\hat{a} \tag{3}$$

Were $\hat{a}_0\,[n \times m]$ and $\hat{a}\,[l \times m]$ are estimated MMLR parameters. The residual matrix $\mathbf{E_O}\,[n \times m]$ of this MMLR model is given by:

$$\mathbf{E_O} = (\mathbf{O} - \hat{\mathbf{O}}) \tag{4}$$

To reproduce the observed temporal variability and spatial dependency, a multivariate normal distribution having error variances $\mathbf{H_O}$ and correlation matrix $\mathbf{C_O}$ equal to that of the residual matrix $\mathbf{E_O}$ is used to obtain generated residuals $\tilde{\mathbf{E}}_\mathbf{O}\,[n \times m]$. Generated residuals $\tilde{\mathbf{E}}_\mathbf{O}$ are then added to the downscaled probability matrix to obtain the generated continuous probability matrix $\tilde{\mathbf{O}}$ as:

$$\tilde{\mathbf{O}} = \hat{\mathbf{O}} + \tilde{\mathbf{E}}_\mathbf{O} \tag{5}$$

Then, to transform the matrix $\tilde{\mathbf{O}}$ to a downscaled binary series $\dot{\mathbf{O}}$, Jeong et al. (2012c) employed a first-order Markov chain model. Let $\dot{\mathbf{O}}_{ij}$ be a [0 or 1] downscaled binary value of precipitation occurrence $\dot{\mathbf{O}}$ at a location j and on a day i. The value of $\dot{\mathbf{O}}_{ij}$ is then written as below:

$$\dot{O}_{ij} = \begin{cases} 1, & \text{if } \tilde{O}_{ij} \geq \mathbf{\Phi}^{-1}(j)[1 - p_{01}(j)] \text{ and } \dot{O}_{i-1\,j} = 0 \\ 1, & \text{if } \tilde{O}_{ij} \geq \mathbf{\Phi}^{-1}(j)[1 - p_{11}(j)] \text{ and } \dot{O}_{i-1\,j} = 1 \\ 0, & \text{otherwise} \end{cases} \tag{6}$$

where $\Phi(j)$ is the normal cumulative distribution function having mean and standard deviation equal to that of the time series of $\tilde{O}$ at the site j. p_{01} is the probability of a wet day following a dry day and p_{11} is the probability of a wet day following a wet day. These transition probabilities p_{11} and p_{01} are estimated separately for each observation site. Jeong et al. (2012) mentioned that the transformed binary series $\dot{\mathbf{O}}$ cannot represent the original multisite cross-correlation. For this reason, they employed empirical relationships of cross-correlations between binary series ($\varphi(j,s)$) and continuous series $(\zeta(j,s))$ at any locations j and s using a simple power function expressed as:

$$\zeta(j,s) = c \times \varphi(j,s)^d \tag{7}$$

The parameters c and d in Eq.(7) have been estimated by minimising RMSE among all $m(m-1)/2$ pairs of cross site correlation coefficients in the observed binary series and transformed binary series $\dot{\mathbf{O}}$.

3.1.2 Precipitation amount

The gamma distribution has been fitted to rainfall amounts in a number of studies (Stephenson et al. 1999; Giorgi et al. 2001; Yang et al. 2005). Thus, before developing a regression-based precipitation amount model, Jeong et al. (2012) employed the Anscombe transformation $R_{ij} = Y_{ij}^{1/3}$ on a day i at a site j, to transform the precipitation amount vector $\mathbf{Y_j}$ for a site j into a normal distribution (Terrell 2003; Yang et al. 2005). Then, using the MMLR model, the transformed precipitation amount matrix $\mathbf{R}[n \times m]$ can be modeled using the following equation:

$$\hat{\mathbf{R}} = \hat{b}_0 + \mathbf{X}\hat{b} \tag{8}$$

where $\hat{\mathbf{R}}\,[n \times m]$ is the downscaled deterministic series of Anscombe residuals matrix. Constant term matrix $\hat{b}_0\,[n \times m]$ and the parameter matrix $\hat{b}\,[k \times m]$ are estimated MMLR parameters using the OLS method. The residual matrix of the deterministic series of daily precipitation amounts $\mathbf{E_R}\,[n \times m]$ can be described by:

$$\mathbf{E_R} = (\mathbf{R} - \hat{\mathbf{R}}) \tag{9}$$

Thereafter, the residual matrix $\tilde{\mathbf{E}}_\mathbf{R}\,[n \times m]$ is generated from multivariate normal distribution having error variances $\mathbf{H_R}$ and correlation matrix $\mathbf{C_R}$ equal to that of the residual matrix $\mathbf{E_R}$. To reproduce at-site variances and multisite cross-correlations the generated residual matrix $\tilde{\mathbf{E}}_\mathbf{R}$ is added to the matrix $\hat{\mathbf{R}}$ as follows:

$$\tilde{\mathbf{R}} = \hat{\mathbf{R}} + \tilde{\mathbf{E}}_\mathbf{R} \tag{10}$$

The generated precipitation amounts are calculated as:

$$\tilde{Y}_{ij} = \tilde{R}_{ij}^3 \tag{11}$$

Thereby, the generated precipitation series in $\dot{\mathbf{Y}}$ are obtained by calculating the product of the generated precipitation occurrence and the generated precipitation amount $[\dot{Y}_{ij} = \dot{O}_{ij} \times \tilde{Y}_{ij}]$. Finally, because the generated series in $\dot{\mathbf{Y}}$ present in general different statistical properties than those of the observed precipitation amount series, and because the residual matrix $\mathbf{E_R}$ of each site may be

not normally distributed, Jeong et al. (2012) adopted a probability distribution mapping technique to adjust generated precipitation amount.

3.2. BMAR model

In most applications, regression based models are performed to reproduce the mean or the central part of predictands conditional on a set of selected predictors. The resulting model defines a mapping from predictors to predictand variables. This mapping is more suitable if predictions are generated from a deterministic function that is corrupted by a normally distributed noise process with constant variance (Cannon 2008).

For precipitation, the normality assumption might not be feasible on short time scales. At a daily time scale, precipitations are more skewed and commonly modeled with a Gamma distribution (Stephenson et al. 1999; Giorgi et al. 2001; Yang et al. 2005). To handle such situations, the GLM extends linear regression to model conditional mean of variables that may follow a wide class of distributions, such as the Gamma distribution (Coe and Stern 1982; Stern and Coe 1984; Chandler and Wheater 2002). However, the Gamma distribution may not be flexible enough to capture all rainfall amount behaviors and can be heavy tailed at some sites. Wan et al. (2005) showed that a mixed exponential distribution outperformed the Gamma distribution in a part of Canada. In general, other alternatives are needed to model extreme amounts such us Weibull (WEI) distribution or Generalized Pareto (GP) distribution. Note that these two distributions cannot be used directly in a GLM. For this purpose, vector GLMs (VGLMs) have been proposed (Yee and Stephenson 2007). Instead of the conditional mean of a distribution, an appropriate probability density function (PDF) is selected, and then a linear regression model is employed where outputs are vectors of parameters corresponding to this selected PDF. Thus, in a

probabilistic regression framework, VGLM is able to build the whole conditional distribution (Kleiber et al. 2012). In addition, it has a particular advantage in downscaling applications where it is able to recapture the variability of the process.

3.2.1. Bernoulli-GP regression

According to the above literature review, four distributions were considered for precipitation amount: Gamma, mixed Exponential, GP and WEI. The modular structure employed in this paper allows selecting a suitable distribution for each station. Nevertheless, we simplify the procedure by using only one distribution that satisfies the fitting for all stations. To select the suitable distribution, we compare the performance of each one in reproducing the observed distribution of precipitation amount on wet days for the nine precipitation stations. This type of comparisons could be performed by examining the Q-Q plots. Based on the Q-Q plot visualisation (see Figure 2), for all stations, all the considered distributions seem to be able to correctly model precipitation amounts. However, the GP distribution systematically outperforms the other three models on the upper tail of the distribution. Due to the importance of reproducing extremes in downscaling applications, GP is chosen in this paper. Therefore, a mixed Bernoulli–GP distribution is employed in this paper to model precipitation series that includes both occurrences and amounts in a single distribution. The Bernoulli–GP PDF is given by:

$$f(y;\rho,\alpha,\beta)=\begin{cases} 1-\rho & \text{if} \quad y = 0 \\ \rho\left[1-\left(1+\beta\frac{y}{\alpha}\right)^{-1/\beta}\right] & \text{if} \quad y > 0 \end{cases} \tag{12}$$

where y is the precipitation amount, α ($\alpha > 0$) and β (where $1+\beta y/\alpha > 0$) are respectively the scale and the shape parameters of the zero-adjusted GP model, and ρ ($0 \le \rho \le 1$) is the probability of precipitation.

Using the VGLM, the parameters of this Bernoulli-GP distribution are considered to change from one day to another according to the value of large-scale atmospheric predictors. Only the shape parameter β_j for a site j is fixed in time, to guarantee the convergence of the maximum likelihood estimates. For the parameter of the probability of precipitation occurrences we adopt a logistic regression which is written as:

$$\rho_j(t) = \frac{1}{1+\exp\left[-c_j^T x(t)\right]} \tag{13}$$

where $\rho_j(t)$ is the probability of precipitation occurrence at a site j on a day t, c_j is the coefficient of the logistic model, and $x(t)$ is the value of the predictors at the day t. The scale parameters $\alpha_j(t)$ are given by:

$$\alpha_j(t) = \exp\left[d_j^T x(t)\right] \tag{14}$$

where d_j are the coefficients of the model. Hence, the conditional Bernoulli-GP density function for the precipitation $y_j(t)$ on a day t and at site j is given by:

$$f_{tj}[y_j(t) \mid x(t)] = \begin{cases} 1-\rho_j(t) & \text{if} \quad y_j(t) = 0 \\ \rho_j(t)\left[1-\left(1+\beta_j \dfrac{y_j(t)}{\alpha_j(t)}\right)^{-1/\beta_j}\right] & \text{if} \quad y_j(t) > 0 \end{cases} \tag{15}$$

Figure 3.a shows the steps involved in training the proposed Bernoulli-GP regression model given the calibration data. The coefficients c_j, d_j and β_j for all sites are obtained following the method of maximum likelihood by minimizing the negative log predictive density (NLPD) cost function (Haylock et al. 2006; Cawley et al. 2007; Cannon 2008):

$$\mathscr{L}_j = \sum_{t=1}^{n} \log\left\{ f_j\left[y_j(t) \mid x(t) \right] \right\} \text{ for } \quad j = 1, \ldots, m \tag{16}$$

via the simplex search method of Lagarias et al. (1999). This is a direct search method that does not use numerical or analytic gradients.

3.2.2. Conditional simulation using a latent multivariate autoregressive Gaussian field.

Once the proposed Bernoulli-GP regression model has been trained, it can be used to estimate the Bernoulli-GP parameters ($\rho_j(t), \alpha_j(t)$ and β_j) for each site j and for a given day t when we have the AOGCM predictors. Then, it is possible to create synthetic predicted series of precipitations by sampling in the obtained Bernoulli-GP distribution for each day. In this step, it is important to maintain realistic spatio-temporal intermittence of precipitation at multiple sites. In this paper, the dependence structure between multisite precipitations is reproduced by assuming a multivariate first-order autoregressive model (MAR(1)) for a multivariate latent Gaussian process $z(t) = \left[z_1(t), \ldots, z_m(t) \right]$, through the following equations:

$$z_j(t) = h_{tj}(y_j(t)) \text{ where } h_{tj}(y_j(t)) = \Phi^{-1}[F_{tj}(y_j(t))] \tag{17}$$

Where Φ is the standard normal cumulative distribution function and F_{tj} is the Bernoulli-GP cumulative density function at time t and site j. Figure 3.b shows the steps involved in obtaining the latent multivariate Gaussian variables over the calibration period. Based on the

fitted conditional Bernoulli-GP parameters, the Bernoulli-GP cumulative density function is used to express precipitation amounts as cumulative probabilities ranging from 0 to 1. In order to map $z_j(t)$ onto the full range of the normal distribution, the cumulative probabilities $F_{ij}(y_j(t))$ are randomly drawn from a uniform distribution on $[0, 1-\rho(t)]$ for dry days. Finally, to obtain the set of the latent variables, data are normalized by applying the standard normal inverse cumulative density function to the series of cumulative probabilities.

Let $\mathbf{Z_t} = (z_{1t}, z_{2t}, \ldots, z_{mt})^T$ denote the obtained latent Gaussian vector of z values at the m sites at time $t = 1, 2, \ldots n$ after the normalisation step. The latent multivariate first-order autoregressive model for $\mathbf{Z_t}$ is given by:

$$\mathbf{Z_t} = \mathbf{A}\mathbf{Z_{t-1}} + \mathbf{B}\boldsymbol{\varepsilon}_\mathbf{t} \tag{18}$$

where $\mathbf{A}$ and $\mathbf{B}$ are $(m \times m)$ parameter matrices, and $\boldsymbol{\varepsilon}_\mathbf{t}$ is a random $(m \times 1)$ noise vector with a standard multivariate normal distribution. The method of moment estimators of the MAR(1) model are given by Bras and Rodríguez-Iturbe (1985):

$$\hat{\mathbf{A}} = L_1 L_0^{-1} \tag{19}$$

$$\hat{\mathbf{B}}\hat{\mathbf{B}}^\mathbf{T} = L_0 - L_1 L_0^{-1} L_1^T \tag{20}$$

where L_0 is the sample lag-0 cross covariance matrix and L_1 is the sample lag-1 covariance matrix. L_0 and L_1 can be estimated in a pairwise manner. An element (k, s) in L_0, can be estimated by noting that the joint distribution of the z-variables at sites k and s is a bivariate normal, and the correlation coefficient is the only unknown parameter. The elements of L_1 can be

estimated using this similar procedure. The matrix $\hat{\mathbf{A}}$ can be obtained from Eq.(19) and the matrix $\hat{\mathbf{B}}$ can be obtained from Equation (20) using for example Cholesky decomposition (Rasmussen 2013).

Then, for a new day t' from the validation period, it is possible to randomly generate $z(t') = [z_1(t'), \ldots, z_m(t')]$ using the fitted MAR(1) model. Figure 4 illustrates how the Bernoulli-GP regression model and the MAR(1) model are combined to produce one simulation at a day t'. The value of the synthetic precipitation time series $\hat{y}_j(t')$ is given by $F_{t'j}^{-1}[u_j(t')]$, where $u_j(t')$ are obtained by applying the standard normal cumulative distribution to the generated $z_j(t')$ at site j using the MAR(1) model and $F_{t'j}^{-1}$ is the inverse cumulative Bernoulli-GP distribution whose parameters are obtained from the probabilistic regression model on day t' and at site j.

3.3 Quality assessment of downscaled precipitation

Data between 1991 and 2000 are used to assess the downscaling quality. In a first validation approach based on statistical criteria, BMAR results are compared to those obtained using the MMLR model and the hybrid model of Jeong et al. (2012). Two statistical criteria are used:

$$ME = \frac{1}{n}\sum_{t=1}^{n}\left(y_{obs_t} - y_{est_t}\right) \tag{21}$$

$$RMSE = \sqrt{\frac{1}{n}\sum_{t=1}^{n}\left(y_{obs_t} - y_{est_t}\right)^2} \tag{22}$$

where n refers to the number of observations, y_{obs_t} denotes the observed value, y_{est_t} is the estimated value, and t refers the day. The mean error (ME) is a measure of accuracy, whereas the

root mean square error (RMSE) is given by an inverse measure of the accuracy and must be minimized.

In a second validation approach, several precipitation indices defined in Table 3 are considered. For precipitation amounts, five indices are considered: the mean precipitation of wet day (MPWD), the 90^{th} percentile of daily precipitation (P90), the maximum 1-day precipitation (PX1D), the maximum 3-day precipitation (PX3D), and the maximum 5-day precipitation (PX5D). For precipitation occurrences, three indices are considered: the maximum number of consecutive wet days (WRUN), the maximum number of consecutive dry days (DRUN) and the number of wet days (NWD). The BMAR, the hybrid and the MMLR models are then compared by calculating the RMSE for each of the climatic indices for all sites.

4. Results

The BMAR model has been trained for the calibration period (1960-1990), using precipitation data series from the nine stations and the 40 predictors obtained by the PCA. All the coefficients c_j, d_j and β_j for each site were set following the maximum likelihood estimator. Once the parameters of the conditional Bernoulli-GP distribution ($\rho_j(t), \alpha_j(t)$) have been estimated for each day t and for each site j over the calibration period, all the obtained conditional marginal distributions were used to obtain the latent variables $z(t)$ as shown in Figure 3.b and to fit the parameters of MAR(1) model. Finally, all the fitted BMAR parameters where used to generate precipitation series during the validation period (1991-2000) as shown in Figure 4. Figure 5 shows an example of the obtained result using the BMAR model for precipitation at Chelsea station during the year 1991. Figure 5.a shows the estimated series of the probability of precipitation occurrences, and Figure 5.b shows both synthetic and observed precipitation series.

We can see that the BMAR model provides interesting results for both precipitation amounts and precipitation occurrences.

Application results of the BMAR model are compared to those of both the hybrid and MMLR models. To explain the abilities of the stochastic weather generating scheme in the hybrid model, the MMLR model here is employed without stochastic variation. Note that, for the MMLR model, the wet day was determined when the deterministic series of the daily probability of precipitation occurrence by the MMLR occurrence model was larger than the threshold value of 0.5. The BMAR and the hybrid models give probabilistic predictions, thus, for stability and robustness of both BMAR and the hybrid model, 100 realizations are generated of the precipitation series.

For each station, values of RMSE and ME for the three models are given in Table 4. The RMSE and ME for both BMAR and hybrid were calculated using the conditional mean for each day. From Table 4 it can be seen that, for all stations, BMAR shows the best performance, since it has lower RMSE and close to zero ME compared to both the hybrid and MMLR. These results demonstrate the effectiveness of the conditional Bernoulli-GP regression component in BMAR to adequately replicate the observed series. Table 4 indicates also that the hybrid model performs better than MMLR in terms of both RMSE and ME. This result is expected due to the fact that the MMLR model is in reality biased because zero precipitation amounts were included to calibrate the MMLR amount model. In addition, the anscombe residuals R from the observed precipitation amount may not be exactly normally distributed. For this reason, the hybrid model employs a probability mapping technique to correct this bias. However, in terms of ME and RMSE, BMAR not only performs better than the hybrid model but also it has the advantage of its automatic aspect of mapping in the conditional distribution of precipitation using its probabilistic

regression component. Thus, there is no need to rely on transformation steps or on bias correction procedures (such us a probability mapping technique) when evaluating the BMAR model.

Figure 6 summarises the RMSE of downscaled precipitation climatic indices for each model over the nine weather stations during the validation period (1991–2000). The RMSE of both the BMAR and hybrid models are calculated using the mean of 100 realisations. For all precipitation amounts indices, it can be seen that in terms of RMSE, BMAR performs better than the two other models for all stations. Therefore, the use of the GP distribution allows the BMAR model to better reproduce observed monthly characteristics of precipitation amounts. It can also be noted that the hybrid model gives better results compared to the MMLR model for all stations, except for Pmax90 indices for which it improves the results only for stations 2-Cedars, 4-Drummondville, 5-Donnacona and 7-Bagotville A.

Results of downscaled precipitation occurrence indices are presented in Figure 6.f. for WRUN indices. These results indicate that both the BMAR and hybrid models outperformed the MMLR model in terms of the RMSE for WRUN indices over the all stations. On the other hand, for the same indices BMAR is slightly better than the hybrid. Although, Figure 6.g shows that for all stations BMAR outperformed the two others models in terms of RMSE of NWD indices, and the hybrid gives better results than MMLR. Finally, for DRUN indices, the same conclusion can be deduced from Figure 6.h, except for Nicolet station, where the hybrid model is slightly better than the BMAR model. Thereby, based on WRUN, NWD and DRUN, it can be concluded that the logistic regression in the BMAR model does an overall good job in representing the monthly characteristics of precipitation amounts.

To evaluate the ability of the multivariate autoregressive component in BMAR to reproduce the observed dependence structures in both time and space, scatter plots (Figure 7a-b) of the lag-0 and lag-1 cross-correlation of modeled versus observed precipitations are plotted for the three models. The correlation values of both BMAR and hybrid models are obtained using the mean of the correlation values calculated from a 100 realisations. For lag-0 cross-correlation, points correspond to all 36 combinations of pairs of stations, while for lag-1 cross-correlation, points correspond to all 81 combinations because lag-1 cross-correlations are generally not symmetric. It can be seen that MMLR generally overestimates the cross-correlation of both lag-0 and lag-1, with an RMSE equal to 0.3184 for lag-0 cross-correlation and 0.1063 for lag-1 cross-correlation. MMLR gives the poorest results compared to BMAR and hybrid. This finding is expected since MMLR is not a multisite model. Figure 7.a shows that the BMAR and hybrid models preserve the lag-0 cross-correlation adequately. On the other hand, BMAR outperformed the hybrid in terms of RMSE. Finally, from Figure 7.b it can be seen that BMAR reproduces more adequately the lag-1 cross-correlation than the hybrid model and values of RMSE confirm this result, since they are equal to 0.0495 for BMAR and 0.0795 for the hybrid. In fact, the hybrid model, by its construction, is only able to take into account the lag-1 autocorrelation, unlike BMAR which is assumed to preserve the full lag-1 cross-correlation.

Finally, joint probabilities of the events that two sites are both dry or both wet on a given day are displayed in Figure 8. The BMAR adequately simulates these joint probabilities and outperforms both hybrid and MMLR models that provide overestimates of these probabilities. When dealing with “wet”, the three models provide underestimates. Nevertheless, the BMAR gives better results. In reality, as described in Section 3.2.2, in the step of obtaining the latent variables, the cumulative probabilities $F_{tj}(y_j(t))$ are randomly drawn from a uniform distribution on [0,

$1-\rho(t)$]. This implies that this part of the joint distribution is free from having any spatial correlation structure. However, the autoregressive component can indirectly reproduce a part of the spatial dependence structure in this part of the joint distribution, because a value of the generated latent variable for dry days depends on previous days that may depend on generated values in other sites. Nevertheless, to circumvent this problem Ben Alaya et al. (2014) separated the two processes of the occurrence and amount by considering a latent variable for each process. However, as proposed in the present paper, taking into account the two processes simultaneously makes the model more parsimonious, since the number of the latent variables is reduced.

5. Discussions

In general, regression-based downscaling mapping from coarse-scale predictors reproduces the mean of the process conditionally to the selected independent predictors. As a consequence, the variability of the regression is always smaller than the initial variability. Moreover, spatial dependency among multisite local predictand variables is not reproduced accurately by regression mapping from large-scale predictors (Burger and Chen 2005; Jeong et al. 2013). In this study, cross-site correlations of multisite precipitations are obviously over-estimated using the MMLR model and this over-estimation is evident that one cannot reproduce local-scale spatial dependency by simply using coarse NCEP/NCAR predictors. Therefore, the hybrid model of Jeong et al. (2012) provided a statistical generation procedure based on a randomization approach, in order to reproduce the unexplained temporal variability and the cross-site correlation of precipitation occurrence and amount among the observation sites. However, this hybrid procedure is based on a static noise model and failed to represent local changes in total precipitation variability in a climate change simulation.

Therefore, this study proposed the BMAR model which employs a stochastic generation procedure that considers only the dependency structures. Indeed, temporal variability can be preserved using the conditional distributions through the probabilistic regression. This attractive characteristic of the proposed BMAR model allows the model to correctly reproduce the observed temporal variability. Thereby, the elimination of the marginal effect helps to model and understand effectively the spatio-temporal dependency structures, as it has no relationship with the marginal behavior. To this end, the biggest challenge in the proposed method is to uniformly generate a multivariate distribution in the open interval (0,1) that preserves the spatio-temporal intermittences of several variables after the elimination of marginal distribution effects. Then, the estimation can be obtained by applying the inverse cumulative distribution function using the conditional distributions. Hence, the proposed solution can be considered as similar to a copula based framework (Chebana and Ouarda 2007; El Adlouni and Ouarda 2008). A copula is a multivariate distribution whose marginals are uniformly distributed on the interval [0,1]. In the proposed method the generation of random variables in the open interval (0, 1) is carried out through the latent Gaussian variables obtained by applying the transformation $h(\bullet)$ (Eq.(17)) to the multisite precipitation data. This same transformation is introduced in a Gaussian copula. Nevertheless, in a Gaussian copula, latent variables are modeled through a multivariate Gaussian distribution. In the present work, they are modeled through a multivariate Gaussian autoregressive model in order to include the spatio-temporal dependences, more precisely, the lag-1 cross-correlation. The Gaussian copula is employed by Ben Alaya et al. (2014) in a probabilistic regression model to preserve dependence structures in a multisite and multivariable downscaling perspective. Nevertheless, this approach is limited to preserve only the lag-0 cross-correlation. Therefore, the proposed BMAR model can be considered as an extension to the

Gaussian copula regression model framework to account for the lag-1 cross-correlation when the marginal distributions are specified using the Bernoulli-GP distribution.

As a direct consequence of the elimination of the marginal distribution effect when preserving dependence structures in underlying BMAR, it is straightforward to include the observed series of other variables and to extend the model to multivariable tasks. The extension of the BMAR by adding variables other than precipitation would require that appropriate distributions must be identified and incorporated into the VGLM. However, the stochastic generator component procedures remain the same. For example, for the temperature variable the normal distribution could be chosen, and for a normally distributed noise process with non-constant variance, the conditional density regression for the temperature variable would have two outputs: one for the conditional mean and one for the conditional variance.

The NCEP/NCAR data are used for calibration and validation of BMAR model. Even if NCEP data are complete and physically consistent, since they are basically interpolations of observational data based on dynamical model, they are subject to model biases (Hofer et al. 2012). NCEP variables which are not assimilated, but generated by the parameterizations based on dynamical model can significantly deviate from real weather. The use of such variables for the calibration and validation of empirical downscaling techniques may induce a significant deviation of the modeled relationships predictors/predictands from the reality. Thus, this makes evaluation of downscaling techniques more difficult. Therefore, the selection of appropriate large-scale atmospheric predictor variables for the proposed BMAR requires comprehensive consideration. In this way, studying the sensitivity of the BMAR model to NCEP predictors is important, not only for a better selection of predictors but also for a more realistic elaboration of future climate scenarios.

6. Conclusions

A Bernoulli-GP multivariate autoregressive model is proposed in this paper for simultaneously downscaling AOGCM predictors to daily precipitation at several sites. The BMAR relies on a probabilistic modeling framework in order to predict the conditional distribution of precipitation at a daily time scale using a VGLM applied to the discrete-continues Bernoulli-GP distribution. Prediction parameters of the Bernoulli-GP distribution allow: (i) modeling precipitation occurrences and precipitation amounts at the same time, (ii) dealing with the problem of non-normality of precipitation data and (iii) reproducing observed temporal variability. To allow a realistic representation of relationships between stations at both time and space, stochastic generators procedures where applied to the VGLM using a latent multivariate autoregressive Gaussian field.

The developed model was then applied to generate daily precipitation series at nine stations located in the southern part of the province of Quebec (Canada). NCEP reanalysis data were used as predictors in order to assess the potential of the method, although the final objective is to use AOGCM predictors. Application results of the BMAR model were compared to those obtained using the MMLR and the hybrid model. Results show that the BMAR model gives the best performance compared to the two models in terms of RMSE and ME. Moreover, the comparison based on precipitation indices show that the BMAR model is more able to reproduce precipitation amounts and occurrence characteristics on a seasonal basis. In addition, BMAR gives better preservation of the relationships between multisite precipitation at both time and space.

Model evaluations suggest that the BMAR model is capable of generating series with realistic spatial and temporal variability. In addition, the proposed model performed better than a multisite

hybrid regression-stochastic generator model for most verification statistics. The BMAR model may be a useful tool for multisite precipitation downscaling based on AOGCM data.

Acknowledgments

We gratefully acknowledge the comments of the Editor Joseph Barsugli, and two reviewers, Francesco Serinaldi and anonymous reviewer. We acknowledge Eva Mekis from Environment Canada for providing observed data sets of rehabilitated precipitation. The authors would like to acknowledge also the Data Access and Integration (DAI, see http://loki.qc.ec.gc.ca/DAI/login-e.php) team for providing the predictors data and technical support. The DAI data download gateway is made possible through collaboration among the Global Environmental and Climate Change Centre (GEC3), the Adaptation and Impacts Research Section (AIRS) of Environment Canada, and the Drought Research Initiative (DRI).

References

Ailliot, P., C. Thompson and P. Thomson (2009). "Space–time modelling of precipitation by using a hidden Markov model and censored Gaussian distributions." Journal of the Royal Statistical Society: Series C (Applied Statistics) **58**(3): 405-426.

Baigorria, G. A. and J. W. Jones (2010). "GiST: A stochastic model for generating spatially and temporally correlated daily rainfall data." Journal of Climate **23**(22): 5990-6008.

Bárdossy, A. and G. Pegram (2009). "Copula based multisite model for daily precipitation simulation." Hydrology and Earth System Sciences **13**(12): 2299-2314.

Bardossy, A. and E. J. Plate (1992). "Space-time model for daily rainfall using atmospheric circulation patterns." Water Resources Research **28**(5): 1247-1259.

Beecham, S., M. Rashid and R. K. Chowdhury (2014). "Statistical downscaling of multi-site daily rainfall in a South Australian catchment using a Generalized Linear Model." International Journal of Climatology.

Ben Alaya, M. A., F. Chebana and T. Ouarda (2014). "Probabilistic Gaussian Copula Regression Model for Multisite and Multivariable Downscaling." Journal of Climate **27**(9).

Benestad, R. E., I. Hanssen-Bauer and D. Chen (2008). Empirical-statistical downscaling, World Scientific.

Bras, R. L. and I. Rodríguez-Iturbe (1985). Random functions and hydrology, Courier Dover Publications.

Bremnes, J. B. (2004). "Probabilistic forecasts of precipitation in terms of quantiles using NWP model output." Monthly Weather Review **132**(1).

Buishand, T. A. and T. Brandsma (2001). "Multisite simulation of daily precipitation and temperature in the Rhine basin by nearest-neighbor resampling." Water Resources Research **37**(11): 2761-2776.

Bürger, G. (1996). "Expanded downscaling for generating local weather scenarios." Climate Research **7**(2): 111-128.

Burger, G. and Y. Chen (2005). "A Regression-based downscaling of spatial variability for hydrologic applications." Journal of Hydrology **311**: 299-317.

Cannon, A. J. (2008). "Probabilistic multisite precipitation downscaling by an expanded Bernoulli-gamma density network." Journal of Hydrometeorology **9**(6): 1284-1300.

Cannon, A. J. (2009). "Negative ridge regression parameters for improving the covariance structure of multivariate linear downscaling models." International Journal of Climatology **29**(5): 761-769.

Cannon, A. J. (2011). "Quantile regression neural networks: Implementation in R and application to precipitation downscaling." Computers & Geosciences **37**(9): 1277-1284.

Cawley, G. C., G. J. Janacek, M. R. Haylock and S. R. Dorling (2007). "Predictive uncertainty in environmental modelling." Neural Networks **20**(4): 537-549.

Chandler, R. E. and H. S. Wheater (2002). "Analysis of rainfall variability using generalized linear models: a case study from the west of Ireland." Water Resources Research **38**(10): 10-11-10-11.

Charles, S. P., B. C. Bates and J. P. Hughes (1999). "A spatiotemporal model for downscaling precipitation occurrence and amounts." Journal of Geophysical Research: Atmospheres (1984–2012) **104**(D24): 31657-31669.

Chebana, F. and T. B. M. J. Ouarda (2007). "Multivariate L-moment homogeneity test." Water Resources Research **43**(8).

Clark, M., S. Gangopadhyay, L. Hay, B. Rajagopalan and R. Wilby (2004). "The Schaake Shuffle: A Method for Reconstructing Space–Time Variability in Forecasted Precipitation and Temperature Fields." Journal of Hydrometeorology **5**(1): 243-262.

Coe, R. and R. Stern (1982). "Fitting models to daily rainfall data." Journal of Applied Meteorology **21**(7): 1024-1031.

El Adlouni, S. and T. Ouarda (2008). "Study of the joint distribution flow-level by copulas: Case of the Chateauguay river." Canadian Journal of Civil Engineering **35**(10): 1128-1137.

Fasbender, D. and T. B. M. J. Ouarda (2010). "Spatial Bayesian Model for Statistical Downscaling of AOGCM to Minimum and Maximum Daily Temperatures." Journal of Climate **23**(19): 5222-5242.

Friederichs, P. and A. Hense (2007). "Statistical downscaling of extreme precipitation events using censored quantile regression." Monthly Weather Review **135**(6).

Giorgi, F., J. Christensen, M. Hulme, H. Von Storch, P. Whetton, R. Jones, L. Mearns, C. Fu, R. Arritt and B. Bates (2001). "Regional climate information-evaluation and projections." Climate Change 2001: The Scientific Basis. Contribution of Working Group to the Third Assessment Report of the Intergouvernmental Panel on Climate Change [Houghton, JT et al.(eds)]. Cambridge University Press, Cambridge, United Kongdom and New York, US.

Hammami, D., T. S. Lee, T. B. M. J. Ouarda and J. Le (2012). "Predictor selection for downscaling GCM data with LASSO." Journal of Geophysical Research D: Atmospheres **117**(17).

Haylock, M. R., G. C. Cawley, C. Harpham, R. L. Wilby and C. M. Goodess (2006). "Downscaling heavy precipitation over the United Kingdom: A comparison of dynamical and statistical methods and their future scenarios." International Journal of Climatology **26**(10): 1397-1415.

Hofer, M., B. Marzeion and T. Mölg (2012). "Comparing the skill of different reanalyses and their ensembles as predictors for daily air temperature on a glaciated mountain (Peru)." Climate Dynamics **39**(7-8): 1969-1980.

Hutchinson, M. (1995). "Stochastic space-time weather models from ground-based data." Agricultural and Forest Meteorology **73**(3): 237-264.

Huth, R. (1999). "Statistical downscaling in central Europe: Evaluation of methods and potential predictors." Climate Research **13**(2): 91-101.

Huth, R. (2004). Sensitivity of local daily temperature change estimates to the selection of downscaling models and predictors. Boston, MA, ETATS-UNIS, American Meteorological Society.

Huth, R. and L. Pokorná (2004). "Parametric versus non-parametric estimates of climatic trends." Theoretical and Applied Climatology **77**(1): 107-112.

Jeong, D., A. St-Hilaire, T. Ouarda and P. Gachon (2013). "A multivariate multi-site statistical downscaling model for daily maximum and minimum temperatures." Climate Research **54**(2): 129-148.

Jeong, D. I., A. St-Hilaire, T. B. M. J. Ouarda and P. Gachon (2012). "Multisite statistical downscaling model for daily precipitation combined by multivariate multiple linear regression and stochastic weather generator." Climatic Change **114**(3-4): 567-591.

Kalnay, E., M. Kanamitsu, R. Kistler, W. Collins, D. Deaven, L. Gandin, M. Iredell, S. Saha, G. White and J. Woollen (1996). "The NCEP/NCAR 40-year reanalysis project." Bulletin of the American meteorological Society **77**(3): 437-471.

Katz, R. W. and M. B. Parlange (1998). "Overdispersion phenomenon in stochastic modeling of precipitation." Journal of Climate **11**(4): 591-601.

Kistler, R., E. Kalnay, W. Collins, S. Saha, G. White, J. Woollen, M. Chelliah, W. Ebisuzaki, M. Kanamitsu and V. Kousky (2001). "The NCEP-NCAR 50-year reanalysis: Monthly means CD-ROM and documentation." Bulletin-American Meteorological Society **82**(2): 247-268.

Kleiber, W., R. W. Katz and B. Rajagopalan (2012). "Daily spatiotemporal precipitation simulation using latent and transformed Gaussian processes." Water Resources Research **48**(1).

Lagarias, J. C., J. A. Reeds, M. H. Wright and P. E. Wright (1999). "Convergence properties of the Nelder-Mead simplex method in low dimensions." SIAM Journal on Optimization **9**(1): 112-147.

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

Li, C., V. P. Singh and A. K. Mishra (2013a). "A bivariate mixed distribution with a heavy-tailed component and its application to single-site daily rainfall simulation." Water Resources Research **49**(2): 767-789.

Li, C., V. P. Singh and A. K. Mishra (2013b). "Monthly river flow simulation with a joint conditional density estimation network." Water Resources Research **49**(6): 3229-3242.

Maraun, D., F. Wetterhall, A. Ireson, R. Chandler, E. Kendon, M. Widmann, S. Brienen, H. Rust, T. Sauter and M. Themeßl (2010). "Precipitation downscaling under climate change: recent

developments to bridge the gap between dynamical models and the end user." Reviews of Geophysics **48**(3).

Mekis, E. and W. D. Hogg (1999). "Rehabilitation and analysis of Canadian daily precipitation time series." Atmosphere-Ocean **37**(1): 53-85.

Palutikof, J. P., C. M. Goodess, S. J. Watkins and T. Holt (2002). "Generating rainfall and temperature scenarios at multiple sites: Examples from the Mediterranean." Journal of Climate **15**(24): 3529-3548.

Racsko, P., L. Szeidl and M. Semenov (1991). "A serial approach to local stochastic weather models." Ecological modelling **57**(1): 27-41.

Rasmussen, P. (2013). "Multisite precipitation generation using a latent autoregressive model." Water Resources Research **49**(4): 1845-1857.

Robertson, A. W., S. Kirshner and P. Smyth (2004). "Downscaling of daily rainfall occurrence over northeast Brazil using a hidden Markov model." Journal of Climate **17**(22): 4407-4424.

Schoof, J. T. and S. C. Pryor (2001). "Downscaling temperature and precipitation: A comparison of regression-based methods and artificial neural networks." International Journal of Climatology **21**(7): 773-790.

Serinaldi, F. (2009). "A multisite daily rainfall generator driven by bivariate copula-based mixed distributions." Journal of Geophysical Research: Atmospheres (1984–2012) **114**(D10).

Serinaldi, F. and C. G. Kilsby (2014). "Simulating daily rainfall fields over large areas for collective risk estimation." Journal of Hydrology **512**: 285-302.

Stephenson, D. B., K. Rupa Kumar, F. J. Doblas-Reyes, J. F. Royer, F. Chauvin and S. Pezzulli (1999). "Extreme daily rainfall events and their impact on ensemble forecasts of the Indian monsoon." Monthly Weather Review **127**(9): 1954-1966.

Stern, R. and R. Coe (1984). "A model fitting analysis of daily rainfall data." Journal of the Royal Statistical Society. Series A (General): 1-34.

Terrell, G. R. (2003). "The Wilson–Hilferty transformation is locally saddlepoint." Biometrika **90**(2): 445-453.

Villarini, G., B.-C. Seo, F. Serinaldi and W. F. Krajewski (2014). "Spatial and temporal modeling of radar rainfall uncertainties." Atmospheric Research **135**: 91-101.

Von Storch, H. (1999). "On the Use of "Inflation" in Statistical Downscaling." Journal of Climate **12**(12): 3505-3506.

Wan, H., X. Zhang and E. M. Barrow (2005). "Stochastic modelling of daily precipitation for Canada." Atmosphere-Ocean **43**(1): 23-32.

Widmann, M., C. S. Bretherton and E. P. Salathé Jr (2003). "Statistical precipitation downscaling over the northwestern united states using numerically simulated precipitation as a predictor." Journal of Climate **16**(5): 799-816.

Wilby, R., O. Tomlinson and C. Dawson (2003). "Multi-site simulation of precipitation by conditional resampling." Climate Research **23**(3): 183-194.

Wilby, R. L., C. W. Dawson and E. M. Barrow (2002). "sdsm - a decision support tool for the assessment of regional climate change impacts." Environmental Modelling and Software **17**(2): 145-157.

Wilby, R. L., H. Hassan and K. Hanaki (1998). "Statistical downscaling of hydrometeorological variables using general circulation model output." Journal of Hydrology **205**(1): 1-19.

Wilks, D. S. (2010). "Use of stochastic weathergenerators for precipitation downscaling." Wiley Interdisciplinary Reviews: Climate Change **1**(6): 898-907.

Wilks, D. S. and R. L. Wilby (1999). "The weather generation game: a review of stochastic weather models." Progress in Physical Geography **23**(3): 329-357.

Williams, P. M. (1998). "Modelling seasonality and trends in daily rainfall data." Advances in neural information processing systems: 985-991.

Xu, C.-y. (1999). "From GCMs to river flow: a review of downscaling methods and hydrologic modelling approaches." Progress in Physical Geography **23**(2): 229-249.

Yang, C., R. E. Chandler, V. S. Isham and H. S. Wheater (2005). "Spatial-temporal rainfall simulation using generalized linear models." Water Resources Research **41**(11): 1-13.

Yee, T. W. and A. G. Stephenson (2007). "Vector generalized linear and additive extreme value models." Extremes **10**(1-2): 1-19.

Yee, T. W. and C. Wild (1996). "Vector generalized additive models." Journal of the Royal Statistical Society. Series B (Methodological): 481-493.

Table 1. List of the 9 stations used in this study.

No.	Site	Name of station	Latitude (°N)	Longitude (°W)
1	7031360	Chelsea	45.52	-75.78
2	7014290	Cedars	45.3	-74.05
3	7025440	Nicolet	46.25	-72.60
4	7022160	Drummondville	45.88	-72.48
5	7012071	Donnacona 2	46.68	-71.73
6	7066685	Roberval A	48.52	-72.27
7	7060400	Bagotville A	48.33	-71
8	7056480	Rimouski	48.45	-68.53
3	7047910	Seven Island A	50.22	-66.27

Table 2. NCEP predictors on the CGCM3 grid.

No	Predictors	No	Predictors
1	Mean pressure at the sea level	14	Divergence at 500 hPa
2	Wind speed at 1000 hPa	15	Wind speed at 850 hPa
3	Component U at 1000 hPa	16	Component U at 850 hPa
4	Component V at 1000 hPa	17	Component V at 850 hPa
5	Vorticity at 1000 hPa	18	Vorticity at 850 hPa
6	Wind direction at 1000 hPa	19	Geopotential at 850 hPa
7	Divergence at 1000 hPa	20	Wind direction at 850 hPa
8	Wind speed at 500 hPa	21	Divergence at 1000 hPa
9	Component U at 500 hPa	22	Specific humidity at 500 hPa
10	Component V at 500 hPa	23	Specific humidity at 850 hPa
11	Vorticity at 500 hPa	24	Specific humidity at 1000 hPa
12	Geopotential at 500 hPa	25	Temperature at 2m
13	Wind direction at 500 hPa		

Table 3. Definition of the climatic indices used for the performance assessment of downscaled precipitation.

	Indices	Definition	Unite	Scale Time
Precipitation amount	MWD	Mean precipitation of wet day	Mm	Months
	Pmax90	90^{th} percentile of daily precipitation	Mm	Seasons
	PX1D	Maximum 1-days precipitation	Mm	Months
	PX3D	Maximum 3-days precipitation	Mm	Months
	PX5D	Maximum 5-days precipitation	Mm	Months
Precipitation occurrences	WRUN	Maximum number of consecutive wet days	Days	Months
	DRUN	Maximum number of consecutive dry days	Days	Months
	NWD	Number of wet day	Days	Months

Table 4. Quality assessment of the estimated series for the validation period (1991–2000) for BMAR, Hybrid and MMLR. Criteria are ME and RMSE. For the BMAR model Criteria were calculated from median of 100 realisations. Bold indicates the best result.

	ME			RMSE		
	BMAR	Hybrid	MMLR	BMAR	Hybrid	MMLR
Chelsea	**-0.2543**	-1.4191	2.3450	**5.9113**	6.1346	6.4540
Cedars	**-0.0702**	-1.3667	2.6636	**6.4142**	6.6869	7.0977
Nicolet	**0.5824**	-2.1763	2.7449	**6.1249**	7.4483	6.1448
Drummondville	**0.5185**	-1.1187	2.4709	**5.4397**	5.6545	7.0037
Donnacona 2	**0.8032**	-1.2948	2.9283	**5.9166**	6.4117	6.1347
Roberval A	**0.1426**	-1.4866	2.1974	**5.3472**	5.7093	6.9357
Bagotville A	**1.0849**	-0.6332	2.5702	**6.3124**	6.3571	6.9357
Rimouski	**0.0621**	-1.3980	2.0858	**5.0863**	5.5576	5.8011
Seven Island A	**-0.5391**	-1.9487	2.2077	**5.4987**	6.0580	6.0977

Bold means better result.

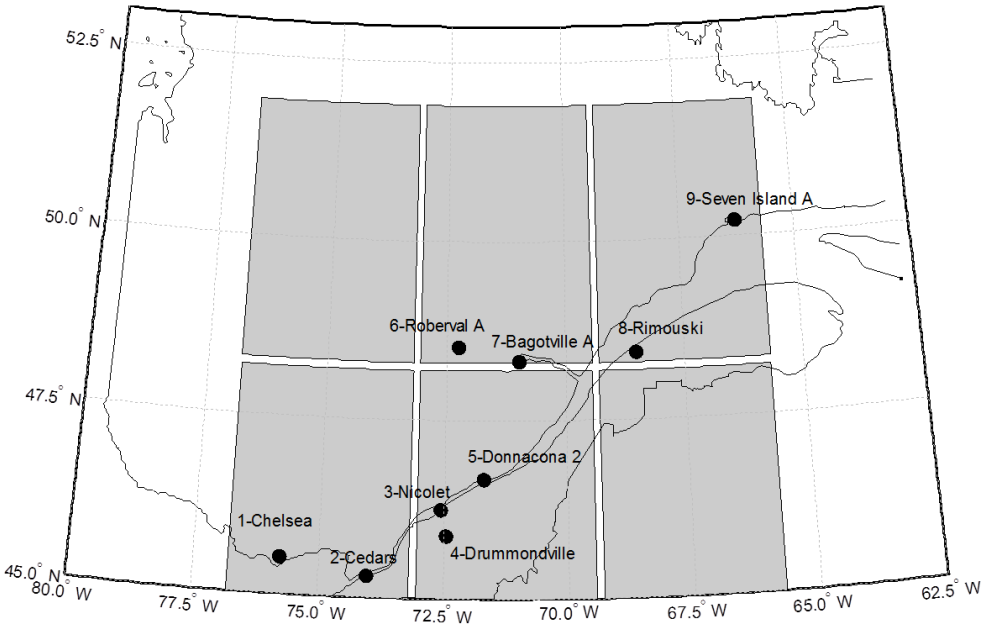

Figure 1. Locations of CGCM3 grid and observation stations of daily precipitation.

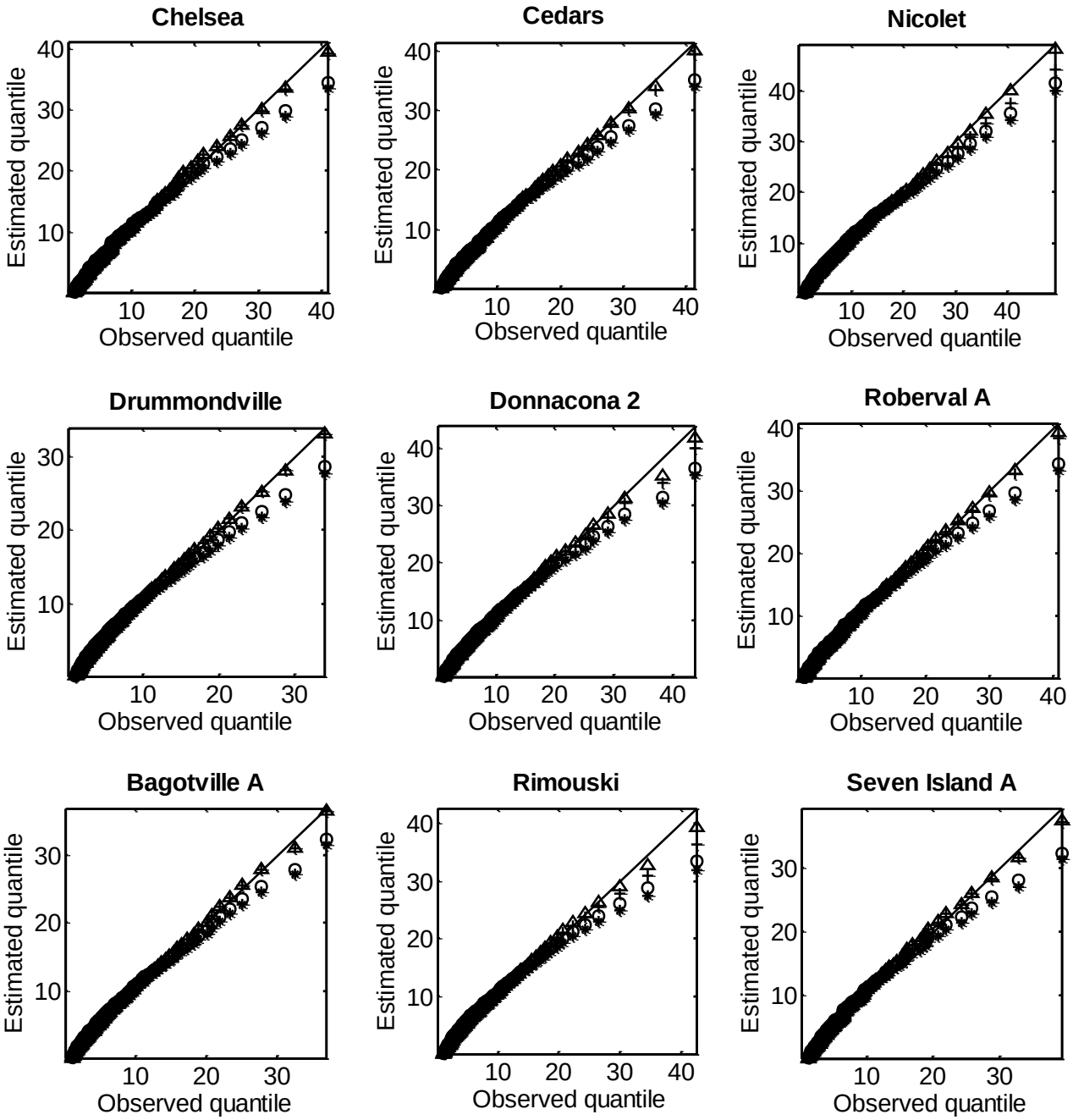

Figure 2. Q–Q plot of observed and modeled quantiles for Gamma distribution (stars), WEI distribution (squares), GP distribution (triangles) and mixed Exponential distribution (plus).

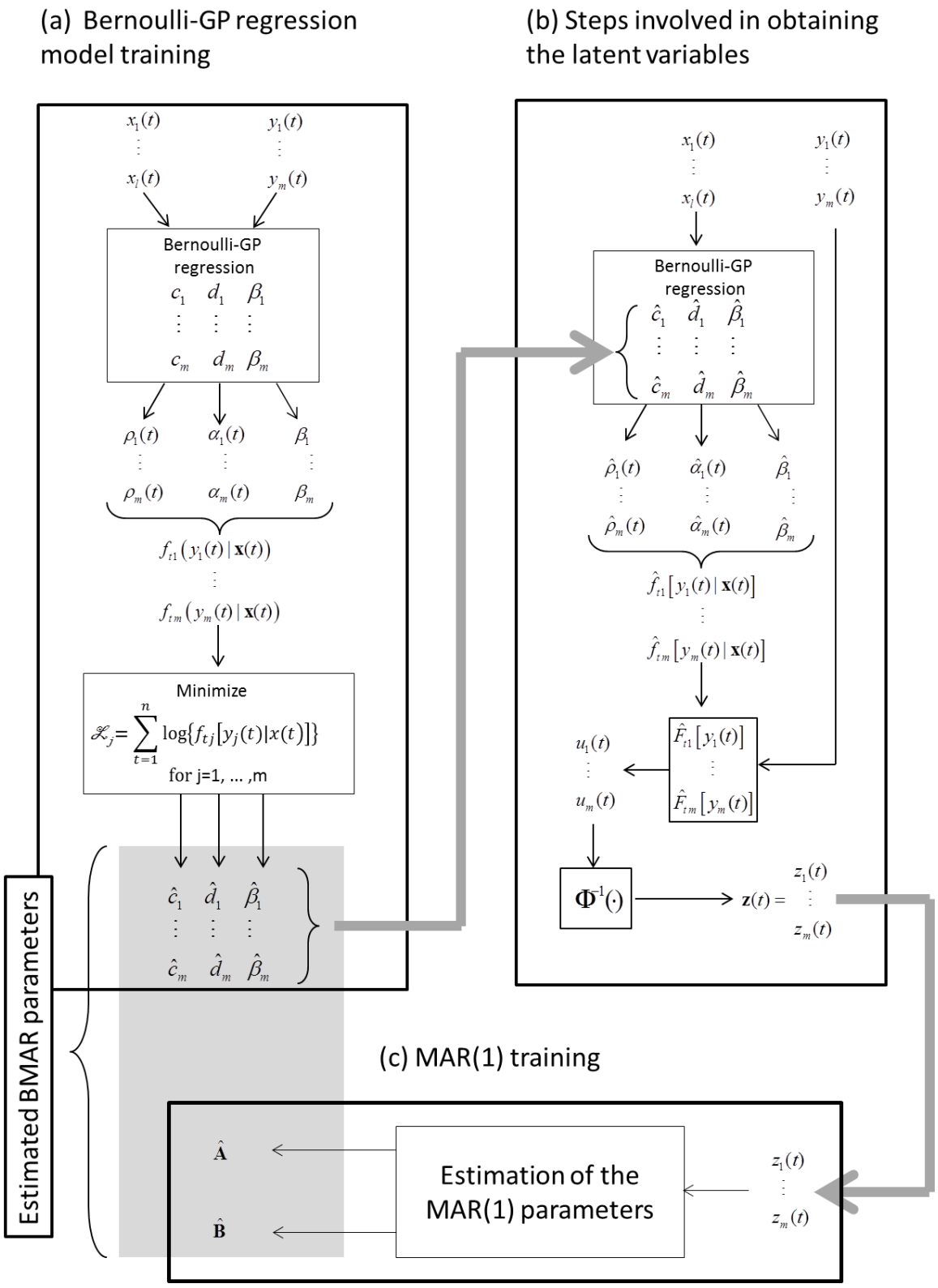

Figure 3. Steps involved in training the BMAR model.

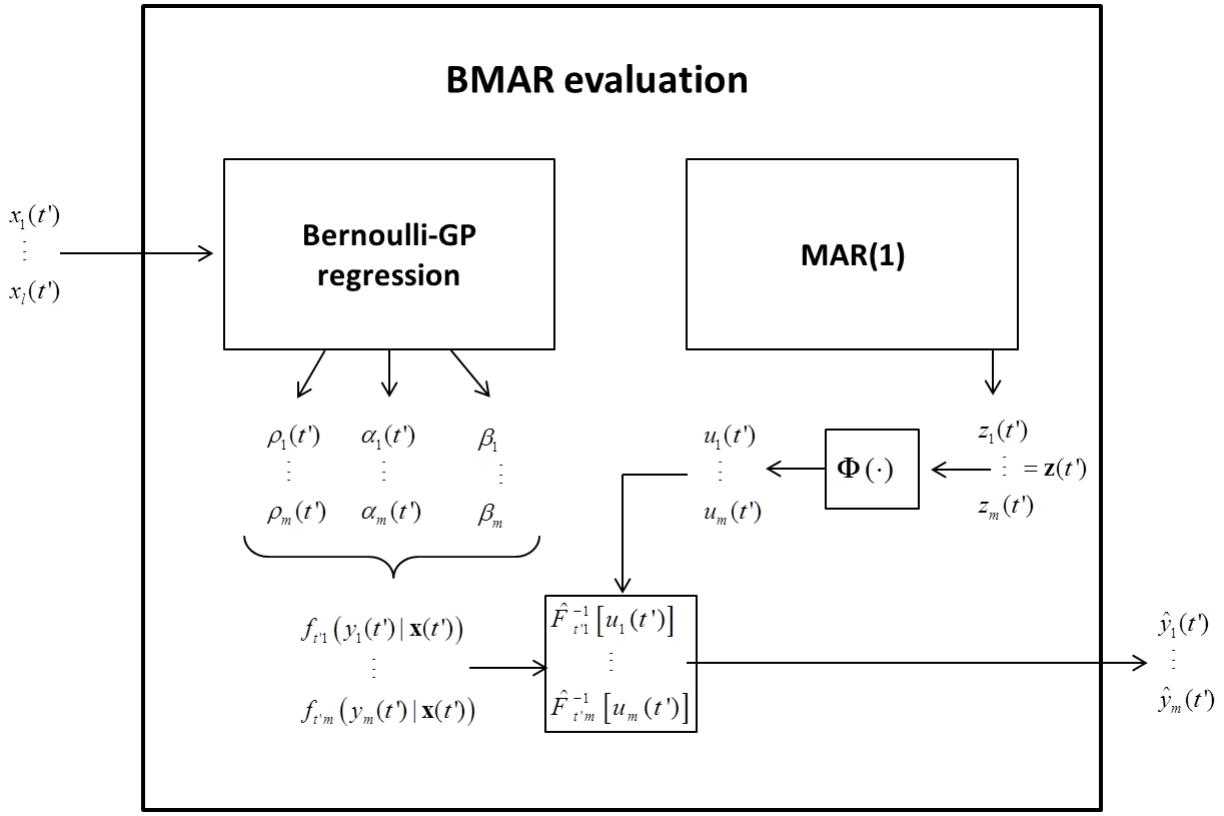

Figure 4. Steps involved in evaluating the BMAR model.

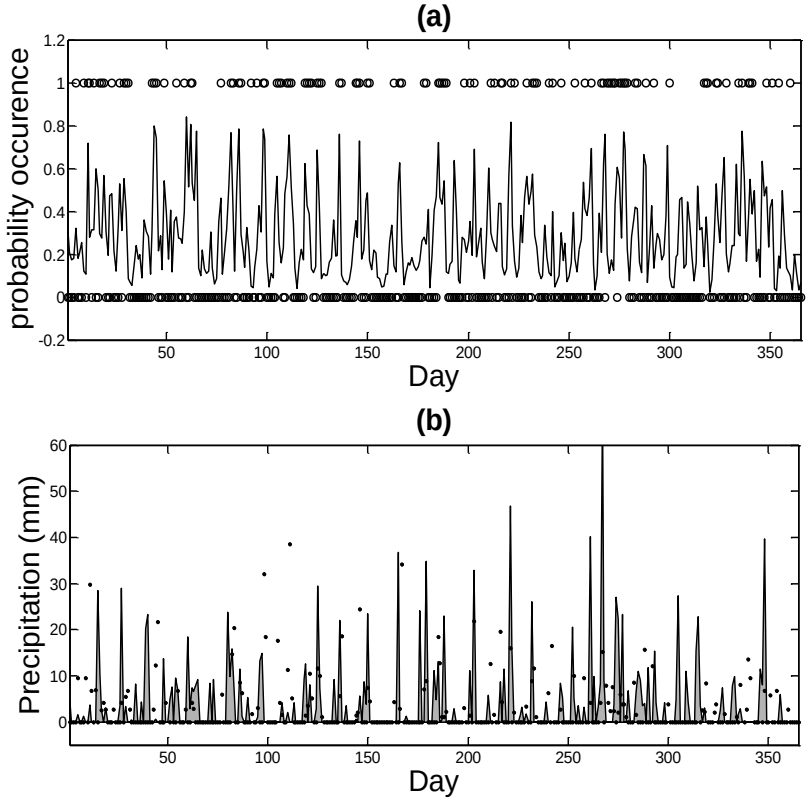

Figure 5. BMAR results for precipitations shown at cedars station during 1991. The probability of precipitation occurrences is shown by the solid line in (a) where circles show the observed precipitation occurrences. Observed precipitation values (dots) and a synthetic precipitation obtained using BMAR model are shown in (b).

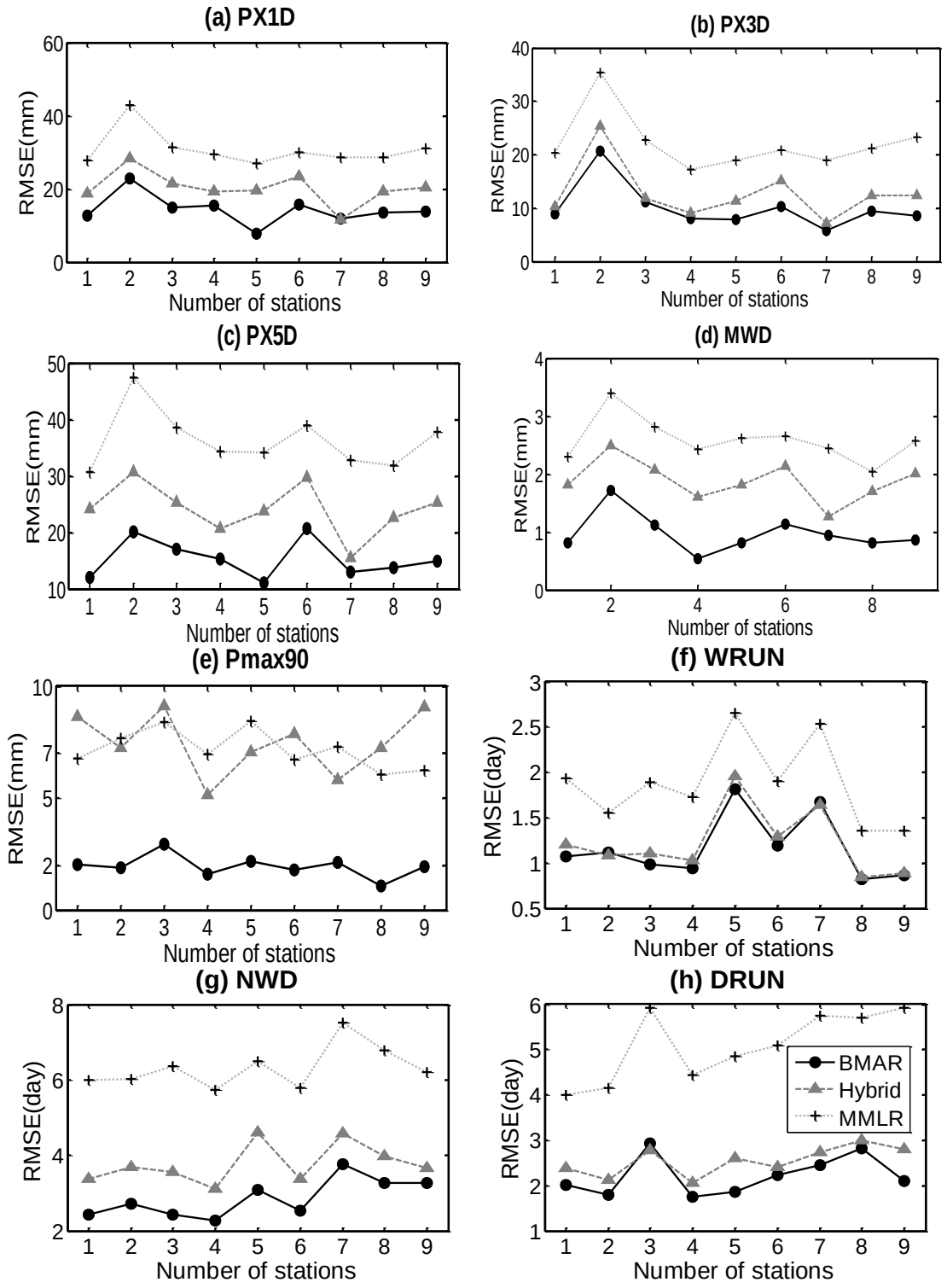

Figure 6. RMSE between observed and estimated climatic indices of downscaled precipitations for the BMAR model, the hybrid model and MMLR model on the nine weather stations during the validation period (1991–2000). RMSE of BMAR and the hybrid model were calculated using the mean of 100 realisations.

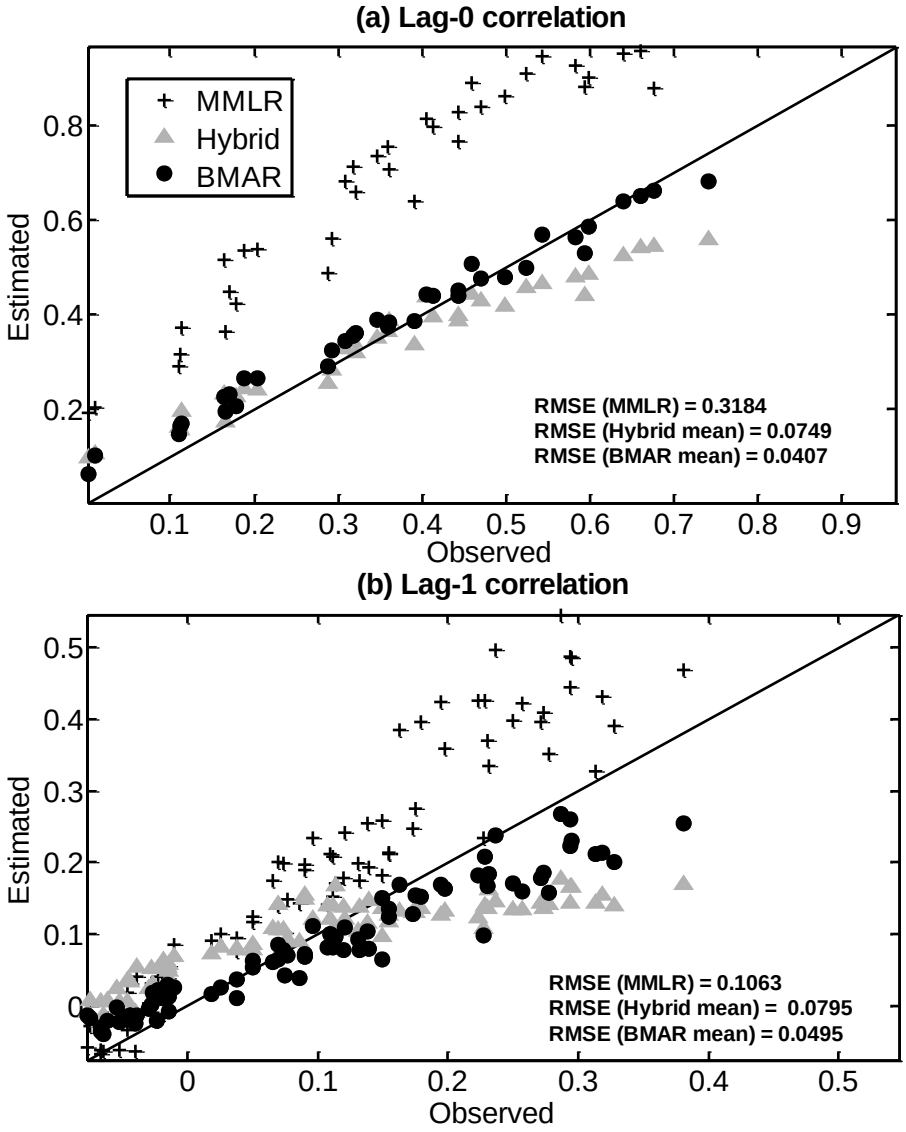

Figure 7. Scatter plots of observed and modeled Lag-0 correlation (a) and Lag-1 correlation for the BMAR model, hybrid model and MMLR model during the validation period. Correlation values of the BMAR and the hybrid model are obtained using the mean of the correlation values calculated from a 100 simulations.

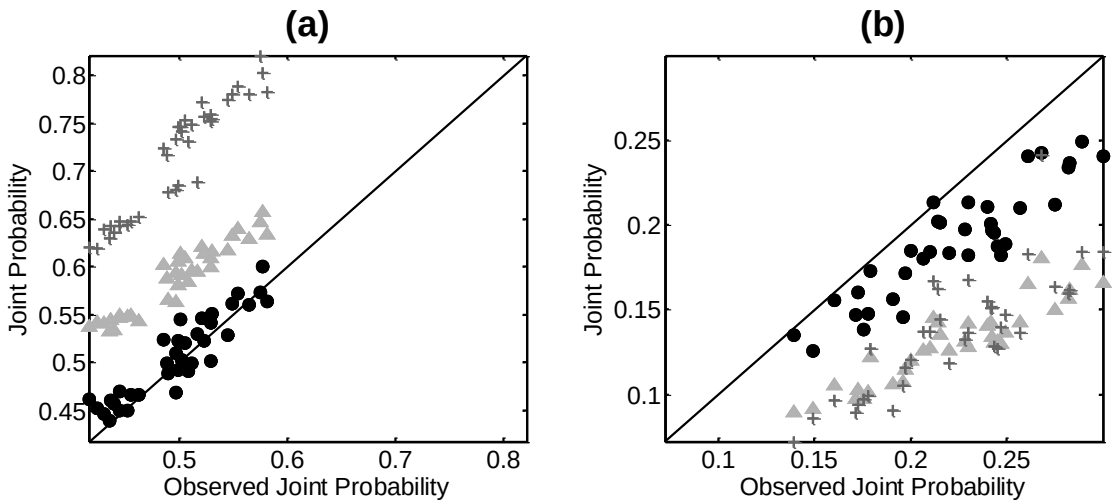

Figure 8. Joint probabilities that station pairs are both dry (a), and both wet (b), on a given day, for the observed and modeled joint probability by BMAR (black dots), MMLR (gray plus) and hybrid (gray triangles) during the validation period. Values of BMAR and hybrid models are obtained using the mean of the joint probability values calculated from 100 simulations.

CHAPITRE 4: MULTISITE AND MULTIVARIABLE STATISTICAL DOWNSCALING USING A GAUSSIAN COPULA QUANTILE REGRESSION MODEL

Multisite and multivariable statistical downscaling using a Gaussian Copula Quantile Regression model

M.A. Ben Alaya[1], F. Chebana[1] and T.B.M.J. Ouarda[2, 1]

[1]INRS-ETE, 490 rue de la Couronne, Québec (QC), Canada G1K 9A9
[2] Institute Center for Water and Environment (iWater), Masdar Institute of Science and Technology
P.O. Box 54224, Abu Dhabi, UAE

[*]**Corresponding author:** Tel: +1 (418) 654 2530#4468
Email: mohammed_ali.ben_alaya@ete.inrs.ca

Accepted November 2, 2015

(Climate Dynamics)

Abstract

Statistical downscaling techniques are required to refine Atmosphere-Ocean Global Climate (AOGCM) data outputs and provide reliable meteorological information such as realistic temporal variability and relationships between sites and variables in a changing climate. To this end, the present paper introduces a modular structure combining two statistical tools of increasing interest during the last years: (i) Gaussian copula and (ii) quantile regression (GCQR). The quantile regression tool is employed to specify the entire conditional distribution of downscaled variables and to address the limitations of traditional regression-based approaches whereas the Gaussian copula is used to describe and preserve the dependence between both variables and sites. A case study based on precipitation and maximum and minimum temperatures from the province of Quebec, Canada, is used to evaluate the performance of the proposed model. Obtained results suggest that this approach is capable of generating series with realistic correlation structures and temporal variability. Furthermore, the proposed model performed better than a classical multisite multivariate statistical downscaling model for most evaluation criteria.

Keywords: Climate downscaling, Gaussian copula, Quantile regression, Temperature, Precipitation, Multisite, Multivariable.

1. Introduction

Atmosphere–ocean general circulation models (AOGCMs) are powerful tools for assessing the evolution of the earth's climate system. However, outputs of the AOGCMs are generally produced on horizontal grids with a low spatial resolution. For the set of global climate models (GCMs) derived from the Coupled Model Intercomparison Projects Phase 3 (CMIP3), the horizontal resolution at 35°N ranged from 103 to 455 km, with an average of 254 km, whereas for the Phase 5 (CMIP5) it ranged from 68 to 342 km with an average of 193 km (Gulizia and Camilloni 2015; Kusunoki and Arakawa 2015). A number of hydro-meteorological applications require the availability of information at a higher resolution. To bridge this resolution gap, downscaling methods have been developed. These methods include dynamic downscaling (DD) and statistical downscaling (SD). DD methods involve regional climate models (RCM) with a high resolution over a limited area, whereas SD methods consider the link between large-scale atmospheric variables (predictors) and local-scale weather variables (predictands) (Wilby et al. 1998). SD methods represent a good alternative to dynamic methods in the case of limited resources, because of their ease of implementation and their low computational requirements (Herrera et al. 2006; Benestad et al. 2008).

Among several weather variables, precipitation and temperature are the most frequently used predictands for downscaling purposes. Generally, these weather variables are collected at various sites where SD models are required to adequately reproduce the observed temporal variability as well as to maintain their spatiotemporal properties at several sites (Cannon 2008a). A poor representation of the temporal variability could lead to a poor representation of extreme events. Furthermore, the adequate representation of spatiotemporal properties of precipitation and

temperature is very important, particularly for hydrological modeling (Lindström et al. 1997; Chen et al. 2015).

Regression-based methods can find a direct link between atmospheric predictors and local predictands and usually perform well for downscaling purposes (Hessami et al. 2008; Jeong et al. 2012b; Jeong et al. 2012a; Chen et al. 2014). However, regression-based methods generally focus mainly on the central part of the distribution and thus they underrepresent the temporal variability (Von Storch 1999; Cawley et al. 2007). In addition, they do not properly reproduce various aspects of the spatial and temporal dependence of the downscaled predictands (Wilby et al. 2003; Harpham and Wilby 2005).

To adequately reproduce the temporal variability, probabilistic regression approaches have made valuable contributions in downscaling applications (Williams 1998; Haylock et al. 2006; Fasbender and Ouarda 2010; Ben Alaya et al. 2015). They provide as an output a complete dynamic distribution function by incorporating the influence of large scale atmospheric predictors on the vector of parameters of the conditional distribution (Cannon 2012). Unlike traditional regression models, a random sampling from the obtained conditional distribution at each forecast step enables the reproduction of a realistic temporal variability. In spite of the advantages of these approaches, they have some drawbacks, such as imposing the homogeneity of the residuals and assuming or selecting a given parametric distribution.

An alternative solution to reproduce the whole conditional distribution consists in using quantile regression which is introduced by Koenker and Bassett (1978) to directly predict individual quantiles of the conditional distribution. Since a more complete summary of a given distribution is provided by its quantiles, quantile regression models allow obtaining the entire response

distribution without assuming any parametric form. In the past decade, application of quantile regression in environmental modeling and climate change impact assessment has grown rapidly. Models based on quantile regression have been introduced to describe effects of meteorological variables on ozone concentration (Baur et al. 2004), to study annual streamflow (Luce and Holden 2009), to forecast wind power (Bremnes 2004b), to estimate hydrological uncertainty (Weerts et al. 2011), to predict flood quantile in a changing climate (Sankarasubramanian and Lall 2003), to model tropical cyclone intensity and trend (Elsner et al. 2008; Jagger and Elsner 2009). In addition, the application of quantile regression has made significant contributions in classical precipitation downscaling (Bremnes 2004a; Friederichs and Hense 2007; Cannon 2011; Tareghian and Rasmussen 2013).

To extend probabilistic approaches to multisite downscaling tasks, Cannon (2008b) employed the methodology used in expanded downscaling (Burger and Chen 2005). Indeed, a constraint is added to the regression cost function to preserve observed covariance. However, Von Storch (1999) suggested that this technique is not realistic because it considers that large-scale atmospheric variability can reproduce all the local variability. Alternatively, Ben Alaya et al. (2014) employed a Gaussian copula simulation procedure to preserve realistic relationships between multisite precipitation and temperature in a probabilistic downscaling framework.

Copula, as a multivariate distribution with uniform margins, describes the dependence structure independently from the marginal distributions (Sklar 1959). An introduction to the copula theory can be found in Joe (2014) and Nelson (2006). The copula approach can be seen as a simple and flexible method to construct parametric descriptions of multivariate non-normally distributed random variables. After numerous successful applications in fields like econometrics, financial

research, risk management, and insurance, copulas have recently become very popular in hydrological applications including frequency analysis, simulation, and geostatistical interpolation (Bárdossy 2006; Renard and Lang 2007; El Adlouni and Ouarda 2008; Kazianka and Pilz 2010; Chebana and Ouarda 2011; Bargaoui and Bardossy 2015). The reader is directed to Schölzel and Friederichs (2008) for a brief overview of the development and applications of copulas in meteorology and climate research where often non-normally distributed random variables, like precipitation, wind speed, cloud cover, humidity, are involved.

Quantile regression and copula can be combined to take advantage of their strengths in hydrometeorology and climate research. In this context, the aim of the present paper is to develop a Gaussian Copula quantile regression (GCQR) model as a new multisite statistical downscaling model integrating the two concepts of quantile regression and Gaussian copula. By specifying the entire conditional distributions using quantile regression, GCQR allows to address the limitations of traditional regression-based approaches. Then, in the simulation step, GCQR uses a Gaussian copula to preserve the dependence between both variables and sites. Note that other ways of combining both concepts have been proposed in the statistical literature, such as Baur (2013), Chen et al. (2009) and Reich (2012). For instance, in the latter, quantile functions of temperature field that evolve over time are obtained using quantile regression, and spatially smoothed using Gaussian Copula.

The present paper is structured as follows. The proposed GCQR model is presented in section 2 as well as the multivariate multisite statistical downscaling model (MMSDM) (Jeong et al. 2013) as a traditional model to compare with. Then, the used datasets of daily precipitation and minimum and maximum temperatures are described in section 3. Thereafter, section 4 presents

results of the application and the comparison with MMSDM and the multivariate multiple linear regression (MMLR) model. Finally, discussion and conclusions are given in sections 5 and 6 respectively.

2. Methodology

The MMSDM and the GCQR models are presented respectively in sections 2.1 and 2.2. The quantile regression approach is presented as well as the simulation procedure using the Gaussian copula.

2.1. Multivariate multisite statistical downscaling model (MMSDM)

We consider q maximum temperature stations, r minimum temperature stations and s precipitation stations. For each station and for a historical period we have daily data for each predictand. Precipitation data can be decomposed into precipitation occurrence and wet-day precipitation amounts. To avoid problems of low daily precipitation values, wet days are defined as days with a precipitation amount larger than 1 mm/day (e.g. Jeong et al. 2012b). The precipitation amount vector $\mathbf{Pam}$ for a given site is not normally distributed (e.g. Stephenson et al. 1999; Yang et al. 2005). Yang et al. (2005) proposed to take $\mathbf{R} = \sqrt[3]{\mathbf{Pam}}$ to obtain normality, where $\mathbf{R}$ is called Anscombe residuals (e.g. Yang et al. 2005; Jeong et al. 2012b). We denote $\mathbf{A}$ the matrix grouping all the predictand vectors $\mathbf{Tmax}$, $\mathbf{Tmin}$, $\mathbf{Poc}$ (precipitation occurrences) and $\mathbf{R}$, of dimension $n \times m$ where $m = q + r + 2 \times s$:

$$\mathbf{A} = \left(\mathbf{Tmax}_1, \ldots, \mathbf{Tmax}_q, \mathbf{Tmin}_1, \ldots, \mathbf{Tmin}_r, \mathbf{R}_1, \ldots, \mathbf{R}_s, \mathbf{Poc}_1 \ldots, \mathbf{Poc}_s\right) \qquad (1)$$

where a column in $\mathbf{A}$ corresponds to values of the predictand series over n days, and a line corresponds to values of all predictands at a given day. The MMLR model is given by:

$$\mathbf{A} = \mathbf{X} \times \mathbf{B} + \mathbf{E} \tag{2}$$

where $\mathbf{X}$ is the matrix of dimension $n \times l$ grouping l independent predictors, $\mathbf{E}$ is the residual matrix of dimension $n \times m$ and $\mathbf{B}$ is a parameter matrix of dimensions $l \times m$. The Ordinary Least Square (OLS) estimate of $\mathbf{B}$ is given by:

$$\mathbf{B} = (\mathbf{X}^{\mathbf{T}}\mathbf{X})^{-1}\mathbf{X}^{\mathbf{T}}\mathbf{A} \tag{3}$$

Then, the deterministic series of predictands can be obtained using the following MMLR equations:

$$\mathbf{A} = \mathbf{X} \times \mathbf{B} \tag{4}$$

For the MMLR model wet days are determined when values of probability of precipitation obtained by MMLR occurrences model are larger than 0.5. To adequately reproduce the temporal variability and relationships between variables and sites, the MMSDM model of Jeong et al. (2013) generates and adds correlated multivariate random errors to this deterministic model. Finally to reproduce the observed precipitation properties, Jeong et al. (2012b) adapted a probability mapping technique and adjusted the generated precipitation amounts using the Gamma distribution.

2.2. Gaussian copula quantile regression model

In this section, we provide a brief description of the linear quantile regression as well as the Gaussian Copula approach. They represent the main tools used in this paper for downscaling precipitation and temperature.

2.2.1. Quantile regression

Efforts in the field of statistics has been devoted to the elaboration of the linear regression model and associated estimation methods by minimizing a sum of squared residuals called least squares method. Since the sample mean may be defined as the solution to the problem of minimizing a sum of squared residuals, the commonly used ordinary least squares regression provides the conditional mean given a set of predictors. However, the sample median may be defined as the solution to the problem of minimizing a sum of absolute residuals. In this regard, median regression, also known as least-absolute-deviations (LAD) regression, minimizes the sum of absolute residuals. Median regression is more robust to outliers than least squares regression. It represents a semiparametric approach as it avoids specific assumptions about the parametric distribution of the error process. The question is then: the median is the 0.5 quantile, why not use other quantiles as well? In other words, if the sample mean and the sample median may be defined as a solution to an appropriate minimization problem, which optimization problem can have as a solution a sample quantile? By seeking the answer to this question, Koenker and Bassett (1978) introduced a new regression technique called quantile regression which provides the conditional quantile of the response variable given a set of predictors. Quantile regression also provides a richer characterization of the data by considering the impact of a covariate on the entire response distribution, not merely its conditional mean or conditional median. Unlike least

squares regression, quantile regression does not assume a homogenous residual variance and does not make any assumption about the parametric form of the response.

Formally, let $\mathbf{x} = (x_1, x_2, \ldots, x_m)$ denote the set of daily predictors and y be a given predictand. The linear conditional quantile regression model is expressed as:

$$Q_p\left(y \mid \mathbf{x}\right) = \mathbf{x}^T \mathbf{b}_p \tag{5}$$

where $\mathbf{b}_p$ is a vector of parameters related to the $p^{th}\,(0 < p < 1)$ quantile $Q_p\left(y \mid \mathbf{x}\right)$ of the conditional distribution of y given $\mathbf{x}$. The intercept of the vector of parameters can be included by adding 1 in the first element of $\mathbf{x}$. An estimate of the parameters $\mathbf{b}_p$ for a given set of observations $(\mathbf{x}(t), y(t))$, $t = 1, \ldots, n$, is given by:

$$\hat{\mathbf{b}}_p = \arg\min_{\mathbf{b}} \sum_{t=1}^{n} \rho_p\left(y(t) - \mathbf{x}(t)^T \mathbf{b}\right) \tag{6}$$

where the function $\rho_p(.)$ is defined as:

$$\rho_p(u) = \begin{cases} u(p-1) & \text{if } u < 0 \\ up & \text{if } u \geq 0 \end{cases} \tag{7}$$

Let $\mathbf{Y}$ be the random matrix of dimension $n \times m$ grouping each predictand at a given site with:

$$\mathbf{Y} = (\mathbf{Tmax}_1, \ldots, \mathbf{Tmax}_q, \mathbf{Tmin}_1, \ldots, \mathbf{Tmin}_r, \mathbf{Pam}_1, \ldots, \mathbf{Pam}_s, \mathbf{Poc}_1, \ldots, \mathbf{Poc}_s) \tag{8}$$

The vector of predictands $\mathbf{Y}$ is different from the vector $\mathbf{A}$, since the latter contains the transformed variables for the precipitation amount **R**. Let $F_{tk}\left[y_k(t) \mid \mathbf{x}(t)\right]$ be the cumulative

distribution function (CDF) at time t of the k^{th} element of $\mathbf{Y}$, where $k = 1,\ldots,m$. Our goal in this step of the calibration is to specify the entire conditional probability density function (PDF) $f_{tk}\left[y_k(t) \mid \mathbf{x}(t)\right]$ for each day t and for each predictand. Then, drawing values from the obtained conditional distribution at each forecast ensures the reproduction of a realistic temporal variability of downscaled results. The entire conditional PDF distribution $f_{tk}\left[y_k(t) \mid \mathbf{x}(t)\right]$ of the response variable can be specified by all its infinite conditional quantiles. However, in practice, quantile regression can be used to estimate a finite number of quantiles of order *p*, for instance, from *p*= 0.01 to 0.99 by steps of 0.01. Thereafter, the obtained quantile regression estimates can be interpolated to produce the corresponding CDF as a smooth and monotonic increasing curve. More precisely, this curve may be considered as an estimate for the conditional (CDF) $\hat{F}_{tk}\left[y_k(t) \mid \mathbf{x}(t)\right]$. However, as a result of model approximation, the estimated curve is not necessarily monotonic as required of a CDF (Tareghian and Rasmussen 2013). To overcome this problem, we can increasingly sort the estimated values corresponding to the quantile orders.

For precipitation occurrences a standard problem is that the dry-wet dichotomy leads to a Bernoulli process. The parameter of the conditional Bernoulli distribution $\pi(x)$ can be obtained directly using a logistic regression given by:

$$E(y \mid \mathbf{x}) = \pi(\mathbf{x}) = \frac{1}{1+\exp\left[\mathbf{x}^T c\right]} \tag{9}$$

where c is the vector of parameters of the logistic model. Thus the conditional distribution of precipitation occurrences is given by:

$$f\left[y \mid \mathbf{x}\right] = \begin{cases} \pi(x) & if \quad y = 1 \\ 1 - \pi(x) & if \quad y = 0 \end{cases} \tag{10}$$

Then the maximum likelihood method can be used to estimate all the parameters of the vector c for each precipitation station. Note that an alternative way to avoid the split between occurrence and amount process is to use a censored quantile regression approach for a single mixed discrete-continuous distribution of precipitation (Friederichs and Hense 2007; Cannon 2011).

2.2.2. Conditional simulation using Gaussian copula

In the simulation step, the proper reproduction of relationships between sites and variables is very important, for instance, for hydrological applications. To this end, Ben Alaya et al. (2014) proposed a simulation approach based on a Gaussian copula. The latter is more general than the multivariate normal distribution, since it describes the dependence part of the model by allowing the margins to be normal or not, discontinuous or continuous, and not necessary in the same distribution family. In the present paper, the same Gaussian copula procedure is performed to extend quantile regression downscaling model to multisite and multivariable tasks.

A copula $\mathbb{C}$ is a multivariate continuous distribution function where each marginal is a uniform distribution between 0 and 1. A Gaussian copula $\mathbb{C}$ is defined as:

$$\mathbb{C}(w; C) = \Phi_m\left[\Phi^{-1}(w_1), \ldots, \Phi^{-1}(w_m); C\right] \tag{11}$$

where Φ is the standard (univariate) normal CDF and $\Phi_m(w; C)$ is the CDF for a multivariate normal vector w having zero mean and covariance matrix C. Our goal here is to estimate the

Gaussian copula parameter matrix C. It is estimated as the correlation matrix of the latent multivariate Gaussian variables $z(t) = [z_1(t), \ldots, z_m(t)] \approx \mathrm{N}_m(0, C)$ where $z_k(t)$ is given by:

$$z_k(t) = \Phi^{-1}[\hat{F}_{tk}(y_k(t))] \tag{12}$$

where $\hat{F}_{tk}$ is the CDF for the k^{th} element of $\mathbf{Y}$ obtained from quantile regression model at the forecast step *t* from the calibration period. Because precipitation occurrence is a discrete variable, its CDF F_{tk} is discontinuous. Thus, the cumulative probabilities $\hat{F}_{tk}(y_k(t))$ for precipitation occurrences are randomly generated from the uniform distribution on $[0, 1-\pi_k(t)]$ for dry days and $[1-\pi_k(t),\ 1]$ for wet days. Figure 1 demonstrates how quantile regression and Gaussian copula are combined to generate one simulation at a day t' from the validation period. For a new day t' from the validation period it is possible to generate $u = [u_1, \ldots, u_m]$ from Gaussian copula. Then, value $\hat{y}_k(t')$ is given by $F^{-1}_{t'k}[u_k]$, and for precipitation occurrences $\hat{y}_k(t') = 0$ if u_k is less than $1-\pi_k(t')$ and $\hat{y}_k(t') = 1$ if u_k is greater than $1-\pi_k(t')$.

3. Study area and data

Observed daily, maximum and minimum temperatures and precipitations from stations located in the province of Quebec (Canada) are used in our study (see Figure 2). The list of stations is presented in Table 1 for temperatures and in Tbale 2 for precipitation. Nine stations are considered for maximum and minimum temperatures namely: Cedars, Drummondville, Seven Islands, Bagot-ville A, Quebec, Sherbrooke, Maniwaki Airport, La Pocatiére and Mont-joli A. For precipitation data, nine stations are available, namely: Chelsea, Cedars, Nicolet,

Drummondville, Donnacona, Roberval A, Bagotville A, Rimouski and Seven Islands. We note that Cedars, Drummondville, Seven Islands and Bagotville stations, contain both precipitation and temperature data. All predictands series are obtained from Environment Canada weather stations.

Predictors are obtained from the reanalysis product NCEP/NCAR interpolated on the CGCM3 Gaussian grid (3.75 ° latitude and longitude). Six grids covering the predictands stations area are selected (see Figure 2), and 25 NCEP predictors are available for each grid (see Table 3). Thus, a total of 150 daily predictors are available for the downscaling process. To reduce the number of predictors, a principal component analysis (PCA) is performed. The first principal components that preserve more than 97% of the total variance are selected. The data sets cover the period between January, 1st 1961 and December, 31st 2000. This record period is divided into two sub-periods for the calibration (1961-1990) and the validation (1991-2000).

4. Results

The GCQR and MMSDM models were trained for the calibration period and 100 realizations were generated for precipitation and maximum and minimum temperatures series using the procedure described in Section 2 during the validation period (1991-2000). Figure 3 shows the estimated series of daily maximum and minimum temperatures and precipitation using the GCQR model at Chelsea station during the year 1991. For the three predictands, the estimated series are close to the observed series. For temperature variables, the majority of the observations are within the 95% confidence interval.

For model evaluation, a first validation approach is performed based on the mean error (ME), the root mean square error (RMSE), the difference between observed and modeled variances (D), and the percentage of observations in the 95% confidence interval obtained using 100 simulations. This percentage should be close to 95% to ensure that the estimated distribution is relevant. These four criteria allow a direct comparison between observed and simulated series. For GCQR and MMSDM, the RMSE and the ME were calculated using the conditional mean whereas the *D* values were obtained using the mean of *D* values over the 100 realisations. Results of these statistical criteria for GCQR, MMSDM and MMLR without randomisation are shown in Table 4 for temperature variables and Table 5 for precipitations. Based on ME and RMSE, it appears to be no significant difference between the three models for temperature variables (see Table 4). This result is not surprising because temperature is easier to model with the OLS regression model, when the noise process is normally distributed. Nevertheless, GCQR gives better results than the two other modeling approaches in term of *D*. Regarding the downscaled precipitations, GCQR yielded better results for most stations in terms of RMSE, ME and *D*. Moreover, the estimated distributions were in satisfactory agreement with the observed values as approximately the percentage of observed precipitations and temperatures that belong to the 95% confidence interval is very close to 95%. We can also remark that MMSDM generally outperforms the MMLR model in terms of the three criteria *ME, RMSE* and *D*.

To further compare the three modelling approaches, the RMSEs of several climate indices were computed. The definitions of these climate indices are presented in Table 6 for temperature variables and Table 7 for precipitation. These indices reflect the characteristics including the frequency, intensity, and duration of temperature and precipitation extremes (Wilby 1998; Wilby et al. 2002; Gachon et al. 2005; Hessami et al. 2008). The mean of indices values from 100

realisations were used to calculate the RMSEs of these indices for the GCQR and MMSDM. Figure 4 summarizes the results for temperature indices. We can see that GCQR gives better results than both MMLR and MMSDM. Similarly, results for precipitation indices are presented in Figure 5. Five indices were considered for precipitation amounts and three indices for precipitation occurrences. For precipitation amounts indices, results show that GCQR performs better than MMLR and MMSDM for most indices and for most stations. This result indicates the role of quantile regression component to better represent at-site precipitation amount characteristics. Additionally, based on the indices WRUN, NWD and DRUN, the use of the logistic regression allows to adequately replicating observed monthly characteristics of precipitation occurrences. MMLR and MMSDM assume that regression residuals are normally distributed, but this assumption may not be valid even after the application of normalizing transformations. If the only interest is on the central prediction, the non-normality of the model output may not be a serious problem. However, the form of the conditional distribution becomes important when we look to adequately represent extreme values occurring on the tail of the distribution.

Because the daily precipitation distribution is typically heavily skewed, an evaluation based on the quantile-quantile (Q–Q) plots better reveals differences, especially in reproducing the right tail of the precipitation distributions. The Q-Q plots of each model for the nine stations are presented in Figure 6 for the validation period. There appears to be an advantage of the GCQR model for most stations, particularly at Roberval, Rimouski and Seven Island where the extreme right tail of the precipitation distribution seems to be fairly well simulated. Furthermore, GCQR performs well not only for the extreme right tail of the precipitation distribution but also for the

central part. This last finding validates the conclusions from the comparison using the numerical criteria RMSE and ME.

To evaluate the capacity of the three models to reproduce the observed cross-site correlation, the scatter plots of observed and modeled cross-site correlation for each predictand are presented in Figure 7. The mean of the correlation values calculated from a 100 realisations were used to obtain the correlation values of GCQR and MMSDM models. As shown in Figure 7, the MMLR gives the poorest result compared to GCQR and MMSDM and generally overestimates the cross-correlation for all predictands. This result is expected since MMLR is not a multisite model. We can see that GCQR and MMSDM give a good preservation of the cross-site correlation for maximum and minimum temperature and precipitation amount. For precipitation occurrences, both GCQR and MMSDM, underestimate the cross-site correlations, but the GCQR is slightly better, since the value of RMSE is equal to 0.0746 for GCQR and 0.1511 for MMSDM. Similarly, the scatter plots of observed and modeled cross-predictand correlations are shown in Figure 8. Results indicate that GCQR and MMSDM preserve adequately the cross-predictand correlations, and in term of RMSE, GCQR outperform the MMSDM for all stations. It should be noted that MMSDM has difficulty in reproducing the cross-predictand correlations when precipitation amount is present. This result can be explained by the use of the probability mapping step to correct the bias of precipitation. However, there is no need to rely on transformation steps or on bias correction procedures when evaluating the GCQR model. Indeed the mapping in the conditional distribution is automatic using its quantile regression component.

5. Discussions

The MMLR downscaling model gives the mean of the process conditionally to the set of atmospheric predictors. Thus, this model shows difficulty in preserving a realistic temporal variability and cross-correlation among sites and variables. As a solution to this problem, the MMSDM employs a stochastic randomisation procedure by adding spatially correlated random series. Several models based on randomization were developed for climate downscaling. These methods are often called hybrid models, because they combine two components: (i) a deterministic regression component which provides the conditional mean and (ii) an unconditional resampling component to preserve observed weather characteristics at local scale (Harpham and Wilby 2005; Jeong et al. 2012b; Khalili et al. 2013). These hybrid approaches are based on a static noise observed during the calibration of the regression component. Therefore, the part of the variability which is explained by the randomization component does not depend on the predictors, and thus, it is supposed to be constant in a changing climate. For this reason, these hybrid approaches may not represent local change in the temporal variability in a climate change simulation. In this context, the proposed GCQR model can be considered as a hybrid approach. However, it has important advantage compared to traditional hybrid approach regarding the reproduction of the temporal variability in a changing climate. Indeed, the total temporal variability is reproduced in the regression component through quantile regression, and thus it may change in the future according to the large scale atmospheric predictors. However, the dependency structure is reproduced using the Gaussian copula.

In a regionalization framework where temperature and precipitation at ungagged sites are required, it will be necessary to extend the simulation at those ungauged locations. The

regionalization of the GCQR model can be done in two steps. In the first one, parameters of the quantile regression models are regionalized, for example, using the approach proposed by Reich (2012). The resulting model will provide the CDF at ungagged locations for each day. In the second step, through the simulation step, values between 0 and 1 are generated that are spatially correlated at ungauged locations or on the target grid. This can be achieved by simulating a spatial latent process using a kriging model applied to the latent variable $z(t)$ (given in eq. 12) and then applying the standard normal CDF to bring back the simulated values between 0 and 1.

The proposed model was conducted on a relatively small area, and thus only 6 grid points of the climate model that cover the study area were used as predictors. Then, a PCA is performed using all grid points as predictors for each station. As a direct consequence of the underlying modular structure employed here, the adaptation of the model over a large area is straightforward. However, the use of PCA including all grid points can be questioned. As mentioned by Von Storch (1999), synoptic scale atmospheric predictors can reproduce only a part of the local variability of predictands. Given the change of spatial scale between predictors and predictands, one can only expect the predictors to provide information on the smoother part of the physical process (Fasbender and Ouarda 2010). The employment of PCA using all grid points over a large area involves that the gap between the two spatial scales of predictors and predictands becomes more important. This can reduce the amount of the information that could be explained by the predictors at the small scale and the statistical link predictors-predictands should thus be less direct than when using local model outputs. In the current work, the use of the PCA with all grid points can be performed due to the small application area. However, it should be mentioned that, in larger applications, it is more appropriate to use local model output (the corresponding climate model grid point for each station).

Generally, there are two major advantages in modeling using copulas. The first is that copula functions can be adapted flexibly to the data, and there already exists a large body of theoretical models which can describe the individual characteristics of dependency structure. The second is that copula functions allow describing the dependence structure independently from the marginal distributions and thus, using different marginal distributions at the same time without any transformations. It is worth mentioning that the present paper focuses on the second advantage of the copula. The objective of the proposed study is not to determine the best fit using copula, but to consider the Gaussian copula to benefit from the advantages of quantile regression in a multisite and multivariate framework. However, it should be noted that other spatial characteristics could be important, such as the spatial extremes related to tail dependences. In this respect, the Gaussian copula may not be appropriate particularly for precipitation (El Adlouni et al. 2008; Lee et al. 2013; Serinaldi et al. 2014). To overcome this issue, other theoretical copulas can be performed such as the v-transformed copula (Bárdossy and Pegram 2009; AghaKouchak et al. 2010), meta-elliptical copulas (Fang et al. 2002), or the vine copula (Gräler 2014). Thereby, Gaussian copulas may be replaced, for instance, by one of these copulas and more development is required in future research to extend the present study in a more flexible downscaling perspective.

6. Conclusions

A GCQR model is proposed in this paper for the downscaling of AOGCM predictors to daily precipitations and maximum and minimum temperatures at multi-sites. GCQR uses a quantile regression component to specify the entire conditional distribution of each predictand, and a Gaussian copula to preserve the dependence structure between sites and predictands. The

developed model was applied at a set of stations located in the southern part of the province of Quebec (Canada). Our comparison study with classical models (MMLR and MMSDM) suggests that the GCQR model is capable of generating multisite and multivariable series simultaneously and with a realistic temporal variability. For most performance criteria, GCQR results showed better performance than MMLR and MMSDM results.

Acknowledgments

The authors wish to thank the Editor Zhaohua Wu and two reviewers, Alex J. Cannon and anonymous reviewer whose comments contributed to the improvement of the quality of the paper. We also acknowledge Eva Mekis from Environment Canada for providing observed data sets of rehabilitated precipitation. The authors would like to acknowledge also the Data Access and Integration (DAI, see http://loki.qc.ec.gc.ca/DAI/login-e.php) team for providing the predictors data and technical support. The DAI data download gateway is made possible through collaboration among the Global Environmental and Climate Change Centre (GEC3), the Adaptation and Impacts Research Section (AIRS) of Environment Canada, and the Drought Research Initiative (DRI).

7. References

AghaKouchak, A., A. Bárdossy and E. Habib (2010). "Conditional simulation of remotely sensed rainfall data using a non-Gaussian v-transformed copula." Advances in Water Resources **33**(6): 624-634.

Bárdossy, A. (2006). "Copula-based geostatistical models for groundwater quality parameters." Water Resour. Res. **42**(11): W11416.

Bárdossy, A. and G. G. S. Pegram (2009). "Copula based multisite model for daily precipitation simulation." Hydrology and Earth System Sciences **13**(12): 2299-2314.

Bargaoui, Z. K. and A. Bardossy (2015). "Modelling short duration extreme precipitation patterns using Copula and Generalized maximum pseudo-likelihood estimation with censoring." Advances in Water Resources.

Baur, D., M. Saisana and N. Schulze (2004). "Modelling the effects of meteorological variables on ozone concentration—a quantile regression approach." Atmospheric Environment **38**(28): 4689-4699.

Baur, D. G. (2013). "The structure and degree of dependence: A quantile regression approach." Journal of Banking & Finance **37**(3): 786-798.

Ben Alaya, M., F. Chebana and T. Ouarda (2014). "Probabilistic Gaussian Copula Regression Model for Multisite and Multivariable Downscaling." Journal of Climate **27**(9): 3331-3347.

Ben Alaya, M. A., F. Chebana and T. B. Ouarda (2015). "Probabilistic Multisite Statistical Downscaling for Daily Precipitation Using a Bernoulli–Generalized Pareto Multivariate Autoregressive Model." Journal of climate **28**(6): 2349-2364.

Benestad, R. E., I. Hanssen-Bauer and D. Chen (2008). Empirical-statistical downscaling, World Scientific.

Bremnes, J. B. (2004a). "Probabilistic forecasts of precipitation in terms of quantiles using NWP model output." Monthly Weather Review **132**(1).

Bremnes, J. B. (2004b). "Probabilistic wind power forecasts using local quantile regression." Wind Energy **7**(1): 47-54.

Burger, G. and Y. Chen (2005). "A Regression-based downscaling of spatial variability for hydrologic applications." Journal of Hydrology **311**: 299-317.

Cannon, A. J. (2008a). Multivariate statistical models for seasonal climate prediction and climate downscaling, The University Of British Columbia.

Cannon, A. J. (2008b). "Probabilistic multisite precipitation downscaling by an expanded Bernoulli-gamma density network." Journal of Hydrometeorology **9**(6): 1284-1300.

Cannon, A. J. (2011). "Quantile regression neural networks: Implementation in R and application to precipitation downscaling." Computers & Geosciences **37**(9): 1277-1284.

Cannon, A. J. (2012). "Neural networks for probabilistic environmental prediction: Conditional density estimation network creation and evaluation (cadence) in r." Computers & Geosciences **41**: 126-135.

Cawley, G. C., G. J. Janacek, M. R. Haylock and S. R. Dorling (2007). "Predictive uncertainty in environmental modelling." Neural Networks **20**(4): 537-549.

Chebana, F. and T. B. Ouarda (2011). "Multivariate quantiles in hydrological frequency analysis." Environmetrics **22**(1): 63-78.

Chen, J., F. P. Brissette and R. Leconte (2014). "Assessing regression-based statistical approaches for downscaling precipitation over North America." Hydrological Processes **28**(9): 3482-3504.

Chen, J., F. P. Brissette and X. J. Zhang (2015). "Hydrological Modeling Using a Multisite Stochastic Weather Generator." Journal of Hydrologic Engineering: 04015060.

Chen, X., R. Koenker and Z. Xiao (2009). "Copula-based nonlinear quantile autoregression." The Econometrics Journal **12**(s1): S50-S67.

El Adlouni, S., B. Bobée and T. Ouarda (2008). "On the tails of extreme event distributions in hydrology." Journal of Hydrology **355**(1): 16-33.

El Adlouni, S. and T. Ouarda (2008). "Study of the joint distribution flow-level by copulas: Case of the Chateauguay river." Canadian Journal of Civil Engineering **35**(10): 1128-1137.

Elsner, J. B., J. P. Kossin and T. H. Jagger (2008). "The increasing intensity of the strongest tropical cyclones." Nature **455**(7209): 92-95.

Fang, H.-B., K.-T. Fang and S. Kotz (2002). "The meta-elliptical distributions with given marginals." Journal of Multivariate Analysis **82**(1): 1-16.

Fasbender, D. and T. B. M. J. Ouarda (2010). "Spatial Bayesian Model for Statistical Downscaling of AOGCM to Minimum and Maximum Daily Temperatures." Journal of Climate **23**(19): 5222-5242.

Friederichs, P. and A. Hense (2007). "Statistical downscaling of extreme precipitation events using censored quantile regression." Monthly Weather Review **135**(6).

Gachon, P., A. St-Hilaire, T. B. M. J. Ouarda, V. Nguyen, C. Lin, J. Milton, D. Chaumont, J. Goldstein, M. Hessami, T. D. Nguyen, F. Selva, M. Nadeau, P. Roy, D. Parishkura, N. Major, M. Choux and A. Bourque (2005). "A first evaluation of the strength and weaknesses of statistical downscaling methods for simulating extremes over various regions of eastern Canada " Sub-component, Climate Change Action Fund (CCAF), Environment Canada **Final report**(Montréal, Québec, Canada): 209.

Gräler, B. (2014). "Modelling skewed spatial random fields through the spatial vine copula." Spatial Statistics **10**: 87-102.

Gulizia, C. and I. Camilloni (2015). "Comparative analysis of the ability of a set of CMIP3 and CMIP5 global climate models to represent precipitation in South America." International Journal of Climatology **35**(4): 583-595.

Harpham, C. and R. L. Wilby (2005). "Multi-site downscaling of heavy daily precipitation occurrence and amounts." Journal of Hydrology **312**(1): 235-255.

Haylock, M. R., G. C. Cawley, C. Harpham, R. L. Wilby and C. M. Goodess (2006). "Downscaling heavy precipitation over the United Kingdom: A comparison of dynamical and statistical methods and their future scenarios." International Journal of Climatology **26**(10): 1397-1415.

Herrera, E., T. B. M. J. Ouarda and B. Bobée (2006). "Downscaling methods applied to Atmosphere-Ocean General Circulation Models (AOGCM)." Méthodes de désagrégation Appliquées aux Modèles du Climat Global Atmosphère-Océan (MCGAO) **19**(4): 297-312.

Hessami, M., P. Gachon, T. B. M. J. Ouarda and A. St-Hilaire (2008). "Automated regression-based statistical downscaling tool." Environmental Modelling & Software **23**(6): 813-834.

Jagger, T. H. and J. B. Elsner (2009). "Modeling tropical cyclone intensity with quantile regression." International Journal of Climatology **29**(10): 1351.

Jeong, D., A. St-Hilaire, T. Ouarda and P. Gachon (2013). "A multivariate multi-site statistical downscaling model for daily maximum and minimum temperatures." Climate Research **54**(2): 129-148.

Jeong, D. I., A. St-Hilaire, T. B. M. J. Ouarda and P. Gachon (2012a). "Comparison of transfer functions in statistical downscaling models for daily temperature and precipitation over Canada." Stochastic Environmental Research and Risk Assessment **26**(5): 633-653.

Jeong, D. I., A. St-Hilaire, T. B. M. J. Ouarda and P. Gachon (2012b). "Multisite statistical downscaling model for daily precipitation combined by multivariate multiple linear regression and stochastic weather generator." Climatic Change **114**(3-4): 567-591.

Kazianka, H. and J. Pilz (2010). "Copula-based geostatistical modeling of continuous and discrete data including covariates." Stochastic Environmental Research and Risk Assessment **24**(5): 661-673.

Khalili, M., V. T. Van Nguyen and P. Gachon (2013). "A statistical approach to multi-site multivariate downscaling of daily extreme temperature series." International Journal of Climatology **33**(1): 15-32.

Koenker, R. and G. Bassett (1978). "Regression quantiles." Econometrica: journal of the Econometric Society: 33-50.

Kusunoki, S. and O. Arakawa (2015). "Are CMIP5 Models Better than CMIP3 Models in Simulating Precipitation over East Asia?" Journal of Climate **28**(14): 5601-5621.

Lee, T., R. Modarres and T. Ouarda (2013). "Data-based analysis of bivariate copula tail dependence for drought duration and severity." Hydrological Processes **27**(10): 1454-1463.

Lindström, G., B. Johansson, M. Persson, M. Gardelin and S. Bergström (1997). "Development and test of the distributed HBV-96 hydrological model." Journal of Hydrology **201**(1-4): 272-288.

Luce, C. H. and Z. A. Holden (2009). "Declining annual streamflow distributions in the Pacific Northwest United States, 1948–2006." Geophysical Research Letters **36**(16).

Reich, B. J. (2012). "Spatiotemporal quantile regression for detecting distributional changes in environmental processes." Journal of the Royal Statistical Society: Series C (Applied Statistics) **61**(4): 535-553.

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.

Sankarasubramanian, A. and U. Lall (2003). "Flood quantiles in a changing climate: Seasonal forecasts and causal relations." Water Resources Research **39**(5).

Schölzel, C. and P. Friederichs (2008). "Multivariate non-normally distributed random variables in climate research - Introduction to the copula approach." Nonlinear Processes in Geophysics **15**(5): 761-772.

Serinaldi, F., A. Bárdossy and C. G. Kilsby (2014). "Upper tail dependence in rainfall extremes: would we know it if we saw it?" Stochastic Environmental Research and Risk Assessment **29**(4): 1211-1233.

Sklar, A. (1959). Fonctions de répartition à n dimensions et leurs marges.

Stephenson, D. B., K. Rupa Kumar, F. J. Doblas-Reyes, J. F. Royer, F. Chauvin and S. Pezzulli (1999). "Extreme daily rainfall events and their impact on ensemble forecasts of the Indian monsoon." Monthly Weather Review **127**(9): 1954-1966.

Tareghian, R. and P. F. Rasmussen (2013). "Statistical downscaling of precipitation using quantile regression." Journal of Hydrology **487**: 122-135.

Von Storch, H. (1999). "On the Use of "Inflation" in Statistical Downscaling." Journal of Climate **12**(12): 3505-3506.

Weerts, A., H. Winsemius and J. Verkade (2011). "Estimation of predictive hydrological uncertainty using quantile regression: examples from the National Flood Forecasting System (England and Wales)." Hydrology and Earth System Sciences, 15,(1).

Wilby, R. (1998). "Statistical downscaling of daily precipitation using daily airflow and seasonal teleconnection indices." Climate Research **10(3)**.

Wilby, R., O. Tomlinson and C. Dawson (2003). "Multi-site simulation of precipitation by conditional resampling." Climate Research **23**(3): 183-194.

Wilby, R. L., C. W. Dawson and E. M. Barrow (2002). "sdsm - a decision support tool for the assessment of regional climate change impacts." Environmental Modelling and Software **17**(2): 145-157.

Wilby, R. L., H. Hassan and K. Hanaki (1998). "Statistical downscaling of hydrometeorological variables using general circulation model output." Journal of Hydrology **205**(1): 1-19.

Williams, P. M. (1998). "Modelling seasonality and trends in daily rainfall data." Advances in neural information processing systems: 985-991.

Yang, C., R. E. Chandler, V. S. Isham and H. S. Wheater (2005). "Spatial-temporal rainfall simulation using generalized linear models." Water Resources Research **41**(11): 1-13.

Table 1. List of the nine temperature stations used in this study.

No.	Name of station	Latitude (°N)	Longitude (°W)
1	Cedars	45.30	74.05
2	Drummondville	45.88	72.48
3	Seven Islands	50.22	66.27
4	Bagotville A	48.33	71.00
5	Quebec	46.79	71.38
6	Sherbrooke A	45.43	71.68
7	Maniwaki Airport	46.27	75.99
8	La Pocatière	47.36	70.03
9	Mont-Joli A	48.60	68.22

Table 2. List of the nine precipitation stations used in this study.

No.	Name of station	Latitude (°N)	Longitude (°W)
1	Chelsea	45.52	75.78
2	Cedars	45.30	74.05
3	Nicolet	46.25	72.60
4	Drummondville	45.88	72.48
5	Donnacona	46.69	71.73
6	Roberval A	48.52	72.27
7	Bagitville A	48.33	71.00
8	Rimouski	48.45	68.53
9	Seven Islands	50.22	66.27

Table 3. NCEP predictors on the CGCM3 grid.

No	Predictors	No	Predictors
1	mean pressure at the sea level	14	Divergence at 500 hPa
2	Wind speed at 1000 hPa	15	Wind speed at 850 hPa
3	Component U at 1000 hPa	16	Component U at 850 hPa
4	Component V at 1000 hPa	17	Component V at 850 hPa
5	Vorticity at 1000 hPa	18	Vorticity at 850 hPa
6	Wind direction at 1000 hPa	19	Geopotential at 850 hPa
7	Divergence at 1000 hPa	20	Wind direction at 850 hPa
8	Wind speed at 500 hPa	21	Divergence at 1000 hPa
9	Component U at 500 hPa	22	Specific humidity at 500 hPa
10	Component V at 500 hPa	23	Specific humidity at 850 hPa
11	Vorticity at 500 hPa	24	Specific humidity at 1000 hPa
12	Geopotential at 500 hPa	25	Temperature at 2m
13	Wind direction at 500 hPa		

Table 4. Quality assessment of the estimated series for GCQR, MMLR and MMSDM during the validation period (1991–2000) for the nine temperature stations. Criteria are ME, RMSE, differences between observed and modeled variance D and the percentage of observations in 95% confidence interval. For the GCQR and MMSDM models, ME and RMSE criteria were calculated from the conditional mean.

		ME (°C)		RMSE (°C)		$D(°C^2)$		95% CI	
		Tmax	Tmin	Tmax	Tmin	Tmax	Tmin	Tmax	Tmin
Cedars	GCQR	0.55	**0.21**	3.28	3.70	**-0.50**	**-0.50**	**95**	**94**
	MMLR	0.55	0.22	3.28	3.70	8.91	9.21	--	--
	MMSDM	0.55	0.22	3.30	3.71	-3.11	-5.66	96	97
Drummondville	GCQR	0.47	0.49	3.31	4.08	**-3.40**	**6.13**	**96**	**95**
	MMLR	0.47	0.49	3.31	4.08	7.22	18.15	--	--
	MMSDM	0.47	0.49	3.32	4.10	4.25	8.18	92	93
Seven-Island	GCQR	0.19	-0.53	3.17	3.59	**6.74**	**2.33**	**94**	**93**
	MMLR	0.19	-0.53	3.17	**3.58**	16.36	13.42	--	--
	MMSDM	0.19	-0.53	3.17	3.60	-22.49	6.02	97	92
Bagotville A	GCQR	-0.05	**0.13**	3.55	3.84	**1.95**	**-0.15**	**95**	**96**
	MMLR	-0.05	0.14	3.53	3.84	15.42	12.76	--	--
	MMSDM	-0.05	0.14	3.53	3.85	31.17	-21.55	97	98
Quebec	GCQR	0.33	0.16	3.22	3.40	**-2.43**	**1.09**	**95**	**92**
	MMLR	0.33	0.17	3.22	3.39	9.53	10.99	--	--
	MMSDM	0.33	0.16	3.24	3.39	6.33	-8.35	89	90
Sherbrooke A	GCQR	0.79	0.17	3.55	4.27	**-0.63**	-1.39	**95**	94
	MMLR	0.79	0.18	3.54	4.26	10.59	12.39	--	--
	MMSDM	0.78	0.17	3.56	4.27	-6.01	**-4.65**	93	94
Maniwaki Airport	GCQR	0.33	-0.06	3.39	4.10	**3.67**	**-9.88**	97	**95**
	MMLR	0.32	-0.04	3.39	4.10	11.50	4.89	--	--
	MMSDM	0.31	-0.05	3.41	4.12	8.28	7.67	**94**	96
La Pocatière	GCQR	0.82	-0.23	3.35	3.54	**1.39**	**-1.55**	96	**94**
	MMLR	0.82	-0.22	3.35	3.53	10.64	9.78	--	--
	MMSDM	0.81	-0.23	3.37	3.55	-2.45	-16.63	96	95
Mont-Joli	GCQR	0.74	0.40	3.23	3.23	**1.88**	**-0.69**	**93**	**92**
	MMLR	0.74	0.42	3.53	3.22	13.46	8.79	--	--
	MMSDM	0.74	0.41	3.24	3.23	-8.76	-30.43	90	91

Table 5. Quality assessment of the estimated series for GCQR, MMLR and MMSDM during the validation period (1991–2000) for the nine precipitation stations. Criteria are ME, RMSE, differences between observed and modeled variance D and the percentage of observations in 95% confidence interval. For the GCQR and MMSDM models, ME and RMSE criteria were calculated from the conditional mean. Bold means better result.

	ME(mm)			RMSE(mm)			D(mm^2)			95%CI	
	GCQR	MMLR	MMSDM	GCQR	MMLR	MMSDM	GCQR	MMLR	MMSDM	GCQR	MMSDM
Chelsea	**-0.18**	2.29	0.90	**5.82**	6.41	5.90	**0.92**	37.94	14.47	**95**	96
Cedars	**0.07**	2.59	0.97	**6.38**	7.04	6.49	**9.03**	45.99	23.69	**94**	93
Nicolet	**1.05**	2.75	1.34	7.41	7.73	**7.15**	**25.82**	57.42	38.59	**91**	89
Drummondville	**0.41**	2.44	0.99	**5.35**	6.14	5.58	**10.48**	34.54	11.96	92	92
Donnacona	**0.83**	2.88	1.35	**6.17**	6.97	6.30	**14.47**	43.85	21.21	**94**	90
Roberval A	**0.49**	2.15	0.83	**5.29**	5.82	5.36	**4.03**	31.39	8.17	**95**	94
Bagotville A	**0.72**	2.55	1.29	**6.35**	6.92	6.40	22.88	43.67	**20.37**	92	**93**
Rimouski	**0.32**	2.07	0.91	5.42	5.79	**5.32**	**7.09**	32.24	9.36	96	96
Seven Islands	**-0.35**	2.17	0.78	**5.33**	6.07	5.59	**0.89**	34.19	11.18	97	93

Table 6. Definition of climatic indices used for the performance assessment of downscaled temperatures.

Indices	Definition	Unite	Scale Time
DTR	Mean of diurnal temperature range	°C	Season
FSL	Frost season length: Days between 5 consecutive T_{mean}<0 ℃ and 5 consecutive T_{mean}> 0℃	Days	Years
GSL	Growing Season length: Days between 5 consecutive T_{mean}<5 ℃ and 5 consecutive T_{mean}> 5℃	Days	Years
FR-Th	Days with freeze and thaw (T_{max}> 0°C,T_{min}<0°C)	Days	Months
Tmax90	90th percentile of daily maximum temperature	°C	seasons
Tmin90	90^{th} percentile of daily minimum temperature	°C	seasons

$$T_{mean} = \frac{T_{max} + T_{min}}{2}$$

Table 7. Definition of climatic indices used for the performance assessment of downscaled precipitations.

Indices	Definition	Unite	Scale Time
MWD	Mean precipitation of wet day	Mm	Months
Pmax90	90^{th} percentile of daily precipitation	Mm	Seasons
PX1D	Maximum 1-days precipitation	Mm	Months
PX3D	Maximum 3-days precipitation	Mm	Months
PX5D	Maximum 5-days precipitation	Mm	Months
WRUN	Maximum number of consecutive wet days	Days	Months
DRUN	Maximum number of consecutive dry days	Days	Months
NWD	Number of wet day	Days	Months

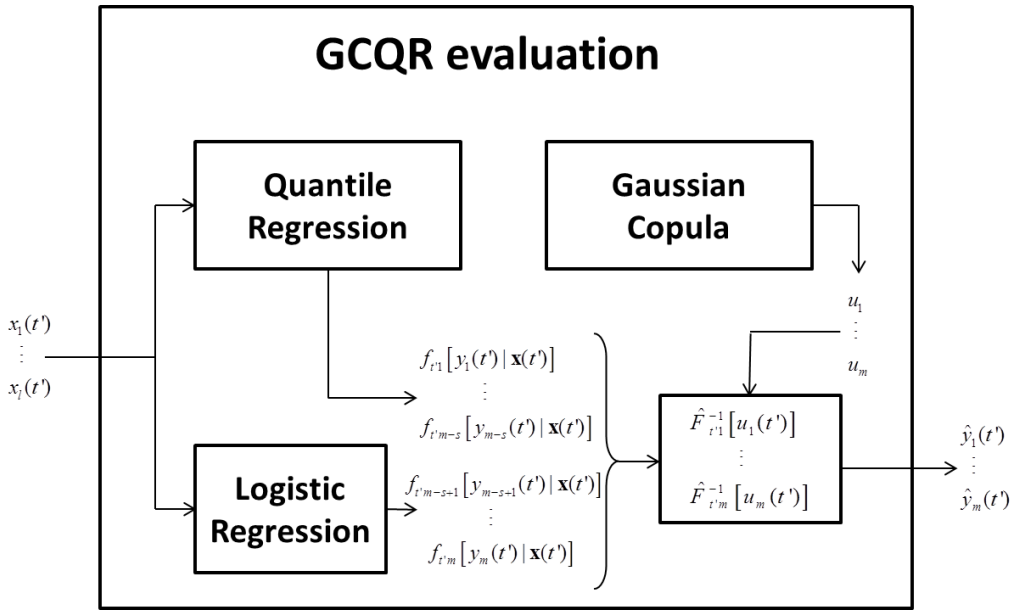

Figure 1. Simulation procedure of the GCQR model, using the probabilistic regression model and the Gaussian copula.

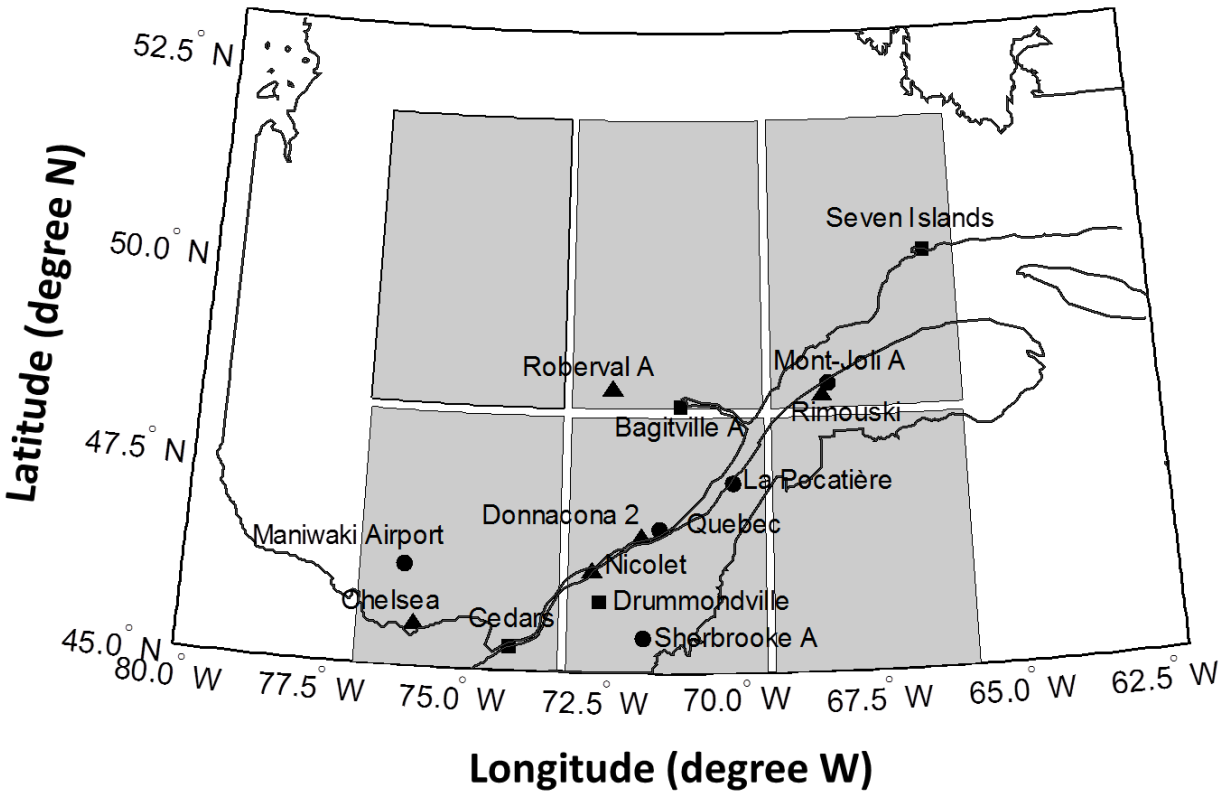

Figure 2. Location of CGCM3 grid and observation stations of daily precipitation and daily maximum and minimum temperature. Cedars, Drummondville, Seven Islands and Bagotville stations, contain both precipitation and temperature data. These four stations are represented by squares. The other temperature stations are illustrated by circles and the other precipitation stations are illustrated by triangles.

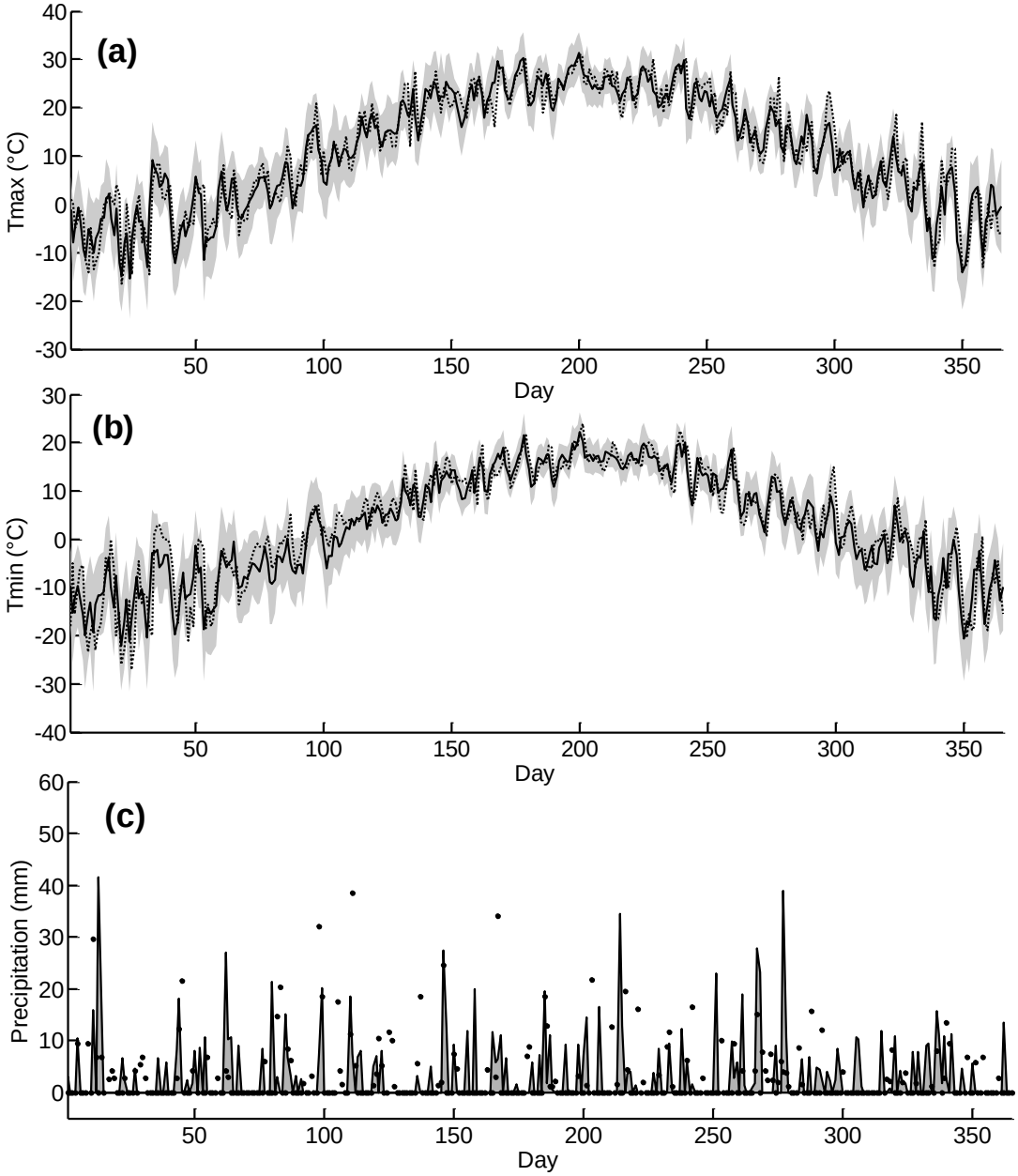

Figure 3. GCQR result for Chelsea station during 1991. Time series of (a) maximum temperature, (b) minimum temperature and (c) precipitation. In (a) and (b) The conditional mean is shown by the solid line, the observed series is shown by the dashed line, and the 95% confidence interval obtained from the estimated standard deviations is illustrated with the gray shaded area. In (c) the simulated precipitation is shown by vertical gray lines and the observed precipitation values are shown by black dots.

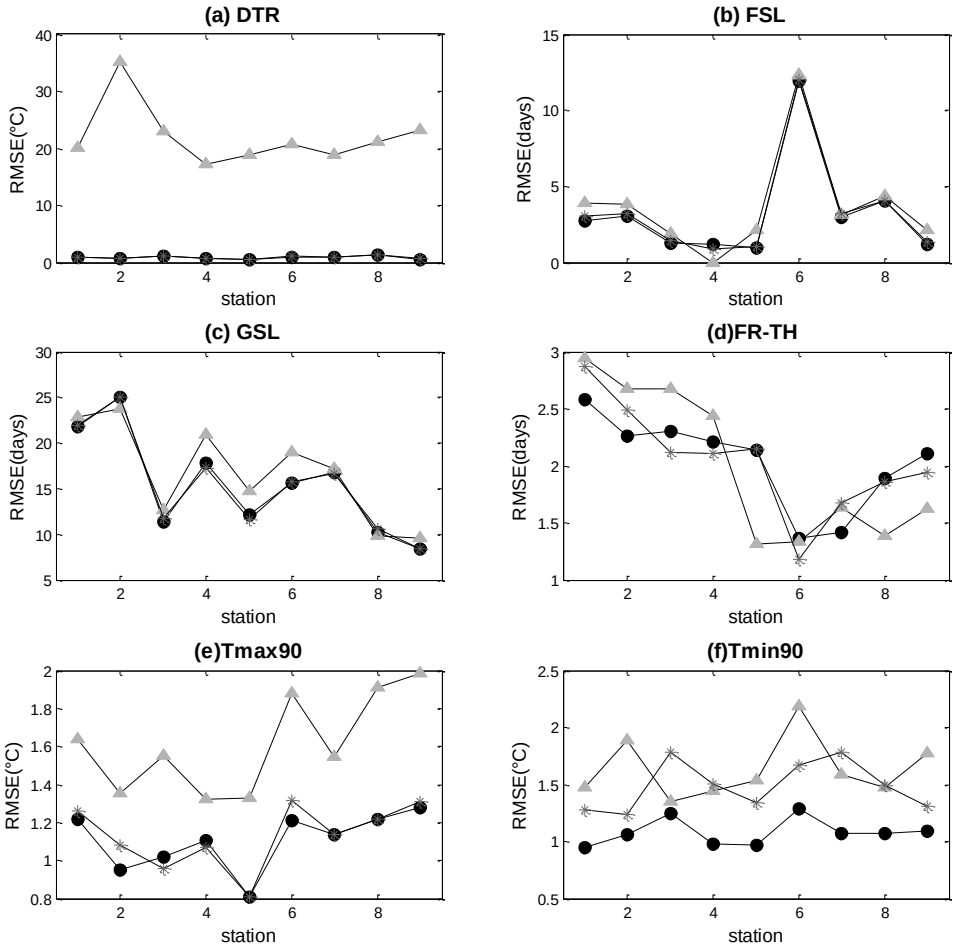

Figure 4. RMSE between observed and estimated climatic indices of downscaled temperatures for GCQR (gray dots), MMSDM (gray stars) and MMLR (gray triangles) models on the nine temperature stations during the validation period (1991–2000). RMSE of GCQR and MMSDM models were calculated using the mean of 100 realisations.

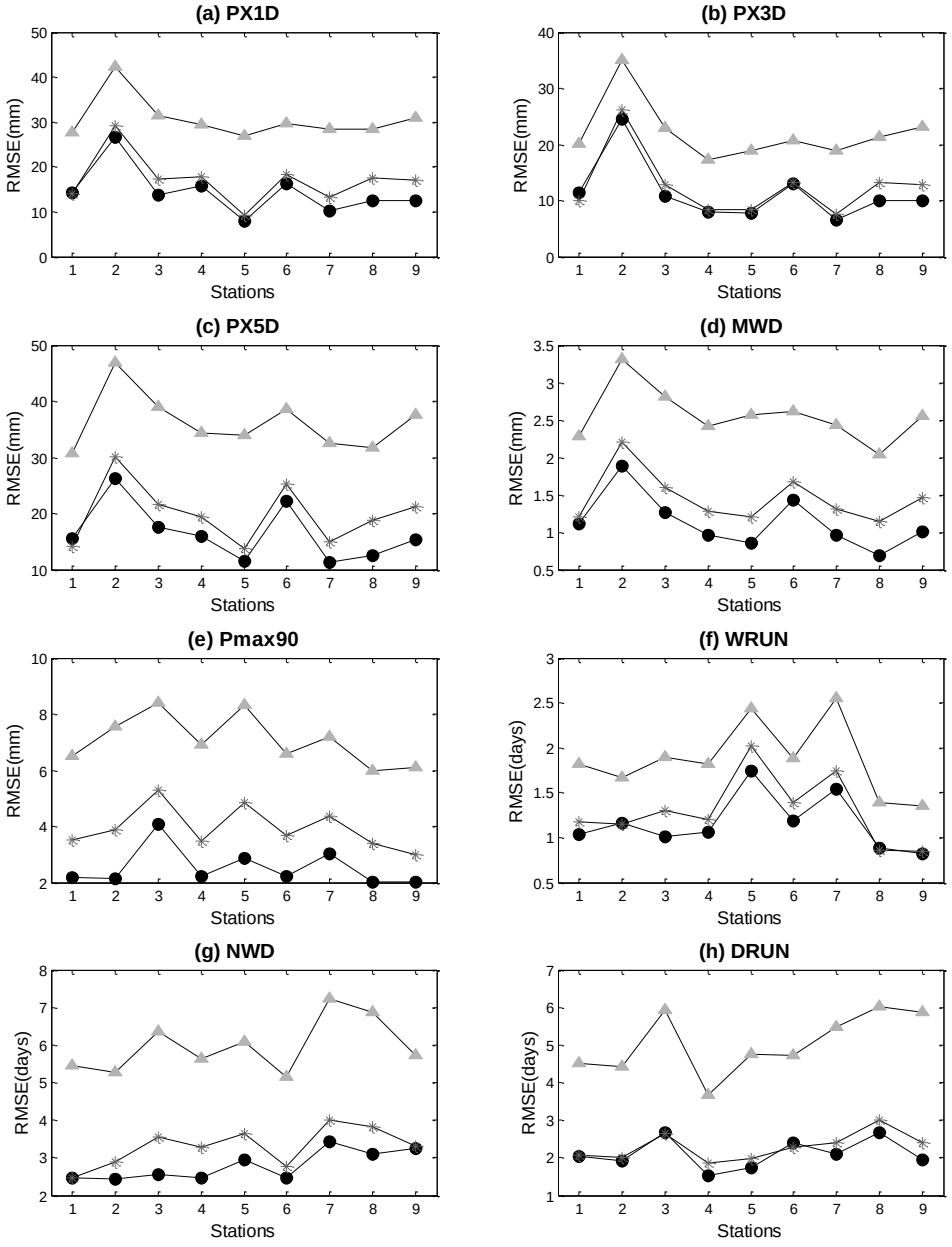

Figure 5. RMSE between observed and estimated climatic indices of downscaled precipitations for GCQR (black dots), MMSDM (gray stars) and MMLR (gray triangles) models on the nine precipitatyion stations during the validation period (1991–2000). RMSE of GCQR and MMSDM models were calculated using the mean of 100 realisations.

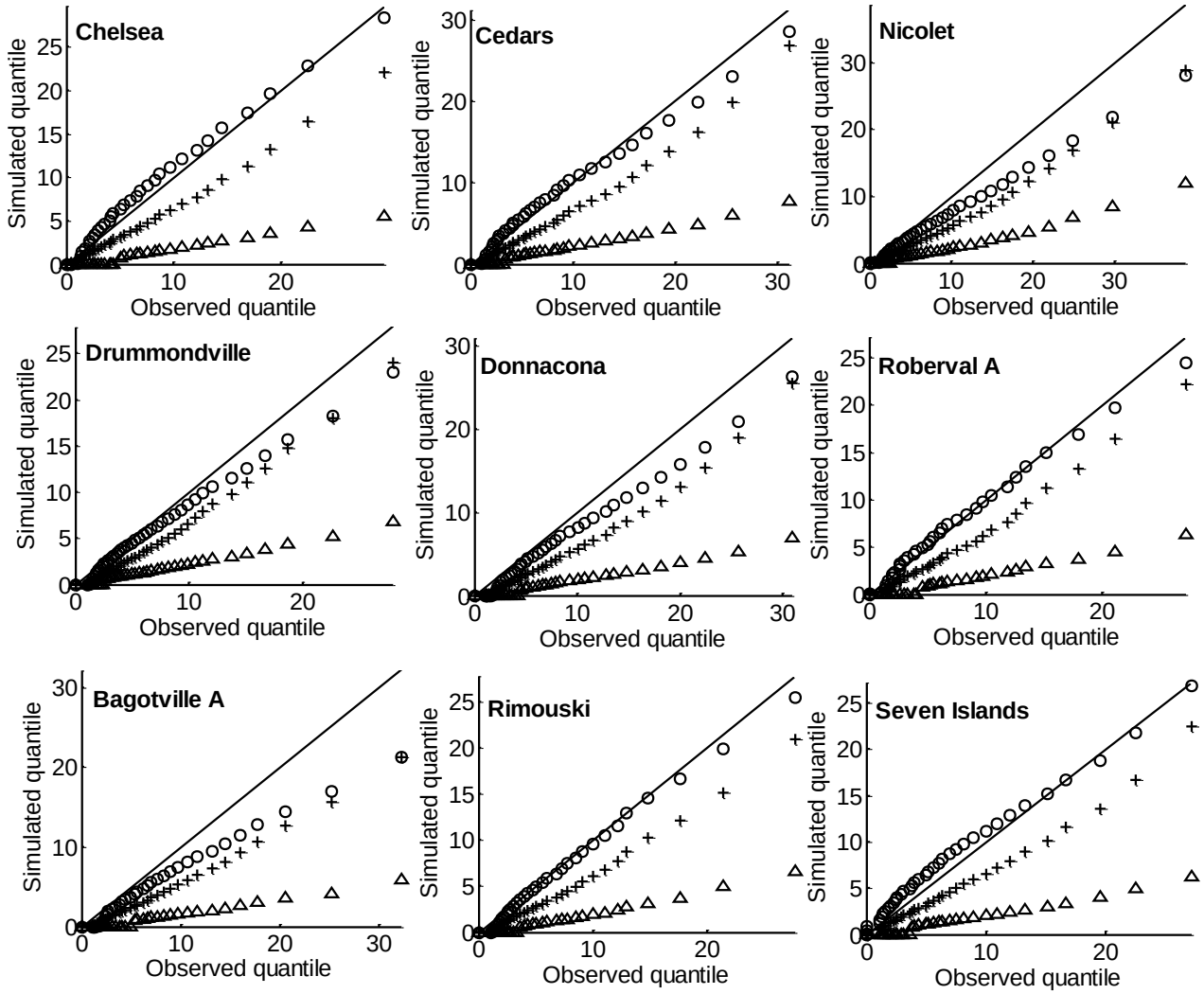

Figure 6. Q–Q plot comparison of observed precipitation vs simulated ones with GCQR (circles), MMSDM (plus) and MMLR (triangles) for the nine stations in the validation period.

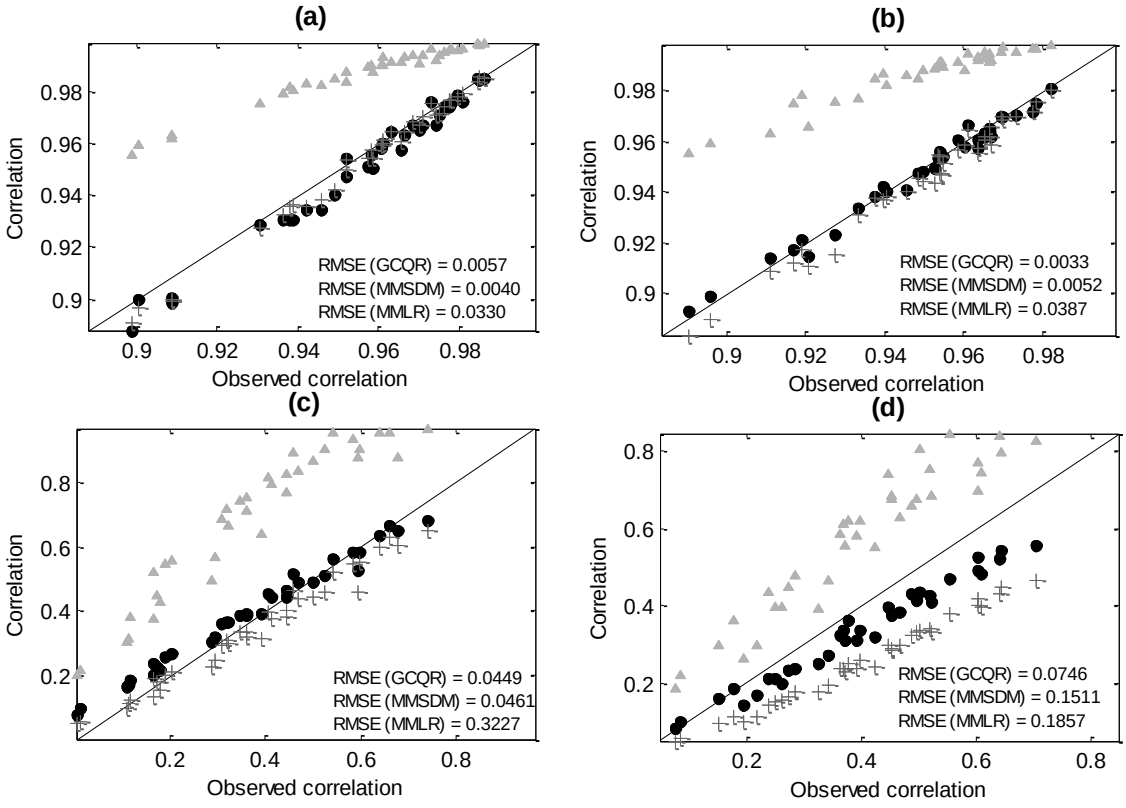

Figure 7. Scatter plot of observed and modeled cross-site correlations by GCQR (black dots), MMLR (gray triangle) and MMSDM (gray plus) for maximum temperature (a), minimum temperature (b), precipitation amount (c) and precipitation occurrence (d). Correlation values of GCQR and MMSDM models are obtained using the mean of the correlation values calculated from 100 simulations.

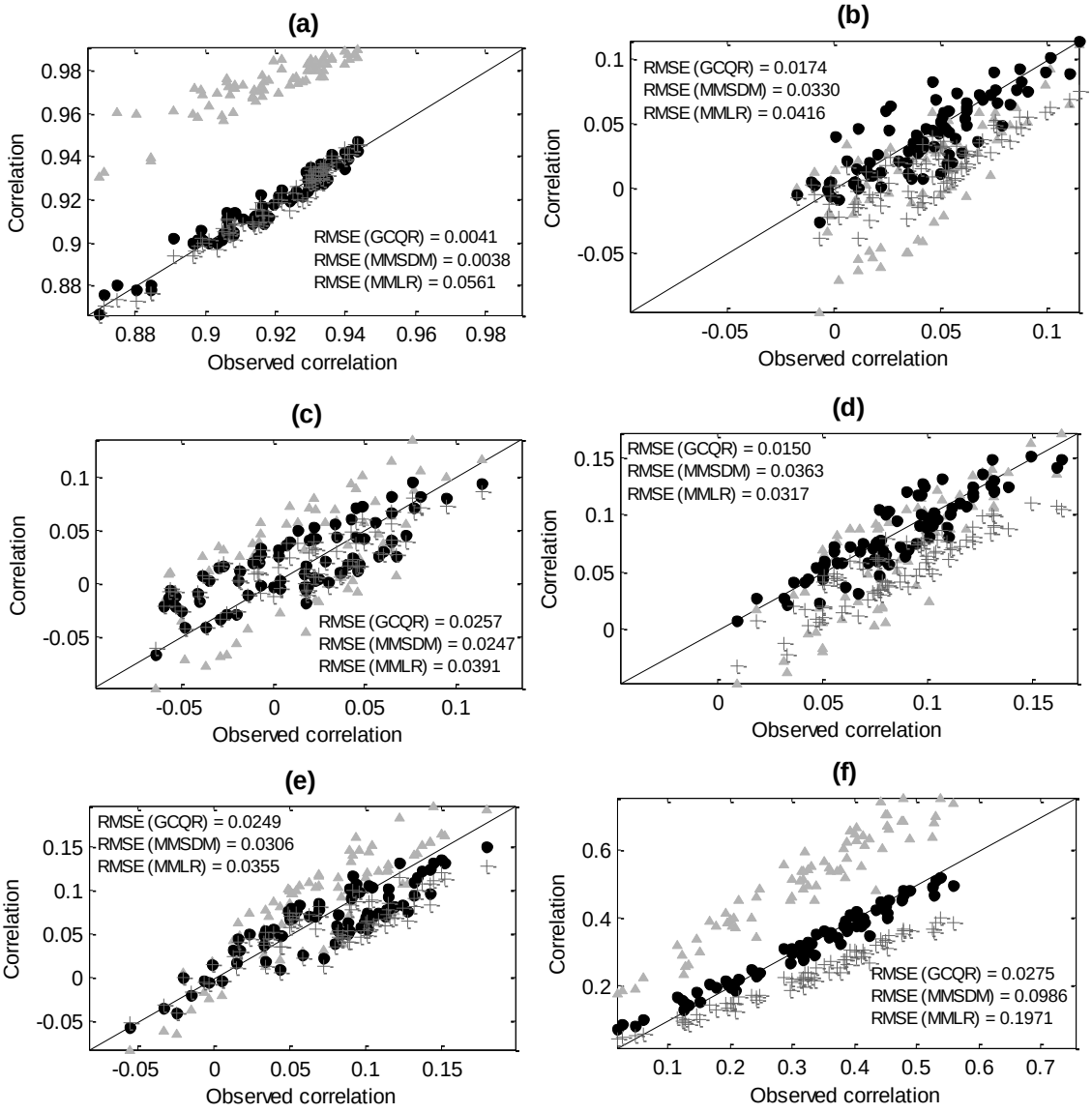

Figure 8. Scatter plot of observed and modeled correlations by GCQR (black dots), MMLR (gray triangle) and MMSDM (gray plus) for *Tmax-Tmin* (a), *Tmax-Pam* (b), *Tmax-Poc* (c), *Tmin-Pam* (d), *Tmin-Poc* (e) and *Pam-Poc* (f). Correlation values of GCQR and MMSDM models are obtained using the mean of the correlation values calculated from 100 simulations.

CHAPITRE 5: NON-GAUSSIAN SPATIOTEMPORAL SIMULATION OF MULTISITE DAILY PRECIPITATIONS: A DOWNSCALING FRAMEWORK

Non-Gaussian spatiotemporal simulation of multisite daily precipitations: downscaling framework

M. A. Ben Alaya[1], F. Chebana[1] and T.B.M.J. Ouarda[2, 1]

[1]*INRS-ETE, 490 rue de la Couronne, Québec (QC),*

Canada G1K 9A9

[2] Institute Center for Water and Environment (iWATER), *Masdar Institute of science and technology, P.O. Box 54224, Abu Dhabi, UAE*

[*]**Corresponding author:** Tel: +1 (418) 654 2530#4468

Email: mohammed_ali.ben_alaya@ete.inrs.ca

Submitted October 12, 2015

(Climate Dynamics)

Abstract:

Probabilistic regression approaches for downscaling daily precipitation are very useful. They provide the whole conditional distribution at each forecast step which ensures the preservation of a realistic temporal variability. The question addressed in this paper is: How to simulate spatiotemporal characteristics of multisite daily precipitation from probabilistic regression models? Recent publications point out to the complexity of multisite properties of daily precipitation and highlight the need of using a non-Gaussian flexible tool. This work proposes a fair compromise between simplicity and flexibility avoiding model misspecification. A suitable nonparametric bootstrapping (NB) technique is adopted. A downscaling model which merges a vector generalized linear model (VGLM as a probabilistic regression tool) and the proposed bootstrapping technique is introduced to simulate realistic multisite precipitation series. The model is applied to data sets from the southern part of the province of Quebec, Canada. It is shown that the model is capable of reproducing both at-site properties and the spatial structure of daily precipitations. Results indicate the superiority of the proposed NB technique, over a multivariate autoregressive Gaussian framework (i.e. Gaussian copula).

Keywords: Statistical downscaling, Vector generalized linear model, Multisite daily precipitation, Copula, Multivariate autoregressive Gaussian field, Binary entropy, Non parametric bootstrapping.

1. Introduction

Atmosphere–ocean general circulation models (AOGCMs) are the initial source of information for assessing the evolution of the earth's climate system. However, the spatial resolution of AOGCMs is too coarse for regional and local climate studies. The above limitation has led to the development of downscaling techniques. These techniques include dynamical downscaling which includes a set of physically based limited area models (Eum et al. 2012), and statistical downscaling which identifies a statistical link between large scale atmospheric variables (predictors) and local variables (predictands) (Benestad et al. 2008). Among a number of weather variables, precipitation poses the largest challenges from a downscaling perspective. In several hydro-climatic studies, precipitation is shown to be the most dominating weather variable to explicitly affect the water resources systems. Precipitation data are generally collected at various sites, and downscaling techniques are required to adequately reproduce the observed temporal variability and to maintain the consistency of the spatiotemporal properties of precipitation at several sites. Properly reproducing the temporal variability in downscaling applications is very important in order to adequately represent extreme events. Furthermore, maintaining realistic relationships between downscaled temperature and precipitation is particularly important for a number of applications such as hydrological modelling (Lindström et al. 1997).

Several Statistical downscaling techniques have been developed in the literature. These methods can be divided into three main approaches: stochastic weather generators (Wilks and Wilby 1999), weather typing (Conway et al. 1996) and regression methods (Hessami et al. 2008; Jeong et al. 2012a). Classical regression methods are commonly used because of their ease of implementation and their low computational requirement but they have several inadequacies. First and most importantly, they generally provide only the mean or the central part of the

predictands and thus they underrepresent the temporal variability (Cawley et al. 2007). Second, they do not adequately reproduce various aspects of the spatial and temporal dependence of the variables (Harpham and Wilby 2005).

In this regard, probabilistic regression approaches have provided useful contributions in downscaling applications to accurately reproduce the observed temporal variability. Probabilistic regression approaches include: the Bayesian formulation (Fasbender and Ouarda 2010), quantiles regression (Bremnes 2004; Friederichs and Hense 2007; Cannon 2011) and regression models where outputs are parameters of the conditional distribution such us the vector form of the generalized linear model (VGLM), the vector form of the generalized additive model (VGAM) (Yee and Wild 1996; Yee and Stephenson 2007) and the conditional density estimation network (CDEN) (Williams 1998; Li et al. 2013b). Probabilistic regression approaches enable the definition of a complete dynamic univariate distribution function. In the case of VGLM, VGAM and CDEN, the output of the model is a vector of parameters of a distribution which depends on the predictors values. In addition to the location parameter (namely the mean) , the scale and shape parameters can vary according to the updated values of atmospheric predictors and thus allowing for a better control and fit of the dispersion, skewness and kurtosis. Therefore, simulation of downscaled time series with a realistic temporal variability is achieved by drawing random numbers from the modeled conditional distribution at each forecast step (Williams 1998; Haylock et al. 2006). In this respect, the problem that arises is how to extend probabilistic regression approaches in multisite downscaling tasks.

Operationally, the multi-site replicates of the field predictands are readily obtained in the simulation stage. Generally, generating from a probabilistic regression model can be achieved by drawing random numbers from the standard uniform distribution and then applying the inverse

cumulative distribution function of the parent distribution obtained from the probabilistic regression model. We must keep in mind that, the parameters of the parent distribution change at each forecast step based on the updated values of large-scale atmospheric predictors. To obtain spatially correlated simulations, we need to simulate standard uniform random variables that are correlated. Thus, generating from a multivariate distribution on the unit cube (i.e, with uniform margins) could solve the issue. Such a multivariate distribution is called copula. Copula functions allow describing the dependence structure independently from the marginal distributions and thus, using different marginal distributions at the same time without any transformations. During the last decade, the application of copulas in hydrology and climatology has grown rapidly. An introduction to the copula theory is provided in Joe (1997) and Nelsen (2013). The reader is directed to Genest and Chebana (2015) and Salvadori and De Michele (2007) for a detailed review of the development and applications of copulas in hydrology including frequency analysis, simulation, and geostatistical interpolation (Bárdossy and Li 2008; Chebana and Ouarda 2011; Requena et al. 2015; Zhang et al. 2015). In recent years, copula functions have been widely used to describe the dependence structure of climate variables and extremes (AghaKouchak 2014; Guerfi et al. 2015; Hobæk Haff et al. 2015; Mao et al. 2015; Vernieuwe et al. 2015).

To extend the probabilistic regression approach to multisite downscaling, Ben Alaya et al. (2014) proposed a Gaussian copula procedure. Nevertheless, this approach does not take into account cross-correlations lagged in time and thus it cannot reproduce the short term autocorrelation properties of downscaled series such us the lag-1 cross-correlation. To solve this issue Ben Alaya et al. (2015) employed a multivariate autoregressive field as an extension to the Gaussian copula to account for the lag-1 cross-correlation. On the other hand, a careful examination of the dependence structure in hydrometeorological processes reveals that the meta-Gaussian

framework is very restrictive and cannot account for features like asymmetry and heavy tails and thus cannot realistically simulate the multisite dependency structure of daily precipitations (El Adlouni et al. 2008; Bárdossy and Pegram 2009; Lee et al. 2013).

To exploit this knowledge for precipitation simulation, Li et al. (2013a) and Serinaldi (2009) considered copulas to introduce non-Gaussian temporal structures at a single site. Bargaoui and Bárdossy (2015) employed a bivariate copula to model short duration extreme precipitation. For multisite precipitation simulation, Bárdossy and Pegram (2009) and AghaKouchak et al. (2010) introduced non-Gaussian spatial tail dependency structures by simulating precipitations from a v-transformed normal copula proposed by Bárdossy (2006). Other theoretical models of copula can also be used to reproduce this spatial tail dependency properties such as metaelliptical copulas (Fang et al. 2002) or using vine copula (Gräler 2014).

In the case of precipitation simulation it would be useful to implement a spatiotemporal flexible copula that allows simultaneously modelling both temporal and spatial dependency. To our best knowledge, such a copula has not been exploited in the hydrometeorological literature including for downscaling, except for the multivariate autoregressive meta-Gaussian copula. Nevertheless, in the statistical literature Smith (2014) employed a vine copula to achieve this end. In the last decade, vine copulas emerged as a new efficient technique in econometrics. Vine copula uses pair copula building blocks offering a flexible way to capture the inherent dependency patterns of high dimensional data sets, with regard to their symmetries, strength of dependence and tail dependency. On the other hand, the full specification of a vine copula model is not straightforward, since it requires the choice of a tree structure of the vine copula, the copula families for each pair copula term and their corresponding parameters (Czado et al. 2013). In addition, the application for spatial and temporal structure dependency greatly increases the

number of parameters which would unquestionably make the model less parsimonious and increase the associated uncertainty.

It is almost axiomatic that we need to select a flexible copula that provides both temporal and spatial dependence. However, the key question, in our problem is not to simulate dependence structure from a copula that gives best fit to the data. Instead, the question is: "how to extract information about the data dependence structure, and how to preserve it in the simulation step?". Before considering this question we should first know where we can find this information. The data rank matrix may be considered as the support set for the empirical copula. We recall that the data ranks are the statistics retaining the greatest amount of information about the data dependence structure (Oakes 1982; Genest and Plante 2003; Song and Singh 2010). In this context, information about the data dependence structure can be reproduced in the simulation step by resampling using the data ranks (Vinod and López-de-Lacalle 2009; Vaz de Melo Mendes and Leal 2010; Srivastav and Simonovic 2014). In this respect, the aim of the present paper is to propose a new approach to maximize the amount of information about the dependence structure that is preserved in the simulation step from a probabilistic regression downscaling model. Hence, instead of using a flexible copula, a simple non-parametric bootstrapping technique is employed. The procedure consists in generating uniform random series between 0 and 1 and then sorting them according to their observed ranks. Therefore, the observed ranks are preserved which consequently allows preserving a large amount of spatiotemporal dependence structure. The resulting multisite precipitation downscaling model involves a new hybrid procedure merging a parametric probabilistic regression model (the VGLM) and a non-parametric bootstrapping (NB) technique. The introduced bootstrapping technique represents a fair

compromise between simplicity and flexibility to generate realistic multisite properties of precipitation from a probabilistic regression model.

The paper is structured as follows: after this introduction, the proposed hybrid multisite VGLM-NB model is described. An application to a case study of daily data sets from the province of Quebec is carried out. The model validation is done using statistical characteristics such as mean, standard deviation, dependence structure (both spatial and temporal), precipitation indices and an entropy-based congregation measure. Obtained results are compared to those corresponding to a VGLM-MAR which is a VGLM combined with multivariate autoregressive (MAR) Gaussian field. Finally discussions and conclusions are given.

2. Study area and data

Observed daily precipitations from nine Environment Canada weather stations located in the province of Quebec (Canada) are used in this study (see Figure 1). The list of stations is presented in Table 1. Predictor variables are obtained from the reanalysis product NCEP/NCAR interpolated on the CGCM3 Gaussian grid (3.75 ° latitude and longitude). Six grids covering the predictand stations area are selected (see Figure 1), and 25 NCEP predictors are available for each grid (see Table 2). Thus, a total of 150 daily predictors are available for the downscaling process. To reduce the number of predictors, a principal component analysis (PCA) is performed. The first principal components that preserve more than 97% of the total variance are selected. The data sets cover the period between January, 1st 1961 and December, 31st 2000. This record period is divided into two periods for the calibration (1961-1980) and the validation (1981-2000).

3. Methodology

In this section, the proposed VGLM-NB model is presented. The corresponding probabilistic framework is presented with a description of the conditional Bernoulli-Generalized Pareto regression model and the proposed nonparametric bootstrapping technique.

3.1. Vector generalized linear model

The precipitation amount distribution, at a daily time scale, tends to be strongly skewed, and is commonly assumed to be gamma distributed (Stephenson et al. 1999; Giorgi et al. 2001; Yang et al. 2005). In a regression perspective, the generalized linear model (GLM) extends classical regression to handle the normality assumption of the model output. Here the output may follow a range of distributions that allow the variance to depend on the mean such us the exponential distribution family and particularly the Gamma distribution (Coe and Stern 1982; Stern and Coe 1984; Chandler and Wheater 2002). Nevertheless, recent findings suggest that the gamma distribution can be unsuitable for modeling precipitation extremes since it is very restrictive and cannot account for features like heavy tails. Therefore, to treat this issue other options have been proposed in the literature particularly the generalized Pareto (GP) and the Weibull (WEI) distributions (Ashkar and Ouarda 1996; Serinaldi and Kilsby 2014). However, due to the fact that the variance does not depend on the mean, these two distributions cannot be used in a GLM. Vector generalized linear models (VGLMs) have been developed to handle this inadequacy (Yee and Stephenson 2007). Instead of the conditional mean only, VGLM provides the entire response distribution by employing a linear regression model where the outputs are vectors of parameters of the selected conditional distribution (Kleiber et al. 2012). Moreover, in downscaling applications, VGLM has a particular advantage since it allows reproducing a realistic temporal

variability of the downscaled results by drawing values from the obtained conditional distribution at each forecast step.

The structure of the proposed model allows considering a suitable distribution for each station. Among several options proposed in the literature, Gamma, mixed exponential, GP and WEI are the most commonly used and are therefore considered in the current work to represent the precipitation amount on wet days. However, for the sake of simplicity, only one distribution that provides a good overall fit for all stations is selected. In our study, the examination of the Q-Q plots presented in Figure 2 reveals that all these distributions fit fairly well the precipitation amounts. However, the GP distribution is chosen since it is more successful in reproducing the upper tails. Therefore, a mixed Bernoulli–GP distribution with a vector of parameters $p = (\rho, \alpha, \beta)$ is considered to represent the whole precipitation distribution that includes both occurrences and amounts in a single distribution. The vector of parameters includes the probability of precipitation ρ which is the parameter of the Bernoulli process, and the scale $\alpha\ (\alpha > 0)$ and shape β (where $1 + \beta y / \alpha > 0$ and y represents the precipitation values) are parameters of the zero adjusted GP distribution. Using the VGLM, these parameters are considered to vary for a given day *t* according to the value of large-scale atmospheric predictors $x(t)$. However, only the shape parameter β is fixed to guarantee the convergence of the maximum likelihood estimates. For the parameter of the probability of precipitation occurrences we adopt a logistic regression which is expressed as:

$$\rho(t) = \frac{1}{1 + \exp\left[-a^T x(t)\right]} \tag{1}$$

where a is the coefficient of the logistic model. The scale parameters $\alpha(t)$ are modeled using an exponential link written as:

$$\alpha(t) = \exp\left[b^T x(t)\right] \tag{2}$$

where b is the coefficient of the model. Thus, the conditional Bernoulli-GP density function for the precipitation $y(t)$ on a day t is expressed as:

$$f_t[y(t) \mid x(t)] = \begin{cases} 1-\rho(t) & \text{if} \quad y(t) = 0 \\ \rho(t)\left[1-\left(1+\beta\dfrac{y(t)}{\alpha(t)}\right)^{-1/\beta}\right] & \text{if} \quad y(t) > 0 \end{cases} \tag{3}$$

The coefficients a, b and β are obtained following the method of maximum likelihood by minimizing the negative log predictive density (NLPD) cost function (Haylock et al. 2006; Cawley et al. 2007; Cannon 2008):

$$\mathscr{L} = \sum_{t=1}^{T} \log\left\{ f_t\left[y(t) \mid x(t)\right]\right\} \tag{4}$$

via the simplex search method of Lagarias et al. (1999). This is a direct search method that does not use numerical or analytic gradients.

Now, consider a calibration period of length T and precipitation series at several sites $j = 1, 2, \ldots, m$. The proposed VGLM regression can be trained separately for each precipitation variables y_j at the site j, and thus to obtain the estimated parameters $\hat{p}_j(t)$ and the conditional distributions $\hat{f}_{tj}(y_j \mid x(t))$ for each day $t = 1, 2, \ldots, T$. Figure 3a shows the steps involved for estimating the parameters of the VGLM models.

3.2. Non parametric bootstrapping technique

These dynamic marginal distributions obtained from the VGLM models can be coupled with a random field with uniform marginals. Thus, in simulation, generation of the multi-site replicates of the precipitation field is readily achieved by generating properly associated multivariate variants between 0 and 1 with uniform margins, which are back-transformed to synthetic field predictands by applying the inverse cumulative density function. To address this point, hidden multivariate variants $u(t) = [u_1(t), \ldots, u_d(t)]$ uniformly distributed between 0 and 1 are extracted where $u_j(t)$ for $j = 1, \ldots, m$ are obtained from the following equation:

$$u_j(t) = \hat{F}_{tj}(y_j(t)) \tag{5}$$

where $\hat{F}_{tj}$ is the cumulative density function at time t for site *j* obtained from the VGLM model. Figure 3b shows the steps involved in obtaining the hidden multivariate variants over the calibration period. First, the VGLM can be evaluated during the calibration period separately for each station. This will allow obtaining the entire conditional distribution for each day from the calibration period. Then the obtained conditional CDFs can be applied to their corresponding predictand values to express precipitation as cumulative probabilities ranging from 0 to 1. In order to map $u_j(t)$ onto the full range of the uniform distribution between 0 and 1, the cumulative probabilities $F_{tj}(y_j(t))$ are randomly drawn from a uniform distribution on $[0, 1-\rho(t)$] for dry days. The resulting data matrix $u(t)$ represents values between 0 and 1 that contain the unexplained information by the VGLM model including spatial dependence structures and long term and short term temporal structures. As mentioned in the introduction, the rank matrix **R** of the obtained data matrix can be used in the simulation to preserve this information. The idea

consists in generating multivariate random variables from the uniform distribution with the same dimension as the matrix **R**, and then ordering each column according to the corresponding column in **R**.

Finally the synthetic precipitation series during the validation period can be obtained from the VGLM-NB model using the following three steps.

(i) Randomly generate multivariate random variables from the uniform distribution with same dimension as the matrix $\mathbf{R}$ during the validation period.

(ii) Sort each column of the obtained matrix in step (i) according to the corresponding column in **R**.

(iii) Apply the inverse cumulative Bernoulli-GP distribution expressed in Equation (3) for each site j and for each forecast day *t* from the validation period to the sorted matrix obtained in step (ii).

4. Results

The VGLM-NB model was trained for the calibration period (1960-1980), using precipitation data series from the nine stations and the 40 predictors obtained by the PCA. Once the parameters of the conditional Bernoulli-GA distribution ($\rho_j(t), \alpha_j(t)$ and $\beta_j(t)$) have been estimated for each day t and for each site j over the calibration period, all the obtained conditional marginal distributions were used to obtain the hidden variables $u(t)$ and then to calculate the rank data matrix **R**. Finally, for each of the nine sites, 100 daily precipitations series were generated during the validation period (1981-2000) using VGLM-NB described in Section 3. To assess the performance of the proposed VGLM-NB model, we compare it to VGLM-MAR which is a

downscaling model using the same mixed Bernoulli-Generalized Pareto distribution and extended to multisite tasks using a first order multivariate autoregressive random field framework (Ben Alaya et al. 2015).

The models are evaluated and compared through the following criteria: the mean errors (ME), the root mean squared errors (RMSE), the differences between observed and estimated variances (D) and the false alarm rate (FAR) for binary predictions. RMSE and ME where calculated using the conditional means of 100 realisations, whereas the differences between observed and modeled variances where calculated using the mean variance values of the 100 simulations. Table 3 shows values of the obtained criteria. Generally, the two compared models give similar results in terms of RMSE and ME. On the other hand the VGLM-NB outperformed the VGLM-MAR in reproducing the temporal variability, since it gives better results at 6 stations in terms of *D*. For precipitation occurrences, a categorical measure of performance is considered based on false alarm rates, and results show that VGLM-NB has fewer false alarm rates over all stations. This result shows that, although both VGLM-NB and VGLM-MAR are trained using the same probabilistic regression component (the Bernoulli-generalized Pareto regression model), the non-parametric bootstrapping technique leads to better at-site results than the MAR approach..

In a second validation approach, a set of several precipitation indices that reflect precipitation variability on a seasonal and monthly basis are considered. Five indices related to precipitation amounts are considered: the mean precipitation of wet days (MPWD), the 90^{th} percentile of daily precipitation (Pmax90), the maximum 1-day precipitation (PX1D), the maximum 3-day precipitation (PX3D), and the maximum 5-day precipitation (PX5D). In addition, three other indices are considered for precipitation occurrences: the maximum number of consecutive wet days (WRUN), the maximum number of consecutive dry days (DRUN) and the number of wet

days (NWD). All indices are calculated on a monthly time scale, whereas the P90max is calculated on a seasonal time scale. The RMSE values of these indices (Table 4) show that VGLM-NB performs better than VGLM-MAR for all indices, except for the 90^{th} percentile of daily precipitation.

To evaluate the ability of the models to simulate spatially realistic precipitation fields, Figure 4 compares the distribution of observed and downscaled daily average precipitations over the 9 stations for VGLM-NB, VGLM-MAR and univariate VGLM without multisite extension. The comparison with the univariate VGLM is beneficial to identify the real gain contributed by the two multisite components of VGLM-NB and VGLM-MAR. The observed and modeled CDFs are presented in Figure 4a and the Q-Q plots for quantiles corresponding to non-exceeded probabilities ranging between 0.01 and 0.99 with a step of 0.01 in Figure 4b. Results indicate that the performance of VGLM-NB in reproducing the distribution of daily average precipitation is satisfactory compared to VGLM and VGLM-MAR. Both VGLM and VGLM-MAR underestimate the higher precipitation amounts and overestimates the lower precipitation amounts. Although VGLM-NB slightly overestimates observed quantiles, it tends to fairly well reproduce low and high values.

Figure 5 shows scatterplots between observed and modeled lag-0 and lag-1 cross-correlations for all station pairs considering only wet days during the validation period. Lag-0 cross-correlation is presented in Figure 5.a and lag-1 cross-correlation in Figure 5.b. The correlation values for each model are obtained using the mean of the correlation values calculated from the 100 realisations. For lag-0 cross-correlation, the points correspond to all 36 combinations of pairs of stations, while for lag-1 cross-correlation points correspond to all 81 combinations because lag-1 cross-correlations are generally not symmetric. Figure 5.a shows that observed values of lag-0 cross-

correlation range between -0.02 and 0.65. VGLM-NB gives better preservation of lag-0 cross-correlation than both VGLM-MAR and traditional VGLM. Because VGLM is not a multisite model, it gives the poorest performances and generally underestimates lag-0 cross-correlations. Figure 5b indicates that, for the lag-1 cross-correlation, observed values range between -0.1 and 0.28. For VGLM-NB the performance in reproducing lag-1 cross correlation is less good than the on corresponding to lag-0 cross correlation. However, this performance seems to be always better than the other two models.

To further evaluate the multisite performance, Figure 6a presents observed and modeled log odds ratios for the VGLM-NB, VGLM-MAR and univariate VGLM at all stations. A log-odds ratio between a pair of stations *i* and *j* is expressed as:

$$LOR_{i,j} = \ln\left[\frac{p00_{i,j}\, p11_{i,j}}{p10_{i,j}\, p01_{i,j}}\right], \tag{6}$$

where $p00_{i,j}$, $p11_{i,j}$, $p10_{i,j}$, $p01_{i,j}$ are the joint probabilities of no rain at either one of the two stations, rain at both stations, rain at station *i* and no rain at station *j*, and finally no rain at station *i* and rain at station *j*, respectively. The log odds ratio provides a measure of the spatial correlation between precipitation occurrences at each pair of stations where higher values indicate better defined spatial dependence (Mehrotra et al. 2004; Mehrotra and Sharma 2006). Figure 6a indicate that the VGLM-NB model provides very close correspondence with observed log odds ratios and gives better results than the two other models. VGLM-MAR outperforms the univariate VGLM but its results are less accurate than VGLM-NB, especially when the observed correlations are high.

The dynamics of flood events are strongly related to the simultaneous occurrence of extreme precipitation at several sites. A pairwise correlation is often used for the specification of multisite precipitation models (this is the case of the VGLM-MAR). On the other hand multisite properties of extreme precipitation could be related to higher-order correlations than a traditional pairwise correlation (Serinaldi et al. 2014). In this respect, a diagnostic based on higher order correlation between extreme precipitations is necessary but often ignored. To this end, Bárdossy and Pegram (2009) introduced the binary entropy as a measure of dependence in a given triplet. This measure overcomes a pairwise validation in order to look effectively at the high-order dependence properties. The entropy theory was first formulated by (Shannon 1948) to provide a measure of information contained in a set of data. To calculate the binary entropy, we first fix a given quantile threshold to divide each precipitation series into binary sets by allocating 0 to the lower partition defined by the threshold and 1 otherwise. At each day, the eight possible states of a given binary triple can be defined using the set $\{i, j, k\}$ for $i, j, k = 0,1$. Then, the eight binary probabilities $p(i, j, k)$, for $i, j, k = 0,1$ can be calculated over all days from the validation period. For example, $p(1,1,1)$ represent the probability that all three binary sets on a given day are simultaneously equal to 1, and $p(0,0,0)$ that they are all equal to 0. The binary entropy H can be computed as

$$H = -\sum_{i,j,k=0}^{1} p(i, j, k) \ln(p(i, j, k)). \quad (7)$$

Hence, the lower the entropy is, the stronger will be the association between the variables at a given threshold. Figure 7 shows scatter plots of observed and modeled binary entropy for

precipitation occurrences (Figure 7a) and at three quantile thresholds: 0.75 (Figure 7b), 0.90 (Figure 7c) and 0.975 (Figure 7d). Points correspond to all combinations of stations triplets.

It can be seen from Figure 7a that simulated precipitation occurrences using both VGLM and VGLM-MAR data exhibit higher binary entropy values than observed data. Similar results were found for binary entropy corresponding to the quantile thresholds 0.75, 0.90 and 0.95. This result indicates that the Gaussian dependence structure is not enough to capture the stronger association of extreme precipitation. It is clear that the VGLM-NB is closer to the data across the range of H than the VGLM-MAR model, indicating that non-parametric bootstrapping simulation is an improvement over the multivariate autoregressive Gaussian framework. In reality, this result is expected, since the VGLM-MAR captures the spatial structure by modeling a combination of bivariate relationships using the Gaussian copula. Improving the capture of spatial structure using parametric models requires the application of high-dimensional copulas such us a vine copula.

5. Discussions

Unlike the VGLM-MAR, an attractive characteristic of the proposed VGLM-NB is that pairwise correlations are not used for the model definition. Indeed, the employed non-parametric bootstrapping technique does not model dependency structures but mimics the observed data ranks to preserve the unexplained multisite properties by the VGLM. As it is the case for most resampling methods (Ouarda et al. 1997; Buishand and Brandsma 1999; Buishand and Brandsma 2001; Mehrotra and Sharma 2009; Lee et al. 2012), this approach is data driven, non-parametric and thus avoiding any model misspecification when preserving multisite properties. However, while resampling models suffer from the inability to generate values that are more extreme than those observed, the probabilistic regression component of the proposed hybrid model allows

overcoming this drawback. Indeed, regression methods and resampling techniques can be combined to take advantage of their strengths for downscaling tasks. For this purpose, a widely used approach consists in using resampling or randomisation methods to address the inability of the traditional regression component to preserve the temporal variability and multisite properties (Jeong et al. 2012b; Jeong et al. 2013; Khalili et al. 2013). These hybrid approaches are based on a static noise observed during the calibration of the regression component. Therefore, the part of the variability which is explained by the randomization component does not depend on the predictors, and thus, it is supposed to be constant in a changing climate. For this reason, this traditional hybrid structure may not represent local change in the temporal variability in a climate change simulation. Hence, the hybrid structure employed here to describe the VGLM-NB (as well as the VGLM-MAR), allows the temporal variability to be reproduced in the regression component (using the VGLM component) and thus it may change in the future according to the large scale atmospheric predictors.

6. Conclusions

A VGLM-NB model is proposed in this paper for simultaneously downscaling AOGCM predictors to daily multisite precipitation. The VGLM-NB relies on a probabilistic modeling framework in order to predict the conditional Bernoulli-Generalized Pareto distribution of precipitation at a daily time scale. A non-parametric bootstrapping technique is proposed to preserve a realistic representation of relationships between sites at both time and space. The developed model was then applied to generate daily precipitation series at nine stations located in the southern part of the province of Quebec (Canada). Model evaluations suggest that the VGLM-NB model is capable of generating series with realistic spatial and temporal variability. The developed model can be easily applied to other variables such as temperature and wind speed

making it a valuable tool not only for downscaling purposes but also for environmental and climatic modelling, where often non-normally distributed random variables are involved.

7. References

AghaKouchak, A. (2014). "Entropy–copula in hydrology and climatology." Journal of Hydrometeorology **15**(6): 2176-2189.

AghaKouchak, A., A. Bárdossy and E. Habib (2010). "Conditional simulation of remotely sensed rainfall data using a non-Gaussian v-transformed copula." Advances in Water Resources **33**(6): 624-634.

Ashkar, F. and T. B. Ouarda (1996). "On some methods of fitting the generalized Pareto distribution." Journal of Hydrology **177**(1): 117-141.

Bárdossy, A. (2006). "Copula-based geostatistical models for groundwater quality parameters." Water Resour. Res. **42**(11): W11416.

Bárdossy, A. and J. Li (2008). "Geostatistical interpolation using copulas." Water Resour. Res. **44**(7): W07412.

Bárdossy, A. and G. G. S. Pegram (2009). "Copula based multisite model for daily precipitation simulation." Hydrology and Earth System Sciences **13**(12): 2299-2314.

Bargaoui, Z. K. and A. Bárdossy (2015). "Modeling short duration extreme precipitation patterns using copula and generalized maximum pseudo-likelihood estimation with censoring." Advances in Water Resources **84**: 1-13.

Ben Alaya, M. A., F. Chebana and T. Ouarda (2014). "Probabilistic Gaussian Copula Regression Model for Multisite and Multivariable Downscaling." Journal of Climate **27**(9).

Ben Alaya, M. A., F. Chebana and T. B. Ouarda (2015). "Probabilistic Multisite Statistical Downscaling for Daily Precipitation Using a Bernoulli–Generalized Pareto Multivariate Autoregressive Model." Journal of climate **28**(6): 2349-2364.

Benestad, R. E., I. Hanssen-Bauer and D. Chen (2008). Empirical-statistical downscaling, World Scientific.

Bremnes, J. B. (2004). "Probabilistic forecasts of precipitation in terms of quantiles using NWP model output." Monthly Weather Review **132**(1).

Buishand, T. A. and T. Brandsma (1999). "Dependence of precipitation on temperature at Florence and Livorno (Italy)." Climate Research **12**(1): 53-63.

Buishand, T. A. and T. Brandsma (2001). "Multisite simulation of daily precipitation and temperature in the Rhine basin by nearest-neighbor resampling." Water Resources Research **37**(11): 2761-2776.

Cannon, A. J. (2008). "Probabilistic multisite precipitation downscaling by an expanded Bernoulli-gamma density network." Journal of Hydrometeorology **9**(6): 1284-1300.

Cannon, A. J. (2011). "Quantile regression neural networks: Implementation in R and application to precipitation downscaling." Computers & Geosciences **37**(9): 1277-1284.

Cawley, G. C., G. J. Janacek, M. R. Haylock and S. R. Dorling (2007). "Predictive uncertainty in environmental modelling." Neural Networks **20**(4): 537-549.

Chandler, R. E. and H. S. Wheater (2002). "Analysis of rainfall variability using generalized linear models: a case study from the west of Ireland." Water Resources Research **38**(10): 10-11-10-11.

Chebana, F. and T. B. Ouarda (2011). "Multivariate quantiles in hydrological frequency analysis." Environmetrics **22**(1): 63-78.

Coe, R. and R. Stern (1982). "Fitting models to daily rainfall data." Journal of Applied Meteorology **21**(7): 1024-1031.

Conway, D., R. Wilby and P. Jones (1996). "Precipitation and air flow indices over the British Isles." Climate Research **7**: 169-183.

Czado, C., E. C. Brechmann and L. Gruber (2013). Selection of vine copulas. Copulae in Mathematical and Quantitative Finance, Springer**:** 17-37.

El Adlouni, S., B. Bobée and T. Ouarda (2008). "On the tails of extreme event distributions in hydrology." Journal of Hydrology **355**(1): 16-33.

Eum, H.-I., P. Gachon, R. Laprise and T. Ouarda (2012). "Evaluation of regional climate model simulations versus gridded observed and regional reanalysis products using a combined weighting scheme." Climate Dynamics **38**(7-8): 1433-1457.

Fang, H.-B., K.-T. Fang and S. Kotz (2002). "The meta-elliptical distributions with given marginals." Journal of Multivariate Analysis **82**(1): 1-16.

Fasbender, D. and T. B. M. J. Ouarda (2010). "Spatial Bayesian Model for Statistical Downscaling of AOGCM to Minimum and Maximum Daily Temperatures." Journal of Climate **23**(19): 5222-5242.

Friederichs, P. and A. Hense (2007). "Statistical downscaling of extreme precipitation events using censored quantile regression." Monthly Weather Review **135**(6).

Genest, C. and F. Chebana (2015). "Copula modeling in hydrologic frequency analysis." In Handbook of Applied Hydrology (V.P. Singh, Editor) **McGraw-Hill, New York,** (in press).

Genest, C. and J. F. Plante (2003). "On Blest's measure of rank correlation." Canadian Journal of Statistics **31**(1): 35-52.

Giorgi, F., J. Christensen, M. Hulme, H. Von Storch, P. Whetton, R. Jones, L. Mearns, C. Fu, R. Arritt and B. Bates (2001). "Regional climate information-evaluation and projections." Climate Change 2001: The Scientific Basis. Contribution of Working Group to the Third Assessment Report of the Intergouvernmental Panel on Climate Change [Houghton, JT et al.(eds)]. Cambridge University Press, Cambridge, United Kongdom and New York, US.

Gräler, B. (2014). "Modelling skewed spatial random fields through the spatial vine copula." Spatial Statistics **10**: 87-102.

Guerfi, N., A. A. Assani, M. Mesfioui and C. Kinnard (2015). "Comparison of the temporal variability of winter daily extreme temperatures and precipitations in southern Quebec (Canada) using the Lombard and copula methods." International Journal of Climatology.

Harpham, C. and R. L. Wilby (2005). "Multi-site downscaling of heavy daily precipitation occurrence and amounts." Journal of Hydrology **312**(1): 235-255.

Haylock, M. R., G. C. Cawley, C. Harpham, R. L. Wilby and C. M. Goodess (2006). "Downscaling heavy precipitation over the United Kingdom: A comparison of dynamical and statistical methods and their future scenarios." International Journal of Climatology **26**(10): 1397-1415.

Hessami, M., P. Gachon, T. B. M. J. Ouarda and A. St-Hilaire (2008). "Automated regression-based statistical downscaling tool." Environmental Modelling & Software **23**(6): 813-834.

Hobæk Haff, I., A. Frigessi and D. Maraun (2015). "How well do regional climate models simulate the spatial dependence of precipitation? An application of pair-copula constructions." Journal of Geophysical Research: Atmospheres **120**(7): 2624-2646.

Jeong, D., A. St-Hilaire, T. Ouarda and P. Gachon (2012a). "Comparison of transfer functions in statistical downscaling models for daily temperature and precipitation over Canada." Stochastic Environmental Research and Risk Assessment **26**(5): 633-653.

Jeong, D., A. St-Hilaire, T. Ouarda and P. Gachon (2013). "A multivariate multi-site statistical downscaling model for daily maximum and minimum temperatures." Climate Research **54**(2): 129-148.

Jeong, D. I., A. St-Hilaire, T. B. M. J. Ouarda and P. Gachon (2012b). "Multisite statistical downscaling model for daily precipitation combined by multivariate multiple linear regression and stochastic weather generator." Climatic Change **114**(3-4): 567-591.

Joe, H. (1997). Multivariate models and multivariate dependence concepts, CRC Press.

Khalili, M., V. T. Van Nguyen and P. Gachon (2013). "A statistical approach to multi-site multivariate downscaling of daily extreme temperature series." International Journal of Climatology **33**(1): 15-32.

Kleiber, W., R. W. Katz and B. Rajagopalan (2012). "Daily spatiotemporal precipitation simulation using latent and transformed Gaussian processes." Water Resources Research **48**(1).

Lagarias, J. C., J. A. Reeds, M. H. Wright and P. E. Wright (1999). "Convergence properties of the Nelder-Mead simplex method in low dimensions." SIAM Journal on Optimization **9**(1): 112-147.

Lee, T., R. Modarres and T. Ouarda (2013). "Data-based analysis of bivariate copula tail dependence for drought duration and severity." Hydrological Processes **27**(10): 1454-1463.

Lee, T., T. B. Ouarda and C. Jeong (2012). "Nonparametric multivariate weather generator and an extreme value theory for bandwidth selection." Journal of Hydrology **452**: 161-171.

Li, C., V. P. Singh and A. K. Mishra (2013a). "A bivariate mixed distribution with a heavy-tailed component and its application to single-site daily rainfall simulation." Water Resources Research **49**(2): 767-789.

Li, C., V. P. Singh and A. K. Mishra (2013b). "Monthly river flow simulation with a joint conditional density estimation network." Water Resources Research **49**(6): 3229-3242.

Lindström, G., B. Johansson, M. Persson, M. Gardelin and S. Bergström (1997). "Development and test of the distributed HBV-96 hydrological model." Journal of Hydrology **201**(1-4): 272-288.

Mao, G., S. Vogl, P. Laux, S. Wagner and H. Kunstmann (2015). "Stochastic bias correction of dynamically downscaled precipitation fields for Germany through Copula-based integration of gridded observation data." Hydrology and Earth System Sciences **19**(4): 1787-1806.

Mehrotra, R. and A. Sharma (2006). "A nonparametric stochastic downscaling framework for daily rainfall at multiple locations." Journal of Geophysical Research: Atmospheres (1984–2012) **111**(D15).

Mehrotra, R. and A. Sharma (2009). "Evaluating spatio-temporal representations in daily rainfall sequences from three stochastic multi-site weather generation approaches." Advances in Water Resources **32**(6): 948-962.

Mehrotra, R., A. Sharma and I. Cordery (2004). "Comparison of two approaches for downscaling synoptic atmospheric patterns to multisite precipitation occurrence." Journal of Geophysical Research: Atmospheres (1984–2012) **109**(D14).

Nelsen, R. B. (2013). An introduction to copulas, Springer Science & Business Media.

Oakes, D. (1982). "A model for association in bivariate survival data." Journal of the Royal Statistical Society. Series B (Methodological): 414-422.

Ouarda, T. B. M. J., J. W. Labadie and D. G. Fontaine (1997). "Indexed sequential hydrologic modeling for hydropower capacity estimation." Journal of the American Water Resources Association **33**(6): 1337-1349.

Requena, A. I., I. Flores, L. Mediero and L. Garrote (2015). "Extension of observed flood series by combining a distributed hydro-meteorological model and a copula-based model." Stochastic Environmental Research and Risk Assessment: 1-16.

Salvadori, G. and C. De Michele (2007). "On the use of copulas in hydrology: theory and practice." Journal of Hydrologic Engineering **12**(4): 369-380.

Serinaldi, F. (2009). "A multisite daily rainfall generator driven by bivariate copula-based mixed distributions." Journal of Geophysical Research: Atmospheres (1984–2012) **114**(D10).

Serinaldi, F., A. Bárdossy and C. G. Kilsby (2014). "Upper tail dependence in rainfall extremes: would we know it if we saw it?" Stochastic Environmental Research and Risk Assessment **29**(4): 1211-1233.

Serinaldi, F. and C. G. Kilsby (2014). "Simulating daily rainfall fields over large areas for collective risk estimation." Journal of Hydrology **512**: 285-302.

Shannon, C. (1948). "A mathematical theory of communication." Bell Syst Tech J **27**(3): 379–423.

Smith, M. S. (2014). "Copula modelling of dependence in multivariate time series." International Journal of Forecasting.

Song, S. and V. P. Singh (2010). "Meta-elliptical copulas for drought frequency analysis of periodic hydrologic data." Stochastic Environmental Research and Risk Assessment **24**(3): 425-444.

Srivastav, R. K. and S. P. Simonovic (2014). "Multi-site, multivariate weather generator using maximum entropy bootstrap." Climate Dynamics **44**(11-12): 3431-3448.

Stephenson, D. B., K. Rupa Kumar, F. J. Doblas-Reyes, J. F. Royer, F. Chauvin and S. Pezzulli (1999). "Extreme daily rainfall events and their impact on ensemble forecasts of the Indian monsoon." Monthly Weather Review **127**(9): 1954-1966.

Stern, R. and R. Coe (1984). "A model fitting analysis of daily rainfall data." Journal of the Royal Statistical Society. Series A (General): 1-34.

Vaz de Melo Mendes, B. and R. P. C. Leal (2010). "Portfolio management with semi-parametric bootstrapping." Journal of Risk Management in Financial Institutions **3**(2): 174-183.

Vernieuwe, H., S. Vandenberghe, B. De Baets and N. E. Verhoest (2015). "A continuous rainfall model based on vine copulas." Hydrology and Earth System Sciences Discussions **12**(1): 489-524.

Vinod, H. D. and J. López-de-Lacalle (2009). "Maximum entropy bootstrap for time series: the meboot R package." Journal of Statistical Software **29**(5): 1-19.

Wilks, D. S. and R. L. Wilby (1999). "The weather generation game: a review of stochastic weather models." Progress in Physical Geography **23**(3): 329-357.

Williams, P. M. (1998). "Modelling seasonality and trends in daily rainfall data." Advances in neural information processing systems: 985-991.

Yang, C., R. E. Chandler, V. S. Isham and H. S. Wheater (2005). "Spatial-temporal rainfall simulation using generalized linear models." Water Resources Research **41**(11): 1-13.

Yee, T. W. and A. G. Stephenson (2007). "Vector generalized linear and additive extreme value models." Extremes **10**(1-2): 1-19.

Yee, T. W. and C. Wild (1996). "Vector generalized additive models." Journal of the Royal Statistical Society. Series B (Methodological): 481-493.

Zhang, Q., M. Xiao and V. P. Singh (2015). "Uncertainty evaluation of copula analysis of hydrological droughts in the East River basin, China." Global and Planetary Change **129**: 1-9.

Table 1. List of the 9 stations used in this study.

No.	Site	Name of station	Latitude (°N)	Longitude (°W)
1	7031360	Chelsea	45.52	-75.78
2	7014290	Cedars	45.3	-74.05
3	7025440	Nicolet	46.25	-72.60
4	7022160	Drummondville	45.88	-72.48
5	7012071	Donnacona 2	46.68	-71.73
6	7066685	Roberval A	48.52	-72.27
7	7060400	Bagotville A	48.33	-71
8	7056480	Rimouski	48.45	-68.53
9	7047910	Seven Island A	50.22	-66.27

Table 2. NCEP predictors on the CGCM3 grid.

No	Predictors	No	Predictors
1	Mean pressure at the sea level	14	Divergence at 500 hPa
2	Wind speed at 1000 hPa	15	Wind speed at 850 hPa
3	Component U at 1000 hPa	16	Component U at 850 hPa
4	Component V at 1000 hPa	17	Component V at 850 hPa
5	Vorticity at 1000 hPa	18	Vorticity at 850 hPa
6	Wind direction at 1000 hPa	19	Geopotential at 850 hPa
7	Divergence at 1000 hPa	20	Wind direction at 850 hPa
8	Wind speed at 500 hPa	21	Divergence at 1000 hPa
9	Component U at 500 hPa	22	Specific humidity at 500 hPa
10	Component V at 500 hPa	23	Specific humidity at 850 hPa
11	Vorticity at 500 hPa	24	Specific humidity at 1000 hPa
12	Geopotential at 500 hPa	25	Temperature at 2m
13	Wind direction at 500 hPa		

Table 3. Quality assessment of the estimated series for the validation period (1981–2000) for VGLM-NB and VGLM-MAR. Statistics are ME and RMSE, Differences between observed and modeled variances (D) and false alarm ratio FAR.

Number of station		(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)
RMSE	VGLM-NB	**7.340**	**7.176**	7.294	5.539	**6.060**	**5.493**	**5.494**	5.496	6.471
	VGLM-MAR	7.377	7.225	**6.916**	**5.187**	6.285	5.603	5.608	**5.368**	**6.295**
ME	VGLM-NB	**0.024**	-0.295	-0.318	-0.464	**-1.030**	-0.305	-1.056	**-0.248**	**0.133**
	VGLM-MAR	0.434	**-0.277**	**-0.302**	**-0.410**	-1.043	**-0.304**	**-0.902**	-0.280	0.483
D	VGLM-NB	-19.55	**6.48**	**-1.05**	**4.42**	**17.08**	9.09	**19.25**	**2.79**	-11.60
	VGLM-MAR	**-17.41**	10.70	9.24	9.40	21.74	**8.95**	19.38	7.59	**-6.24**
FAR	VGLM-NB	**0.358**	**0.356**	**0.318**	**0.314**	**0.332**	**0.373**	**0.339**	**0.366**	**0.378**
	VGLM-MAR	0.398	0.377	0.346	0.331	0.350	0.412	0.373	0.410	0.410

Bold character means better result.

Table 4. RMSE of precipitation indices for the validation period (1981–2000) for both VGLM-NB and VGLM-MAR.

	Indices	VGLM-NB	VGLM-MAR
Precipitation amount	PX1D (mm)	**23.334**	33.424
	PX3D (mm)	**20.312**	35.999
	PX5D (mm)	**20.599**	38.53
	Pmax90 (mm)	3.6920	**3.4567**
	MWD (mm)	**1.4413**	2.0179
Precipitation occurences	WRUN (days)	**1.9579**	2.1016
	DRUN (days)	**3.2365**	4.6395
	NWD (days)	**4.0239**	4.7337

Bold character means better result.

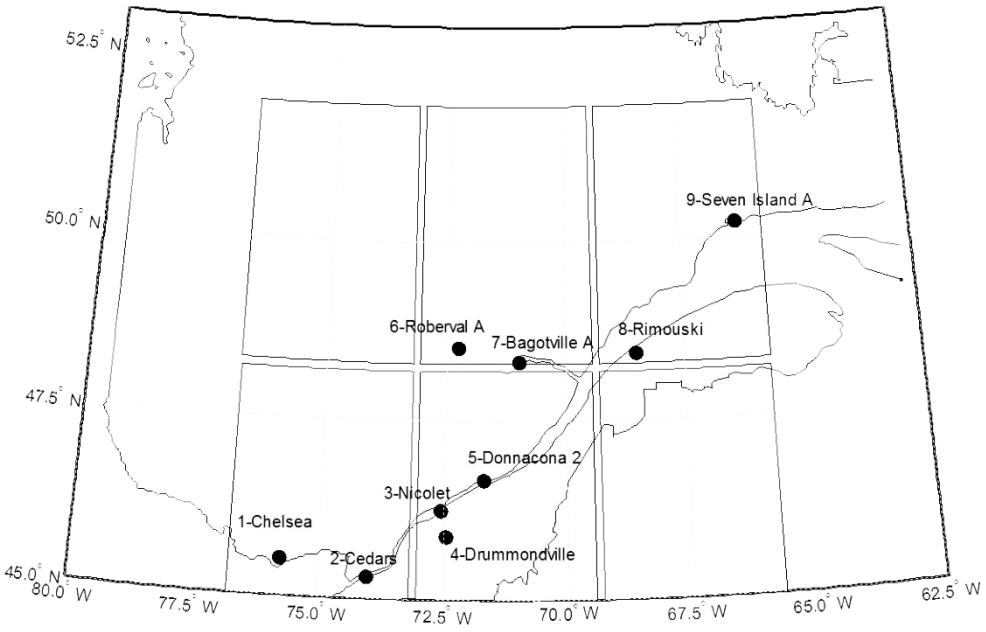

Figure 1. The locations of precipitation stations and CGCM3 grid.

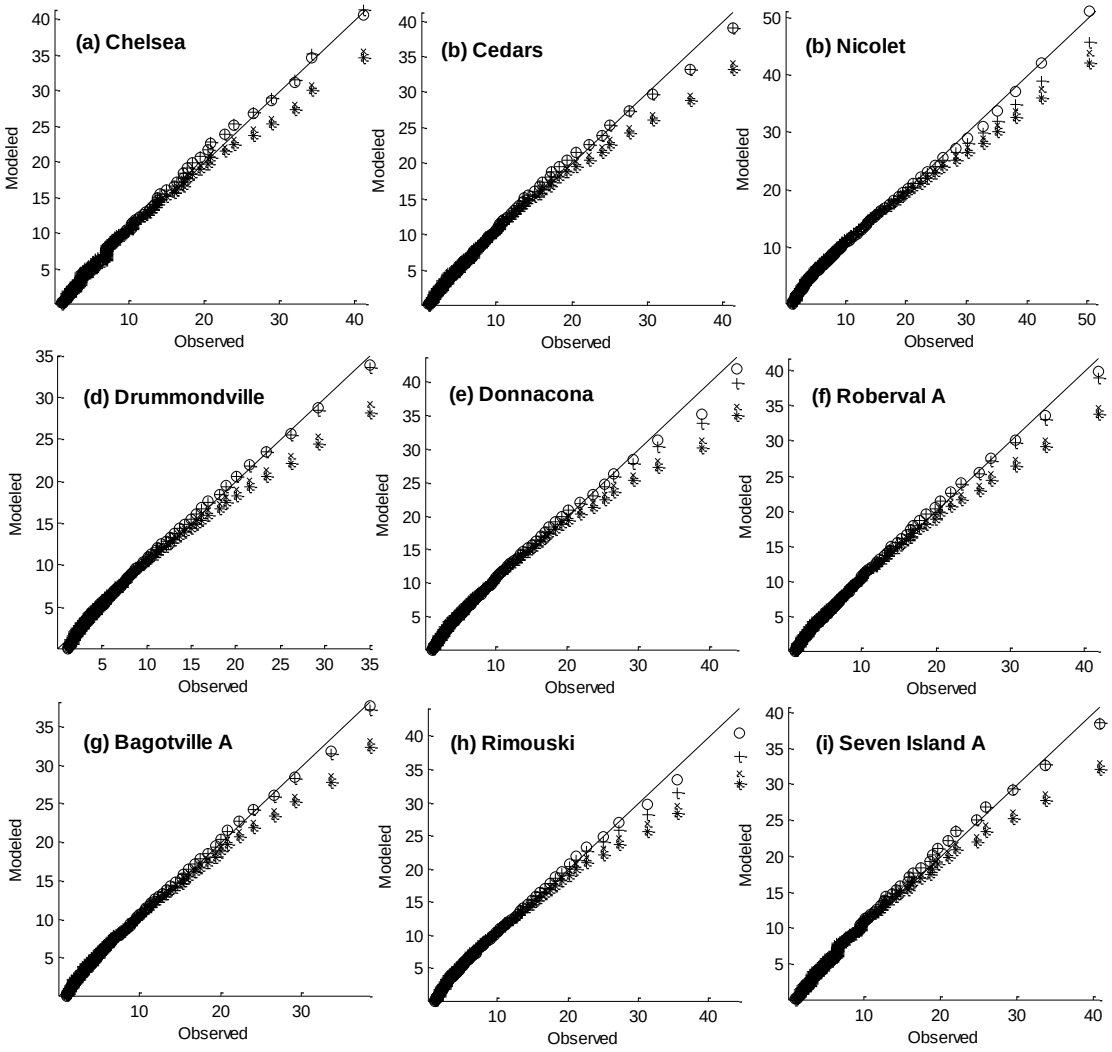

Figure 2. Q–Q plot of observed and modeled quantiles for Gamma distribution (stars), WEI distribution (x-mark), GP distribution (circles) and mixed Exponential distribution (plus)**.**

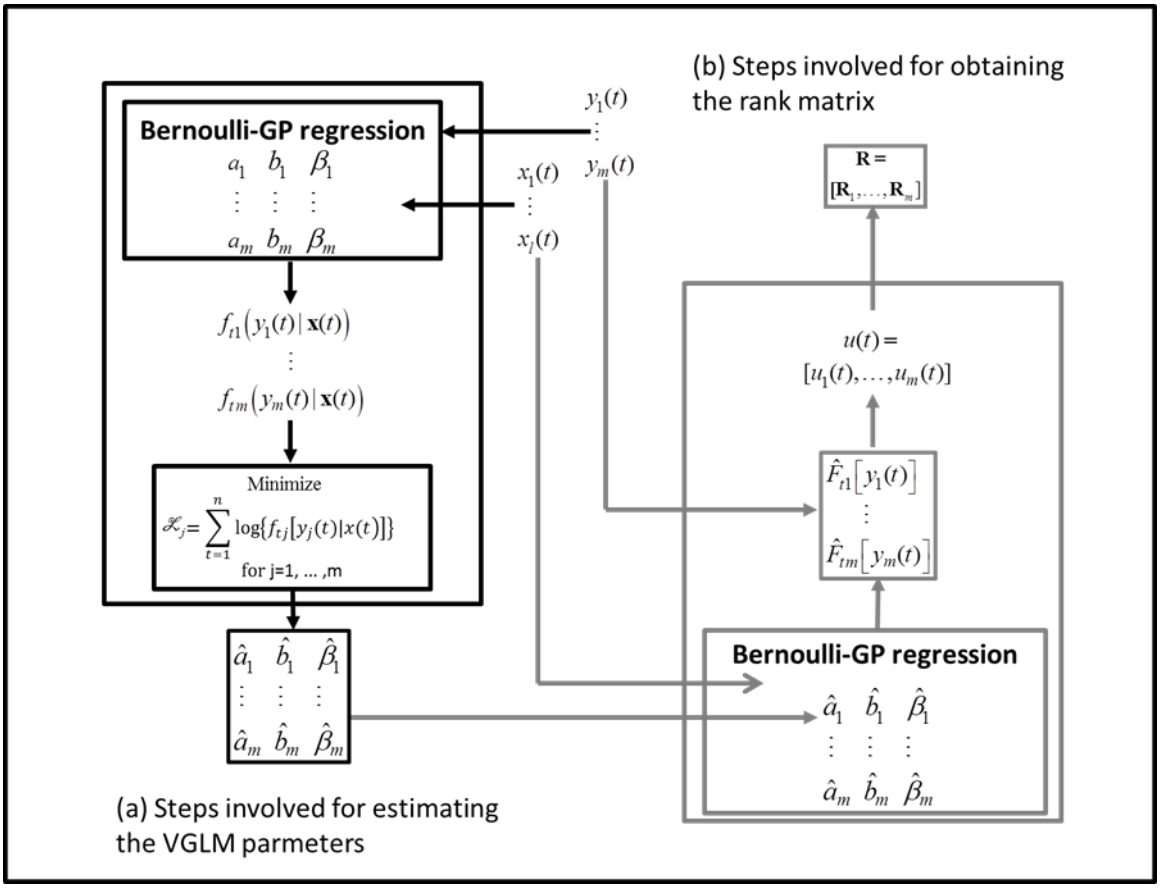

Figure 3. Steps involved for estimating the VGLM prameters (a) and obtaining the rank matrix (b).

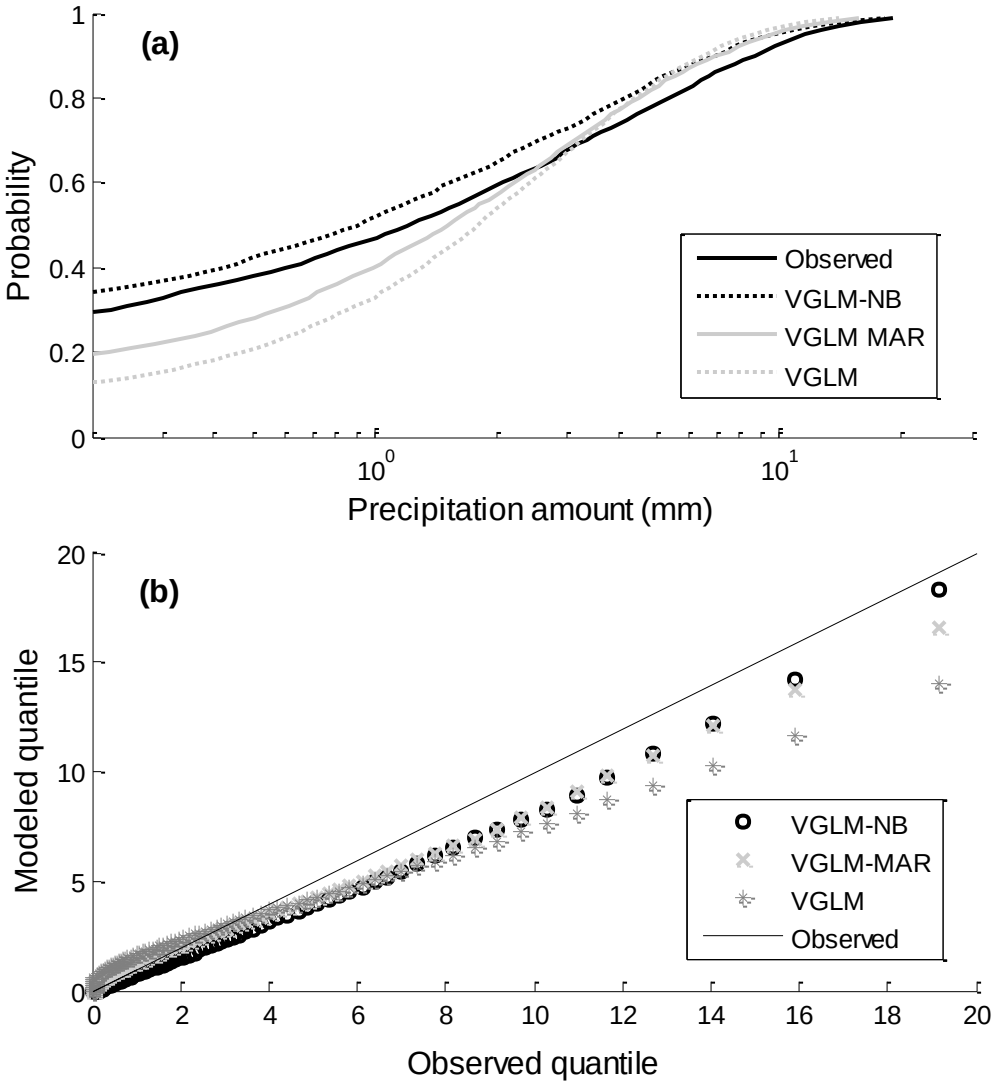

Figure 4. Observed and predicted daily average precipitation over the nine stations. The CDF is presented in (a) and the Q-Q plots in (b).

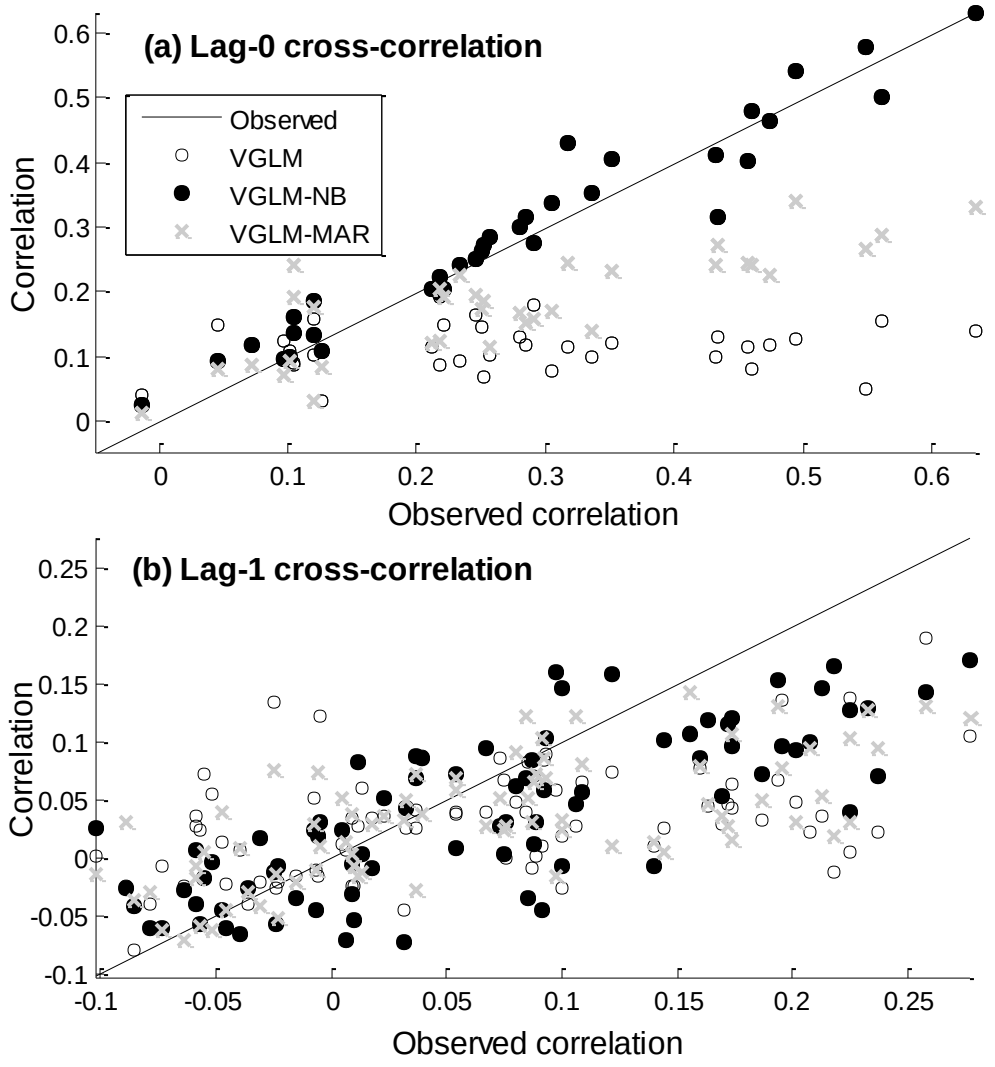

Figure 5. Scatter plots of observed and modeled lag-0 cross-correlation (a) and lag-1 cross-correlation during the validation period. Correlation values are obtained using the mean of the correlation values calculated from 100 simulations.

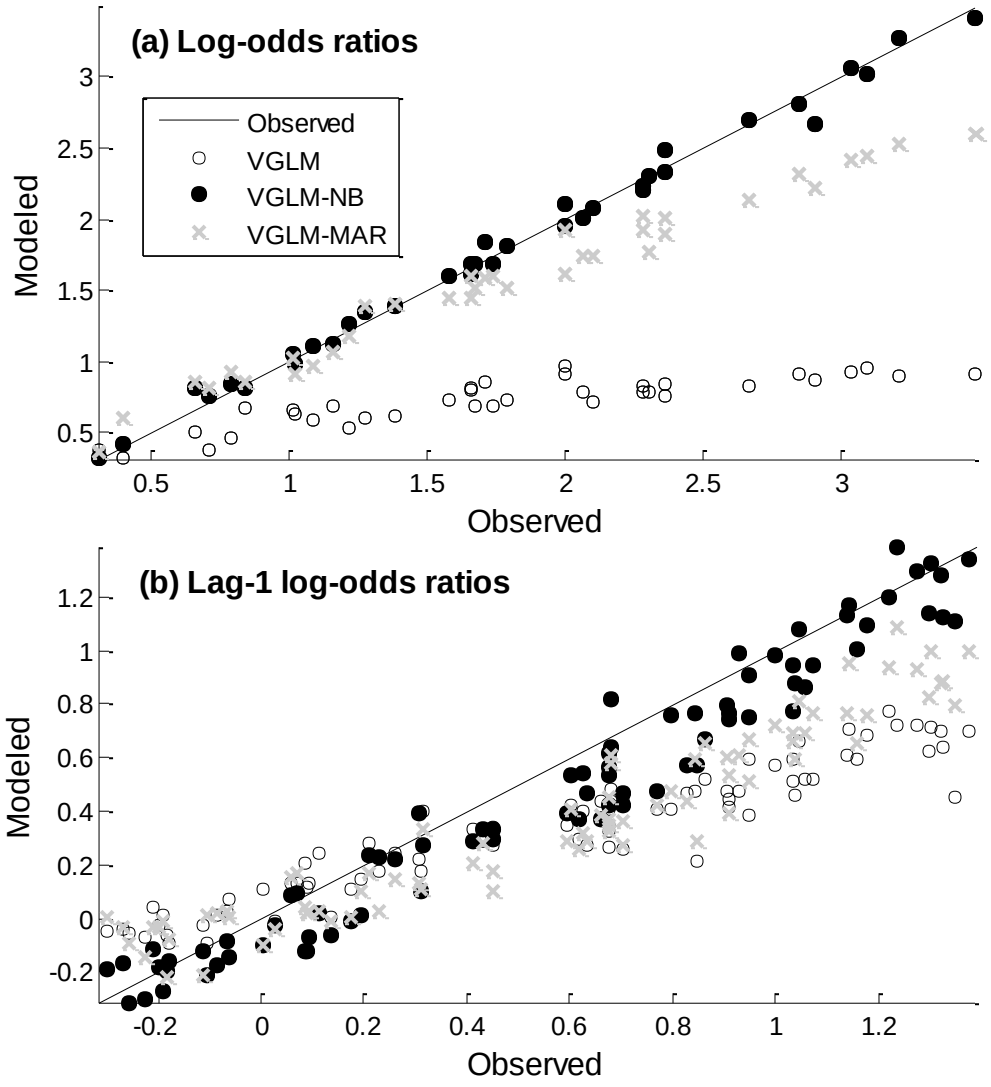

Figure 6. Scatter plots of observed and modeled log odds ratios (a) and lag-1 log odds ratios during the validation period. Values are obtained using the mean values from 100 simulations.

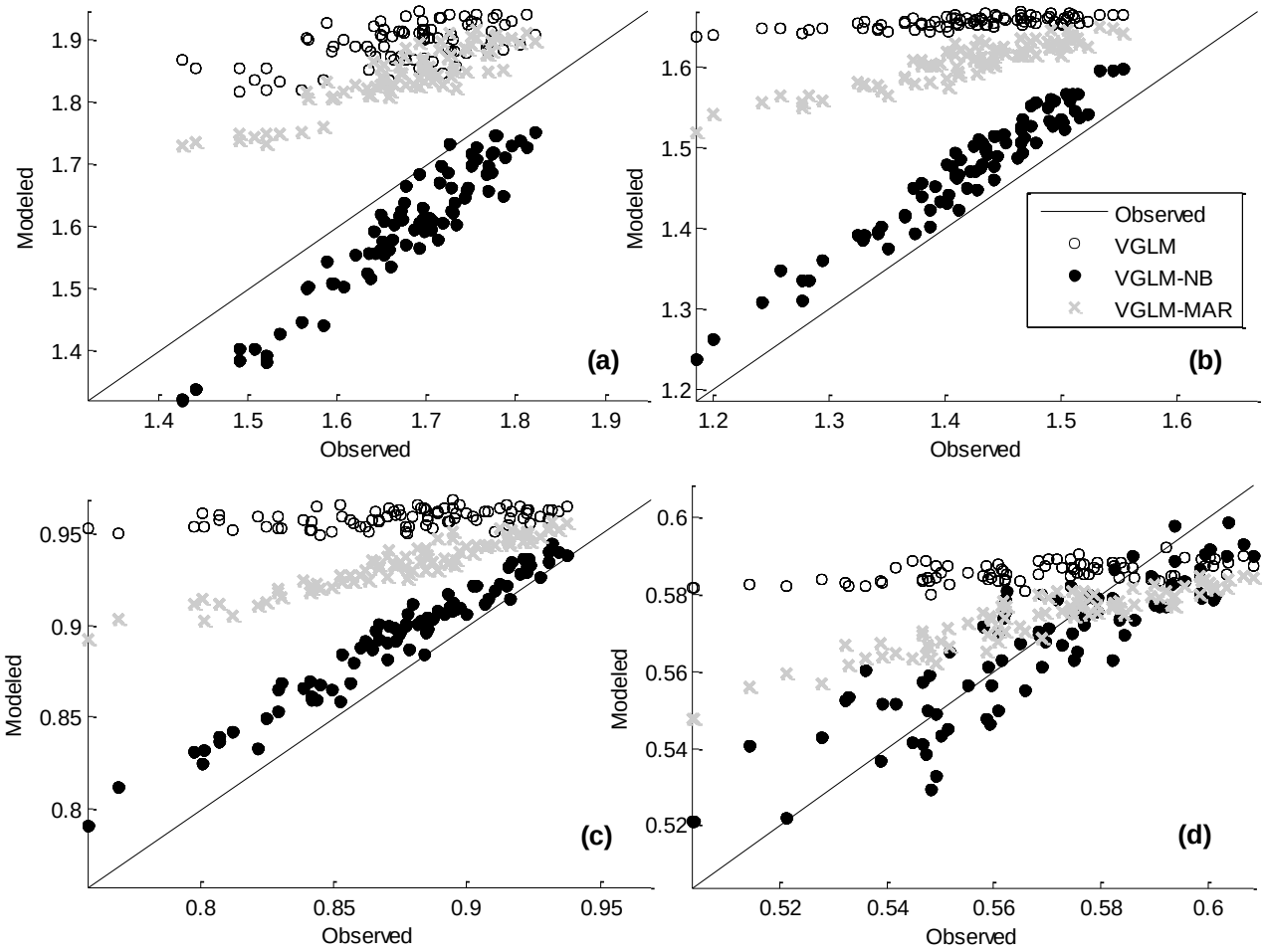

Figure 7. Scatter plots of observed and modeled binary entropy for precipitation occurrences (a), and at three quantile thresholds: 0.75 (b), 0.90 (c) and 0.95 (d). Points correspond to all combinations of triplets of stations.

CHAPITRE 6: QUANTILE REGRESSION MULTIVARIATE AUTOREGRESSIVE MODEL FOR DOWNSCALING MULTISITE DAILY PRECIPITATIONS

Quantile regression multivariate autoregressive model for downscaling multisite daily precipitations

B. A. Mohamed Ali[1], F. Chebana[1] and T.B.M.J. Ouarda[2, 1]

[1]*INRS-ETE, 490 rue de la Couronne, Québec (QC),*

Canada G1K 9A9

[2] Institute Center for Water and Environment (iWATER), *Masdar Institute of science and technology, P.O. Box 54224, Abu Dhabi, UAE*

[*]**Corresponding author:** Tel: +1 (418) 654 2530#4468

Email: moahammed_ali.ben_alaya@ete.inrs.ca

December 2015

(To be submitted)

Abstract

A quantile regression multivariate autoregressive (QRMAR) model is proposed in this paper for multisite statistical downscaling of daily precipitations. The proposed model can be considered as a probabilistic regression-based downscaling model with a stochastic generator component. In a probabilistic framework, QRMAR employs a quantile regression model to reproduce the conditional distribution of precipitation amount and a logistic regression model to reproduce the at site precipitation occurrences. As a stochastic generators component, the QRMAR employs a latent multivariate autoregressive Gaussian field to preserve spatiotemporal properties of precipitation at multiple sites. The proposed model is applied for the downscaling of AOGCM data to daily precipitation in the southern part of Quebec, Canada. Results of the study indicate the superiority of the proposed model over a multivariate multiple linear regression (MMLR) model and a multisite hybrid statistical downscaling procedure that combines MMLR and a stochastic generator schemes.

Keywords: Statistical downscaling, Quantile regression, Multisite daily precipitation, Multivariate autoregressive Gaussian field.

1. Introduction

Stochastic weather generators (WG) are statistical models aimed to produce realistic random sequences of atmospheric variables such as precipitation, temperature and wind speeds. Development of WG models play an important role in hydrological applications and water resources management under future conditions (Wilks 1999). In many hydroclimatic studies, precipitation is the most dominant weather variable that is explicitly affecting the water resources systems. This weather variable is collected at various sites, and precipitation models are required to adequately reproduce the observed temporal variability and to maintain consistency of spatiotemporal properties of precipitation at several sites.

The implementation of the weather generator can be accomplished in one of two different ways: (i) single site precipitation generator, where the precipitation is generated to reproduce local properties independently at each site without the influence of other sites (ii) multisite generator where the precipitation is generated to include the influence of other sites taking into account the spatial correlation. Irrespective of the type of implementation the way of the implementation, precipitation models can be classified as (i) parametric (Kleiber et al. 2012), (ii) non-parametric (Zorita and Von Storch 1999; Mehrotra and Sharma 2006) and (ii) hybrid or semi-parametric (Semenov et al. 2002).

Parametric methods are indeed very useful but they have several inadequacies. First and most importantly, they do not adequately reproduce various aspects of the spatial and temporal dependence. Second, they are not easily transportable to other sites due to the site-specific assumptions made regarding the probability distributions of the variables. The non-parametric models on the other hand do not make such assumptions, but rather shuffle the data itself, to

reproduce the characteristics of the observed data. However, a limitation of these non-parametric models is that they do not produce new values but merely reshuffle the historical data to generate new weather sequences. In spite of considerable progress, the precipitation models proposed in the literature are found to be far from being universally accepted among the researchers and practitioners. Comparing the various models presented above is a difficult, if not impossible, task since they have been validated on different datasets, and for different scientific purpose: downscaling, prediction or simulation. This study focuses on the development of multi-site multivariate weather generator for downscaling purpose.

Downscaling techniques have been developed to refine Atmosphere-Ocean Global Climate Models (AOGCMs) data and to provide information at more relevant scales. These techniques include dynamic downscaling, which uses regional climate models (RCM) over a limited area, and statistical downscaling which considers statistical relationships between large-scale variables (predictors) and small-scale variables (predictands) and provide climate information at the equivalent of point climate observations (Wilby et al. 2002). Statistical downscaling techniques represent a good alternative to dynamic methods in the case of limited resources, because of their ease of implementation and their low computational requirements (Benestad et al. 2008). Wilks (2012) provided a detail review and merits of parametric models for single, as well as multisite statistical downscaling of climate variables. Maraun et al. (2010) provided an overview of downscaling precipitation techniques. Unlike nonparametric approaches, parametric methods can be easily adjusted for downscaling purposes. Wilks (2010) mentioned that these adjustments can be accomplished in two ways: (i) through imposed changes in the corresponding monthly statistics, (ii) or by controlling the generator parameters by daily variations in simulated atmospheric circulation. In this context, Williams (1998) successfully described seasonal

variations in precipitation using a conditional density network model were an artificial network (ANN) is employed to model parameters of a mixed Bernoulli-gamma distribution. This same modular structure is used by Haylock et al. (2006) to downscale precipitation at single sites in the United Kingdom. Precipitation occurrence and wet-day precipitation amounts can be specified by the same model using a mixed distribution that includes both dry and wet days such as the Bernoulli-gamma, Poisson-gamma, or Bernoulli-Generalized Pareto distribution (Ben Alaya et al. 2015a).

An alternative approach to assuming a parametric distribution of precipitation, involves estimating point values of individual quantiles of the conditional distributions directly, using a quantile regression model. Quantile regression is introduced by Koenker and Bassett (1978) to provide a more complete picture of the relationships between variables. In addition quantile regression overcomes some of the limitations of the standard regression models, such as the assumption of homogenous residual variance, and do not make any assumption about the error distribution. Quantile regression models have been successfully applied in the environmental and meteorological sciences to global temperature change (Koenker and Schorfheide 1994), to modelling the effects of meteorological variables on ozone concentration (Baur et al. 2004), to wind power forecasting (Bremnes 2004b) and ecological modelling (Cade and Noon 2003). In addition, the application of quantile regression has made significant contributions in precipitation downscaling (Bremnes 2004a; Friederichs and Hense 2007; Cannon 2011; Tareghian and Rasmussen 2013).

To extend quantile regression approach to multisite downscaling tasks, Ben Alaya et al. (2015b) proposed a Gaussian Copula Quantile Regression (GCQR) model. However, this model does not take into account cross-correlations lagged in time and thus it could not reproduce the short term

autocorrelation properties of precipitation series. To this end, the aim of this paper is to propose a new probabilistic regression-based downscaling model for daily precipitations that extend the quantile regression model to multisite downscaling tasks. Precisely, the proposed approach is a quantile regression multivariate autoregressive (QRMAR) model. The QRMAR provides the at site conditional distribution of precipitation through AOGCM predictors using a quantile regression model. Then, QRMAR employed a latent multivariate autoregressive Gaussian field (Rasmussen 2013; Serinaldi and Kilsby 2014; Villarini et al. 2014; Ben Alaya et al. 2015a) to extend the quantile regression model to a multisite task. This last component allows the QRMAR model to reproduce the observed spatial relationships between sites (such as the observed lag-0 and lag-1 cross-correlations), and to randomly generate realistic synthetic precipitation series.

The present paper is structured as follows: after a brief presentation of the multisite hybrid statistical downscaling model of Jeong et al. (2012) as a classical model to compare, the proposed QRMAR model is presented. The QRMAR model is then applied to the case of daily precipitations in the southern part of the province of Quebec, Canada. Results are compared with those obtained using a multisite hybrid model of Jeong et al. (2012) and the (MMLR) model. Finally a discussion and conclusion are provided.

2. Data and study area

The study area is located in Quebec, in the latitudes between 45 ° N and 60 ° N and the longitudes between 60 ° W and 80 ° W. Seven series of observed daily precipitations (see Figure 1) are considered as predictands. These series, are provided by Environment Canada's hydro-meteorological network, have been rehabilitated by Mekis and Hogg (1999) and cover the period from 1 January 1961 to 31 December 2000. Table 1 reports their names and their latitude-

longitude locations. The reanalysis data from the National Center for Environmental Prediction (NCEP)/ National Center for Atmospheric Research (NCAR) over the period 1961-2000 (Kalnay et al. 1996; Kistler et al. 2001) are used to evaluate the potential of the downscaling method. NCEP / NCAR data are averaged on a daily basis from 6-hour data on the original regular grid of 2.5 ° latitude and longitude. Obtained predictors are then linearly interpolated on the CGCM3 Gaussian grid (3.75 ° latitude and longitude) and normalized to the reference period 1961-1990. The study area is covered by six grid points (see Figure 1), and for each grid point, 25 NCEP predictors are provided (see Table 2). For each day, 150 predictors are thus available. In order to reduce the number of predictors, a principal component analysis (PCA) is employed and the first components that preserve more than 97% of the variance of the original NCEP predictors are then preserved as predictor variables. Finally, data from 1961 to 1990 are used for the calibration, whereas data from 1991 to 2000 are used for the validation.

3. Methodology

The multisite hybrid downscaling model of Jeong et al. (2012) and the QRMAR model are presented in section 3.1 and section 3.2 respectively.

3.1. Multisite hybrid statistical downscaling of Jeong et al. (2012)

Let $\mathbf{Y}$ denote a multivariate predictand variables matrix of dimension $n \times m$ and $\mathbf{X}$ a multiple atmospheric predictor variables matrix of dimension $n \times l$. The MMLR model is expressed as:

$$\mathbf{Y} = \mathbf{X} \times \mathbf{W} + \mathbf{E} \tag{1}$$

Where $\mathbf{E}$ is the residual matrix of dimension $n \times m$ and $\mathbf{W}$ is a parameter matrix which can be estimated using the Ordinary Least Squares (OLS) method given by:

$$\hat{\mathbf{W}} = (\mathbf{X}^{\mathbf{T}}\mathbf{X})^{-1}\mathbf{X}^{\mathbf{T}}\mathbf{Y} \quad (2)$$

3.1.1. Precipitation occurrences

A standard problem for precipitation data is that the responses are the product of two processes, an occurrence process which decides whether or not there is any precipitation on a particular day, and an amount process. Let $\mathbf{O}[n \times m]$ be the observed binary matrix of precipitation occurrence. An element O_{ij} of the matrix $\mathbf{O}$,. For a given day $t = 1, 2, \ldots, n$, and a given site $j = 1, 2, \ldots, m$, is equal to 1 for a wet day and 0 for a dry day. The matrix of the downscaled deterministic series of daily precipitation occurrence probabilities $\hat{\mathbf{O}}$ is modeled using the MMLR equation. Then, to reproduce the observed temporal variability and spatial dependency, the residual matrix $\mathbf{E_O}[n \times m]$ of this MMLR model is modeled using a multivariate normal distribution. Therefore, residuals $\tilde{\mathbf{E}}_{\mathbf{O}}[n \times m]$ are generated and added to the downscaled probability matrix to obtain the generated continuous probability matrix $\tilde{\mathbf{O}}$. Then, Jeong et al. (2012) employed a first-order Markov chain model to transform the matrix $\tilde{\mathbf{O}}$ to a downscaled binary series $\dot{\mathbf{O}}$. Finally, Jeong et al. (2012) mentioned that continued transformed series $\dot{\mathbf{O}}$ cannot represent the original binary multisite cross-correlation. For this reason, they employed empirical relationships of cross-correlations between binary series and continuous series using a simple power function.

3.1.2 Precipitation amount

Before developing a classical regression-based precipitation amount model, Jeong et al. (2012) transformed the precipitation amount vector $\mathbf{Y_j}$ for a site j to a normal distribution by employing the Anscombe transformation $R_{tj} = Y_{tj}^{1/3}$ (Terrell 2003; Yang et al. 2005). Then, the downscaled

deterministic series of Ascombre residual matrix $\hat{\mathbf{R}}[n \times m]$ is modeled using the MMLR model. To reproduce at-site variances and multisite cross-correlations in the generated Ascombre residual matrix $\tilde{\mathbf{R}}$, the residual matrix $\tilde{\mathbf{E}}_{\mathbf{R}}[n \times m]$ is generated from multivariate normal distribution and added to the matrix $\hat{\mathbf{R}}$. Therefore generated precipitation amounts are calculated as $\tilde{Y}_{tj} = \tilde{R}_{tj}^3$. Thereby, the generated precipitation series in $\dot{\mathbf{Y}}$ are obtained by calculating the product of the generated precipitation occurrence and the generated precipitation amount $[\dot{Y}_{ij} = \dot{O}_{ij} \times \tilde{Y}_{ij}]$. Finally, because the generated series in $\dot{\mathbf{Y}}$ present in general different statistical properties than those of the observed precipitation amount series, and because the residual matrix $\mathbf{E}_{\mathbf{R}}$ of each site may be not normally distributed, Jeong et al. (2012) adopted a probability distribution mapping technique to adjust generated precipitation amount using the gamma distribution. For more detail about this hybrid downscaling precipitation occurrences and amounts models see (Jeong et al. 2012).

3.2. QRMAR model

Our objective in this paper is to derive the most suitable probability density function (PDF) conditioned on the set of large scale atmospheric predictors. Then, downscaled time series with a realistic temporal variability can be generated by sampling from the obtained conditional distributions at each forecast day. The probabilistic framework for the QRMAR model is presented with a description of how the conditional discrete-continued distribution of precipitation is obtained from the quantile regression model and the logistic regression model. Then, a simulation procedure is presented using a latent multivariate autoregressive Gaussian field to reproduce the dependence structure of precipitations at multiple sites.

3.2.1. Quantile regression

As minimizing a sum of squared errors leads to an estimate of the conditional mean, minimizing the mean absolute errors leads to an estimate of the conditional median. Quantile regression is a generalization of median regression models, by giving an estimate to the conditional quantile of the predictive distribution instead of the conditional median. Instead of minimising the mean absolute errors, Koenker and Bassett (1978) applied asymmetric weights to positive/negative errors using a pinball loss function to compute conditional quantiles of the predictive distribution. The pinball loss function is given by:

$$\rho_p(u) = \begin{cases} u(p-1) & \text{if } u < 0 \\ up & \text{if } u \geq 0 \end{cases} \tag{3}$$

Where $1 < p < 0$. Given a set of potential daily predictor variables $\mathbf{x} = (x_1, x_2, \ldots, x_m)$, a given predictand y, and $\mathbf{b}_p$ a vector of parameters, the linear regression equation for the p^{th} quantile $Q_p(y \mid \mathbf{x})$ of the conditional distribution of y given $\mathbf{x}$ is expressed as:

$$Q_p(y \mid \mathbf{x}) = \mathbf{x}^T \mathbf{b}_p \tag{4}$$

For a given set of observations $(\mathbf{x}_i, y_i)$, $i = 1, \ldots, n$, the vector of parameters $\mathbf{b}_p$ can be estimated by minimizing the quantile regression error function given by:

$$e_p = \sum_{i=1}^{n} \rho_p\left(y_i - Q_p(y \mid \mathbf{x})\right) \tag{5}$$

Quantile regression is semiparametric as it avoids assumptions about the parametric distribution of the error process, does not assume homogenous residual variance and does not make any assumption about the error distribution. Quantile regression also provides a richer

characterization of the data, allowing us to consider the impact of a covariate on the entire conditional distribution of y, not merely its conditional mean or conditional median.

3.2.2. Precipitation amount model using quantile regression

To obtain the whole conditional distribution of precipitation on a given day, quantile regression model can be applied to produce quantiles from 0.01 to 0.99 by steps of 0.01 for each station and for each day. The obtained sample can be considered as sample from the target conditional distribution. Then, a parametric distribution can be assumed to represent this target distribution whose parameters can be estimated using the maximum likelihood method. Concerning the choice of the parametric conditional distribution, the normality assumption might not be feasible on shorter time scales for precipitation amounts. On daily time scale, precipitation amount becomes more skewed, and is commonly modeled with a gamma (GA) distribution. However, the GA distribution may be not flexible enough to capture all precipitation amount behaviors and can be heavy tailed at some sites. It should be noted that several other options have been suggested in the literature, such as the Exponential (EXP) distribution, the Generalized Pareto (GP) distribution with threshold parameter fixed to zero and the Weibull (WEI) distribution. Depending on the value of its shape parameter, WEI distribution can have either an apparent heavy or bounded tail (Furrer and Katz 2008). However, according to the extreme value theory GP distribution is the asymptotic distribution for over threshold exceedances (Coles et al. 2001).

Hereafter, for the choice of the most appropriate distribution among the four options GA, EXP, GP and WEI, a preliminary analysis is made for each 10957 day from the calibration period and for each site. In this preliminary analysis we have applied the quantile regression model to produce quantiles from 0.01 to 0.99 by steps of 0.01 for each station and for each day from the

calibration period. The obtained sample can be considered as sample from the conditional distribution for a given day. Then, the maximum likelihood is used to estimate parameters of the four considered conditional distributions. To select the suitable distribution, the Kolmogorov-Smirnov test is applied for each day and for each station. Kolmogorov-Smirnov (K-S) test is usually performed to determine if a random sample could have the hypothesized continuous cumulative distribution, for a desired significance level, against the alternative that the empirical cumulative distribution of the sample is unequal to the hypothesized cumulative distribution. The decision to reject the null hypothesis (sample from the hypothesized distribution) occurs when the significance level equals or exceeds the P-value of the Kolmogorov-Smirnov statistic. Figure 2 shows an example of the obtained quantile regression results at Nicolet station for the days 04-02-1961, 04-05-1961, 04-08-1961 and 04-11-1961. For the day 04-02-1961 and at 5% significance level the two distribution GAM and WEI are rejected, and for the day 04-05-1961 both GPD and EXP distributions are rejected, for the day 04-08-1961 the four distributions are not rejected, and finally for the day 04-11-1961 all distributions are rejected. The structure of the model allows selecting a suitable distribution for each site. However, aiming at making the model as simple as possible, we look for a common distribution yielding a satisfying fitting for all series. Results for each day from the calibration period are summarized in Table 2 as number of failures to reject the null hypothesis (sample from a given distribution) for each station and distribution at 5% significance level. Overall, GA and WEI yield the highest number of failures to reject, but the GA distribution is slightly better. Therefore the GA distribution is chosen.

3.2.3. The mixed discrete-continues distribution model for precipitation

For precipitation occurrences, a standard problem is to consider a Bernoulli process that describes the dry-wet dichotomy. In this case, a logistic regression model could be employed to obtain the parameter of the conditional Bernoulli distribution $\pi(x)$ given by:

$$E(y \mid \mathbf{x}(t)) = \pi(t) = \frac{1}{1+\exp\left[\mathbf{x}(t)^T c\right]} \tag{6}$$

where c is the vector of parameters of the logistic model which are set following the method of maximum likelihood for each precipitation station. Hence the conditional distribution of precipitation occurrences is expressed by:

$$f\left[y \mid \mathbf{x}\right] = \begin{cases} \pi(t) & if \quad y = 1 \\ 1-\pi(t) & if \quad y = 0 \end{cases} \tag{7}$$

Finally, the whole conditional distribution of precipitation is defined as a single mixed Bernoulli-GA distribution that describes both precipitation occurrences and precipitation amounts. The conditional Bernoulli-Gamma PDF $f_{tj}[y_j(t) \mid x(t)]$ of precipitation $y_j(t)$ for a site $j = 1, 2, \ldots, m$ and a day t is expressed as:

$$f_{tj}[y_j(t) \mid x(t)] = \begin{cases} 1-\pi_j(t) & \text{if} \quad y_j(t) = 0 \\ \dfrac{\pi_j(t)\ (y_j(t)/\beta_j(t))^{\alpha_j(t)-1} \exp(-y_j(t)/\beta_j(t))}{\beta_j(t)\,\Gamma(\alpha_j(t))} & \text{if} \quad y_j(t) > 0 \end{cases} \tag{8}$$

Where $\Gamma(\bullet)$ is the gamma function, $\alpha_j(t)\ (\alpha_j(t) \geq 0)$ and $\beta_j(t)\ (\beta_j(t) \geq 0)$ are respectively the shape and the scale parameters of the conditional gamma distribution of the precipitation amounts on wet day obtained from quantile regression models following the methodology described in

3.2.2, and $\pi_j(t)\,(0 \le \pi_j(t) \le 1)$ is the probability of precipitation occurrences obtained from the logistic model described in 3.2.3.

3.2.4. Multivariate autoregressive Gaussian fields

Maintaining realistic spatial dependency and short term autocorrelation properties between multisite precipitation series is particularly important in hydrological applications. To this end, a multivariate first-order autoregressive model (MAR(1)) for a multivariate latent Gaussian process $u(t) = [u_1(t),\ldots,u_m(t)]$, is considered using the following equations:

$$u_j(t) = \mathbf{\Phi}^{-1}[F_{tj}(y_j(t))] \tag{9}$$

Where $\mathbf{\Phi}$ is the standard normal cumulative distribution function and F_{tj} is the Bernoulli-GA cumulative density function at time t and site j obtained from quantile regression and logistic regression models. The cumulative probabilities $F_{tj}(y_j(t))$ is employed to express precipitation series as cumulative probabilities ranging from 0 to 1. Then $u_j(t)$ are obtained by applying the standard normal inverse cumulative density function to the series of cumulative probabilities $F_{tj}(y_j(t))$. Let $\mathbf{U_t} = (u_{1t}, u_{2t},\ldots,u_{mt})^T$ denote the obtained multivariate latent Gaussian vector of z values at the m sites at time $t = 1,2,\ldots n$ after the normalisation step. To take into account the spatial dependences and the short term autocorrelation of precipitation series, $\mathbf{U_t}$ is modeled using a MAR(1) model expressed as:

$$\mathbf{U_t} = \mathbf{A}\mathbf{U_{t-1}} + \mathbf{B}\boldsymbol{\varepsilon_t} \tag{10}$$

where $\mathbf{A}$ and $\mathbf{B}$ are $(m \times m)$ parameter matrices, and $\boldsymbol{\varepsilon}_t$ is a random $(m \times 1)$ noise vector with a standard multivariate normal distribution. Parameters $\mathbf{A}$ and $\mathbf{B}$ are estimated using the moment estimators proposed by Bras and Rodríguez-Iturbe (1985):

$$\hat{\mathbf{A}} = L_1 L_0^{-1} \tag{11}$$

$$\hat{\mathbf{B}}\hat{\mathbf{B}}^{\mathbf{T}} = L_0 - L_1 L_0^{-1} L_1^{T} \tag{12}$$

where L_0 is the sample lag-0 cross covariance matrix and L_1 is the sample lag-1 covariance matrix of $\mathbf{Z}_t$. In the Equation (12), $\hat{\mathbf{B}}$ can be obtained using for example Cholesky decomposition.

The synthetic precipitation series during the validation period can be obtained from the QRMAR model using two following steps.

(i) Randomly generate multivariate autoregressive random field using the MAR(1) of pamaeters $\hat{\mathbf{A}}$ and $\hat{\mathbf{B}}$ during the validation period.

(ii) Applying the inverse cumulative Bernoulli-GA distribution expressed in Equation (8) to the generated variables in the step (i) for each site and for each forecast day from the validation period.

4. Results

The QRMAR model has been trained for the calibration period (1960-1990), using precipitation data series from the seven stations and the 40 predictors obtained by the PCA. Once the parameters of the conditional Bernoulli-GA distribution ($\pi_j(t), \alpha_j(t)$ and $\beta_j(t)$) have been estimated for each day t and for each site j over the calibration period, all the obtained

conditional marginal distributions were used to obtain the latent variables $u(t)$ and to fit the parameters of MAR(1) model. Finally, all the fitted QRMAR parameters where used to generate precipitation series during the validation period (1991-2000). To assess the performance of the proposed QRMAR model, we compare it to hybrid and MMLR models. The MMLR model here is employed without stochastic variation compared to the hybrid model, and the wet day was determined when the series of the daily probability of precipitation occurrence obtained by the MMLR was larger than the threshold value of 0.5. For stability and robustness of both QRMAR and the hybrid, 100 realizations are generated of the precipitation series for each model.

Table 4 shows values of the mean errors (ME) and the root mean squares errors (RMSE), where the conditional mean of 100 realisations is used for both QRMAR and hybrid. The QRMAR statistical downscaling model does better than both the hybrid and MMLR in terms of ME and RMSE, and generally the hybrid model shows a better performance than the MMLR. The MMLR model shows best performance only for the two stations Nicolet and Donnacona 2 in term of ME. Because there is a significant random component in both the QRMAR and the hybrid models, it can be difficult to appreciate differences in model performance using the two criteria RMSE and ME. The MMLR model shows the poorest results. This result is expected due to the fact that the MMLR model is in reality biased because zero precipitation amounts were included to calibrate the MMLR amount model. In addition, the anscombe residuals R from the observed precipitation amount may not be exactly normally distributed. For this reason, the hybrid model employs a probability mapping technique to correct this bias. However, QRMAR not only performs better than the hybrid model but also it has the advantage of its automatic aspect of mapping in the conditional distribution of precipitation using the quantile regression component.

Thus, there is no need to rely on transformation steps or on bias correction procedures (such us a probability mapping technique) when evaluating the QRMAR model.

To evaluate the ability of the multivariate autoregressive component in QRMAR to reproduce the observed dependence structures in both time and space, scatter plots of the lag-0 and lag-1 cross-correlation of modeled versus observed precipitations are plotted for the three models in Figure 3. The correlation values of both QRMAR and hybrid models are obtained using the mean of the correlation values calculated from the 100 realisations. For lag-0 cross-correlation, points correspond to all 21 combinations of pairs of stations, while for lag-1 cross-correlation, points correspond to all 49 combinations because lag-1 cross-correlations are generally not symmetric. Because MMLR is not a multisite model, it gives the poorest performances and generally overestimates the cross-correlation of both lag-0 and lag-1. Figure 3 shows that the QRMAR and hybrid models preserve the lag-0 cross-correlation adequately. On the other hand, QRMAR reproduces more adequately the lag-1 cross-correlation than the hybrid model. In fact, the hybrid model, by its construction, is only able to take into account the lag-1 autocorrelation, unlike QRMAR which is assumed to preserve the full lag-1 cross-correlation.

5. Discussions

The downscaling problem as is tackled in this paper can be viewed as a regression problem, where we try to predict climate variables at small scale from climate variables at synoptic scale. Thereby, the QRMAR model may be considered as an extension of the quantile regression model to the multisite context, to reproduce the dependence structure of precipitation in both space and time. However, due to the large literature that addresses the precipitation modelling in general, the downscaling problem in the case of precipitation may be treated differently. In this respect,

precipitation downscaling problem may be viewed as an adjustment problem of existed precipitation models in general to account for large scale climate drivers (GCM precipitation, SLP, wind speed, etc.). Wilks 2010 suggested that these adjustments can be accomplished in two ways: (i) through imposed changes in the corresponding monthly statistics, (ii) or by controlling the precipitation model parameters by daily variations in simulated atmospheric circulation (Wilks 2010). In this context, most of probabilistic regression downscaling models focus on the second way (Cannon 2008; Ben Alaya et al. 2014; Ben Alaya et al. 2015a). Indeed, large scale climate drivers are employed as exogenous variables to explain parameters of an assumed conditional distribution, and by this way parametric precipitation models are routinely adapted for downscaling purposes. Thus, the proposed QAMAR model presents a new alternative in the context of probabilistic regression framework, an alternative that cannot be considered as a routine adaptation of an existed model. The Quantile regression component in the QRMAR model, requires no strong assumptions about the output distribution. To be clear, in this work, the choice of the gamma distribution to represent the precipitation amount was made only to simplify the model. Indeed, another option can be used, and in a more general way, the gamma distribution can be replaced by a non-parametric distribution such as a kernel distribution, and thus making the method more general and applicable to other variables such as temperature or wind speed. However, even if a choice of a parametric distribution was made, there will always be the advantage that no direct relationship is imposed between predictors and parameters of the assumed conditional distribution, and thus offering a more flexibility and freedom to the model output.

In our comparison study, the proposed QR model showed better performance than the MMLR and the hybrid model. These methods are often called hybrid models, because they combine two

components: (i) a deterministic regression component which provides the conditional mean and (ii) an unconditional resampling component to preserve observed weather characteristics at local scale (Harpham and Wilby 2005; Jeong et al. 2012b; Khalili et al. 2013). These hybrid approaches are based on a static noise observed during the calibration of the regression component. Therefore, the part of the variability which is explained by the randomization component does not depend on the predictors, and thus, it is supposed to be constant in a changing climate. For this reason, these hybrid approaches may not represent local change in the temporal variability in a climate change simulation. In this context, the proposed QRMAR model can be considered as a hybrid approach. However, it has important advantage compared to traditional hybrid approach regarding the reproduction of the temporal variability in a changing climate. Indeed, the total temporal variability is reproduced in the regression component through quantile regression, and thus it may change in the future according to the large scale atmospheric predictors.

6. Conclusions

A QRMAR model is proposed in this paper for simultaneously downscaling AOGCM predictors to daily precipitation at several sites. The QRMAR relies on a probabilistic modeling framework in order to predict the conditional Bernoulli-gamma distribution of precipitation at a daily time scale using quantile regression model and logistic regression model. To allow a realistic representation of relationships between stations at both time and space, stochastic generators procedures where applied using a latent multivariate autoregressive Gaussian field. The developed model was then applied to generate daily precipitation series at seven stations located in the southern part of the province of Quebec (Canada). Model evaluations suggest that the QRMAR model is capable of generating series with realistic spatial and temporal variability. The

developed model requires no strong assumptions and it can be easily applied to other variables such as temperature and wind speed making it a valuable tools not only for downscaling purposes but also for environmental and climate modelling.

Acknowledgments

We acknowledge Eva Mekis from Environment Canada for providing observed data sets of rehabilitated precipitation. The authors would like to acknowledge also the Data Access and Integration (DAI, see http://loki.qc.ec.gc.ca/DAI/login-e.php) team for providing the predictors data and technical support.

References

Baur, D., M. Saisana and N. Schulze (2004). "Modelling the effects of meteorological variables on ozone concentration—a quantile regression approach." Atmospheric Environment **38**(28): 4689-4699.

Ben Alaya, M. A., F. Chebana and T. Ouarda (2014). "Probabilistic Gaussian Copula Regression Model for Multisite and Multivariable Downscaling." Journal of Climate **27**(9).

Ben Alaya, M. A., F. Chebana and T. B. Ouarda (2015a). "Probabilistic Multisite Statistical Downscaling for Daily Precipitation Using a Bernoulli–Generalized Pareto Multivariate Autoregressive Model." Journal of climate **28**(6): 2349-2364.

Ben Alaya, M. A., F. Chebana and T. B. M. J. Ouarda (2015b). "Multisite and multivariable statistical downscaling using a Gaussian copula quantile regression model." Climate Dynamics: 1-15.

Benestad, R. E., I. Hanssen-Bauer and D. Chen (2008). Empirical-statistical downscaling, World Scientific.

Bras, R. L. and I. Rodríguez-Iturbe (1985). Random functions and hydrology, Courier Dover Publications.

Bremnes, J. B. (2004a). "Probabilistic forecasts of precipitation in terms of quantiles using NWP model output." Monthly Weather Review **132**(1).

Bremnes, J. B. (2004b). "Probabilistic wind power forecasts using local quantile regression." Wind Energy **7**(1): 47-54.

Cade, B. S. and B. R. Noon (2003). "A gentle introduction to quantile regression for ecologists." Frontiers in Ecology and the Environment **1**(8): 412-420.

Cannon, A. J. (2008). "Probabilistic multisite precipitation downscaling by an expanded Bernoulli-gamma density network." Journal of Hydrometeorology **9**(6): 1284-1300.

Cannon, A. J. (2011). "Quantile regression neural networks: Implementation in R and application to precipitation downscaling." Computers & Geosciences **37**(9): 1277-1284.

Coles, S., J. Bawa, L. Trenner and P. Dorazio (2001). An introduction to statistical modeling of extreme values, Springer.

Friederichs, P. and A. Hense (2007). "Statistical downscaling of extreme precipitation events using censored quantile regression." Monthly Weather Review **135**(6).

Furrer, E. M. and R. W. Katz (2008). "Improving the simulation of extreme precipitation events by stochastic weather generators." Water Resources Research **44**(12).

Haylock, M. R., G. C. Cawley, C. Harpham, R. L. Wilby and C. M. Goodess (2006). "Downscaling heavy precipitation over the United Kingdom: A comparison of dynamical and statistical methods and their future scenarios." International Journal of Climatology **26**(10): 1397-1415.

Jeong, D. I., A. St-Hilaire, T. B. M. J. Ouarda and P. Gachon (2012). "Multisite statistical downscaling model for daily precipitation combined by multivariate multiple linear regression and stochastic weather generator." Climatic Change **114**(3-4): 567-591.

Kalnay, E., M. Kanamitsu, R. Kistler, W. Collins, D. Deaven, L. Gandin, M. Iredell, S. Saha, G. White and J. Woollen (1996). "The NCEP/NCAR 40-year reanalysis project." Bulletin of the American meteorological Society **77**(3): 437-471.

Kistler, R., E. Kalnay, W. Collins, S. Saha, G. White, J. Woollen, M. Chelliah, W. Ebisuzaki, M. Kanamitsu and V. Kousky (2001). "The NCEP-NCAR 50-year reanalysis: Monthly means CD-ROM and documentation." Bulletin-American Meteorological Society **82**(2): 247-268.

Kleiber, W., R. W. Katz and B. Rajagopalan (2012). "Daily spatiotemporal precipitation simulation using latent and transformed Gaussian processes." Water Resources Research **48**(1).

Koenker, R. and G. Bassett (1978). "Regression quantiles." Econometrica: journal of the Econometric Society: 33-50.

Koenker, R. and F. Schorfheide (1994). "Quantile spline models for global temperature change." Climatic Change **28**(4): 395-404.

Maraun, D., F. Wetterhall, A. Ireson, R. Chandler, E. Kendon, M. Widmann, S. Brienen, H. Rust, T. Sauter and M. Themeßl (2010). "Precipitation downscaling under climate change: recent

developments to bridge the gap between dynamical models and the end user." Reviews of Geophysics **48**(3).

Mehrotra, R. and A. Sharma (2006). "A nonparametric stochastic downscaling framework for daily rainfall at multiple locations." Journal of Geophysical Research: Atmospheres (1984–2012) **111**(D15).

Mekis, E. and W. D. Hogg (1999). "Rehabilitation and analysis of Canadian daily precipitation time series." Atmosphere-Ocean **37**(1): 53-85.

Rasmussen, P. (2013). "Multisite precipitation generation using a latent autoregressive model." Water Resources Research **49**(4): 1845-1857.

Semenov, M. A., E. M. Barrow and A. Lars-Wg (2002). "A stochastic weather generator for use in climate impact studies." User Manual, Hertfordshire, UK.

Serinaldi, F. and C. G. Kilsby (2014). "Simulating daily rainfall fields over large areas for collective risk estimation." Journal of Hydrology **512**: 285-302.

Tareghian, R. and P. F. Rasmussen (2013). "Statistical downscaling of precipitation using quantile regression." Journal of Hydrology **487**: 122-135.

Terrell, G. R. (2003). "The Wilson–Hilferty transformation is locally saddlepoint." Biometrika **90**(2): 445-453.

Villarini, G., B.-C. Seo, F. Serinaldi and W. F. Krajewski (2014). "Spatial and temporal modeling of radar rainfall uncertainties." Atmospheric Research **135**: 91-101.

Wilby, R. L., C. W. Dawson and E. M. Barrow (2002). "sdsm - a decision support tool for the assessment of regional climate change impacts." Environmental Modelling and Software **17**(2): 145-157.

Wilks, D. S. (1999). "Multisite downscaling of daily precipitation with a stochastic weather generator." Climate Research **11**: 125-136.

Wilks, D. S. (2010). "Use of stochastic weathergenerators for precipitation downscaling." Wiley Interdisciplinary Reviews: Climate Change **1**(6): 898-907.

Wilks, D. S. (2012). "Stochastic weather generators for climate-change downscaling, part II: multivariable and spatially coherent multisite downscaling." Wiley Interdisciplinary Reviews: Climate Change **3**(3): 267-278.

Williams, P. M. (1998). "Modelling seasonality and trends in daily rainfall data." Advances in neural information processing systems: 985-991.

Yang, C., R. E. Chandler, V. S. Isham and H. S. Wheater (2005). "Spatial-temporal rainfall simulation using generalized linear models." Water Resources Research **41**(11): 1-13.

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

Table 1. List of the 9 stations used in this study.

No.	Site	Name of station	Latitude (°N)	Longitude (°W)
1	7025440	Nicolet	46.25	-72.60
2	7022160	Drummondville	45.88	-72.48
3	7012071	Donnacona 2	46.68	-71.73
4	7066685	Roberval A	48.52	-72.27
5	7060400	Bagotville A	48.33	-71
6	7056480	Rimouski	48.45	-68.53
7	7047910	Seven Island A	50.22	-66.27

Table 2. NCEP predictors on the CGCM3 grid.

No	Predictors	No	Predictors
1	Mean pressure at the sea level	14	Divergence at 500 hPa
2	Wind speed at 1000 hPa	15	Wind speed at 850 hPa
3	Component U at 1000 hPa	16	Component U at 850 hPa
4	Component V at 1000 hPa	17	Component V at 850 hPa
5	Vorticity at 1000 hPa	18	Vorticity at 850 hPa
6	Wind direction at 1000 hPa	19	Geopotential at 850 hPa
7	Divergence at 1000 hPa	20	Wind direction at 850 hPa
8	Wind speed at 500 hPa	21	Divergence at 1000 hPa
9	Component U at 500 hPa	22	Specific humidity at 500 hPa
10	Component V at 500 hPa	23	Specific humidity at 850 hPa
11	Vorticity at 500 hPa	24	Specific humidity at 1000 hPa
12	Geopotential at 500 hPa	25	Temperature at 2m
13	Wind direction at 500 hPa		

.Table 3. Failures to reject over 10957 tests for each station.

	Marginal distribution			
Stations	GAM	WEI	GPD	EXP
Nicolet	**8377**	8235	5686	4720
Drummondville	**9563**	9353	4922	3276
Donnacona 2	9332	**9491**	6766	5276
Roberval A	9153	**9271**	6455	5389
Bagotville A	9535	**9608**	6679	5292
Rimouski	**8566**	8348	5094	3766
Seven Island A	9201	**9228**	5671	3933
Sum	**63727**	63534	41273	31652

Bold character denotes the distribution with the highest number of failures to reject.

Table 4. Quality assessment of the estimated series for the validation period (1991–2000) for QRMAR, Hybrid and MMLR. Criteria are ME and RMSE. For the QRMAR model Criteria were calculated from median of 100 realisations. Bold indicates the best result.

	ME (mm)			RMSE (mm)		
	QRMAR	Hybrid	MMLR	QRMAR	Hybrid	MMLR
Chelsea	**-0.16**	-1.41	2.34	**5.80**	6.13	6.45
Cedars	**0.08**	-1.36	2.66	**6.42**	6.68	7.09
Nicolet	**1.11**	-2.17	2.74	7.61	7.44	**6.14**
Drummondville	**0.51**	-1.11	2.47	**5.47**	5.65	7.00
Donnacona 2	**0.78**	-1.29	2.92	6.21	6.41	**6.13**
Roberval A	**0.08**	-1.48	2.19	**5.31**	5.70	6.93
Bagotville A	0.93	**-0.63**	2.57	**6.30**	6.35	6.90
Rimouski	**0.24**	-1.39	2.08	**5.39**	5.55	5.80
Seven Island A	**-0.33**	-1.94	2.20	**5.52**	6.05	6.09

Bold means better result.

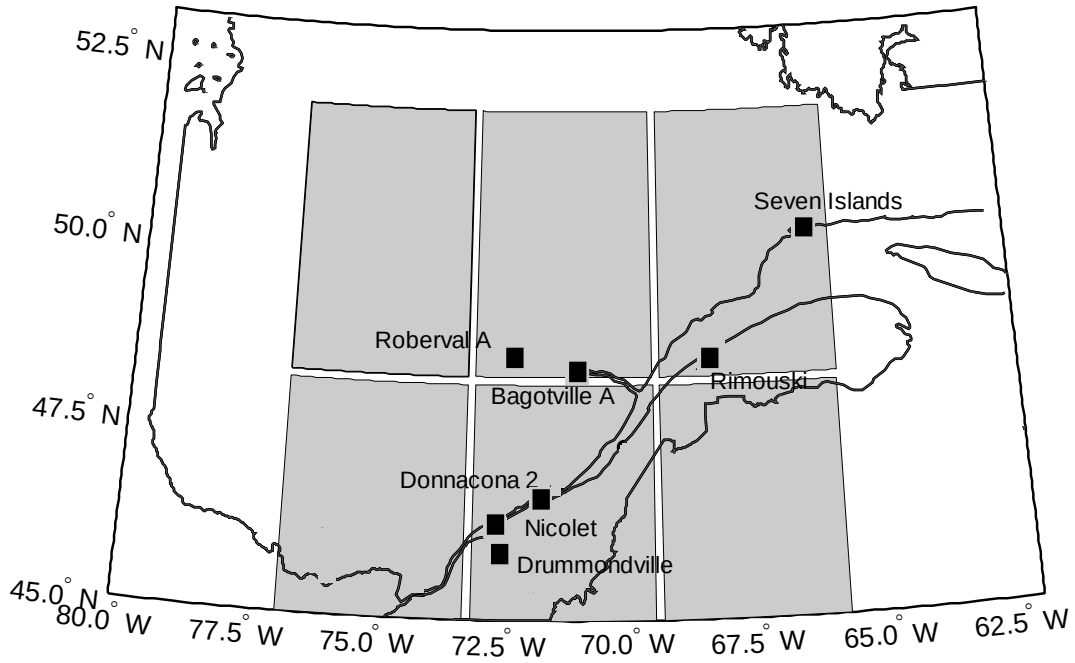

Figure 1. Locations of CGCM3 grid and observation stations of daily precipitation.

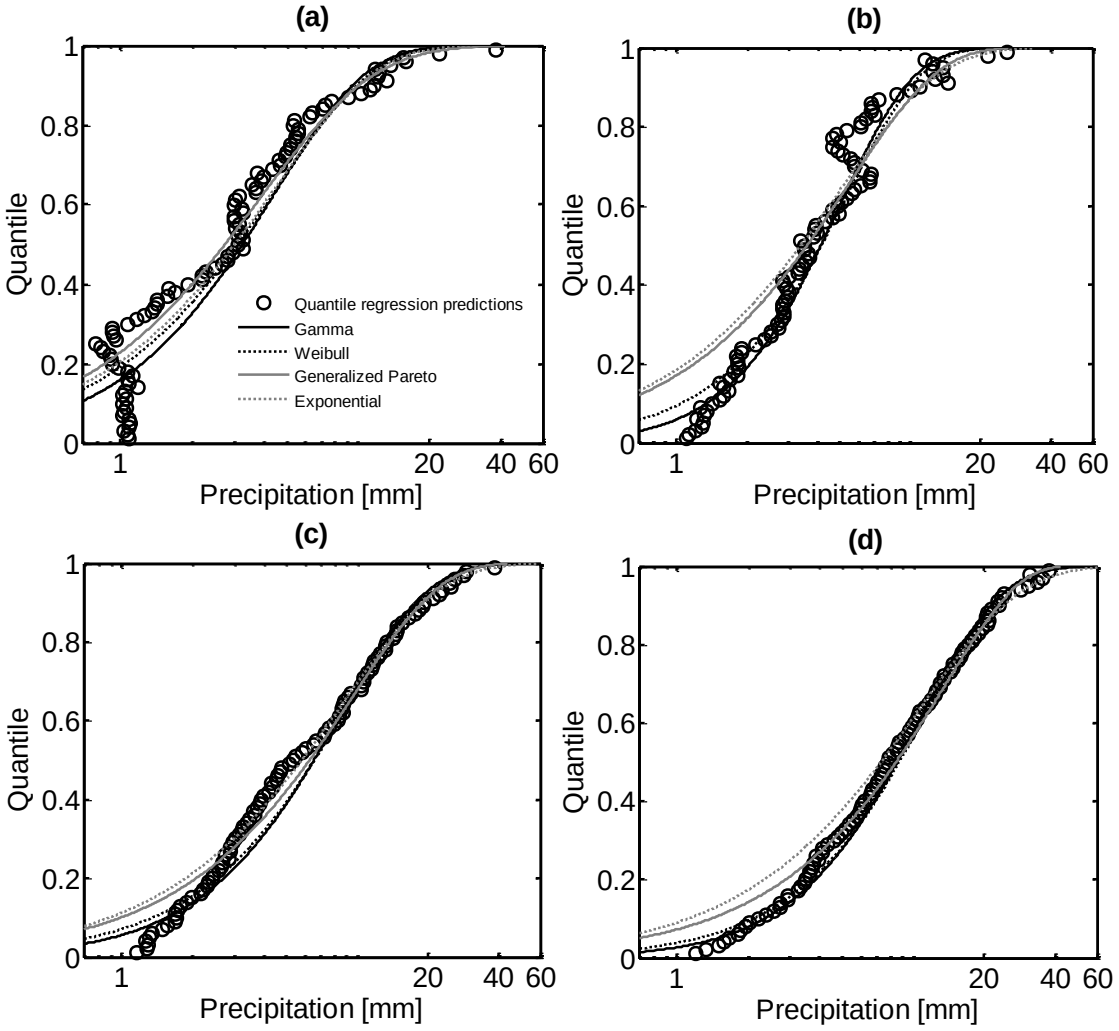

Figure 2. Quantile regression predictions at Nicolet station for 02-02-1961 (a), 02-05-1961 (b), 02-08-1961 (c) and 02-11-1961 (c).

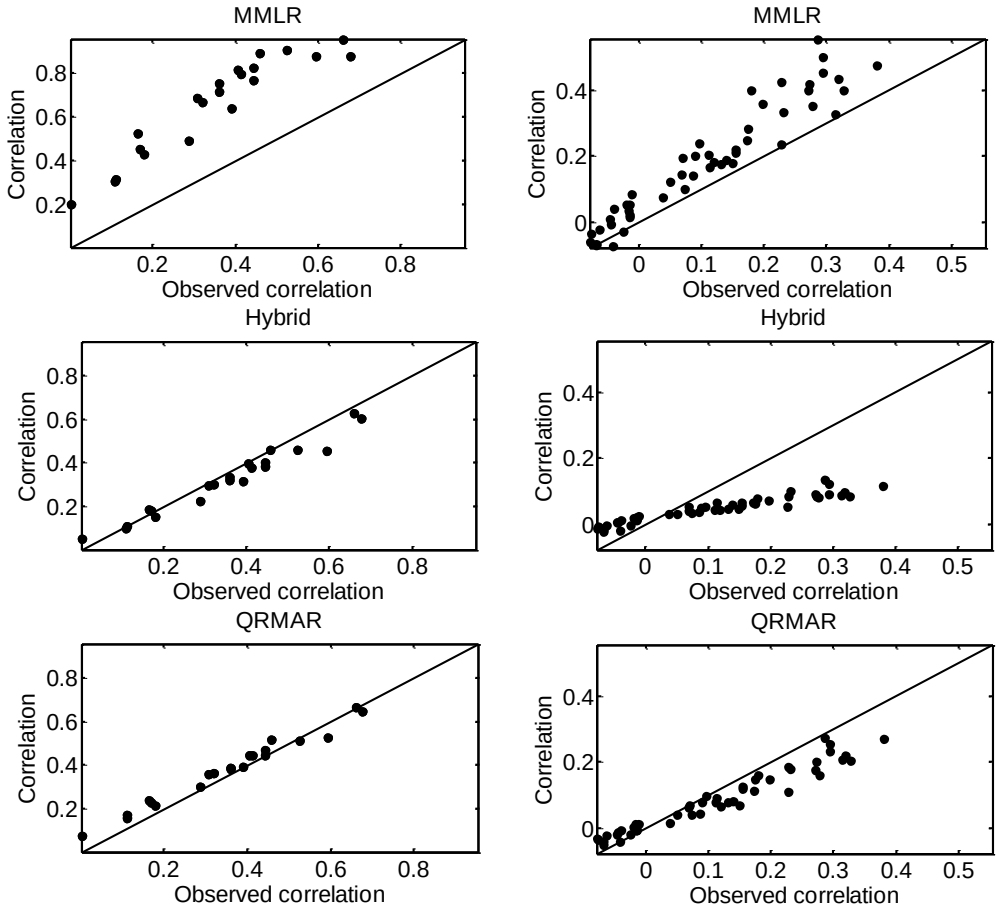

Figure 3. Scatter plots of observed and modeled Lag-0 correlation (left columns) and Lag-1 correlation (right columns) for the QRMAR model, hybrid model and MMLR model during the validation period. Correlation values of the QRMAR and the hybrid model are obtained using the mean of the correlation values calculated from a 100 simulations.

CHAPITRE 7: APPLICATION OF SPATIAL BAYESIAN MODEL FOR DOWNSCALING DAILY TEMPERATURES AND COMPARISON WITH TWO PROBABILISTIC REGRESSION APPROACHES

Application of a spatial Bayesian model for daily temperature downscaling and comparison with two probabilistic regression approaches

M. A. Ben Alaya[1], D. Fasbender[2], T.B.M.J. Ouarda[3, 1] and F. Chebana[1]

[1]*INRS-ETE, 490 rue de la Couronne, Québec (QC),*

Canada G1K 9A9

[2]*European Commission, Joint Research Centre,*

Italy

[3]*Masdar Institute of science and technology*

P.O. Box 54224, Abu Dhabi, UAE

Corresponding author: Tel: +1 (418) 654 2530#4468

Email: mohammed_ali.ben_alaya@ete.inrs.ca

Submitted December 15, 2015

Abstract

In the present paper a spatial Bayesian model (SBM) is adapted and applied for the downscaling of AOGCM data to minimum and maximum daily temperatures in the province of Quebec, Canada. The model is proposed in order to circumvent the inability of classic regression methods to produce the observed temporal variability and to provide spatial estimations at ungauged sites. In this method, the monthly mean of the prior distribution is modeled using local characteristics in a geographical regression model (GRM). The model relies on a Bayesian framework for combining a temperature spatial model reflecting monthly local patterns with the daily fluctuations induced by the atmospheric predictors. Reanalysis products are used in this work to assess the potential of the proposed method. The adopted model is compared with two probabilistic regression approaches namely the vector generalized linear model (VGLM) and the probabilistic quantile regression (QR) model based on their application to a data base of 22 gauged sites over a large area from the province of Quebec, Canada. Results showed that the three models accurately simulate the temporal variability of daily temperature series. In addition, validation results of the SBM based on climatic indices are in sufficient agreement compared to both VGLM and QR model. While VGLM and QR results are slightly better at gauged sites, the SBM has the advantage of providing estimates at ungauged locations.

Keywords: statistical downscaling, spatial Bayesian model, maximum and minimum temperatures, quantile regression, vector generalized linear model.

1. Introduction

Atmosphere-ocean general circulation models (AOGCM) represent the most commonly used method to simulate large-scale climate evolutions and projections. Unfortunately, AOGCM data are generally produced on regular grids with a low horizontal resolution around 2.5 ° longitude and latitude (approximately 250 to 300 km). This coarse resolution is not appropriate and cannot provide the required information to reliably assess the hydrological impacts of climate change (Grotch and MacCracken 1991; Huth and Kyselý 2000). To solve this problem, various downscaling techniques were developed to refine AOGCM data (Benestad et al. 2008). These techniques can be classified into two main methods: dynamic methods and statistical methods (Herrera et al. 2006; Benestad et al. 2008; Maraun et al. 2010).

Dynamic downscaling methods use physically based limited area models with a high resolution. These methods require large computational capabilities and substantial human resources. Statistical downscaling methods establish statistical links between large-scale variables (predictors) and small-scale variables (predictands). One of the first benefits of these methods is their ease of implementation (Jeong et al. 2012a). The large body of literature on statistical downscaling models and techniques can be classified into three main approaches: stochastic weather generators (Wilks 1999; Wilks 2010; Jeong et al. 2012b), weather typing (Conway et al. 1996) and regression methods (Cannon 2009; Hammami et al. 2012; Jeong et al. 2013).

Classical regression models are successfully used in downscaling, but their major drawback is that their estimations are generally limited to gauged sites, and they are unable to extend results to ungauged locations. Another disadvantage of classical regression models is that they generally produce the mean or the central predictions conditionally on the selected predictors. Thus,

regression variability is always lower than the observed one. To adequately reproduce the observed temporal variability of the process, Karl et al. (1990) suggested "inflating" the modeled variance to match the observed. On the other hand, Von Storch (1999) mentioned that inflation is not realistic, and the local scale variation could not be completely explained by synoptic scale atmospheric predictors; instead, randomization of the predictand by adding an explicit resampled noise term is more realistic. As a result of the resampling step, the part of the variability which is explained by the randomization component does not depend on the predictors, and thus, it is supposed to be constant in a changing climate. For this reason, randomisation procedures may not represent local change in the temporal variability in a climate change simulation.

In this regard, probabilistic approaches have made valuable contributions in downscaling applications. They account for the non-explained local variability by predicting the entire conditional distributions at each forecast step (Bates et al. 1998; Bellone et al. 2000; Tebaldi et al. 2004). Probabilistic regression approaches can be adopted using Bayesian formulation (Fasbender and Ouarda 2010), quantile regression (QR) (Bremnes 2004; Friederichs and Hense 2007; Cannon 2011; Ben Alaya et al. 2015b) or regression models where outputs are parameters of the conditional distribution such as the vector form of generalized linear model (VGLM) (Maraun et al. 2011; Ben Alaya et al. 2015a), the vector form of the generalized additive model (VGAM) (Yee and Wild 1996; Yee and Stephenson 2007) and conditional density estimation network (CDEN) (Williams 1998; Cannon 2008; Li et al. 2013). Drawing values from the obtained conditional distribution at each forecast step using a probabilistic regression model ensures the preservation of a realistic temporal variability of downscaled predictand. Unlike traditional randomisation approaches, the temporal variability is reproduced from a pure

regression model, and thus it may change in the future according to the large scale atmospheric predictors.

To make estimation at ungauged locations, Michelangeli et al. (2009) interpolated AOGCM predictors on the gauged locations before the regression. Another solution is to interpolate the downscaled results (Benestad 2002; Hundecha and Bárdossy 2005; Benestad 2007). A third solution is to interpolate the predictands at gauged stations on the target regular grid before the regression (Lim et al. 2007; Baigorria et al. 2008). This last method gives good results and adequately reproduces the observed climate features. However, it suffers from the drawback of high computational requirements which are directly related to the size of the target regular grid. As classical spatial models use the spatial interpolation either before or after the regression model, Fasbender and Ouarda (2010) introduced a promising approach that takes into account the spatial dependence directly in the downscaling process. To this end, Fasbender and Ouarda (2010) proposed a spatial Bayesian model (SBM) for downscaling AOGCM data to maximum and minimum temperatures in the southern part of Quebec, Canada. This model is based on a Bayesian framework to link predictors and predictands through some hidden upscaled predictands that have the same resolution and same location as AOGCM data. The prior information is described using a geographical regression model (GRM) which takes into account the local characteristics (latitude and altitude). This model has the advantage of providing estimations of the entire conditional distribution at ungauged sites. Although this model produces satisfactory results in downscaling maximum and minimum temperature fields, it suffers from some shortcomings. First, this Bayesian model was developed and tested on a relatively small area (about 1000x1000 km ^{2}), and the estimated model parameters cannot be used for large surfaces. Second, the low number of stations used for calibration, failed to take into account other

local features than latitude and altitude to express the GRM prior. In this respect, the present work has two objectives:

(i) To adapt the SBM proposed by Fasbender and Ouarda (2010) over a large area and to improve the GRM model by adding two other covariates: the longitude and the distance from the coast.

(ii) To compare SBM, VGLM and QR model for daily maximum and minimum temperature downscaling over an area covering a large part of the province of Quebec, Canada.

The paper is structured as follows: after a brief description of the study area and the used data, the SBM, VGLM and QR models are described. Then, the quality assessment method is presented. Thereafter the three models are applied to the case study and their results are compared. Finally discussions and conclusions are given.

2. Study area and data

The study area is located in Quebec, in the latitudes between 45°N and 60°N and the longitudes between 60 ° W and 80 ° W (see Figure 1). Daily maximum and minimum temperature series at 22 stations obtained from Environment Canada are considered as predictands (see Figure 1). These series cover the period between 1 January 1961 and 31 December 2000 and were homogenized by Vincent et al. (2002). Table 1 shows the annual statistics of these series.

The reanalysis products NCEP / NCAR (Kalnay et al. 1996; Kistler et al. 2001) are used as predictors to evaluate the potential of the downscaling methods. All the NCEP / NCAR data are averaged on a daily basis from six hourly data and then linearly interpolated to match the

CGCM3 grid (Scinocca et al. 2008). 24 grid points covering the study area are available (see Figure 1), and for each grid point, 25 NCEP predictors are provided (see Table 2). For each day, 600 variables are available for the downscaling process. A principal component analysis (PCA) is performed to reduce the number of NCEP predictors. Only the first principal components which retain 95% of the original predictor variability are retained as uncorrelated predictors. The total data set is divided into two independent sets: a calibration period between 1961 and 1990 and a validation period between 1991 and 2000. It is worth mentioning that there are more recent predictors developed by Environment Canada and covering longer periods (cf. http://ccds-dscc.ec.gc.ca/index.php?page=dst-sdi). These predictors can be used for future works.

3. Methodology

In the following subsections, the SBM, the VGLM and the QR models as well as the quality assessment criteria are presented.

3.1. Spatial Bayesian model

For a given day, we aim to estimate maps of maximum and minimum temperatures for the study area from values of NCEP predictors outputs for this day. We denote the r uncorrelated predictors retained from the PCA by:

$$\mathbf{X} = \left(X_1 \quad \ldots \quad X_r \right)^T \tag{1}$$

where T denotes the vector transpose. Consider now the target grid for downscaling process. It is a regular grid finer than the AOGCM grid and covering the whole study area. We denote $\mathbf{a}_i$ the coordinates of a specific point on this grid, and $\mathbf{a}_0$ the coordinates of the target point for the

downscaling process. On a given day, given the vector $\mathbf{X}$, we try to estimate maximum temperature $T_{\max}(\mathbf{a}_0)$ and minimum temperature $T_{\min}(\mathbf{a}_0)$ at the site $\mathbf{a_0}$. Let us define the vector of predictands $\mathbf{Y}_0$ for the location $\mathbf{a_0}$, and the vector of predictands $\mathbf{Y_i}$ for the location $\mathbf{a}_i$,

$$\mathbf{Y_0} = \left(T_{\max}(\mathbf{a}_0) \quad T_{\min}(\mathbf{a}_0) \right)^T , \tag{2}$$

$$. \mathbf{Y_i} = \left(T_{\max}(\mathbf{a_i}) \quad T_{\min}(\mathbf{a_i}) \right)^T \tag{3}$$

In a probabilistic context, we look to estimate the conditional distribution $f(\mathbf{Y}_0 \mid \mathbf{X})$. The Bayesian method proposed by Fasbender and Ouarda (2010) is applied in this work to estimate $f(\mathbf{Y_0} \mid \mathbf{X})$. The application of this method for all locations of the target grid, allows estimating all the local conditional distributions $f(\mathbf{Y_i} \mid \mathbf{X})$. This allows estimating the two maps of maximum and minimum temperatures. In a Bayesian approach, all the unknown parameters are considered as random variables. To reflect the randomness of an unknown parameter, the available information on this parameter is modeled by a prior probability distribution. Then, using Bayes theorem, this prior distribution is updated by the observations to finally obtain a posterior distribution.

a. Geographical regression model

The prior information is specified for each month using a geographical regression model (GRM) which takes into account the longitude, latitude, altitude and the distance from the coast at the location $\mathbf{a_0}$ (see Benestad et al. 2008). One denotes by α the parameter vector of the prior distribution and by m_0 its mean vector. In this case the GRM model is given by:

$$m_{0;d}(\alpha) = \alpha_{1;d} + \alpha_{2;d}(\lambda - m_\lambda) + \alpha_{3;d}(\varphi - m_\varphi) + \alpha_{4;d}\sqrt{D} + \alpha_{5;d}H + \varepsilon_{m;d}\ , d = 1,\ldots,12 \quad (4)$$

where (i) $d = 1,\ldots,12$ corresponds to the current month, (ii) λ, φ, D, and H are respectively the longitude, latitude, distance from the coast and the altitude of the target location, (iii) m_φ and m_λ are respectively the averages of the latitudes and longitudes observed at weather stations, and (iv) $\alpha_{1;d},\ldots,\alpha_{5;d}$ are the sub-vectors of the vector α. The parameters are adjusted using the ordinary least squares (OLS) method. The parameters $\alpha_{5;d}$ are imposed in accordance with the International Standard Atmosphere (ISO 2533:1975; see the ISO Web site at http://www.iso.org/), they are defined so that the temperature decreases with altitude by 6.5 °C / km, which corresponds to the standard adiabatic gradient of temperature. OLS estimators are asymptotically multivariate Gaussian. One denotes by $\mathbf{\Sigma}_\alpha$ the covariance matrix of the vector α. To respect the seasonal patterns observed at meteorological stations, the GRM is estimated and corrected by adding the errors $\varepsilon_{m;d}$ using the inverse distance interpolation method. The parameters for this interpolation are estimated using a "leave-one-out" procedure for each month. The observed variance of $\varepsilon_{m;d}$ is chosen here as the variance of a Gaussian semivariogram model and $\mathbf{\Sigma}_m$ is defined as the diagonal covariance matrix constructed using this Gaussian semivariogram model (Fasbender and Ouarda 2010).

b. Spatial model for the temperature

A joint spatial model for both daily maximum and minimum temperatures is estimated using a linear model of co-regionalization (LMC) (see Delfiner 2009). To account for the smooth spatial variation of temperature, this spatial model combines a Gaussian model and a nugget model. The

spatial distances were calculated using the UTM (Universal Transverse Mercator) coordinates. The use of UTM coordinates avoids the spatial distortions due to high latitudes. Now we assume that each AOGCM grid point $\mathbf{A_i}$ corresponds to a spatial area. Let us define the upscaled random vector of temperatures by:

$$\mathbf{Z} = \left(T_{\max}(\mathbf{A_1}) \quad \ldots \quad T_{\max}(\mathbf{A_n}) \quad T_{\min}(\mathbf{A}_1) \quad \ldots \quad T_{\min}(\mathbf{A_n}) \right)^T, \tag{5}$$

where $T_{\max}(\mathbf{A_i})$ and $T_{\min}(\mathbf{A_i})$ are expressed as follows:

$$T_{\max}(\mathbf{A_i}) = \frac{1}{|\mathbf{A_i}|} \int_{\mathbf{A_i}} T_{\max}(\mathbf{a}) \mathrm{d}\mathbf{a} \ , \tag{6}$$

$$T_{\min}(\mathbf{A_i}) = \frac{1}{|\mathbf{A_i}|} \int_{\mathbf{A_i}} T_{\min}(\mathbf{a}) \mathrm{d}\mathbf{a} \ . \tag{7}$$

where $|\mathbf{A_i}|$ is the area of $\mathbf{A_i}$. The vector of the upscaled predictands $\mathbf{Z}$ is related to the vector $\mathbf{X}_0$ by their spatial positions, and is used here to take into account the spatial dependence of temperature during the downscaling process. The relationship between $\mathbf{Z}$ and $\mathbf{X}_0$ is described by a regularized covariance function (Goovaerts 2008) which is calculated using the LMC model. Thus, we can estimate a series of $\mathbf{Z}$ covering the calibration period. The GRM and the two equations (6) and (7) are used to ensure a more consistent estimate of the monthly mean $m_{\mathbf{Z}}(\alpha)$ of the vector $\mathbf{Z}$.

c. Multivariate Multiple Linear Regression

Once the series of $\mathbf{Z}$ are estimated for the calibration period from historical data, we identify

their relationship with the predictors **X** using the MMLR given by:

$$\mathbf{Z} = \beta_{0,d} + \beta_{1,d}\mathbf{X} + \varepsilon \qquad d = 1,\ldots 12, \tag{8}$$

Where β is the vector of the model parameters, $\beta_{0,d}$ is the sub-vector of β corresponding to the monthly intercept, $\beta_{1,d}$ is a $2 \times r$ matrix with columns corresponding to the predictors $\mathbf{Y}$ and ε is a residual vector with zero mean and covariance matrix Σ_ε.

d. Posterior estimation

The SBM combines three models: (i) the GRM that is used to estimate the prior means of maximum and minimum temperatures, (ii) the joint spatial model for both maximum and minimum temperature that determines the relationship between predictands and the upscaled predictands using the regularized covariance function and (iii) the MMLR model which determines the relationship between the upscaled predictands and atmospheric predictors. This sub-section shows how these three models are combined to estimate the posterior distribution.

The upscaled predictand vector $\mathbf{Z}$ and the original predictors (NCEP predictors) are located in the same location on the AOGCM grid, and also represent the same spatial resolution. Thus, the information provided by the predictands $\mathbf{Y}_0$ on the predictors $\mathbf{X}$ can be neglected compared to the information provided by the upscaled predictands. The vector $\mathbf{Z}$ can be considered as a vector of hidden variables allowing the transfer of information from predictors to predictands. Thereafter, independence between the predictors $\mathbf{X}$ and the predictands $\mathbf{Y_0}$ conditionally to the vector $\mathbf{Z}$ is assumed:

$$\mathbf{Y}_0 \perp \mathbf{X} \,|\, \mathbf{Z}. \tag{9}$$

In addition, the vector β is supposed to influence only the link between $\mathbf{X}$ and $\mathbf{Z}$, and α is assumed to influence only the prior distributions. Furthermore, α and β are supposed to be independent. With all these assumptions, and using the Bayes theorem, the posterior distribution is written as:

$$f\left(\mathbf{Y}_0 \mid \mathbf{X}\right) \propto \iiint \frac{f\left(\mathbf{Z} \mid \mathbf{X};\beta\right)}{f\left(\mathbf{Z} \mid \alpha\right)} f\left(\mathbf{Y}_0, \mathbf{Z} \mid \alpha\right) f\left(\alpha\right) f\left(\beta\right) \mathrm{d}\mathbf{Z}\, \mathrm{d}\alpha\, \mathrm{d}\beta . \tag{10}$$

More details about this development are provided in Fasbender and Ouarda (2010).

To take into account the seasonal effects, all distributions in equation (10) are modeled monthly. Generally, the monthly temperature data follows approximately a Gaussian distribution. Thus, distributions of $f\left(\mathbf{Y}_0, \mathbf{Z} \mid \beta\right)$, $f\left(\mathbf{Z} \mid \mathbf{X};\beta\right)$ and $f\left(\mathbf{Z} \mid \alpha\right)$ are assumed to be Gaussian. Also, all distributions in equation (10) are considered stationary. However, non-stationary fluctuations could be induced by the non-stationary evolution of predictors. By considering these hypotheses Fasbender and Ouarda (2010) showed that $(\mathbf{Y}_0, \mathbf{Z})$ follows a multivariate Gaussian distribution with mean $\mathbf{M}(\alpha, \beta)$ and variance-covariance matrix $\mathbf{S}$ given by:

$$\begin{cases} \mathbf{S}^{-1} = \boldsymbol{\Sigma}^{-1} + \begin{pmatrix} 0 & 0 \\ 0 & \boldsymbol{\Sigma}_{\varepsilon}^{-1} - \boldsymbol{\Sigma}_{\mathbf{Z}}^{-1} \end{pmatrix} \\ \mathbf{M}\left(\alpha, \beta\right) = \mathbf{S}\left(\boldsymbol{\Sigma}^{-1} \begin{pmatrix} m_0\left(\alpha\right) \\ m_{\mathbf{Z}}\left(\alpha\right) \end{pmatrix} + \begin{pmatrix} 0 \\ \boldsymbol{\Sigma}_{\varepsilon}^{-1}(\beta_{0,d} + \beta_{1,d})\mathbf{X} - \boldsymbol{\Sigma}_{\mathbf{X}}^{-1} m_{\mathbf{Z}}\left(\alpha\right) \end{pmatrix} \right) \end{cases}, \tag{11}$$

Where Σ is the covariance matrix of $\mathbf{Y}_0$ and $\mathbf{Z}$, $\boldsymbol{\Sigma}_{\mathbf{Z}}$ is the sub-matrix of Σ corresponding only to $\mathbf{Z}$, $\boldsymbol{\Sigma}_{\varepsilon}$ is the covariance matrix of MMLR residuals. $m_0(\alpha)$ and $m_{\mathbf{Z}}(\alpha)$ are respectively the

mean vectors of $\mathbf{Y}_0$ and $\mathbf{Z}$.

Fasbender and Ouarda (2010) showed that it was not necessary to find an analytical expression for the posterior distribution $f(\mathbf{Y}_0 | \mathbf{X})$ because it is possible to estimate all properties of this distribution (e.g. standard deviation, mode, median,…) from a sampling algorithm. The procedure consists first in drawing samples $\tilde{\beta}_i$ from $f(\beta)$ using the MMLR model of section 3.1.c, $\tilde{\alpha}_i$ from $f(\alpha)$, and $(\tilde{\varepsilon}_{m;d})_i$ using the GRM in Section 3.1.a. Then, using the samples $\tilde{\alpha}_i$ and $(\tilde{\varepsilon}_{m;d})_i$, we calculate the perturbed values $(\tilde{m}_0(\tilde{\alpha}_i))_i$ and $(\tilde{m}_{\mathbf{Z}}(\tilde{\alpha}_i))_i$ using the GRM and equations (6) and (7) of Section 3.1.b. In the next step we use the samples $\tilde{\beta}_i$ and $(\tilde{m}_0(\tilde{\alpha}_i))_i$, to draw a sample $(\tilde{\mathbf{Y}}_0, \tilde{\mathbf{Z}})_i$ from the Gaussian distribution with the parameters given in equation (11). Finally, we repeat these steps until the number of simulations is reached.

3.2. Vector generalized linearized linear model

In most applications, traditional regression models are performed to describe the conditional mean given a set of selected predictors. For this reason, they underrepresent the temporal variability of the model outputs. To accurately reproduce the observed temporal variability, VGLM allows modeling the whole conditional distribution. Instead of the conditional mean only, a suitable parametric probability density function (PDF) is assumed for the predictand and then parameters of this distribution are considered to vary at each forecast step according to the predictor values. In the current work, the Gaussian distribution is assumed for temperature variables (Dorling et al. 2003; Ben Alaya et al. 2014). Thereby, for a given predictand $y(t)$ on a day t assumed to be normally distributed, the VGLM would have two outputs: one for the

conditional mean $\mu(t)$ and one for the conditional variance $\sigma^2(t)$. In this case, the VGLM is described by:

$$\mu(t) = \mathrm{E}\left[y(t) \mid x(t)\right] = a^T x(t) \tag{12}$$

$$\sigma(t) = \exp\left[b^T x(t)\right] \tag{13}$$

where $x(t)$ represents the value of the predictors (obtained from PCA) on day t, and the coefficients a and b are VGLM parameters. Then, the conditional normal PDF of $y(t)$ for a day t is given by:

$$f_t\left[y(t) \mid x(t)\right]) = \frac{1}{\sqrt{2\pi\sigma^2(t)}} \exp\left[-\frac{(y(t)-\mu(t))^2}{2\sigma^2(t)}\right] \tag{14}$$

The model parameters are set following the method of maximum likelihood by minimizing the negative log predictive density (NLPD) cost function (Haylock et al. 2006; Cawley et al. 2007; Cannon 2008):

$$\mathscr{L}_k = \sum_{t=1}^{n} \log\left\{ f_{tk}\left[y_k(t) \mid x(t)\right]\right\} \tag{15}$$

VGLM parameters are estimated for each month and for each station separately.

3.3. Probabilistic quantile regression model

Quantile regression **(**QR) provides an alternative solution to reproduce the whole conditional distribution by directly estimating point values of the individual quantiles of the conditional distribution. QR models generalize models for the conditional median to provide the conditional

quantile. Instead of minimising the mean absolute errors, Koenker and Bassett (1978) applied asymmetric weights to positive/negative errors using a pinball loss function to compute conditional quantiles of the predictive distribution. The pinball loss function is given by:

$$\rho_p(u) = \begin{cases} u(p-1) & \text{if } u < 0 \\ up & \text{if } u \geq 0 \end{cases} \tag{16}$$

Where $1 < p < 0$. Given a set of potential daily predictor variables $x(t)$, a given predictand $y(t)$, and a vector of parameters c_p, the linear regression equation for the p^{th} quantile $Q_{tp}\left(y(t) \mid x(t)\right)$ of the conditional distribution of $y(t)$ given $x(t)$ is expressed as:

$$Q_{tp}\left(y(t) \mid x(t)\right) = c_p^T x(t) \tag{17}$$

For a given set of observations $(x(t), y(t))$, $t = 1, \ldots, n$, the vector of parameters c_p can be obtained by minimizing the QR error function given by:

$$e_p = \sum_{t=1}^{n} \rho_p\left(y(t) - c_p^T x(t)\right) \tag{18}$$

QR is a semi-parametric approach as it avoids assumptions about the parametric distribution of the error process and does not assume homogenous residuals. To obtain the whole conditional distribution of temperature on a given day, the QR model can be applied to produce quantiles from 0.01 to 0.99 by steps of 0.01 for each station and for each day (Tareghian and Rasmussen 2013; Ben Alaya et al. 2015b). The obtained sample can be considered as a sample from the target conditional distribution. Therefore, a non-parametric empirical distribution can be assumed to represent the conditional cumulative distribution function. Thus, simulation of downscaled

temperature is carried out by drawing random values from the obtained conditional distribution at each forecast step. In our case study, QR parameters are estimated for each month and for each station separately.

3. 4. Quality assessment of downscaled results

For assessing the downscaling quality, observed temperature series at the 22 stations between 1991 and 2000 are used. A first evaluation is performed by considering a direct comparison between observed and simulated series using three statistical criteria, which are the mean error (ME), the root mean square error (RMSE) and the percentage of observed data in the 95% confidence interval (CI). The ME is a measure of accuracy and should be close to 0, the RMSE is an inverse measure of the accuracy and must be minimized, whereas the percentage of observations in the 95% CI should be close to 95% to ensure that the estimated distribution is relevant. In the second validation approach, we consider the RMSE values of the monthly mean and the monthly standard deviation and a set of monthly and seasonal climate indices that have been proposed in northern climates. The definitions of these climate indices are presented in Table 3. These indices are chosen to evaluate the performance of downscaling temperature models (Gachon et al. 2005; Hessami et al. 2008). The evaluation consists in comparing the concordance between these indices obtained using the downscaling model and those observed at gauged sites. The RMSE is used to measure this concordance.

4. Results

4.1. Calibration results of the spatial Bayesian model

Local characteristics (longitude, latitude, altitude and distance from the coast) at the 22 gauged

stations were used to calibrate the GRM. Altitude data were provided by a numerical terrain model (NTM) covering the study area with around 9 Km resolution (see Figure 2). GRM residuals were then interpolated in space and added using inverse distance interpolation. The power parameters of the inverse distance interpolation were estimated monthly, using a leave-one-out procedure. The variance of this interpolation was modeled as a Gaussian variogram function with a variance equal to the variance of the residuals of the GRM and a range equal to 100 km. The uncertainty of the GRM was considered proportional to the distance to the nearest weather station. A cross-validation was used to validate the GRM. Figure 3 shows the comparison for each month between the modeled standard deviation provided by the GRM and the observed RMSE calculated using cross-validation. One can see that the two lines are very close and thus, the standard deviations of the GRM are consistent with the RMSE of the cross-validation.

The GRM was applied to the study area. Figure 4 shows the effect of the distance from the coast on the prior mean of maximum and minimum temperatures. One may note that there is a seasonal effect provided by the distance from the coast. Indeed, during the summer, temperature increases according to the distance from the coast, however, in winter we can see a contrary effect of smaller importance. Figure 4.e and Figure 4.f show that, during the summer, the effect of the distance from the coast on the maximum temperature is greater than the effect on the minimum temperature. These results show that the distance from the coast cannot be ignored, and considering this local characteristic in the GRM could certainly improve the downscaled results.

Figure 5 shows the effect of the geographical position on the prior mean of the maximum and minimum temperatures for the months of January, April, July and October. This figure shows that temperature variations, according to the geographical position, are more important in winter

than in summer. We can also see that the longitude effect is more marked in winter, and during the other seasons it is very small and negligible compared to the latitude effect.

The joint spatial model for daily maximum and minimum temperatures was estimated for each month using the LMC model. Figure 6 shows the fitted multivariate semivariogram model for both maximum and minimum temperatures. Then, the upscaled predictands were estimated using the spatial model according to the theory of regularized covariance functions (Goovaerts 2008), and their means prior were calculated by the GRM. Subsequently, the MMLR model of equation (8) was adjusted using these upscaled predictands and the uncorrelated predictors obtained from the PCA.

4. 2. Application Results

The spatial Bayesian model was applied to the study area using NCEP predictors. Four dates corresponding to certain observed anomalies were chosen for the illustrations. The obtained temperature maps for these dates are shown in Figure 7. The temperatures maps estimated for January 3rd 1981 are shown in Figure 7a and 7b. This day is characterized by very low temperatures in the southern part of the province of Quebec. We can see that the model was able to reproduce the observed anomaly on this day. Figure 7c and 7d show the temperature maps estimated for January 17th 1996, which was characterized by fairly high temperatures in the low latitudes. One can see that using information from predictors, the Bayesian model was able to reproduce the high temperatures recorded on this day. Similarly for August 24th 1994, the Bayesian approach was able to correctly estimate the very low maximum temperatures (Figures 7.e and 7.f). The last two figures (Figure 7.g and 7.h) show the estimated maps for August 24th 1983, for which the Bayesian model gives good results to correctly estimate the very high

recorded temperatures.

Results were then validated by comparing the observed and simulated series at the 22 gauged stations. Series of maximum and minimum temperatures were simulated by the SBM, the VGLM and the QR models during the validation period (1991-2000). For maximum and minimum temperatures respectively, Table 4 and Table 5 show values of the mean errors (ME), the root mean squares errors (RMSE), and the percentage of observations in the 95% confidence interval. RMSE and ME were calculated using the conditional means of 100 simulations. Generally, for both maximum and minimum temperatures, the three models provide satisfactory results based on ME and RMSE (on a daily basis). However, the QR model is found to be slightly more accurate. Based on the percentage of observations in the 95% CI, it appears that no significant difference between the three models for both maximum and minimum temperature was detected. Approximately, the percentage of observed temperatures that belong to the 95% confidence interval is very close to 95% for each model, which ensures that all models are able to reproduce the daily fluctuations of maximum and minimum temperatures in Quebec.

To assess the three models' ability to replicate observed temperature variability on a seasonal basis, Table 6 shows the RMSE between observed and simulated monthly mean (MM) and monthly standard deviation (MSD) for both maximum and minimum temperatures for the three models. In general, the QR model outperformed the two other models, although the three models were able to reproduce MM and MSD for all sites. Regarding maximum temperature, for both MM and MSD, the QR model shows lower RMSE values than VGLM and SBM at 10 of the 22 stations. In addition, for minimum temperatures, the QR model exhibits lower RMSE values than the two other models at 12 stations for the MM and 15 stations for the MSD.

To further compare the three modelling approaches, the RMSEs of climate indices presented in Table 3 were computed. The means of the indices values from the 100 realisations were used to calculate the RMSEs of these indices for each models. Table 7 summarizes the obtained RMSE values. We can note that the QR model generally provides the best performance in terms of all indices except for FRTH for which the VGLM provides the lower RMSE values. In addition, we can see that both VGLM and QR models are slightly better than those of the SBM in terms of RMSEs of climate indices.

5. Discussions

In the SBM, upscaled predictands are considered as a vector of hidden variables allowing the transfer of information from predictors to original predictands. They are related to the large-scale predictors by the MMLR model. On the other hand, they are linked to the original predictands by their spatial positions where their relationship is described using the theory of regularized covariance functions which was calculated using the LMC. This allows the SBM to provide estimations at any location in the study area. Since atmospheric predictors and uspcaled predictands share the same spatial resolution, the statistical link between the predictors and the upscaled predictands should be more direct than when using the at-site original predictands. In the SBM, the prior information is modeled using the GRM to reflect the monthly spatial dependence using local characteristics. Therefore, the posterior distribution reflects both monthly local patterns by the GRM and daily large-scale fluctuations induced by atmospheric predictors. The SBM is adapted over a large area covering most of the province of Quebec, Canada. As, the data set used in this paper (22 stations and 24 AOGCM grid points) is larger than the one used by Fasbender and Ouarda (2010) (9 stations and 6 grid points), the GRM proposed by (Fasbender and Ouarda 2010) (which is limited to the latitude and the altitude) is improved by adding the

longitude and the distance from the coast.

In our comparison study, results at gauged stations show that VGLM and QR models are slightly better than those obtained using the SBM. This finding is not surprising since the VGLM and QR models determine a direct relationship between the predictors and the 22 predictands, unlike the SBM which introduces information from predictors by their links through the upscaled predictands. Generally, both VGLM and QR downscaling models give very good results at stations. But these models are very specific to observed predictands at gauged sites and are unable to provide estimations at ungauged locations. In addition, observed differences in results with the SBM seem to be insignificant. Therefore, the loss of precision when using the SBM at gauged sites can be neglected compared to the advantage of providing estimates at ungauged locations.

6. Conclusions

In the present paper the SBM is adapted and applied in order to downscale NCEP data to maximum and minimum temperatures over a large area from the province of Quebec, Canada. The present work focused on using reanalysis products in the spatial downscaling model in order to assess the potential of the proposed method. However, the final objective of downscaling applications is to use AOGCM data. This model directly takes into account the spatial dependence of predictands in the downscaling process by employing some hidden upscaled predictands which have the same resolution and the same location as AOGCM models. This model relies on a Bayesian framework to combine information provided from three different models, (i) the LMC as a joint spatial model for both minimum and maximum temperatures, (ii) the GRM to estimate the mean of the prior distribution using local characteristics and (iii) the

MMLR model which determines the relationship between the upscaled predictands and atmospheric predictors. Results of the SBM at 22 gauged stations from the province of Quebec are then compared with those obtained using two probabilistic regression approaches namely the VGLM and the QR model. Results show that all three models accurately simulate daily temperature series at the 22 stations with a proper preservation of the temporal variability. In addition, validation results of SBM based on climatic indices are in sufficient agreement compared to both VGLM and QR models. Indeed, even if both VGLM and QR models lead to slightly better results at observed sites, the SBM has the advantage of directly producing temperature maps.

Atmospheric variable scales from NCEP do not include local or regional effects on temperature (i.e. influenced by surface conditions). In this respect, the use of spatial model and regularization of covariance functions allows to take into account other predictors at a third scale. This possibility might be of great interest for downscaling applications since the method would not only account for the change of spatial resolution but also for multiple scales or multiple AOGCM or RCM models at once. Future research efforts can focus on these factors and key components. In addition, the proposed GRM could be improved by including other local factors of surface conditions, for instance, the orientation of the local slope and/or valley (ex. main rivers as Saguenay, St. Lawrence, and others).

Acknowledgments

The authors would like to acknowledge the Data Access Integration Team for providing the data and technical support. The Numerical Terrain Model (NTM) was provided by the Department of Natural Resources.

7. References

Baigorria, G. A., J. W. Hansen, N. Ward, J. W. Jones and J. J. O'Brien (2008). "Assessing predictability of cotton yields in the southeastern United States based on regional atmospheric circulation and surface temperatures." Journal of Applied Meteorology and Climatology 47(1): 76-91.

Bates, B. C., S. P. Charles and J. P. Hughes (1998). "Stochastic downscaling of numerical climate model simulations." Environmental Modelling and Software 13(3-4): 325-331.

Bellone, E., J. P. Hughes and P. Guttorp (2000). "A hidden Markov model for downscalling synoptic atmospheric patterns to precipitation amounts." Climate Research 15(1): 1-12.

Ben Alaya, M. A., F. Chebana and T. Ouarda (2014). "Probabilistic Gaussian Copula Regression Model for Multisite and Multivariable Downscaling." Journal of Climate 27(9).

Ben Alaya, M. A., F. Chebana and T. B. Ouarda (2015a). "Probabilistic Multisite Statistical Downscaling for Daily Precipitation Using a Bernoulli–Generalized Pareto Multivariate Autoregressive Model." Journal of climate 28(6): 2349-2364.

Ben Alaya, M. A., F. Chebana and T. B. M. J. Ouarda (2015b). "Multisite and multivariable statistical downscaling using a Gaussian copula quantile regression model." Climate Dynamics: 1-15.

Benestad, R. (2002). "Empirically downscaled temperature scenarios for northern Europe based on a multi-model ensemble." Climate Research 21(2): 105-125.

Benestad, R. E. (2007). "Novel methods for inferring future changes in extreme rainfall over Northern Europe." Climate Research 34(3): 195.

Benestad, R. E., I. Hanssen-Bauer and D. Chen (2008). Empirical-statistical downscaling, World Scientific.

Bremnes, J. B. (2004). "Probabilistic forecasts of precipitation in terms of quantiles using NWP model output." Monthly Weather Review 132(1).

Cannon, A. J. (2008). "Probabilistic multisite precipitation downscaling by an expanded Bernoulli-gamma density network." Journal of Hydrometeorology 9(6): 1284-1300.

Cannon, A. J. (2009). "Negative ridge regression parameters for improving the covariance structure of multivariate linear downscaling models." International Journal of Climatology 29(5): 761-769.

Cannon, A. J. (2011). "Quantile regression neural networks: Implementation in R and application to precipitation downscaling." Computers & Geosciences 37(9): 1277-1284.

Cawley, G. C., G. J. Janacek, M. R. Haylock and S. R. Dorling (2007). "Predictive uncertainty in environmental modelling." Neural Networks 20(4): 537-549.

Conway, D., R. Wilby and P. Jones (1996). "Precipitation and air flow indices over the British Isles." Climate Research 7: 169-183.

Delfiner, P. (2009). Geostatistics: modeling spatial uncertainty, Wiley-Interscience.

Dorling, S. R., R. J. Foxall, D. P. Mandic and G. C. Cawley (2003). "Maximum likelihood cost functions for neural network models of air quality data." Atmospheric Environment 37(24): 3435-3443.

Fasbender, D. and T. B. M. J. Ouarda (2010). "Spatial Bayesian Model for Statistical Downscaling of AOGCM to Minimum and Maximum Daily Temperatures." Journal of Climate 23(19): 5222-5242.

Friederichs, P. and A. Hense (2007). "Statistical downscaling of extreme precipitation events using censored quantile regression." Monthly Weather Review 135(6).

Gachon, P., A. St-Hilaire, T. B. M. J. Ouarda, V. Nguyen, C. Lin, J. Milton, D. Chaumont, J. Goldstein, M. Hessami, T. D. Nguyen, F. Selva, M. Nadeau, P. Roy, D. Parishkura, N. Major, M. Choux and A. Bourque (2005). "A first evaluation of the strength and weaknesses of statistical downscaling methods for simulating extremes over various regions of eastern Canada " Sub-component, Climate Change Action Fund (CCAF), Environment Canada Final report(Montréal, Québec, Canada): 209.

Goovaerts, P. (2008). "Kriging and semivariogram deconvolution in the presence of irregular geographical units." Mathematical Geosciences 40(1): 101-128.

Grotch, S. L. and M. C. MacCracken (1991). "The use of general circulation models to predict regional climatic change." Journal of Climate 4(3): 286-303.

Hammami, D., T. S. Lee, T. B. M. J. Ouarda and J. Le (2012). "Predictor selection for downscaling GCM data with LASSO." Journal of Geophysical Research D: Atmospheres 117(17).

Haylock, M. R., G. C. Cawley, C. Harpham, R. L. Wilby and C. M. Goodess (2006). "Downscaling heavy precipitation over the United Kingdom: A comparison of dynamical and statistical methods and their future scenarios." International Journal of Climatology 26(10): 1397-1415.

Herrera, E., T. B. M. J. Ouarda and B. Bobée (2006). "Downscaling methods applied to Atmosphere-Ocean General Circulation Models (AOGCM)." Méthodes de désagrégation Appliquées aux Modèles du Climat Global Atmosphère-Océan (MCGAO) 19(4): 297-312.

Hessami, M., P. Gachon, T. B. M. J. Ouarda and A. St-Hilaire (2008). "Automated regression-based statistical downscaling tool." Environmental Modelling & Software 23(6): 813-834.

Hundecha, Y. and A. Bárdossy (2005). "Trends in daily precipitation and temperature extremes across western Germany in the second half of the 20th century." International Journal of Climatology 25(9): 1189-1202.

Huth, R. and J. Kyselý (2000). "Constructing site-specific climate change scenarios on a monthly scale using statistical downscaling." Theoretical and Applied Climatology 66(1-2): 13-27.

Jeong, D., A. St-Hilaire, T. Ouarda and P. Gachon (2012a). "Comparison of transfer functions in statistical downscaling models for daily temperature and precipitation over Canada." Stochastic Environmental Research and Risk Assessment 26(5): 633-653.

Jeong, D., A. St-Hilaire, T. Ouarda and P. Gachon (2013). "A multivariate multi-site statistical downscaling model for daily maximum and minimum temperatures." Climate Research 54(2): 129-148.

Jeong, D. I., A. St-Hilaire, T. B. M. J. Ouarda and P. Gachon (2012b). "Multisite statistical downscaling model for daily precipitation combined by multivariate multiple linear regression and stochastic weather generator." Climatic Change 114(3-4): 567-591.

Kalnay, E., M. Kanamitsu, R. Kistler, W. Collins, D. Deaven, L. Gandin, M. Iredell, S. Saha, G. White and J. Woollen (1996). "The NCEP/NCAR 40-year reanalysis project." Bulletin of the American meteorological Society 77(3): 437-471.

Karl, T. R., W.-C. Wang, M. E. Schlesinger, R. W. Knight and D. Portman (1990). "A method of relating general circulation model simulated climate to the observed local climate. Part I: Seasonal statistics." Journal of Climate 3(10): 1053-1079.

Kistler, R., E. Kalnay, W. Collins, S. Saha, G. White, J. Woollen, M. Chelliah, W. Ebisuzaki, M. Kanamitsu and V. Kousky (2001). "The NCEP-NCAR 50-year reanalysis: Monthly means CD-ROM and documentation." Bulletin-American Meteorological Society 82(2): 247-268.

Koenker, R. and G. Bassett (1978). "Regression quantiles." Econometrica: journal of the Econometric Society: 33-50.

Li, C., V. P. Singh and A. K. Mishra (2013). "Monthly river flow simulation with a joint conditional density estimation network." Water Resources Research 49(6): 3229-3242.

Lim, Y. K., D. Shin, S. Cocke, T. LaRow, J. T. Schoof, J. J. O'Brien and E. P. Chassignet (2007). "Dynamically and statistically downscaled seasonal simulations of maximum surface air temperature over the southeastern United States." Journal of Geophysical Research: Atmospheres (1984–2012) 112(D24).

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.

Maraun, D., F. Wetterhall, A. Ireson, R. Chandler, E. Kendon, M. Widmann, S. Brienen, H. Rust, T. Sauter and M. Themeßl (2010). "Precipitation downscaling under climate change: recent developments to bridge the gap between dynamical models and the end user." Reviews of Geophysics 48(3).

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

Scinocca, J., N. McFarlane, M. Lazare, J. Li and D. Plummer (2008). "The CCCma third generation AGCM and its extension into the middle atmosphere." Atmos. Chem. Phys 8(23): 7055-7074.

Tareghian, R. and P. F. Rasmussen (2013). "Statistical downscaling of precipitation using quantile regression." Journal of Hydrology 487: 122-135.

Tebaldi, C., L. O. Mearns, D. Nychka and R. L. Smith (2004). "Regional probabilities of precipitation change: A Bayesian analysis of multimodel simulations." Geophysical Research Letters 31(24).

Vincent, L. A., X. Zhang, B. Bonsal and W. Hogg (2002). "Homogenization of daily temperatures over Canada." Journal of Climate 15(11): 1322-1334.

Von Storch, H. (1999). "On the Use of "Inflation" in Statistical Downscaling." Journal of Climate 12(12): 3505-3506.

Wilks, D. S. (1999). "Multisite downscaling of daily precipitation with a stochastic weather generator." Climate Research 11: 125-136.

Wilks, D. S. (2010). "Use of stochastic weathergenerators for precipitation downscaling." Wiley Interdisciplinary Reviews: Climate Change 1(6): 898-907.

Williams, P. M. (1998). "Modelling seasonality and trends in daily rainfall data." Advances in neural information processing systems: 985-991.

Yee, T. W. and A. G. Stephenson (2007). "Vector generalized linear and additive extreme value models." Extremes 10(1-2): 1-19.

Yee, T. W. and C. Wild (1996). "Vector generalized additive models." Journal of the Royal Statistical Society. Series B (Methodological): 481-493.

Table 1. Annual mean (MA) and standard deviation (SD) of the 22 gauged stations for maximum temperature (Tmax) and minimum temperature (Tmin).

				Tmax		Tmin	
Station	Lat	Long	Elevation	MA	SD	MA	SD
Inukjuak	58.47	-78.08	25	-2.9	13.45	-10.23	13.98
Kuujjuaq	58.1	-68.42	39	-1.08	14.13	-10.10	13.92
Kuujjuarapik	55.28	-77.77	10	0.06	13.74	-8.89	14.04
Schefferville	54.8	-66.82	522	-0.12	13.76	-9.61	14.34
Sept-Iles A	50.22	-66.27	55	5.53	10.98	-3.64	11.85
Natashquan	50.18	-61.82	11	5.63	10.10	-3.36	11.39
Chibougamau Chapais	49.77	-74.53	387	4.90	13.76	-6.74	14.40
Gaspe	48.78	-64.48	33	8.61	11.34	-2.76	10.70
Mont-Joli A	48.6	-68.22	52	7.56	11.79	-0.74	10.46
Amos	48.57	-78.13	310	6.48	13.74	-4.20	13.69
Causapscal	48.37	-67.23	168	8.13	12.52	-3.54	12.05
Bagotville A	48.33	-71	159	7.88	13.40	-2.39	12.69
Val-D'Or A	48.05	-77.78	337	7.20	13.52	-3.46	13.38
La Tuque	47.4	-72.78	152	9.54	13.09	-3.19	13.25
La Pocatiere	47.35	-70.03	31	8.81	12.15	-0.38	11.21
Québec	46.78	-71.38	74	9.18	12.56	-0.05	11.46
ManiwakiAirport	46.3	-76	200	9.90	12.64	-1.99	12.50
Drummondville	45.88	-72.48	82	10.49	12.61	0.89	12.11
McTavish (Montréal)	45.5	-73.58	73	11.14	12.29	3.49	11.52
Sherbrooke A	45.43	-71.68	241	10.34	12.13	-1.06	11.81
Lennoxville	45.37	-71.82	181	11.10	12.17	-0.20	11.93
Les Cedres	45.3	-74.05	47	10.77	12.25	2.37	11.72

Table 2. NCEP predictors on the CGCM3 grid.

No	Predictors	No	Predictors
1	Mean pressure at the sea level	14	Divergence at 500 hPa
2	Wind speed at 1000 hPa	15	Wind speed at 850 hPa
3	Component U at 1000 hPa	16	Component U at 850 hPa
4	Component V at 1000 hPa	17	Component V at 850 hPa
5	Vorticity at 1000 hPa	18	Vorticity at 850 hPa
6	Wind direction at 1000 hPa	19	Geopotential at 850 hPa
7	Divergence at 1000 hPa	20	Wind direction at 850 hPa
8	Wind speed at 500 hPa	21	Divergence at 1000 hPa
9	Component U at 500 hPa	22	Specific humidity at 500 hPa
10	Component V at 500 hPa	23	Specific humidity at 850 hPa
11	Vorticity at 500 hPa	24	Specific humidity at 1000 hPa
12	Geopotential at 500 hPa	25	Temperature at 2m
13	Wind direction at 500 hPa		

Table 3. Definitions of the climatic indices used for the performance assessment of downscaled temperatures.

Indices	Definition	Time scale
DTR (°C)	Mean of diurnal temperature range	Season
FSL (day)	Frost season length: Days between 5 consecutive T_{mean}<0 ℃ and 5 consecutive T_{mean}> 0℃	Years
GSL (day)	Growing Season length: Days between 5 consecutive T_{mean}<5 ℃ and 5 consecutive T_{mean}> 5℃	Years
FR-Th (day)	Days with freeze and thaw (T_{max}> 0°C,T_{min}<0°C)	Months
Tmax90 (°C)	90th percentile of daily maximum temperature	Season
Tmin90 (°C)	90th percentile of daily minimum temperature	Season

$$T_{mean} = \frac{T_{max} + T_{min}}{2}$$

Table 4. Quality assessment of downscaled maximum temperature series during the validation period (1991–2000) for SBM, VGLM and QR models. Criteria are ME, RMSE, and percentage of observations in the 95% confidence interval (% in CI).

Stations	ME (°C)			RMSE (°C)			% in CI		
	SBM	VGLM	QR	SBM	VGLM	QR	SBM	VGLM	QR
Inukjuak	0.48	0.39	**0.35**	3.34	3.32	**3.28**	**0.95**	0.94	0.96
Kuujjuaq	0.60	0.47	**0.43**	3.46	**3.45**	3.48	0.91	0.92	**0.93**
Kuujjuarapik	**0.15**	0.27	0.28	3.38	3.29	**3.32**	0.92	**0.94**	0.92
Schefferville	0.57	**0.28**	0.34	3.59	**3.40**	3.48	0.93	0.92	**0.94**
Sept-Iles A	0.24	0.19	**0.15**	**3.52**	3.54	3.58	**0.95**	0.96	0.93
Natashquan	0.19	**0.13**	0.23	3.58	3.25	**3.23**	0.92	0.91	0.92
Chibougamau Chapais	0.30	0.28	**0.25**	3.47	3.29	**3.28**	0.92	0.93	**0.95**
Gaspe	0.20	**0.15**	0.26	**3.07**	3.19	3.24	**0.95**	0.94	0.97
Mont-Joli A	0.58	0.45	**0.42**	**3.23**	3.29	3.25	0.90	**0.92**	0.89
Amos	0.25	**0.22**	0.29	3.33	**3.30**	3.35	0.91	**0.92**	0.91
Causapscal	0.26	**0.18**	0.32	3.30	3.31	**3.19**	0.91	0.93	0.93
Bagotville A	0.57	0.50	**0.45**	3.48	**3.29**	3.42	0.90	0.92	0.92
Val-D'Or A	**0.03**	0.19	0.15	3.25	3.27	**3.22**	0.91	0.93	**0.94**
La Tuque	**-0.03**	-0.15	-0.09	3.62	**3.55**	3.58	0.89	**0.96**	0.91
La Pocatiere	0.38	-0.12	**-0.11**	3.62	3.50	**3.45**	0.89	**0.94**	0.92
Québec	0.47	0.39	**0.41**	3.74	3.66	**3.61**	**0.93**	0.91	0.91
ManiwakiAirport	0.39	0.38	**0.32**	3.72	3.69	**3.60**	0.91	0.97	**0.94**
Drummondville	**0.19**	0.23	0.24	3.65	3.55	**3.50**	0.93	0.98	**0.95**
McTavish (Montréal)	0.64	0.59	**0.44**	**3.49**	3.51	3.54	0.92	**0.94**	0.93
Sherbrooke A	0.26	**0.03**	0.12	3.97	3.82	**3.72**	**0.94**	0.93	0.92
Lennoxville	0.59	**0.55**	0.58	3.73	3.65	**3.43**	0.93	0.96	**0.95**
Les Cedres	**0.004**	0.23	0.11	**3.46**	3.51	3.54	**0.92**	0.90	0.91

Bold character means best result.

Table 5. Quality assessment of downscaled minimum temperature series during the validation period (1991–2000). Criteria are ME, RMSE, and percentage of observations in the 95% confidence interval (% in CI).

Stations	ME (°C)			RMSE (°C)			% in CI		
	SBM	VGLM	QR	SBM	VGLM	QR	SBM	VGLM	QR
Inukjuak	0.19	**0.11**	0.29	3.65	**3.54**	3.67	0.97	0.96	**0.95**
Kuujjuaq	0.78	0.72	**0.68**	3.49	3.39	**3.22**	0.97	**0.94**	0.93
Kuujjuarapik	0.46	0.41	**0.39**	3.94	**3.74**	3.81	**0.95**	0.94	0.92
Schefferville	0.49	0.45	**0.41**	4.27	4.29	**4.13**	**0.95**	0.92	0.94
Sept-Iles A	**-0.19**	-0.25	-0.21	**3.26**	3.33	3.46	0.98	0.96	**0.95**
Natashquan	0.28	0.31	**0.19**	4.33	**4.13**	4.17	**0.95**	0.91	0.92
Chibougamau Chapais	0.60	0.49	**0.43**	4.02	**3.71**	3.89	**0.95**	0.93	**0.95**
Gaspe	0.28	**0.24**	0.50	3.54	3.34	**3.25**	0.96	0.95	0.95
Mont-Joli A	0.49	**0.23**	0.51	3.26	3.22	**3.16**	0.97	0.94	**0.95**
Amos	0.35	**0.33**	0.38	4.39	**3.95**	4.03	0.90	**0.92**	0.91
Causapscal	**0.06**	0.19	0.23	3.52	3.46	**3.32**	**0.96**	0.93	0.93
Bagotville A	**-0.02**	-0.11	-0.29	3.43	3.39	**3.33**	0.97	**0.95**	0.94
Val-D'Or A	0.16	**0.14**	0.18	**3.16**	3.22	3.18	0.97	0.93	**0.94**
La Tuque	0.24	**0.20**	0.27	3.57	**3.42**	3.44	0.96	0.96	0.93
La Pocatiere	0.73	0.58	**0.53**	4.86	4.56	**4.46**	0.89	**0.94**	0.93
Québec	0.39	0.31	**0.23**	4.20	**4.08**	4.14	**0.95**	0.91	0.91
ManiwakiAirport	0.49	0.41	**0.40**	4.24	4.14	**4.11**	0.94	0.97	0.94
Drummondville	**0.02**	0.23	0.11	4.15	3.96	**3.93**	0.95	0.98	**0.95**
McTavish (Montréal)	0.79	**0.55**	0.62	3.64	3.52	**3.44**	**0.95**	0.94	0.93
Sherbrooke A	1.02	0.81	**0.75**	3.85	3.65	**3.35**	0.95	0.95	0.92
Lennoxville	0.79	0.65	**0.55**	3.69	**3.49**	3.55	0.94	0.96	**0.95**
Les Cedres	**0.03**	0.23	0.12	3.85	**3.65**	3.97	0.92	0.90	**0.93**

Bold character means best result.

Table 6. RMSE between observed and downscaled monthly mean (MM) and monthly standard deviation (MSD) for SBM, VGLM and QR models during the validation period at the 22 gauged stations.

Station	Tmax (°C)						Tmin (°C)					
	MM			MSD			MM			MSD		
	SBM	VGLM	QR	SBM	VGLM	QR	SBM	VGLM	QR	SBM	VGLM	QR
Inukjuak	1.26	1.24	**1.22**	0.89	0.84	**0.83**	1.38	1.35	**1.31**	1.33	1.30	**1.28**
Kuujjuaq	1.18	1.16	**1.15**	0.84	0.83	**0.81**	1.40	1.42	**1.36**	1.22	1.23	**1.20**
Kuujjuarapik	1.02	1.01	1.02	0.78	**0.76**	0.77	1.32	1.31	**1.30**	1.12	1.11	**1.10**
Schefferville	1.23	1.20	**1.18**	0.82	0.84	0.82	1.44	1.43	**1.41**	1.08	1.05	**1.04**
Sept-Iles A	1.16	1.12	1.12	0.83	0.85	**0.81**	**1.19**	1.22	1.20	1.55	1.35	**1.31**
Natashquan	1.06	1.07	**1.05**	0.86	**0.81**	0.83	1.35	**1.29**	1.32	1.04	**0.95**	1.03
Chibougamau Chapais	1.17	1.17	**1.15**	0.80	**0.75**	0.78	1.60	1.58	**1.52**	1.11	1.12	**1.05**
Gaspe	1.09	**1.06**	1.07	0.76	0.74	**0.72**	1.31	**1.26**	1.28	0.94	0.92	**0.86**
Mont-Joli A	1.18	**1.15**	1.16	0.82	0.82	0.84	1.14	**1.12**	1.17	**1.00**	1.03	1.05
Amos	**0.98**	0.99	1.01	0.77	**0.73**	0.75	**1.27**	1.29	1.29	1.10	**1.04**	1.06
Causapscal	0.83	**0.81**	0.82	0.83	0.82	**0.79**	**1.10**	1.15	1.12	1.02	1.04	**0.98**
Bagotville A	**1.21**	1.24	1.22	0.85	0.81	**0.73**	1.17	1.16	**1.13**	1.20	1.17	**1.15**
Val-D'Or A	0.95	0.94	0.94	**0.78**	0.79	0.81	**1.06**	1.09	1.08	1.48	**1.41**	1.43
La Tuque	1.08	1.07	**1.03**	0.88	0.88	**0.86**	1.18	**1.08**	1.15	0.99	0.95	**0.93**
La Pocatiere	**1.20**	1.22	1.21	0.85	0.84	**0.82**	1.54	1.52	**1.51**	1.23	1.21	1.21
Québec	**1.20**	1.23	1.21	0.88	**0.86**	0.87	1.37	1.34	**1.27**	0.94	0.93	**0.90**
Maniwaki Airport	1.25	**1.21**	1.23	0.88	0.82	0.82	1.39	**1.35**	1.40	1.03	1.11	**1.01**
Drummondville	1.13	1.13	**1.12**	**0.82**	0.85	0.88	1.40	1.35	**1.33**	1.03	**1.02**	1.03
McTavish (Montréal)	1.69	1.66	**1.64**	0.94	0.90	**0.89**	1.82	1.81	**1.79**	1.30	1.25	**1.24**
Sherbrooke A	1.27	**1.23**	1.24	0.98	0.96	**0.95**	1.87	1.87	**1.85**	1.12	1.10	**1.08**
Lennoxville	1.40	1.42	**1.38**	**0.90**	0.93	0.92	1.66	**1.56**	1.58	1.08	**1.04**	1.07
Les Cedres	0.97	0.95	**0.94**	0.75	0.71	0.70	1.32	1.33	**1.30**	0.99	0.92	**0.89**

Bold character means best result.

Table 7. RMSE of climate indices over all stations for SBM, VGLM and QR during the validation period (1991–2000).

	SBM	VGLM	QR
FR-TH (day)	2.6823	**2.6318**	2.6336
Tmax90 (°C)	1.4005	1.3832	**1.3732**
Tmin90 (°C)	1.6314	1.6032	**1.5432**
DTR (°C)	0.8809	0.8714	**0.8636**
GSL (day)	17.1859	16.6336	**16.2264**
FSL (day)	2.5577	2.4136	**2.3127**

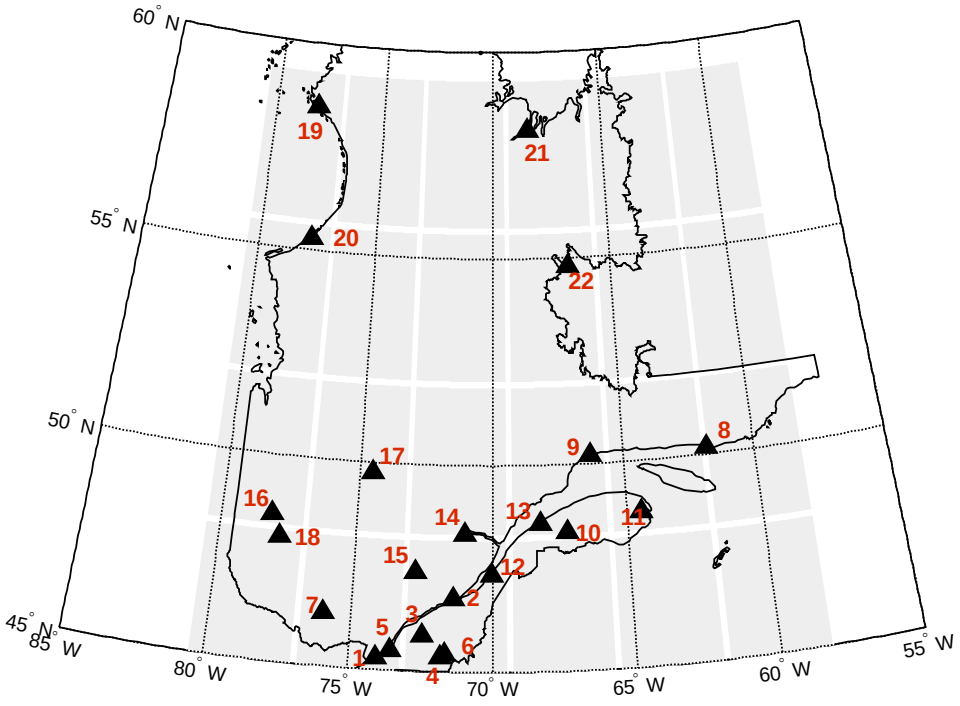

Figure 1. AOGCM grid points and meteorological stations selected for the study area : (1) Cesdres, (2) Quebec, (3) Drummondville, (4) Lennoxville, (5) McTavish (Montréal), (6) Sherbrooke A, (7) ManiwakiAirport, (8) Natashquan, (9) Sept-Iles A, (10) Causapscal, (11) Gaspe, (12) La Pocatière, (13) Mont-Joli A, (14) Bagotville A, (15) La Tuque, (16) Amos, (17) Chibougamau Chapais, (18) Val-D'Or A, (19) Inukjuak, (20) Kuujjuarapik, (21) Kuujjuaq, (22) Schefferville.

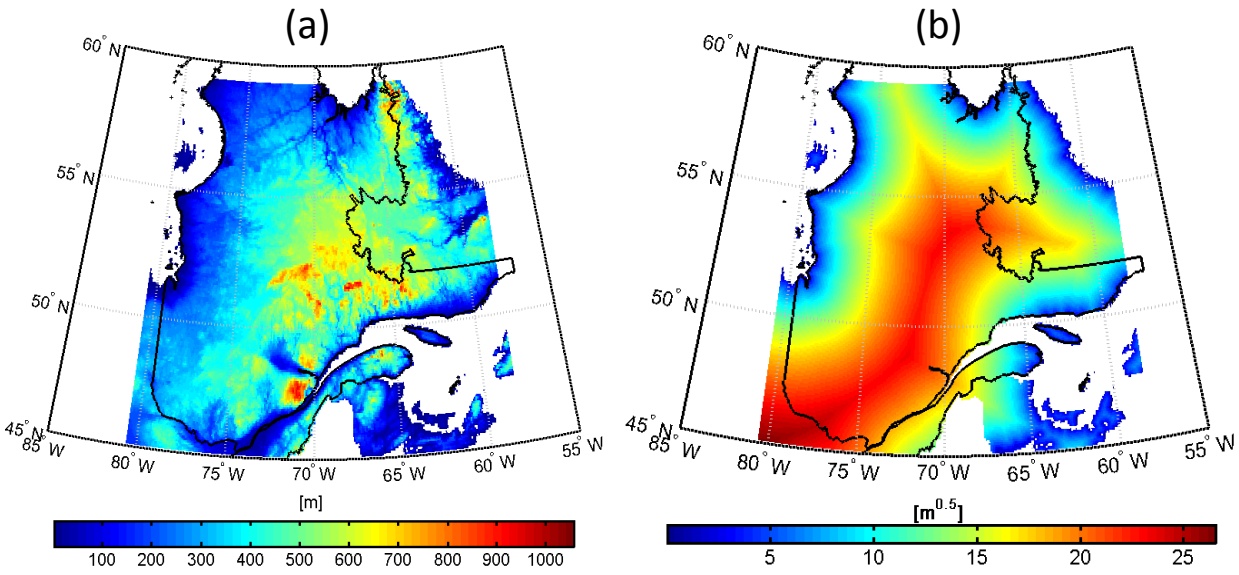

Figure 2. Numerical Terrain Model of the study area (a) and the square root of the distance from the cost (b).

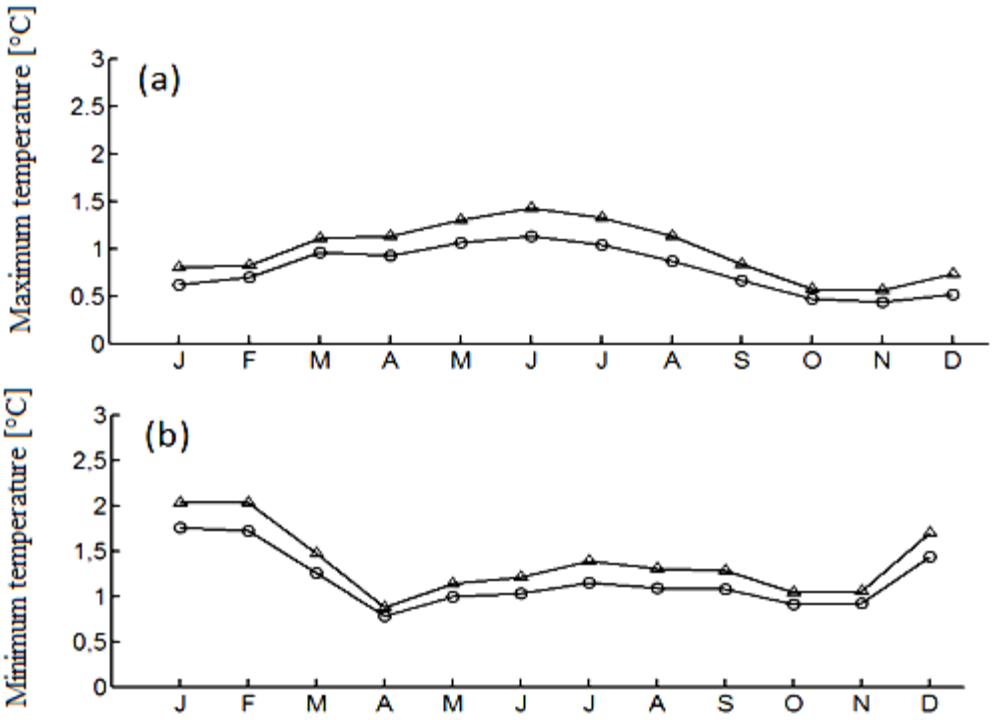

Figure 3. Visual comparison between the estimated standard deviation provided by the GRM (circles) and the observed RMSE computed using a leave-one-out cross validation (triangles) for (a) the maximum temperatures and (b) the minimum temperatures for each month.

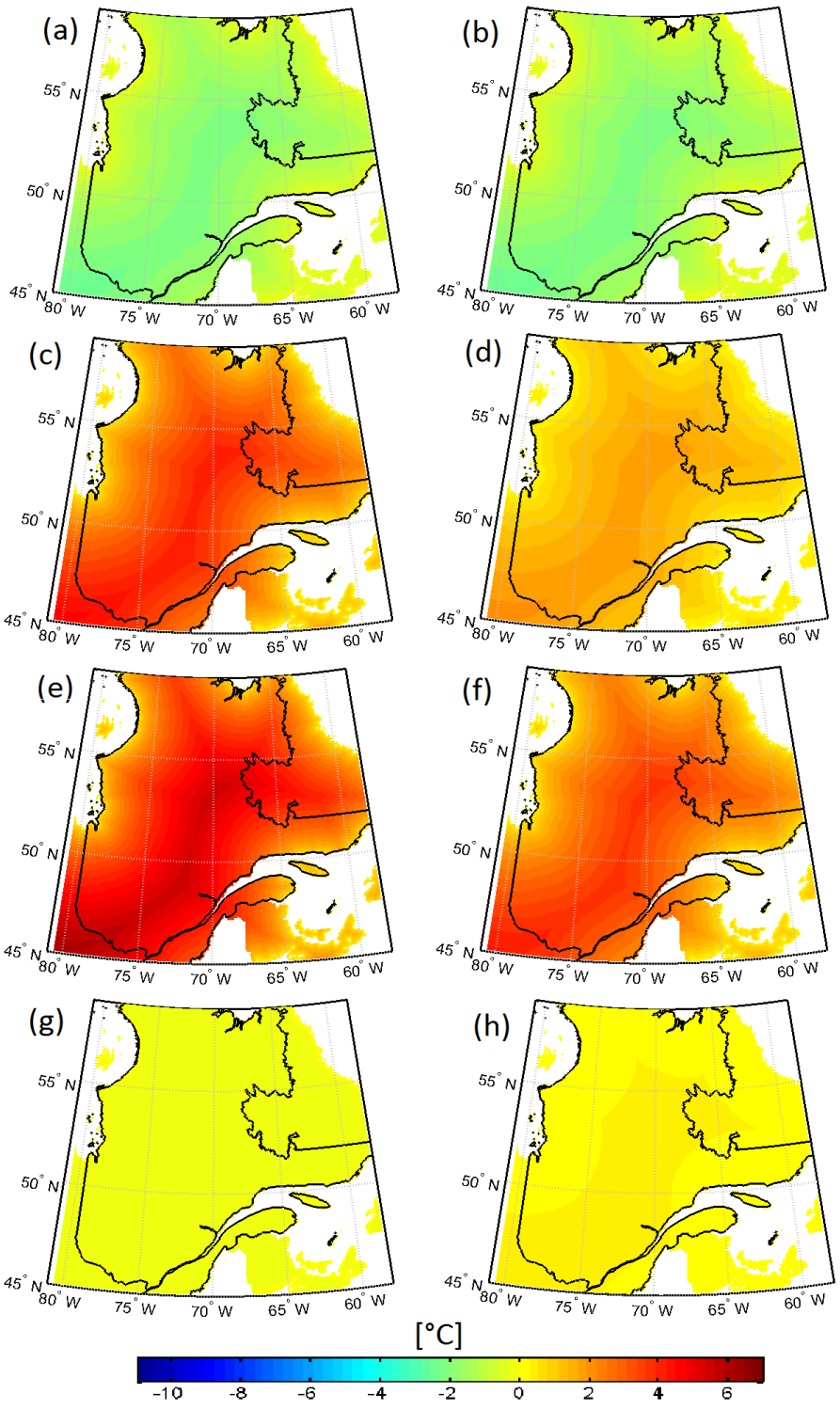

Figure 4. Effect of the square root of the distance from the cost on the prior mean of the maximum temperature (left column) and the minimum temperature (right column) in January (a and b), April (c and d), July (e and f) and October (g and h).

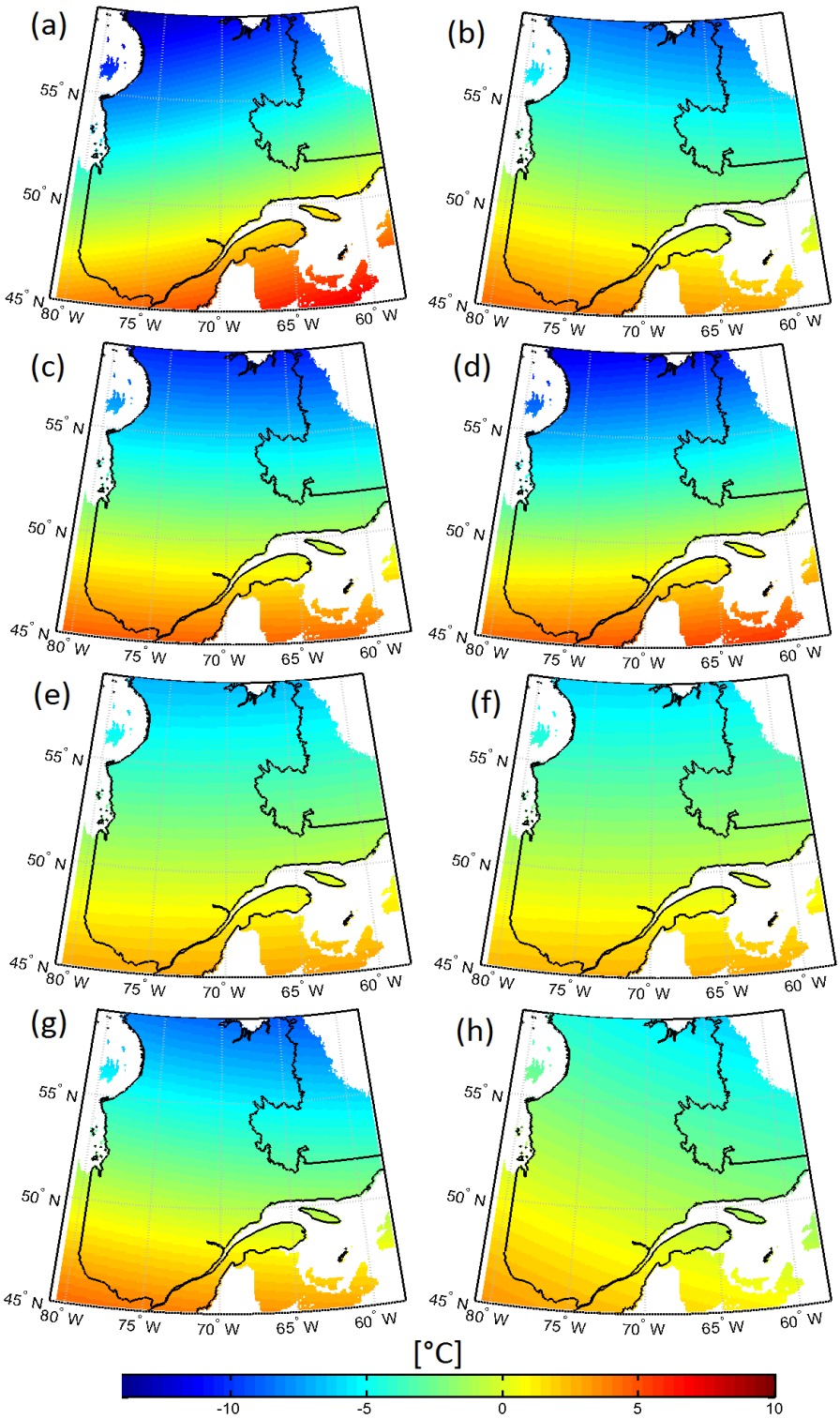

Figure 5. Effect of the geographical position on the prior mean of the maximum temperature (left column) and the minimum temperature (right column) in January (a and b), April (c and d), July (e and f) and October (g and h).

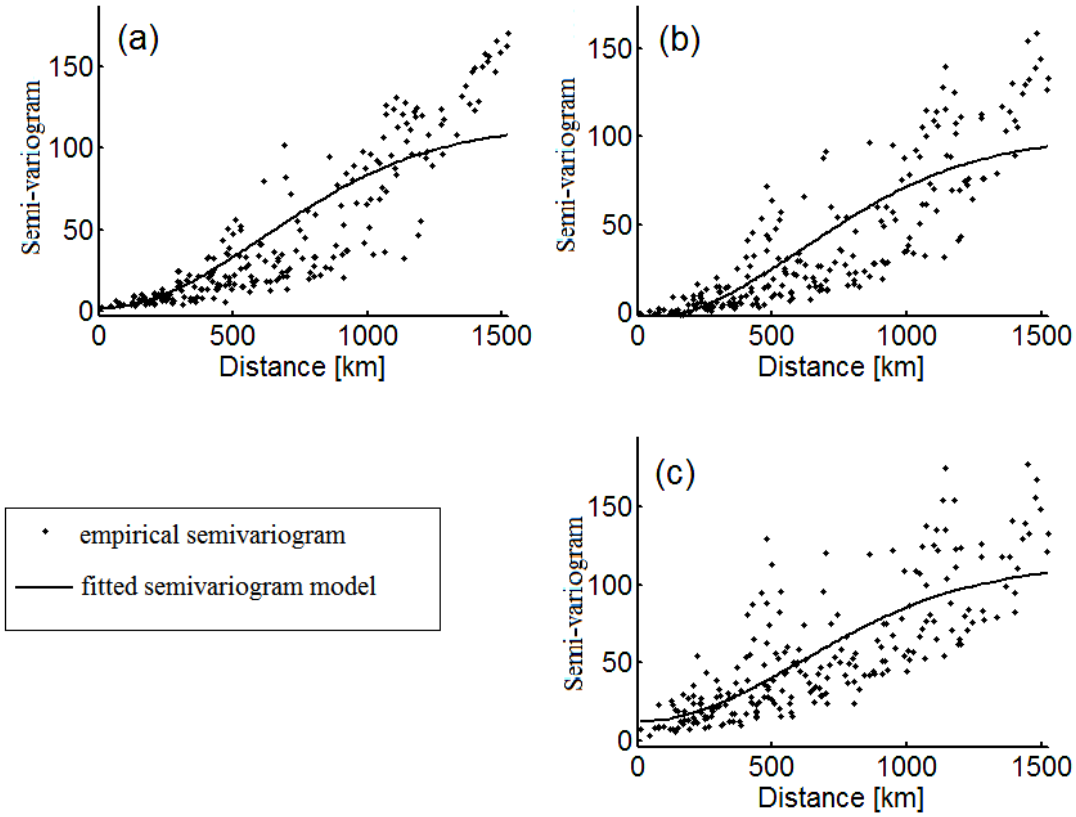

Figure 6. The fitted multivariate semivariogram model (bold lines) and the empirical semivariogram for the maximum temperature (a), the joint maximum–minimum temperatures (b), and the minimum temperatures (c).

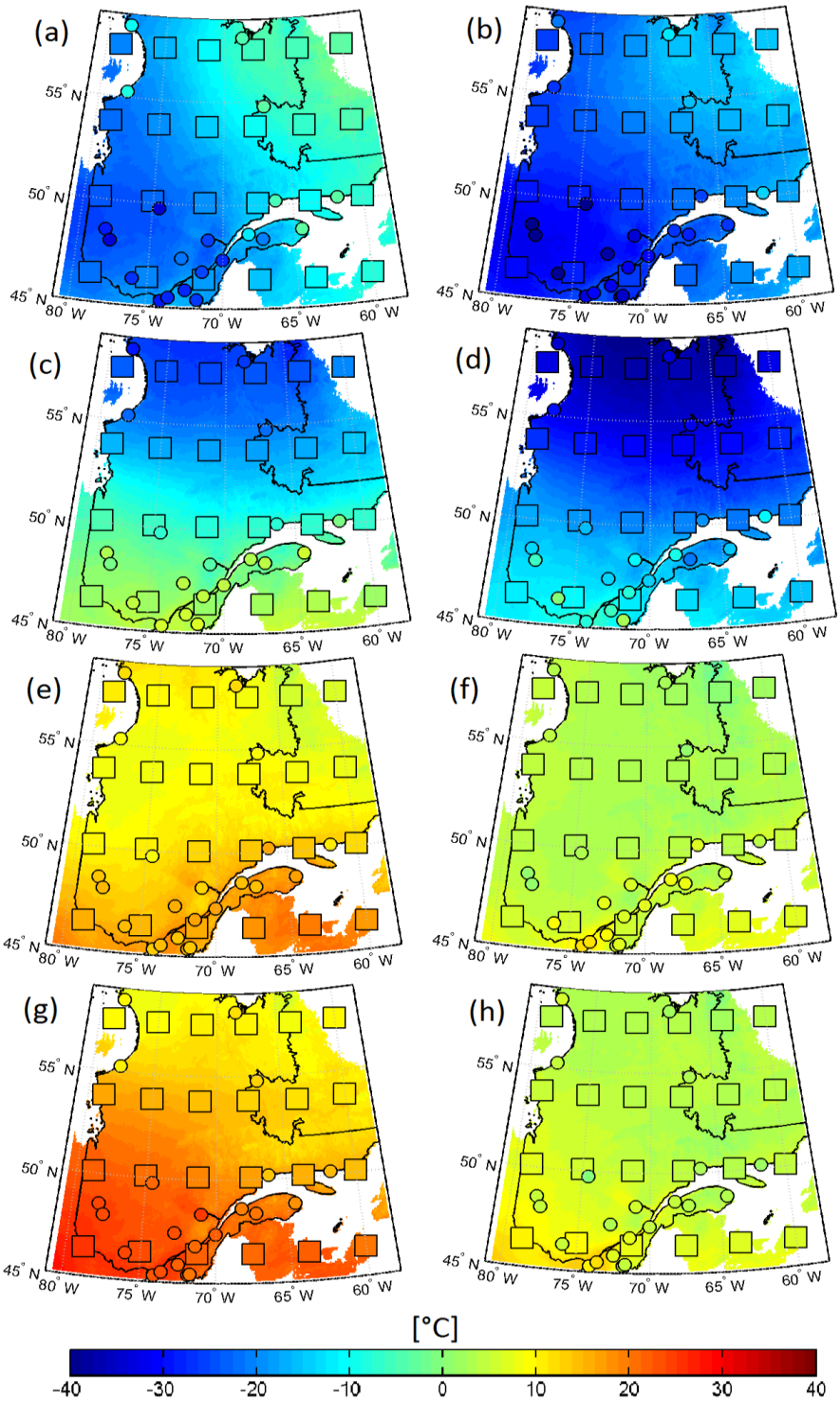

Figure 7. Examples of (left) maximum and (right) minimum temperatures estimation maps for 3 Jan 1981 (a,b); (c),(d) 17 Jan 1996; (e),(f) 24 Aug 1994; 24 Aug 1983. Circles represent the true observations and squares represent the upscaled temperatures on the AOGCM grid points.