Université du Québec
Institut National de la Recherche Scientifique
Centre Eau Terre Environnement

Approches statistiques pour la prédiction de variables hydrologiques extrêmes à des sites non jaugés en l'absence d'hypothèses de linéarité et normalité

Par
Martin Durocher

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

Jury d'évaluation

Président du jury et examinateur interne	Erwan Gloaguen Institut National de la Recherche Scientifique
Examinateur externe	Louis Mediero Université Polytechnique de Madrid
Examinateur externe	Jean-François Quessy Université du Québec à Trois-Rivières
Codirecteur de recherche	Taha B.M.J. Ouarda Institut Masdar de Sciences et Technologies
Directeur de recherche	Fateh Chebana Institut National de la Recherche Scientifique

REMERCIEMENTS

Ce document marque le point culminant de cette thèse de doctorat. Malgré tout le travail personnel, la réussite de ce troisième et dernier volet universitaire n'aurait pu se réaliser sans la disponibilité et la collaboration de plusieurs personnes que je désire ici remercier.

Je tiens en premier lieu à exprimer ma gratitude envers mon directeur, le Professeur Fateh Chebana, ainsi que mon codirecteur, le Professeur Taha B.M.J. Ouarda, pour leur soutien et leur confiance. Je les remercie d'avoir cru en moi depuis le premier jour et de m'avoir donné les outils pour réussir. Leurs conseils ont été précieux et inépuisables.

Je tiens également à exprimer ma reconnaissance à Mathieu Ribatet, Tae Sam Lee, Pierre Gosselin et Stephan Vida pour leurs collaborations sur les différents projets effectués durant le doctorat. Vous côtoyer fut un plaisir.

En général, je tiens à remercier les membres du groupe de recherche en statistique hydrologique à INRS pour leur coopération et leur aide. En particulier, je remercie Hussein Wazneh avec qui j'ai eu la chance d'échanger et de débattre de plusieurs idées.

Je tiens également à remercier mes parents qui ont toujours été derrière moi. Il est impossible pour moi d'énumérer tout ce que je leurs suis redevable.

Je voudrais finalement remercier le Conseil de Recherche en Science Naturelles et de Génie du Canada (CRSNG) d'avoir financé ma thèse.

RÉSUMÉ

L'analyse fréquentielle régionale (AFR) regroupe une panoplie de méthodes statistiques visant à prédire le comportement de variables hydrologiques extrêmes à des sites non jaugés. Les techniques de régression, les méthodes géostatistiques et la classification sont parmi les outils statistiques fréquemment rencontrés dans la littérature. Des méthodologies basées sur ces outils conduisent à des modèles régionaux qui offrent une description simple, mais très utile de la relation entre les variables hydrologiques extrêmes et les caractéristiques physiométéorologiques d'un site. Ces modèles régionaux permettent alors de prédire le comportement de variables d'intérêt à des endroits où aucune information hydrologique n'est disponible. Ces méthodes reposent généralement sur des hypothèses théoriques restrictives, dont la linéarité et la normalité. Ces dernières ne reflètent pas la réalité des phénomènes naturels. Les objectifs généraux de cette thèse sont d'identifier les méthodes affectées par ces hypothèses, évaluer leurs impacts et proposer des améliorations visant à obtenir des représentations plus réalistes et plus justes.

La régression à directions révélatrices (*projection pursuit regression*) est une méthode non paramétrique similaire aux modèles additifs généralisés et aux réseaux de neurones artificiels qui sont considérés en AFR afin de tenir compte de la non-linéarité des processus hydrologiques. Dans une étude comparative, cette thèse montre que la régression à directions révélatrices permet d'obtenir des modèles plus parcimonieux tout en préservant le même pouvoir prédictif que les autres méthodes non paramétriques.

L'analyse des corrélations canoniques (ACC) est employée afin de créer des voisinages à l'intérieur desquels un modèle (e.g. la régression multiple) sert à prédire les variables hydrologiques à des sites non jaugés. Par contre, l'ACC dépend fortement des hypothèses de

normalité et de linéarité. Une nouvelle méthodologie pour délimiter des voisinages est proposée dans cette thèse et utilise la régression à directions révélatrices afin de prédire un point de référence représentant l'information hydrologique et physiométéorologique qui est pertinente à ces regroupements. Les résultats montrent que la nouvelle méthodologie généralise celle de l'ACC, améliore l'homogénéité des voisinages et conduit à de meilleures performances.

En AFR, les techniques de krigeage sur des espaces transformés sont suggérées afin de prédire les variables hydrologiques extrêmes. Cependant, une transformation est requise afin que les variables hydrologiques d'intérêt proviennent approximativement d'une loi normale multidimensionnelle. Cette transformation introduit un biais et conduit à des prédictions sous-optimales. Des solutions ont été proposées, mais n'ont pas été testées en AFR. Cette thèse propose l'approche des copules spatiales et montre que cette approche apporte des solutions satisfaisantes aux problèmes rencontrés avec les techniques de krigeage.

Les processus max-stables sont une formalisation théorique des extrêmes spatiaux et correspondent à une représentation plus fidèle des processus hydrologiques. Par contre, leur caractérisation de la dépendance extrême pose des problèmes techniques qui ralentissent leur adoption. Dans cette thèse, le calcul bayésien approximatif est examiné comme une solution. Les résultats d'une étude de simulations montrent que le calcul bayésien approximatif est supérieur à l'approche standard de la vraisemblance composée. De plus, cette approche s'avère plus appropriée afin de tenir compte des erreurs de spécifications.

ABSTRACT

Regional Frequency Analysis (RFA) regroups several statistical methods in order to predict the behaviour of extreme hydrological variables at ungauged sites. Regression techniques, geostatistical methods and classification are among the most frequent tools found in the RFA literature. Methodologies based on these tools lead to simple, but very useful regional models that describe of the relation between extreme hydrological variables and physio-meteorological characteristics of sites. These regional models allow then to predict the behaviour of variables of interest to locations without actual hydrological information. Generally, these methods rely on restrictive hypothesis including: linearity and normality. The latter are not justified by natural phenomenon properties. Hence, the global objectives of this thesis are to identify these methods based on these hypotheses, to evaluate their impacts and to propose improvements with more accurate representation of the real phenomena.

Projection Pursuit Regression (PPR) is a nonparametric method similar to the generalized additive models and to artificial neural networks, whose have been already considered in RFA to account for the nonlinearity and the non-normality present in hydrological processes. This thesis includes a comparative study that shows that PPR leads to more parsimonious models in comparison to the other nonparametric approaches, without sacrificing predictive performances.

Canonical correlation analysis (CCA) is a technique that unveils the interrelation between two groups of random variables. In RFA, this technique is used to create neighborhoods inside which models (e.g. multiple regressions) can reasonably predict hydrological variables at ungauged locations. However, CCA strongly depends on the hypotheses of linearity and normality. A new methodology for delineating neighborhoods at

ungauged sites is proposed, where PPR is used to predict new neighborhood centers that represents relevant hydrological and physio-meteorological information necessary for pooling together gauged sites. Results show that the new methodology generalizes the CCA approach, improves homogenous properties and predictive performances.

Kriging techniques have been suggested in RFA to predict extreme hydrological variables in transformed spaces (not geographically continues). However, it requires the transformation of the hydrological variables of interest to meet normality requirements. This introduces a bias and produces suboptimal predictions for the kriging estimator. Solutions exist for dealing adequately with such transformations, but they have not been validated in RFA. This thesis investigates the spatial copula approach to this end and shows that the spatial copula framework provides a proper solution to the specific needs of RFA.

Max-stable processes are the formal generalization of the extreme value theory in the case of spatial extremes. They provides more accurate a characterization of the hydrological processes, but the theoretical formulation of the spatial dependence between extreme events remains challenging, which has slow down the adoption of max-stable processes in practical situations. In this thesis, Approximate Bayesian Computing (ABC) is investigated as a solution to these difficulties. A simulation study suggests that ABC is superior to the standard composite likelihood approach. In particular, ABC is found to be more appropriate to deal with model misspecifications.

ARTICLES ET CONTRIBUTIONS

[a] **Durocher, M.**, Chebana, F. Ouarda, T.B.M.J. (2015a) A nonlinear approach to regional flood frequency analysis using projection pursuit regression. Journal of Hydrometeorology (sous press), doi:10.1175/JHM-D-14-0227.1

[b] **Durocher, M.**, Chebana, F. Ouarda, T.B.M.J. (2015b) On the importance of hydrological information to form neighborhoods in regional frequency analysis. (Soumis).

[c] **Durocher, M.**, Chebana, F. Ouarda, T.B.M.J. (2015c) On the prediction of extremes flood quantiles at ungauged locations with spatial copula. (En révision).

[d] **Durocher, M.**, Ribatet, M., Ouarda, T.B.M.J. (2015d) Regional frequency analysis from approximate bayesian computing of max-stable processes. Soumis

L'article [a] évalue la pertinence et la performance de la méthode de régression à directions révélatrices pour la prédiction des quantiles de crues au Québec, Canada. L'idée originale est de F. Chebana. Le développement et l'écriture ont été effectués par M. Durocher. L'article a été révisé par F. Chebana et T.B.M.J. Ouarda.

Le article [b] propose l'utilisation de la régression à directions révélatrices pour déterminer le centre de voisinages délimité à partir de distances hydrologiques. L'idée originale est de M. Durocher. Le développement et la rédaction ont été effectués par M. Durocher. L'article a été révisé par F. Chebana et T.B.M.J. Ouarda.

Le article [c] suggère les copules spatiales afin d'apporter des correctifs aux techniques de krigeage des quantiles de crues. L'idée originale est de T.B.M.J. Ouarda. Le développement et la rédaction ont été effectués par M. Durocher. L'article a été révisé par F. Chebana et T.B.M.J. Ouarda.

Le article [d] utilise le calcul bayésien approximatif afin d'estimer les paramètres de processus max-stables. L'idée originale est de M. Ribatet. Le développement et la rédaction ont été effectués par M. Durocher. L'article a été révisé par T.B.M.J. Ouarda et M. Ribatet.

TABLE DES MATIÈRES

LISTE DES TABLEAUX DE LA SYNTHESE

LISTE DES FIGURES DE LA SYNTHESE

CHAPITRE 1: SYNTHESE

Ce chapitre a pour objectif de faire le lien entre les chapitres subséquents de cette thèse par articles. De plus, les principaux éléments méthodologiques y sont résumés et discutés afin d'offrir une perspective générale des contenues de cette thèse. Les chapitres suivants proposeront ensuite des solutions rigoureuses et détaillées aux problématiques évoquées dans cette synthèse.

1 INTRODUCTION

1.1 CONTEXTE

L'information provenant des stations hydriques et météorologiques est essentielle à l'évaluation des risques rattachés à l'occurrence d'événements hydrologiques extrêmes. Or, la majorité des endroits dans le monde ne sont pas jaugés puisque l'installation et le maintien de stations demandent temps et ressources. Il n'en demeure pas moins nécessaire d'obtenir des informations de qualité sur les processus hydrologiques à des sites non jaugés afin d'effectuer une gestion et une planification optimales des ressources hydriques (Davie, 2008, Shaw *et al.*, 2010).

Plusieurs méthodes sont ainsi proposées afin de transférer à un site cible (non jaugé) l'information recueillie à des sites jaugés. On parle ainsi de régionalisation ou de prédiction à des sites non jaugés. Ces prédictions sont basées sur la relation entre les variables hydrologiques étudiées et les caractéristiques physiométéorologiques disponibles. Un modèle régional exprime ainsi cette relation sous forme mathématique. Une fois calibré à partir des sites jaugés, le modèle régional permet de prédire le comportement d'une variable hydrologique à un site cible à partir de ses caractéristiques physiométéorologiques observées (Cunnane, 1988).

L'analyse fréquentielle régionale (AFR) désigne les méthodes de régionalisation visant la prédiction du risque que représente un événements extrêmes sur la base d'information régionale. En

AFR, le risque d'occurence est représenté sous la forme d'une période de retour qui exprime le temps moyen séparant l'occurrence de deux événements d'une certaine magnitude. Sous l'hypothèse qu'un dépassement survient de façon aléatoire, un seuil ayant une période de retour donnée correspond au quantile de la loi de probabilité de la variable hydrologique (Coles, 2001). Par exemple, on parlera en pratique d'un quantile de période de retour de 10 ou 100 ans et du quantile associé.

Au préalable, une analyse préliminaire est réalisée afin de calculer des variables hydrologiques qui sont décrites par le modèle régional. En AFR, on retrouve deux approches visant à prédire les quantiles de périodes de retour souhaitées. La première approche est la technique de prédiction des quantiles. Dans cette approche, l'analyse préliminaire extrait les quantiles de périodes de retour souhaités aux sites jaugés. Le modèle régional décrit ensuite la relation entre ces quantiles «locaux» et les caractéristiques physiométéorologiques. La deuxième approche est la technique de prédiction des paramètres de la distribution. Pour cette dernière, la loi de probabilités complète d'un site cible est régionalisée par la prédiction de ses paramètres ou celle de statistiques sommaires (e.g. moments ou L-moments). Les quantiles souhaités sont alors calculés à partir de la loi prédite.

Un concept fondamental en AFR est celui de régions homogènes qui consiste à regrouper des sites entre lesquels il est raisonnable de transférer l'information hydrologique. Les méthodes de régionalisation traditionnelles comportent ainsi deux étapes principales : la formation des régions homogènes et l'estimation d'un modèle régional pour ces régions. Toutefois, la subjectivité provenant de la formation des régions homogènes a été critiquée, poussant ainsi l'adoption de modèle régional applicable à l'ensemble des sites disponibles. Cette approche ouvre la porte ainsi à l'utilisation de méthodes qui permettent la prédiction des sites cibles en une seule étape (Laio *et al.*, 2011).

L'AFR a été très fortement étudiée dans le contexte des débits de rivières. Par exemple, une revue de littérature portant sur les méthodes de régionalisation utilisées dans le cadre des crues de rivières est fournie par Ouarda *et al.* (2008b). L'AFR est également fréquemment utilisée afin de

régionaliser d'autres variables hydrologiques comme les précipitations extrêmes (Alila, 1999, Gaál *et al.*, 2008, Renard, 2011, Wallis *et al.*, 2007).

1.2 PROBLÉMATIQUE

De nombreux outils statistiques sont utilisés en AFR. Parmi eux, on notera la régression multiple, l'analyse des corrélations canoniques (ACC), la technique du krigeage et les méthodes de classification. Dans bien des cas, le développement de ces outils suppose la linéarité ou la normalité afin de construire ses bases théoriques. Cependant, ces hypothèses ne sont pas toujours réalistes d'un point de vue pratique, puisque les processus hydrologiques sont naturellement non linéaires (Wittenberg, 1999). D'autre part, relaxer ces hypothèses entraîne des difficultés techniques qui nécessitent le développement d'approches alternatives qui soit adaptées à la réalité de l'AFR.

De façon générale, l'hypothèse de linéarité peut être relaxée par l'utilisation de méthodes non paramétriques, comme les réseaux de neurones artificiels (RNA) qui s'inspirent du fonctionnement du cerveau humain. Bien que flexibles, ces méthodes nécessitent une quantité suffisante d'information pour discerner convenablement un modèle régional sans supposer de forme analytique prédéfinie. Des exemples d'utilisation des RNA en AFR sont donnés par Shu *et al.* (2004) et Dawson *et al.* (2006) avec respectivement 404 et 850 sites. Des jeux de données aussi larges ne sont toutefois pas la norme en AFR. De plus, la complexité de ces structures motive la recherche de nouvelles solutions qui soient plus généralement applicables et dont l'interprétation permet une meilleure illustration des processus hydrologiques (Solomatine *et al.*, 2003). Il est alors nécessaire de développer des méthodes non paramétriques tenant compte de la non-linéarité des processus sans pour autant sacrifier sa compréhension.

En AFR on distingue deux types de régions homogènes, les régions fixes et les régions de type voisinages (Ouarda *et al.*, 2001). Un voisinage est défini comme une région centrée autour d'un site cible qui inclut les sites jaugés les plus près, en un certain sens, de ce point de référence.

Lorsque non jaugé, le site cible ne possède aucune information hydrologique et donc le voisinage ne peut être délimité sur la base de ses propriétés hydrologiques. Par conséquent, une approche fréquente est celle des régions d'influences, ou *regions of influence* (ROI), qui utilise la distance entre les caractéristiques physiométéorologiques des sites afin de former les voisinages (Acreman *et al.*, 1986, Burn, 1990). Toutefois la similarité entre les caractéristiques physiométéorologiques ne garantit pas la similarité des propriétés hydrologiques (Oudin *et al.*, 2010). Une autre approche est celle de l'ACC qui utilise des estimations préliminaires des quantiles souhaité afin de déterminer le centre d'un voisinage autour d'un site non jaugé (Ouarda et al. 2001). Par contre, l'ACC est une méthode linéaire qui est incapable de saisir la non-linéarité dans la relation entre les variables hydrologiques et les caractéristiques physiométéorologiques (He *et al.*, 2011). Ceci affecte la qualité de l'estimation servant à déterminer les voisinages. Des méthodes alternatives permettant d'obtenir plus efficacement ces estimations doivent être étudiées et une plus grande attention doit être portée sur la nature des variables de référence qui représente les centres des voisinages (Ouali *et al.*, 2015).

Chokmani *et al.* (2004) a montré que les méthodes géostatistique peuvent être utilisées afin de prédire les quantiles de crues à l'intérieur d'espaces construits à partir de caractéristiques physiométéorologiques. À ce titre, la technique du krigeage est l'une des meilleures méthodes géostatistiques utilisées en AFR (Castiglioni *et al.*, 2009). En effet, cette technique offre un prédicteur optimal et sans biais lorsque les variables hydrologiques sont distribuées selon une loi normale multidimensionnelle (Schabenberger *et al.*, 2004). Toutefois, les quantiles de crue sont souvent modélisés selon un modèle de puissance (Pandey *et al.*, 1999), ce qui invalide l'assomption de normalité et requière une transformation des variables hydrologiques. Des méthodes intégrant adéquatement les transformations aux techniques de krigeage existent dans la littérature géostatistique (Diggle *et al.*, 1998, Kazianka *et al.*, 2010), mais n'ont pourtant pas été considérées à ce jour en AFR. De telles méthodes devraient alors être validées dans le cadre de l'AFR.

Les méthodes traditionnelles en AFR reconnaissent deux sources d'erreurs: l'erreur due à l'échantillonnage et l'erreur due à l'estimation du modèle régional (Stedinger *et al.*, 1985). L'erreur due

à l'échantillonnage provient de l'utilisation d'une analyse préliminaire dans le but d'obtenir les variables hydrologiques étudiées. Ces erreurs ont des niveaux d'incertitudes inégales puisque la variabilité et le nombre d'événements extrêmes peuvent varier d'un site à l'autre. Les modèles probabilistes représentent une solution qui permet de tenir compte de ces deux sources d'erreur en plus de traiter les extrêmes hydrologiques comme des observations sans passer par le calcul de variables hydrologiques aux sites jaugés.

Les précipitations extrêmes sont des phénomènes météorologiques qui surviennent sur de grandes échelles et couvrent de larges territoires. Par conséquent, un même événement extrême peut provoquer des valeurs extrêmes à plusieurs sites. En pratique, certains modèles probabiliste vont présumer que les valeurs extrêmes sont indépendantes conditionnellement aux lois des sites (Cooley *et al.*, 2006) ou que la dépendance entre elles est décrite par une loi normale multidimensionnelle (Sang *et al.*, 2010). Cependant ces structures ne sont pas adaptées au contexte des extrêmes spatiaux (Davison *et al.*, 2012). Théoriquement, la structure de dépendance entre extrêmes spatiaux correspond à celle de processus max-stables (Haan, 1984, Schlather, 2002). Malgré tout, l'adoption des processus max-stables a tardé à s'établir en AFR en raison des difficultés techniques reliées à la formulation des probabilités jointes (Padoan *et al.*, 2010). En particulier, la fonction de vraisemblance qui est fondamentale dans les techniques d'estimation classiques n'a pas de forme analytique pratique dans les procusses max-stables. Des efforts se doivent d'être accordés au développement de méthodes d'estimation alternatives pour les processus max-stables afin de favoriser leur utilisation en AFR étant donné le spatial en régionalisation.

1.3 OBJECTIFS ET RÉALISATIONS

Les objectifs généraux de cette thèse sont donc :

i. Identifier les difficultés techniques provenant des hypothèses de la non-linéarité ou de la non-normalité dans les approches statistiques utilisés en AFR ;

ii. Proposer et développer de nouvelles méthodes qui tiennent en compte la non-linéarité et de la non-normalité en AFR ;

iii. Évaluer le rendement des nouvelles méthodes proposées en ce qui concerne la performance de prédictions et la qualité de la représentation des processus hydrologiques.

Afin de répondre aux objectifs de cette thèse, quatre articles sont rédigés afin d'apporter une solution aux problématiques qui découlent de ces objectifs généraux et des difficultés techniques provenant de la remise en question d'hypothèses jugées non réalistes. Chaque article s'attaque à une sous-problématique qui est rattachée à un groupe de méthodes précis. Dans l'ordre, les quatre articles abordent les méthodes non paramétriques, les méthodes de formation des régions homogènes de type voisinage, les techniques de krigeages et les processus max-stables. Tous les développements proposés à l'intérieur de ces articles sont appliqués à des cas d'étude dans le but de valider les méthodes étudiées et de les comparer à celles existantes. Cette approche permet de répondre systématiquement aux objectifs généraux de cette thèse pour chaque groupe de méthodes.

Un survol des méthodes rencontrées en AFR est illustré à la Figure 1 et seront discutées dans la section suivante qui fait la revue de littérature de méthodes employées en AFR. Dans cette figure, les méthodes ciblées dans cette thèse sont indiquées en rouges. Notons que les modèles probabilistes sont classés comme des méthodes distinctes puisqu'elles ne nécessitent pas d'analyse préliminaire aux sites comme le font les méthodes traditionnelles en AFR. Les modèles probabilistes sont par contre directement reliés aux techniques de prédiction des paramètres puisque les lois des sites y sont spécifiées et servent à prédire le quantile de période de retour souhaité. Les méthodes traditionnelles en AFR sont classées selon qu'elles utilisent ou non la formation de régions homogènes. Dans le cas « avec régions homogènes », les méthodes indiquées sont des méthodes de formation des régions homogènes à l'intérieur desquelles la régression multiple permet la prédiction des variables hydrologiques. Pour le cas « sans régions homogènes », les méthodes évoquées servent à prédire directement les variables hydrologiques à des sites non jaugés.

Durocher *et al.* (2015a) propose les méthodes de régression à directions révélatrices (RDR) (Friedman *et al.*, 1974) comme technique de prédiction des quantiles afin d'obtenir des équations de régression non linéaires, simples et explicites. Cette méthode n'a jamais été utilisée en AFR et est présentée comme une alternative plus parcimonieuse que les autres méthodes non paramétriques, en particulier les RNA, sans pour autant sacrifier les performances prédictives. La méthode RDR est appliquée aux données de crues du Québéc et est comparée aux autres méthodes AFR déjà utilisées sur le même jeu de données.

La méthode RDR est également considérée afin d'améliorer la formation des régions homogènes de type voisinages dans Durocher *et al.* (2015b). Une nouvelle méthodologie basée sur la généralisation de l'approche de l'ACC est développée et vise à améliorer l'estimation du centre d'un voisinage basé sur des variables de références hydrologiques. L'utilisation de RDR relaxe les hypothèses de normalité et de linéarité assumées par l'ACC. La qualité de cette procédure est validée à partir d'un jeu de données portant sur les crues annuelles du Québec. Cette étude vise de plus à examiner l'impact du choix de différentes variables hydrologiques sur la formation des voisinages. En particulier, les L-moments y sont considérés afin d'améliorer les propriétés d'homogenéités.

Durocher *et al.* (2015c) propose le cadre des copules spatiales afin d'améliorer les prévisions sous-optimales provoquées par la transformation des variables hydrologiques dans les techniques de krigeage. Ce nouveau cadre de travail est plus général et permet en particulier de modéliser l'hétéroscédasticité dans les variables hydrologiques. La comparaison de cette méthode géostatistique avec les méthodes géostatistiques traditionnelles a fait l'objet d'une étude appliquée au jeu de données des crues annuelles du Québec. Ce choix permet de comparer la méthode des copules spatiale avec les autres méthodes de krigeage déjà utilisées sur les mêmes données.

En raison des difficultés numériques reliées à l'estimation des processus max-stables, Durocher *et al.* (2015d) considère le calcul bayésien approximatif, ou *approximate bayesian computing* (ABC) afin d'estimer les paramètres de modèles hiérarchiques bayésiens. Cette méthode a

été proposée par Erhardt et Smith (2012), mais se concentrait exclusivement sur l'estimation de la dépendance. Pour sa part, l'approche proposée par Durocher *et al.* (2015d) considère une modélisation jointe de la dépendance spatiale extrême et des lois marginales aux sites dans le but d'obtenir un modèle bayésien hiérarchique complet. L'approche ABC requiert le choix de statistiques sommaires afin de calibrer l'algorithme d'échantillonnage de la loi a postériori. Ces choix sont validés lors d'études de simulations. De plus, un cas d'étude est réalisé sur des données de précipitations extrêmes pour une petite région de la Californie (États-Unis). Les résultats de la méthode ABC sont comparés à celle de la vraisemblance composée qui est la méthode la plus souvent utilisée pour l'estimation des processus max-stables.

1.4 ORGANISATION DU DOCUMENT

Le reste du chapitre de synthèse est organisé de la façon suivante. Dans la section 2, une revue de littérature portant sur les approches statistiques utilisés en AFR est présentée. Cette revue est organisée suivant la classification des méthodes présentées à la Figure 1. Dans la section 3, les résultats obtenus à l'intérieur des articles font l'objet de remarques et de discussions. Finalement, la section 4 apporte une conclusion et des pistes de recherche en AFR lorsque la non-linéarité et la non-normalité sont considérées.

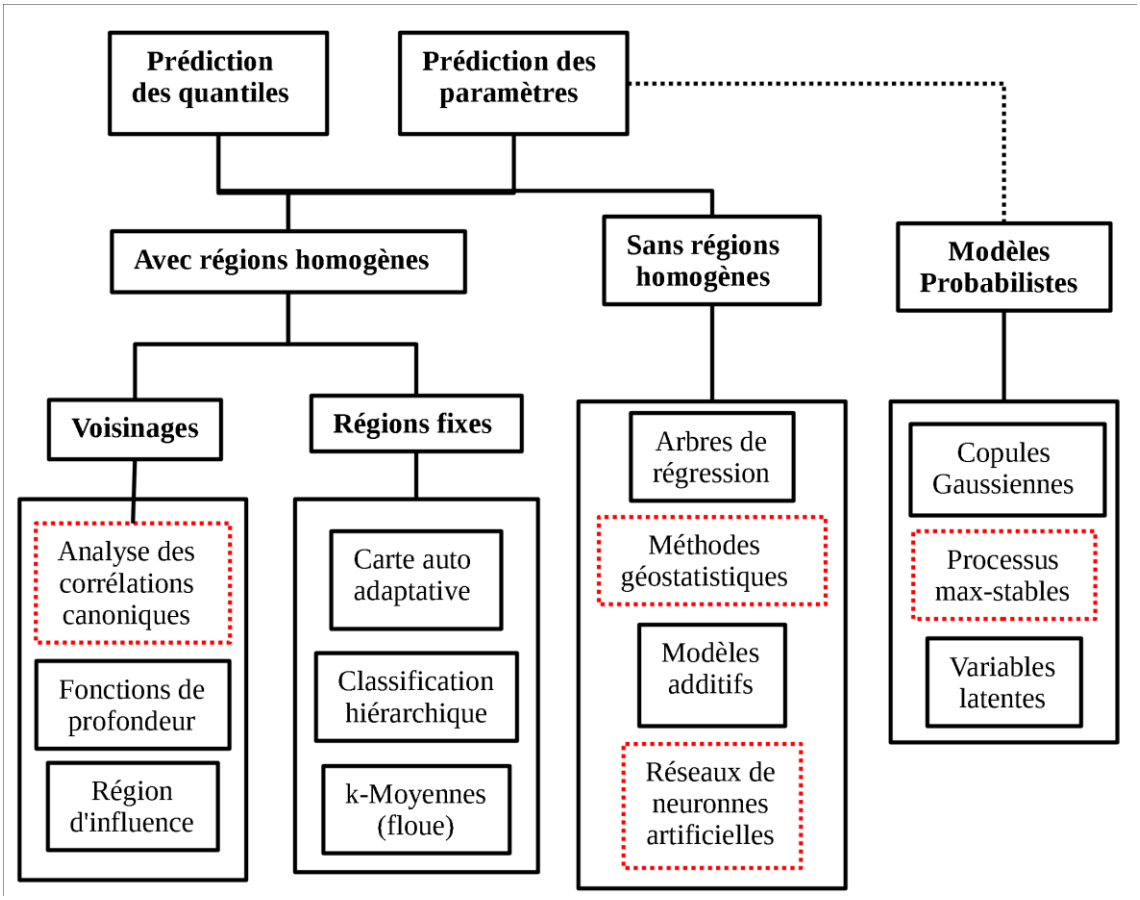

Figure 1: Classification des méthodes présentées dans la revue de littérature. Les méthodes ciblées dans cette thèse sont en rouges.

2 REVUE DE LITTÉRATURE

2.1 MODÈLES RÉGIONAUX AVEC RÉGIONS HOMOGÈNES

Afin de décrire la distribution des pointes de crues d'une rivière, Dalrymple (1960) suppose qu'à l'intérieur d'une région homogène les lois de probabilité de sites $i = 1, 2, \ldots$ sont proportionnelles à une même loi régionale ayant une fonction de répartition, ou courbe de croissance $q(r)$ et où la valeur r désigne une période de retour. Les observations aux sites peuvent ainsi être standardisées par un facteur d'échelle μ_i appelé indice de crue. Les quantiles d'un site sont alors décrits par la relation suivante:

$$Q_i(r) = \mu_i q(r) \qquad (1)$$

Régionaliser le quantile d'une loi consiste alors à établir la courbe de croissance d'une régionale homogène et à prédire l'indice de crue d'un site cible.

La théorie des valeurs extrêmes démontre que les séries de maximums annuelles vont suivre approximativement une loi des valeurs extrêmes, tandis que les valeurs au-dessus d'un seuil suffisamment élevé (série de durée partielle) suivent approximativement une loi de Pareto généralisée (Coles, 2001). Malgré ces approximations théoriques, des considérations pratiques font en sorte que plusieurs autres lois de probabilité sont considérées en AFR. Entre autres, on notera parmi les lois les plus fréquentes: la loi log-normale, la loi de Pearson type III et la loi logistique généralisée (Hosking *et al.*, 1997). L'indice de crue μ_i est généralement la moyenne ou la médiane des maximums annuels (Katz *et al.*, 2002, Kjeldsen *et al.*, 2007). Toutefois, d'autres variables hydrologiques peuvent servir d'indice de crue. Par exemple, le paramètre de dispersion d'une loi de Pareto généralisée a servi dans une étude où les séries de durée partielle ont été préférées aux séries de maximums annuels (J. A. Smith, 1989).

Les hypothèses du modèle d'indice de crue sont équivalentes à considérer constants le coefficient de variation et les moments d'ordre trois ou plus (Gupta *et al.*, 1994). Il est ainsi possible de faire la moyenne des moments empiriques des sites d'une région homogène afin de déterminer la forme de la loi régionale. De la même façon, les moments classiques peuvent être remplacés par les L-moments (Hosking, 1990) qui sont eux-mêmes des statistiques sommaires déduites des moments de probabilité pondérée (Landwehr *et al.*, 1979). La forte présence de ces dernières en AFR est attribuable à leur bonne performance dans le but d'estimer les lois de variables hydrologiques lorsque peu d'observations sont disponibles (Hosking *et al.*, 1985, Madsen *et al.*, 1997).

Les hypothèses du modèle d'indice de crue sont toutefois restrictives et limitent les formes que peuvent prendre les lois de probabilité à des sites. De plus, ces hypothèses ne sont pas toujours vérifiées en pratique (Robinson *et al.*, 1997) et ne reposent sur aucune justification physique (Katz *et al.*, 2002). On notera que pour ce modèle, la relation entre les caractéristiques physiométéorologiques et les variables hydrologiques intervient uniquement au niveau de l'indice de crue. Une approche hiérarchique de formation des régions homogènes a permis à Gabriele *et al.* (1991) de proposer un modèle d'indice de crues avec des moments d'ordre supérieur à trois constants, mais qui acceptent une certaine variation du coefficient de variation. Cette approche est motivée par l'observation d'un lien entre l'aire d'un basin versant et le coefficient de variation. Les auteurs ont reconnu que cette approche présente des avantages prédictifs, mais comporte toutefois des difficultés lorsqu'il est nécessaire de tester l'homogénéité des régions.

Une approche plus flexible que l'indice de crue est celle proposée par Laio *et al.* (2011) qui a utilisé un modèle de régression séparé pour prédire chaque L-moments. Ils admettent ainsi que les L-moments puissent varier séparément, selon les caractéristiques physiométéorologiques. Le choix de la loi de probabilité demeure une question délicate puisque plusieurs lois peuvent ajuster similairement les observations. Pour leur part, Laio *et al.* (2011) ont prédit les quantiles souhaités en considérant simultanément plusieurs lois de probabilités. Des expériences Mont-Carlo sont ainsi utilisées afin d'approximer la distribution des quantiles aux sites cibles. Notons que si une unique

famille de lois de probabilité est sélectionnée pour l'ensemble des sites, les lois aux sites peuvent être régionalisées directement par l'entremise des paramètres. Cette approche représente une technique de prédiction des paramètres qui a été considérée par Haddad *et al.* (2012) dans le cas d'une loi log-Pearson III.

Katz *et al.* (2002) ont noté que la régression dans le domaine hydrologique s'appuie presque exclusivement sur la théorie des moindres carrés, au détriment de la théorie du maximum de vraisemblance, pourtant fondamentale en statistique. Ils attribuent cette prépondérance à l'adoption massive des L-moments. L'emploi de la fonction de vraisemblance comme moyen d'estimation pour les techniques de prédiction des paramètres conduits naturellement à des modèles probabilistes où une structure complète des lois marginales et des probabilités jointes sont spécifiées pour l'ensemble des sites. Ce lien entre les méthodes de prédiction des paramètres et les modèles probabilistes est présenté à la Figure 1 par une ligne hachurée.

L'alternative aux techniques de prédiction des paramètres et au modèle d'indice de crue est celle des techniques de prédiction des quantiles où la variable hydrologique prédite est directement le quantile de période de retour souhaitée (Pandey *et al.*, 1999). Dans cette approche, une période de retour est calculée à chaque site jaugé avant d'être régionalisée. Relativement peu d'études ont comparé la performance entre les techniques de prédiction des quantiles et les techniques de prédiction des paramètres, mais celles disponibles montrent que les deux approches offrent généralement des performances similaires (GREHYS, 1996a, GREHYS, 1996b, Haddad *et al.*, 2012). Par contre, Haddad *et al.* (2012) soutiennent que la technique de prédiction des paramètres a certain avantage:

i. Toutes les périodes de retour sont disponibles et sont assurées d'une croissance graduellement cohérente ;

ii. La mise en commun de l'information régionale avec celles des sites est directe, comme présentée par Micevski et Kuczera (2009).

Une particularité des techniques de régression en AFR est que la variabilité des périodes de retour prédites possède deux sources d'erreurs (Gary D. Tasker, 1980). D'une part, il y a l'erreur η_i qui est due à l'échantillonnage pour un site i. Cette erreur provient du fait que les variables hydrologiques régionalisées sont des statistiques calculées à partir de données brutes, et non des mesures directes. D'autre part, il y a l'erreur δ_i qui découle de l'incertitude sur l'estimation du modèle régional. Une hypothèse souvent rencontrée est l'indépendance entre les sites. Notons que comme les phénomènes météorologiques extrêmes couvrent de vastes territoires, ils affectent simultanément plusieurs sites rapprochés (Coles, 1993). D'un point de vue hydrologique, assumer l'indépendance entre les sites est une hypothèse forte qui risque de ne pas être satisfaite en pratique. Par contre, l'indépendance entre les sites peut être imposée dans la construction d'un jeu de données en incluant uniquement des sites relativement éloignés et pour lesquelles l'hypothèse d'indépendance peut être testée positivement (Hosking *et al.*, 1997). Cette dernière stratégie risque toutefois d'exclure volontairement de l'information utile pouvant améliorer l'estimation du modèle régionale.

Afin de tenir compte des sources d'incertitudes sur les variables hydrologiques et de la corrélation entre les sites, Stedinger *et al.* (1985) ont proposé de considérer l'erreur totale $\varepsilon_i = \eta_i + \delta_i$. En effet, si Σ_η est la matrice de covariance des erreurs η, I la matrice identité et que σ_δ^2 est la variance du modèle régionale, alors la matrice de covariance pour l'erreur totale proposée est:

$$E(\varepsilon'\varepsilon) = \sigma_\delta^2 I + \Sigma_\eta \qquad (2)$$

Il a été démontré que la méthode des moindres carrées généralisées qui utilise directement (2) afin d'estimer les paramètres de modèles de régression multiples est plus efficace que la méthode des moindres carrés ordinaires (Kroll *et al.*, 1998, Stedinger *et al.*, 1985). En pratique, des formules analytiques approximatives pour Σ_η ont été développées afin de spécifier la covariance de chaque paire de sites (Griffis *et al.*, 2007, Hamed *et al.*, 1999). Ces formules dépendent alors de la nature des lois aux sites, de l'incertitude sur l'estimation des variables hydrologiques et de la corrélation croisée

entre les sites. Néamoins la variance σ_δ^2 demeure inconnue et doit être estimée. Une procédure itérative peut alors être utilisée (G. Tasker *et al.*, 1989). De manière alternative, Reis *et al.* (2005) ont proposé un cadre bayésien afin d'obtenir la loi a posteriori de tous les paramètres du modèle régional, incluant les éléments inconnus de la matrice de covariance.

Par ailleurs, la structure du terme d'erreur en (2) assume que le modèle régional suffit à expliquer la relation entre les variables hydrologiques et les caractéristiques physiométéorologiques, puisque les erreurs δ_i sont indépendantes. Toutefois, une corrélation résiduelle peut persister entre les sites en raison d'un manque d'information ou d'un modèle régional incomplet. Dans ce cas, la partie $\sigma_\delta^2 I$ de la matrice de covariance en (2) n'est pas adéquate. Kjeldsen *et al.* (2009) ont ainsi proposé des modèles plus flexibles pour la matrice de covariances et où la corrélation résiduelle est intégrée.

2.2 MÉTHODES DE FORMATION DES RÉGIONS HOMOGÈNES

Les méthodes de régression présentée à la sous-section 2.1 sont effectuées à l'intérieur de régions homogènes, pour lesquelles il existe différentes méthodes de formation. On distingue deux types de régions homogènes; les régions fixes et les voisinages. Une région fixe forme une partition de l'ensemble des sites jaugés disponible et représente donc des sous-ensembles à l'intérieur desquels un site appartient à une unique classe. De l'autre côté, les voisinages sont des régions centrés autour d'un site cible et où un même site jaugé peut appartenir à plusieurs voisinages.

Avec les régions fixes, des méthodes traditionnelles de formation des régions homogènes, ou classification, sont la méthode des résidus (Bhaskar *et al.*, 1989, Laaha *et al.*, 2006), la méthode des k-moyennes (Burn *et al.*, 2000, Wiltshire, 1985) et les méthodes de classification ascendante hiérarchique (Mosley, 1981, Nathan *et al.*, 1990). La difficulté avec ces méthodes est qu'elles proposent un partitionnement basé sur des séparations linéaires. Cette restriction sur la forme des régions homogènes n'est pas appropriée à la complexité des phénomènes étudiés en hydrologie.

Idéalement, les méthodes de classification en AFR devraient permettre des partitions de formes diverses. Pour répondre à ce besoin, la logique floue et les méthodes de classification non linéaire sont nécessaires (Hall *et al.*, 1999).

Contrairement à la logique classique, qui assume qu'un site appartient à une seule classe, la logique floue suppose qu'un site peut appartenir simultanément à plusieurs classes, mais à différents degrés (Rao *et al.*, 2006). Cette approche est plus réaliste dans le contexte de l'AFR puisqu'il est facile d'imaginer qu'un site jaugé à la frontière de deux régions homogènes possède des propriétés propres aux deux régions. La notion de logique floue a été utilisée en combinaison avec la méthode des k-moyennes. Celle-ci a montré sa supériorité sur les méthodes classiques des k-moyennes et les méthodes de classification ascendante hiérarchique (Jingyi *et al.*, 2004, Rao *et al.*, 2006). Malgré l'utilisation de la logique floue, les méthodes des k-moyennes produisent toujours des partitions basées sur des séparations linéaires. En AFR, Basu *et al.* (2014) ont ainsi utilisé une approche par noyaux afin d'obtenir des délimitations non linéaires qui permettent d'améliorer les performances de la méthode des k-moyennes avec logique floue

Les cartes auto-adaptatives sont un cas particulier de RNA qui on été utilisées pour la délimitation de régions homogènes (Hall *et al.*, 1999). Ce type de RNA permet des séparations non linéaires entre les régions. Des études comparatives ont montré que ces méthodes conduisent à des performances supérieures par rapport aux méthodes traditionnelles de classification en AFR (Jingyi *et al.*, 2004, Lin *et al.*, 2006). Par ailleurs, Shu *et al.* (2008) ont proposé un système adaptatif neuroflou d'inférence, qui combine les RNA avec la logique floue dans le but d'intégrer en une seule étape la formation des régions homogènes et la régression des variables hydrologiques. Leur travail montre que l'intégration de ces deux étapes est bénéfique, puisque leur méthode conduit à une meilleure capacité de généralisation, en plus de fournir un mécanisme automatique réalisant simultanément toutes les étapes de l'AFR.

De l'autre côté, une région de type voisinage est centrée autour d'un site cible. De cette façon les voisinages ne sont pas affectés par les limitations causées par des séparations linéaires. Un certain nombre d'études comparatives ont été conduites afin de comparer la performance entre les approches utilisant des régions fixes et les voisinages (GREHYS, 1996a, GREHYS, 1996b, Ouarda *et al.*, 2008a, G. Tasker *et al.*, 1996). Dans ces études, les méthodes de type voisinages se sont avérées plus efficaces. Toutefois, ces études comparatives ne prennent pas en considération l'utilisation des outils comme la logique floue et les cartes auto-adaptatives. À cet effet Basu *et al.* (2014) ont présenté un cas d'étude où l'utilisation de régions fixes a été supérieure à celle de voisinages.

Les distances géographiques et les distances physiographiques, qui sont respectivement basées sur la proximité entre les coordonnées d'un site et les caractéristiques physiométéorologiques, ont été considérées afin de délimiter des voisinages au tour d'un site cible. Dans le cas de débit de rivière, la distance géographique ne garantit pas une similitude hydrologique entre les sites puisque les propriétés physiographiques entre deux basins adjacents peuvent varier drastiquement et conduire à des comportements distincts (Chokmani *et al.*, 2004). Néamoins, certaines études se sont intéressées au choix entre ces deux notions de distance (Eng *et al.*, 2005, Merz *et al.*, 2005). Leurs résultats montrent que dans certaines situations la distance géographique fonctionne aussi bien, sinon mieux que la distance physiographique. Toutefois, de meilleures performances sont obtenues en combinant ces deux types de distances. Par ailleurs, Oudin *et al.* (2010) se sont intéressés aux propriétés hydrologiques des régions homogènes pour des débits de rivière délimitées à partir de distances physiographiques. Leur étude a montré que l'utilisation d'une distance basée sur les propriétés physiographiques ne garantit pas l'homogénéité hydrologique des régions délimitées.

Une méthode simple pour délimiter un voisinage consiste à utiliser les k-sites les plus proches (Haddad *et al.*, 2012, G. Tasker *et al.*, 1996). De façon alternative, il est possible d'attribuer un poids selon la distance séparant un site au site cible (Burn, 1990). Ces poids reflètent alors l'influence de chaque site jaugé et permettent d'attribuer un poids plus important aux sites les plus rapprochés.

Notons que des poids non nuls peuvent être affectés à chaque site, ce qui permet d'éviter les effets de bord occasionnés par l'exclusion de sites à la bordure des voisinages (Chebana *et al.*, 2008).

L'ACC est une méthode statistique qui permet d'étudier la corrélation entre deux groupes de vecteurs multidimensionnels. Ouarda *et al.* (2001) ont indiqué comment utiliser l'ACC en AFR afin de créer des voisinages dans des espaces transformés qui considèrent la distance entre des variables hydrologiques. Une des particularités de cette méthode est la prédiction des centres des voisinages puisque les véritables centres hydrologiques sont inconnus. Une autre approche utilisant des espaces transformés est celle des fonctions de profondeur (Chebana *et al.*, 2008). La notion de fonction de profondeur est un outil qui permet d'ordonner des observations dans un espace multidimensionnel (Tukey, 1975). Un exemple de procédure optimale et automatisée permettant de déterminer les poids et les points référence des voisinages a été présentée par Wazneh *et al.* (2013).

2.3 MODÈLES RÉGIONAUX SANS RÉGIONS HOMOGÈNES

L'approche de Laio *et al.* (2011) suppose que les variables hydrologiques vont varier continuellement selon les caractéristiques physiométéorologiques. Dans ce contexte, la régression multiple impose une relation linéaire qui est irréaliste puisque les processus naturels sont en pratique non linéaires (Wittenberg, 1999). En AFR, l'application des RNA n’est pas uniquement considérée lors de la formation des régions homogènes. Ceux-ci interviennent également à l'étape de prédiction des variables hydrologiques. On retrouve ainsi des adaptations des RNA aux modèles d'indice de crues (Dawson *et al.*, 2006) et aux techniques de prédiction des quantiles (Aziz *et al.*, 2014, Ouarda *et al.*, 2009). Les RNA sont suffisamment flexibles pour approximer toutes formes de relations entre les variables hydrologiques et les caractéristiques physiométéorologiques, à condition d'avoir suffisamment d'informations pour l'identifier à partir des données disponibles (Bishop, 1995). De plus, notons que des études comparatives ont montré que la restriction d'un RNA à des régions plus homogènes est généralement contre-productive (Aziz *et al.*, 2014, Dawson *et al.*, 2006).

La calibration des RNA pose plusieurs problèmes en raison des risques de surajustement et des difficultés à identifier une estimation globalement optimale (Hastie *et al.*, 2009). Ces difficultés sont généralement réglées par l'utilisation de méthodes d'agrégation qui prédisent une variable hydrologique comme une somme pondérée des résultats de plusieurs RNA. Les méthodes d'agrégations s'avèrent plus efficaces que celles des RNA considérées individuellement. Une comparaison des principales méthodes d'agrégation a été réalisée par Shu *et al.* (2004) et a montré clairement les avantages des méthodes d'agrégation. En utilisant l'ACC, Shu *et al.* (2007) montrent qu'une transformation stratégique des données avant l'utilisation d'un RNA permet également d'améliorer l'efficacité de celui-ci.

L'étude de Solomatine *et al.* (2003) a comparé les RNA aux méthodes d'arbres de régression dans le contexte de prédiction des débits normale de crues. Leurs études montrent que les arbres de régression performent de façon similaire aux RNA. Une étude comparative dans le cadre de l'AFR a été réalisée par Schnier *et al.* (2014) qui ont rapporté une légère diminution de la performance des arbres de régression comparativement aux RNA. Toutefois, pour ces deux études comparatives, les auteurs montrent que les arbres de régression permettentent une meilleure compréhension du rôle des variables explicatives que les RNA et donc du processus hydrologique. On notera que pour la régionalisation des débits d'étiage, Laaha *et al.* (2006) ont observé de meilleures performances de la part des arbres de régression comparativement aux méthodes de régression multiple à l'intérieur de régions homogènes.

Les modèles additifs généralisés, ou *generalized additive model* (GAM), forment une famille de méthodes non paramétriques (et non linéaires) qui est fréquemment utilisée dans plusieurs domaines de statistique appliquée (Davis *et al.*, 1998, Dominici *et al.*, 2002, Guisan *et al.*, 2002, Vida *et al.*, 2012), mais qui a reçu peu d'attention en AFR. Chebana *et al.* (2014) ont testé l'efficacité des modèles additifs en AFR et ont montré que cette méthode peut améliorer les performances prédictives par rapport aux méthodes de régression multiple à l'intérieur de voisinages. Au niveau de l'interprétation, les modèles additifs sont avantageux comparés aux RNA puisqu'ils permettent de visualiser

individuellement l'impact de chaque caractéristique physiométéorologique à partir de fonctions explicites.

Les techniques de régression ne sont pas les seules approches permettant la prédiction des variables hydrologiques extrêmes à des sites non jaugés. Les méthodes géostatistiques à l'intérieur d'espaces transformés a également été rencontrée en AFR (Chokmani *et al.*, 2004, Jon Olav Skøien *et al.*, 2007). Une comparaison entre différentes méthodes d'interpolation est présentée par Castiglioni *et al.* (2009) dans le cadre des débits étiages. Pour cette étude, les techniques de krigeage ont mené à des prédictions supérieures aux méthodes d'interpolation conventionnelles comme celles basées sur l'inverse de la distance et la méthode des polygones de Thiessen (Goovaerts, 2000). De plus, ils ont comparé le krigeage ordinaire avec le krigeage universel, mais aucune différence notable n'a été observée. Cependant, Nezhad *et al.* (2010) ont montré dans une autre étude que la technique du krigeage universel a conduit à une amélioration des performances grâce à l'introduction d'une tendance déterministe.

La technique du krigeage physiographique designe l'application des méthodes de krigeage sur des espaces basés sur la transformation des caractéristiques physiométéorologiques. Chokmani *et al.* (2004) ont suggéré d'employé l'ACC et l'analyse en composante principales (ACP) afin de construire ces espaces. Leurs résultats mènent à la conclusion que l'ACC est plus appropriée. Une autre étude entre ces deux méthodes de construction a également été réalisée par Guillemette *et al.* (2009) qui conduit à la même conclusion, mais cette fois dans la prédiction des températures maximales mensuelles des rivières.

Par ailleurs, les rivières à l'intérieur d'un basin versant sont connectées suivant un réseau bien défini qui dépend de sa topographie et pour laquelle une méthodologie permettant le krigeage de variables hydrologiques dans ces circonstances a été présentée par Sauquet *et al.* (2000), puis adaptée au contexte AFR par J. O. Skøien *et al.* (2006). Cette approche est communément désignée par le terme krigeage topographique. Castiglioni *et al.* (2011) ont conduit une étude comparative entre

le krigeage physiographique et topographique pour des débits d'étiages en Italie. Des résultats similaires ont alors été observés entre les deux méthodes avec un léger avantage pour le krigeage physiographique. D'un autre côté Archfield *et al.* (2013) ont comparé le krigeage physiographique et topographique avec la régression multiple à l'intérieur de régions homogènes. Cette étude a montré que les prédictions provenant des techniques de krigeage sont supérieures à celles des techniques de régression et que le krigeage topographique s'est avéré le plus performant.

2.4 MODÈLES PROBABILISTES

Un processus stochastique $Z(\mathbf{x})$ représente un ensemble de variables aléatoires indexé par un vecteur aléatoire $\mathbf{x} \in \mathbb{R}$ de dimension p. Pour décrire l'évolution d'une variable aléatoire, on peut alors utiliser un processus stochastique en fonction de variables explicatives $\mathbf{x}$. La loi de $Z(\mathbf{x})$ pour un $\mathbf{x}$ donné est appelée loi marginale et la loi de deux ou plusieurs variables explicatives $\mathbf{x}_i$ est appelée loi jointe. En particulier, dans le contexte géostatistique, un processus stochastique dont la loi jointe est une loi normale multidimensionnelle est appelé champ aléatoire gaussien.

Les méthodes traditionnelles utilisées en AFR nécessitent une analyse préliminaire qui fournit les variables hydrologiques devant être régionalisées. Par contre, le besoin d'une étape préliminaire peut être évitée en adoptant des modèles probabilistes qui déterminent un processus stochastique directement sur les valeurs extrêmes (Coles *et al.*, 1998). Dans cette thèse, ces modèles sont dit probabilistes puisqu'ils offrent une formulation complète des probabilités des valeurs extrêmes. En AFR, la loi marginale d'un processus stochastique $Z(\mathbf{x})$ représente la loi d'une variable hydrologique à un site associée aux caractéristiques physiométéorologies $\mathbf{x}$. La formulation analytique de la relation entre $\mathbf{x}$ et les paramètres des lois marginales de $Z(\mathbf{x})$ est appelée surface de réponse. Celle-ci a le même rôle qu'un modèle régional pour les techniques de prédiction des paramètres dans l'approche traditionnelle. Ce lien est illustré à la Figure 1 par une ligne hachurée. Ces valeurs

extrêmes sont généralement des maximums annuels, mais peuvent également être les dépassements de seuils (Cooley *et al.*, 2007, Thibaud *et al.*, 2013).

En utilisant une structure hiérarchique comme illustrée à la Figure 2, on observe qu'un modèle probabiliste est plus près du mécanisme physique servant à générer les valeurs extrêmes. De plus on voit à la Figure 2 quatre paliers qui représentent la hiérarchie de l'information: les données journalières, les maximums annuels, les paramètres des lois marginales et les quantiles. Cette hiérarchie est utile afin de représenter la structure de dépendances entre les sites au niveau des différents paliers. La dépendance qui provient de l'existence d'un phénomène météorologique à grande échelle qui crée une dépendance spatiale (Figure 2, étape i) pour les données journalières. Toutefois, deux maximums annuels peuvent ne pas être provoqués par le même événement, ce qui montre que la dépendance spatiale entre valeurs extrêmes est différente (étape h). L'étape g indique également une autre forme de dépendance au niveau des paramètres des lois marginales (Renard, 2011). Cette forme de dépendance, dite résiduelle, indique qu'une partie de la relation entre les paramètres et les caractéristiques physiométéorologiques n'est pas complètement spécifiée par la surface de réponse. Notons également que la Figure 2 ne montre pas directement de dépendance entre les variables hydrologiques, puisque cette relation est induite pas les structures présentes aux différents paliers.

Le recours à des structures hiérarchiques est courant dans les approches bayésiennes (Carlin *et al.*, 2009, Gelman, 2004). Une couverture de leur utilisation pour les données spatiales est présentée par Banerjee *et al.* (2004). Le cadre bayésien est initialement introduit en AFR dans le but d'améliorer l'analyse fréquentielle des sites partiellement jaugés en permettant de combiner l'information régionale avec celle d'un site (Kuczera, 1982, Micevski *et al.*, 2009, Mathieu Ribatet *et al.*, 2007, Seidou *et al.*, 2006).

La principale difficulté des modèles probabilistes est la modélisation de la dépendance spatiale extrême. Pour éviter ce problème, des modèles à variables latentes ont été proposés dans le contexte

des précipitations extrêmes (Cooley *et al.*, 2007, Sang *et al.*, 2009) et dans le cas de débits de rivières (Lima *et al.*, 2010). Pour ces modèles, chaque site jaugé est associé à un processus stochastique dont les lois marginales sont de la famille des valeurs extrêmes, et dont les maximums annuels sont indépendants. Le terme variable latente souligne ici le fait que les paramètres de la surface de réponse ne sont pas directement observés. Les modèles à variables latentes peuvent être considérés à des fins pratiques lorsque l'objectif est strictement de modéliser les lois marginales, mais sont irréalistes dans leur interprétation de la dépendance spatiale extrême. La Figure 3a montre un exemple théorique de la simulation d'un modèle à variables latentes qui souligne le manque de continuité.

Afin d'obtenir des simulations plus réalistes, Sang *et al.* (2010) ont proposé un modèle hiérarchique bayésien dont la dépendance entre les sites est décrite par une copule gaussienne, c'est-à-dire que les lois marginales suivent une loi des valeurs extrêmes et que la structure de dépendance est celle d'un champ aléatoire gaussien. Un exemple théorique de simulations utilisant une copule gaussienne est présenté à la Figure 3b. Par contre, la généralisation de la théorie des valeurs extrêmes aux processus stochastiques indique que les copules gaussiennes ne représentent pas une structure de dépendance appropriée (Davison *et al.*, 2012). La généralisation de la théorie des valeurs extrêmes à des dimensions plus grandes conduit à la définition de copules extrêmes (Joe, 2014). Bien qu'il existe un certain nombre de copules extrêmes bidimensionnelles (Salvadori *et al.*, 2007), pour les extrêmes spatiaux il est nécessaire d'utiliser une copule ayant une dimension égale aux nombres de sites. Dans ce cas, la t-copule extrême, construite comme la transformation de lois de Student, est plus appropriée dans le contexte des modèles probabilistes en AFR (Demarta *et al.*, 2005, Nikoloulopoulos *et al.*, 2009).

En principe, la généralisation de la théorie des valeurs extrêmes pour les processus stochastiques conduit à des modèles basés sur les processus max-stables (Haan, 1984). L'utilisation des processus max-stables en pratiques est relativement récente et comporte encore plusieurs défis. La principale difficulté provient du fait que la vraisemblance des données ne possède pas de forme

analytique pratique et empêche l'emploi des outils statistiques habituels. Toutefois, des représentations pratiques des processus max-stables permettent de les définir comme la superposition de processus stochastiques sous-jacents (Coles, 1993). Ces représentations permettent ainsi de décrire la structure de dépendance des processus max-stables grâce à la nature des processus sous-jacents et d'effectuer des simulations de ces processus max-stables à l'intérieur de délais raisonnables (Schlather, 2002). À cet effet, la Figure 3c montre la réalisation d'un modèle de Smith où les processus sous-jacents sont représentés par des cellules orageuses elliptiques décrites comme la superposions de lois normales (R. L. Smith, 1990). Également, la Figure 3d représente un processus de Brown-Resnick (Kabluchko *et al.*, 2009) qui représente un processus max-stable comme la superposition de processus de Weiner. Comme l'indique la Figure 3, les modèles de Brown-Resnick mènent à une structure moins simpliste que le modèle de Smith. Notons également que l'approche utilisant la t-copule extrême représente un processus max-stable qui inclut le cas particulier où les processus sous-jacents sont des champs aléatoires gaussiens (Opitz, 2013). À ce jour, les processus max-stables ont été utilisés en hydrologie uniquement pour modéliser les précipitations extrêmes (e.g., Neves *et al.*, 2011, Shang *et al.*, 2011, Westra *et al.*, 2011). En particulier, une généralisation du modèle d'indice de crue a été présentée par Wang *et al.* (2014). Cette dernière étude indique comment faire l'estimation du modèle régional en tenant compte de la dépendance entre les sites sous la forme de processus max-stables. Ceci a permis de réduire considérablement l'incertitude comparativement à l'approche traditionnelle des L-moments. Ce gain nécessite toutefois la spécification additionnelle d'une structure de dépendance max-stable pour laquelle les auteurs ont noté qu'une mauvaise spécification de la dépendance peut entraîner d'importants biais.

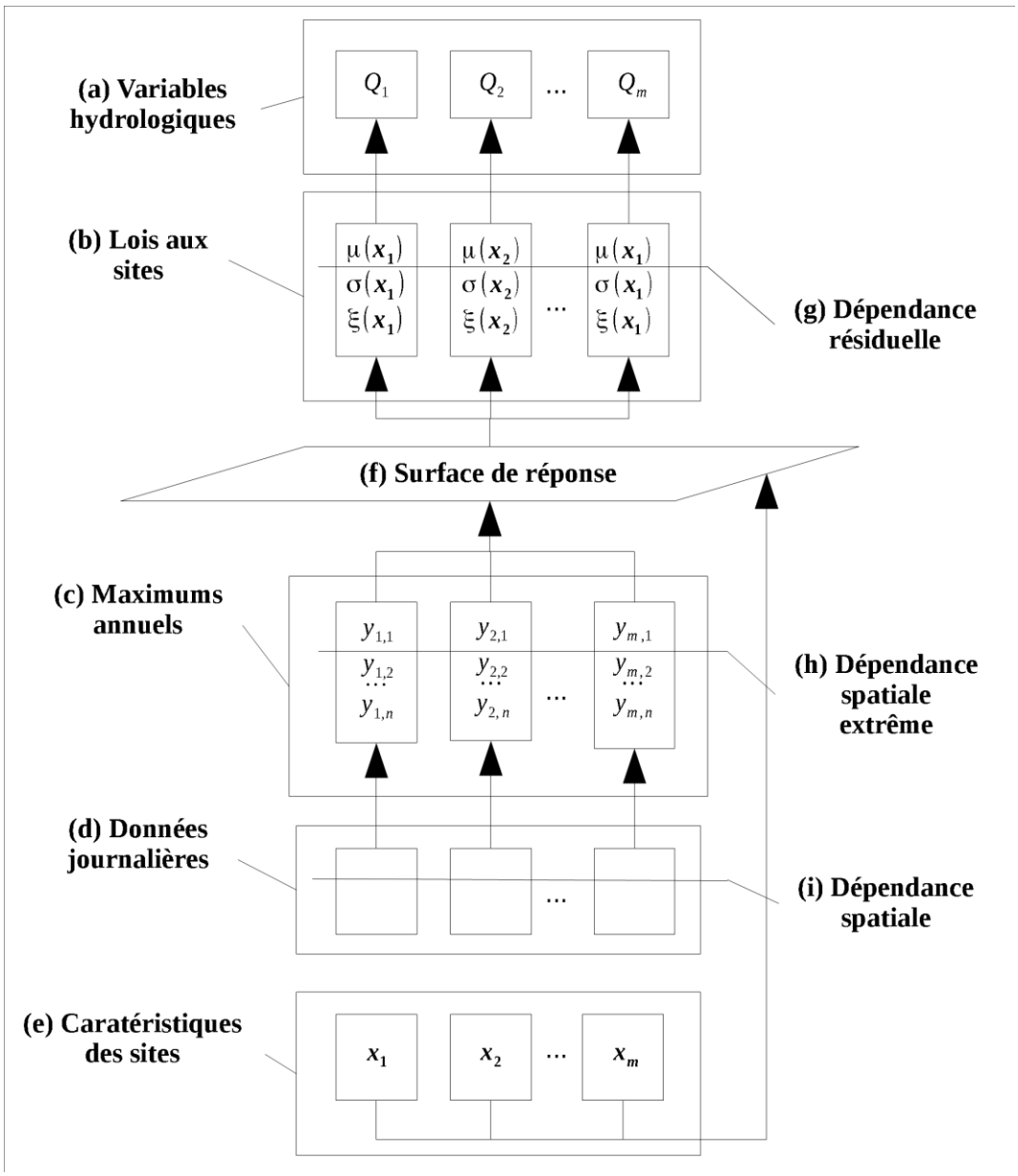

Figure 2: Schéma représentant le méchanisme de génération des données. Les distributions aux sites suivent des lois des valeurs extrêmes généralisées.

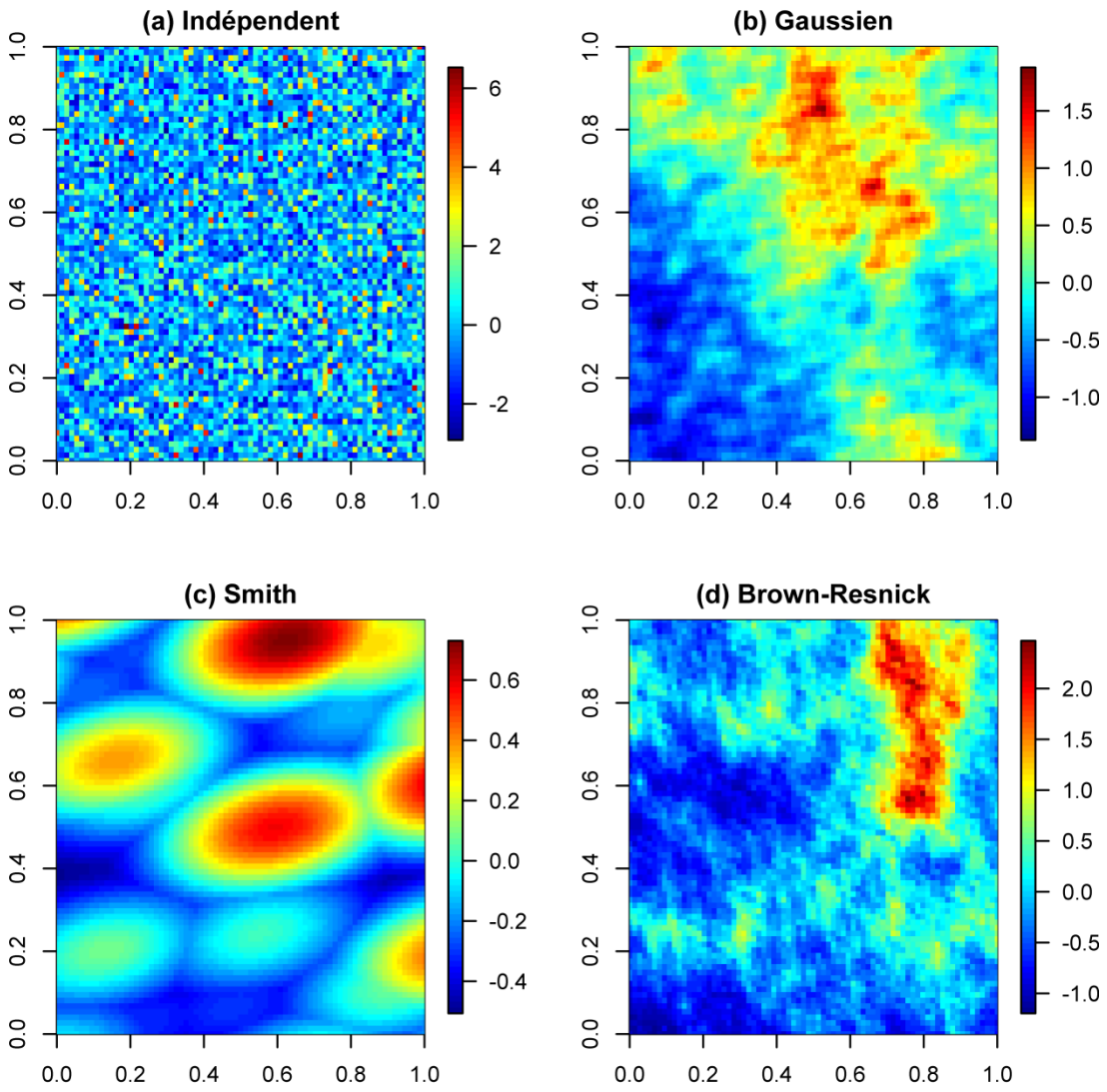

Figure 3: Simulation de processus stochastiques avec différentes structures de dépendances. Les lois marginales sont des lois de Gumbel unitaire.

3 MÉTHODES PROPOSÉES ET PRINCIPAUX RÉSULTATS

Dans cette section nous discutons des méthodes proposées et des principaux résultats obtenus à l'intérieur des articles présentés dans les chapitres 2 à 5. Cette section est divisée en trois sous-sections: choix des données, méthodes non paramétriques et modélisation de la dépendance intersite. En premier lieu, une présentation résume les données utilisées dans les articles. Par la suite, les deux sous-sections suivantes discutent des méthodes ciblées dans cette thèse et souslignent les particularités des méthodes proposées dans cette thèse afin de tenir compte de la non-linéarité et la non-normalité. Finalement les résultats importants de cette thèse sont mis en contexte et discutés dans une optique générale.

3.1 CHOIX DES DONNÉES

Lorsque possibles, des jeux de données dont l'analyse préliminaire des sites jaugés a été réalisée par des études antérieures ont été préférés. Ceci a permis de se concentrer uniquement sur la partie prédiction des sites non jaugés. Cette approche nous assure également que le choix des sites est validé et que les hypothèses habituelles sont testées: homogénéité, stationnarité et indépendance. À ce titre Kouider *et al.* (2002) ont construit une base de données de 151 sites situés dans le sud du Québec, Canada, et pour lesquels une estimation des quantiles de crues de plusieurs périodes de retour sont disponible.

Les données du Québec ont été utilisées dans trois articles et permettent de comparer les méthodes proposées entre elles (Durocher *et al.* 2015a,b,c). Ce jeu de données a de plus l'avantage d'avoir servi à plusieurs études antérieures pour lesquelles les quantiles de crues étaient également les variables hydrologiques d'intérêt. On utilisera par la suite la notation Q100 pour désigner par exemple un quantile de périodes de retour de 100 ans.

Par des études antérieures, différentes sélections des caractéristiques physiométéorologiques ont été considérées pour expliquer les comportements des variables hydrologiques aux sites non

jaugés. Le Tableau 1 énumère la liste complète des caractéristiques disponibles. Une première sélection est proposée par Chokmani *et al.* (2004) qui a retenu 4 caractéristiques sur la base d'une étude de corrélations avec les variables hydrologiques. Les caractéristiques retenues étaient: la pente moyenne du bassin versant (PMBV), la fraction du bassin versant occupé par des lacs (PLAC), la moyenne annuelle des précipitations totales (PTMA) et le nombre de degrés-jours sous zéro degré Celsius (DJBZ). Notons alors la décision d'exclure l'aire du basin (BV) en raison du désir de travailler avec des quantiles de crues standardisés par l'aire du basin.

La majorité des travaux subséquents ont conservé ces 4 variables, mais ont ajouté l'air du bassin versant (BV). Le Tableau 2 présente un résumé des différentes caractéristiques utilisées dans les différentes études abordées dans de cette thèse. La sélection initiale a toutefois été remise en question par Chebana *et al.* (2014) qui a comparé ce choix avec celui obtenu par une méthode pas-à-pas. Cette approche a permis d'identifier de nouvelles caractéristiques sur une base plus objective. La même approche a été adoptée par Durocher *et al.* (2015a). Notons que la combinaison de la méthode pas-à-pas avec les méthodes non paramétriques prend en compte la non-linéarité de la relation avec les variables. On notera des ressemblances entre les caractéristiques sélectionnées par ces deux dernières études (voir Tableau 2). En effet, toutes deux ont identifié la longitude comme caractéristique importante et ont préféré l'utilisation des précipitations liquides (PLMA) aux précipitations totales (PTMA).

Malgré la remise en question du choix des caractéristiques physiométéorologiques, les méthodes proposées à l'intérieur de cette thèse doivent être calibrées sur les mêmes caractéristiques que les études antérieures afin d'assurer une comparaison équitable entre les méthodes. Pour cette raison, Durocher *et al.* (2015a,c) considèrent également les caractéristiques des études antérieures indiquées au Tableau 2b. Par contre, les développements de Durocher *et al.* (2015b) visent spécifiquement les méthodes avec des régions homogènes de type voisinages. Par conséquent, la méthode proposée n'est pas comparée aux méthodes des études antérieures comme le krigeage physiographique et les autres méthodes non paramétriques. Une sélection basée sur les méthodes

pas-à-pas des études antérieures a été préférée puisqu'elle conduit à de meilleurs résultats. De plus, l'approche de formation des régions homogènes basée sur l'ACC a été reproduite avec ces nouvelles caractéristiques physiométéorologiques.

Les données utilisées par Durocher *et al.* (2015d) étaient les maximums annuels de précipitations pour des stations situées en Californie. Pour cette étude portant sur les processus max-stables, il n'était pas possible d'utiliser directement les périodes de retour fournies par des analyses antérieures, puisque les processus max-stables utilisent directement les maximums annuels comme observations. Un ensemble de 39 sites ont été sélectionnés. La taille limitée de cet échantillon a été choisie volontairement en raison de la lourdeur des calculs rattachés à la méthode employée.

Notons qu'a priori, le jeu de données des crues annuelles du Québec, utilisé à l'intérieur des trois autres articles de cette thèse, ne représente pas un cas d'étude appropriée pour les processus max-stables. En effet, ceux-ci représentent un résultat asymptotique basé sur le maximum d'une infinité de processus stochastiques sous-jacents. Or, une grande partie des crues annuelle au Québec sont le résultat de la fonte printanière. Par conséquent, les maximums annuels sont essentiellement le résultat d'un seul et même événement. Les arguments asymptotiques de la théorie des valeurs extrêmes ne s'appliquent donc pas forcément dans ce cas-ci. Ce raisonnement s'applique également au choix des lois marginales, c'est-à-dire que la théorie des valeurs extrêmes n'impose pas une loi des valeurs extrêmes généralisées. Dans les faits, l'analyse fréquentielle des sites menés par Kouider et al. (2002) a révélé que pour une majorité des sites, la loi sélectionnée était de la famille log-normale ou gamma. Néamoins, des études ultérieures sont nécessaires afin de confirmer précisément la pertinence réelle des processus max-stables dans la description des crues annuelles du Québec.

Tableau 1: Liste des caractéristiques des bassins disponibles pour le jeu de données du Québec

Variables	Abbréviation
Aire du bassin (km^2)	BV
Longueur du canal principal (km)	LCP
Pente du canal principal (m/km)	PCP
Pente du bassin versant (°)	PMBV
Fraction du bassin occupé par des forêts (%)	PFOR
Fraction du bassin occupé par des lacs (%)	PLAC
Précipitation total annuelle moyenne (mm)	PTMA
Précipitation liquide annuelle moyenne (mm)	PLMA
Précipitation solide annuelle moyen (cm)	PSMA
Précipitation liquide moyenne de Jul-Dec (mm)	PLME
Niveau de neige en date du 30 mars (cm)	MNS30
Degrée-jours sous 0 Celsius (°/jr)	DJBZ
Longitude	LON
Latitude	LAT

Tableau 2: Caratéristiques physiométéorologiques utilisées pour la régionalisation des données du sud du Québec avec Q100

Articles	Caractéristiques des sites
a. Chokmani et Ouarda (2004)	PMBV, PLAC, PTMA, DJBZ
b.Durocher *et al.* 2015(a,c); Nezhad *et al.* (2010); Shu et Ouarda (2007); Wazneh *et al.* (2013b)	BV, PMBV, PLAC, PTMA, DJBZ
c. Chebana *et al.* (2014)	BV, PLAC, PLMA, LON, LAT
d. Durocher *et al.* (2015a)	BV, LCP, PCP, PLAC, PLMA, LON
e. Durocher *et al.* (2015b)	BV, PLAC, PLMA, LON

3.2 MÉTHODES NON PARAMÉTRIQUES

3.2.1 Méthodes non paramétriques existantes

Des méthodes considérant la non-linéarité entre les variables hydrologiques et les caractéristiques physiométéorologiques ont déjà été proposées en AFR. La prise en charge de la non-linéarité dans un modèle régional se fait généralement par l'entremise de méthodes non paramétriques pour lesquels aucune forme spécifique n'est déterminée a priori. Notons que les méthodes de régression multiple à l'intérieur de voisinages permettent indirectement de tenir compte de la non-linéarité. En effet, en se déplaçant d'un point de référence à un autre, une grande partie des sites utilisés sont les mêmes. Par conséquent, les paramètres du modèle de régression multiple varient graduellement et peuvent être considérés commes des approximations localisées d'un modèle régional appliqué sur l'ensemble des sites étudiés. Une discussion générale de ce type d'approche statistiques se trouve dans Hastie *et al.* (2009) et des approches de types voisinage qui ont été considérées pour les données du Québec sont l'ACC (Chokmani *et al.*, 2004) et les fonctions de profondeurs (Chebana *et al.*, 2008).

Pour les données du Québec, les RNA sont également utilisées comme techniques de prédiction des quantiles par Shu *et al.* (2007). Leurs résultats ont montré que l'ajustement direct d'un seul RNA performe moins bien que la régression multiple à l'intérieur de voisinages. Les améliorations qu'ils ont apportées sont l'utilisation de l'ACC comme pré-traitement des caractéristiques physiométéorologiques et l'utilisation d'une méthode d'agrégation bootstrap, ou *bagging*. Avec ces modifications, l'approche utilisant les RNA ont conduit à des performances supérieures à l'approche de l'ACC traditionnelle, mais similaires à celle des fonctions de profondeur.

Les équations de régression offerte par les méthodes utilisant des voisinages sont simples, mais sont valides uniquement pour un site cible. Afin d'obtenir des modèles non paramétriques possédant une meilleure compréhension du rôle des caractéristiques physiométéorologiques pour l'ensemble des sites, les modèles GAM ont été appliqués aux données du Québec par Chebana *et al.*

(2014). En effet, ceux-ci sont plus pratiques puisqu'ils décrivent explicitement le rôle de chaque caractéristique physiométéorologiques pour l'ensemble de la région d'étude. Ces méthodes ont cependant performé légèrement moins bien que les RNA pour les données du Québec.

3.2.2 Méthodes non paramétriques proposées

La régression à directions révélatrices, ou *projection pursuit regression* (PPR), a été initialement introduite par Friedman *et al.* (1974). Pour illustrer ces modèles, prenons $\mathbf{x}$ un vecteur de plusieurs caractéristiques physiométéorologique et y une variable hydrologique. Un modèle PPR s'écrit alors:

$$y = \mu + \sum_{i=1}^{p} f_i(\alpha_i'\mathbf{x}) + \varepsilon \qquad (2)$$

où μ est la moyenne globale, les f_i sont des fonctions non linéaires de moyenne nulle $E(f_k) = 0$ et les α_i sont des vecteurs de coefficients unitaire $\|\alpha_k\| = 1$, ou directions révélatrices, servant à définir des prédicteurs intermédiaires $\eta_i = \alpha_i'\mathbf{x}$. En général, les modèles PPR peuvent utiliser plusieurs caractéristiques et faire la sommation de p termes $f_i(\alpha_i'\mathbf{x})$. Dans ces situations générales, l'utilisation des PPR est comparable à celle des RNA puisqu'elle partage plusieurs ressemblances. En effet, des études comparatives ont démontré que ces deux méthodes ont des pouvoirs prédictifs similaires et permettent d'approximer un modèle de régression avec la précision voulue (Bishop, 1995, Hwang *et al.*, 1994). Par contre, les PPR partagent également certains désavantages des RNA qui conduisent à des modèles surparamétrisés et sans solution unique. De plus, lorsque le nombre de directions révélatrices α_i est imposant, ce nombre masque le rôle des caractéristiques physiométéorologique à l'intérieur de chaque terms $f_i(\alpha_i'\mathbf{x})$.

Toutefois, l'utilisation faite des PPR dans cette thèse est très différente du cas général. Durocher *et al.* (2015a) ont utilisé les PPR comme techniques de prédiction des quantiles pour les

crues annuelles au Québec. L'objectif était de vérifier si ces modèles pouvaient être utiles en AFR lorsque limités à seulement quelques directions révélatrices. En effet, le cas particulier utilisant une direction unique α peut être traité comme une approche distincte (Antoniadis *et al.*, 2004, Carroll *et al.*, 1997, Hardle *et al.*, 1993). Dans ce cas précis, le prédicteur $\eta = \alpha' \mathbf{x}$ joue le rôle d'un modèle linéaire pour lequel f détermine une transformation inconnue avant estimation (Weisberg *et al.*, 1994). Le modèle PPR à direction unique offre donc une interprétation explicite du modèle régional, similaire à celui de régression classique. Notons qu'un neurone dans un modèle RNA a la même forme qu'un terme $f_i(\alpha_i' \mathbf{x})$ dans un modèle PPR (Hastie *et al.*, 2009). Toutefois, un neurone utilise des fonctions non linéaires f_i spécifiques, avec un nombre de paramètres restreint, tandis que pour un modèle PPR les f_i sont non paramétriques et donc plus flexibles. Cette distinction est importante puisqu'elle indique qu'un modèle RNA réduit à un seul neurone a des applications limitées.

Des modèles PPR ont également été utilisés par Durocher *et al.* (2015b) pour former des voisinages. Cette méthode sera désignée par la suite par RVN, pour *Reference Variables Neighborhoods*. Les approches en AFR utilisant les voisinages devraient comporter trois étapes: (i) identification des centres des voisinages (ii) formation des voisinages et (iii) estimation de modèles régionaux. Le point fondamental de la méthode RVN concerne l'étape (i). Dans cette étape, les PPR sont utilisés dans le but de prédire des variables de référence qui représentent les centres des voisinages. Ces variables de référence peuvent être des variables hydrologiques connues aux sites jaugés, mais inconnues aux sites cibles. Ainsi, la prédiction des variables de référence aux sites non jaugés représente une étape préliminaire importante afin de déterminer le centre inconnu d'un voisinage. À l'opposé, la méthode ROI considère que toutes les variables de références sont des caractéristiques physiométéorologiques qui sont déjà connues. Par conséquent, l'étape (i) pour la méthode ROI est directe et ne demande pas d'analyse préliminaire à la formation des voisinages (ii).

Durocher *et al.* (2015b) ont montré que l'approche RVN généralise celle de de l'ACC. Les avantages sont la considération d'une sélection plus générale des différentes variables de référence pouvant servir à former des voisinages et l'utilisation des PPR afin de tenir compte de la non-linéarité. Durocher *et al.* (2015b) ont présenté un cas d'étude qui illustre la méthode proposée lorsque l'étape (iii) est un modèle d'indice de crue ou un modèle de régression multiple des quantiles. Les variables de référence considérées sont les L-moments des loi des sites gaugés et ce choix vise à obtenir des voisinages qui sont plus homogènes par rapport à la dispersion du L-coefficient de variation (Hosking *et al.*, 1997). Durocher *et al.* (2015b) considère également une situation hybride où les variables de référence comprennent les L-moments, inconnue, ainsi que des caractéristiques physiométéorologiques connues.

3.2.3 Résultats

La performance des modèles régionaux est un point central dans l'évaluation des méthodes proposées en AFR. En relaxant les hypothèses de linéarité, on désire obtenir des approches plus flexibles qui vont mieux refléter la réalité et améliorer l'éfficacité des modèles. La méthode d'exclusion (*leave-one-out cross-validation*) est adoptée comme technique de validation croisée dans trois articles de cette thèse (Durocher *et al.,* 2015a,b,c). Cette méthode d'évaluation considère à tour de rôle chaque site cible comme un site non jaugé. Cette approche conduit à une procédure qui est facilement reproductible. De plus, la méthode d'exclusion a été choisie par les études antérieures servant de comparaison dans cette thèse.

Par ailleurs, pour les méthodes non paramétriques comme PPR, il n'existe pas toujours de mesure de complexité adéquate (e.g. degrée de liberté) comme dans le cas de la régression multiple (Hastie *et al.*, 2009). Ceci empêche la considération des critères d'évaluation comme le critère d'information d'Akaike ou de Schwartz, qui pénalise les modèles sur la base de cette mesure de complexité. La méthode d'exclusion offre une mesure de la qualité réelle des prédictions d'un modèle sans l'utilisation de mesure de complexité.

Les variables hydrologiques d'intérêt dans Durocher *et al.* (2015a) sont les quantiles Q10 et Q100 qui sont standardisées par l'aire du basin. Les résultats de validation croisée utilisant les caractéristiques des études antérieures, comme indiqué au Tableau 2b, ont montré que deux méthodes considérées dans les études antérieures se sont distinguées selon la racine de l'erreur relative moyenne (RERM) avec une valeur d'environ 45% pour Q100. Ces méthodes sont les fonctions de profondeurs (Wazneh *et al.*, 2013) et les RNA (Shu *et al.*, 2007). Pour leur part, les modèles PPR et les modèles additifs ont obtenu des RERM d'environ 48% et la méthode utilisant des voisinages construits à partir de l'ACC a obtenu un RERM de 51%.

Cependant, un aspect n'a pas été pris en compte dans les comparaisons des résultats. Chokmani *et al.* (2004) ont montré l'existence de six sites problématiques qui ont une incidence importante sur le critère RERM. En effet, pour la technique du krigeage physiographique, le RERM passe de 70% à 41% lorsque ces sites sont exclus, soit une différence de 29%. De même, Durocher *et al.* (2015a) ont montré que pour la méthode PPR le retrait de ces six sites réduit le RERM de 14%. À partir des résultats disponibles dans les études antérieures, il n'était pas possible de déterminer si la différence de performance entre les PPR et les autres les méthodes antérieures sont principalement attribuables à une meilleure estimation générale où à une meilleure gestion de ces six sites problématiques. Ces derniers représentent néanmoins une réalité en AFR et les résultats obtenus indiquent comment les méthodes utilisées performent dans ces conditions.

Par ailleurs, l'emploi de la méthode PPR en combinaison avec les méthodes pas-à-pas a réduit le RERM de 8%, pour atteindre 40%. À titre comparatif, avec la combinaison des modèles GAM avec la méthode pas-à-pas de Chebana *et al.* (2014), on trouve un RERM de 42% qui est légèrement supérieurs à celui de PPR. De plus, notons que la méthode PPR utilise une seule fonction non linéaire contrairement aux modèles GAM qui possèdent cinq fonctions non linéaires, une pour chaque caractéristique physiométéorologique. Ceci démontre que la direction révélatrice trouvée dans les modèles PPR est charactéristique du mécanisme de formations des crues et a conduit à une

représentation plus parcimonieuse de l'aspect non linéaire que les modèles GAM, sans pour autant sacrifier les performances.

Dans l'étude menée par Durocher *et al.* (2015b) les comparaisons visaient spécifiquement les méthodes avec voisinages. Les choix faits pour cette étude font en sorte que les résultats ne sont pas directement comparables avec ceux des études antérieures, puisque Durocher *et al.* (2015b) considère un quantile qui n'est pas standardisée par l'air du basin et l'étude utilise uniquement d'autres caractéristiques physiométéorologiques que celles les études antérieures (voir Tableau 2b). Néamoins, les résultats de la méthode RVN pour la même configuration que celle des études antérieures sont présentées ici.

Dans ce cas d'étude, les centres des voisinages sont formés d'une combinaison des caractéristiques physiométéorologiques (BV, PLAC) et de L-moments (L-coefficient de variation et L-coefficient d'asymétrie). Les voisinages sont formés d'un nombre fixe de sites jaugés, déterminé selon la distance euclidienne entre les sites et le centre du voisinage. De plus, cette méthode RVN utilise un modèle de régression mutliple pour prédire Q100. Les résultats de la méthode RVN dans ces conditions conduit à un RERM de 42% qui est plus performant que les méthodes considérées dans les études antérieures. Durocher *et al.* (2015a) ont également évalué la performance des modèles PPR à partir d'un critère de Nash-Sutcliffe (NSH) et ont trouvé une valeur de 71%. Pour la méthode de Durocher *et al.* (2015b) dans les mêmes conditions, on obtient un NSH de 72%, qui est légèrement supérieur. Globalement, ces résultats montrent que l'utilisation des PPR à l'intérieur de la méthode RVN était plus performante que leur utilisation directe comme technique de prédiction des quantiles.

Durocher *et al.* (2015b) ont montré la nécessité de considérer la non-linéarité dans la méthode RVN. En effet, Durocher *et al.* (2015b) ont indiqué que l'ACC, basée sur la linéarité, a moins bien performé que la méthode RVN en termes de REQM et NSH. Néamoins, l'avantage principal des voisinages formés par la méthode RVN est l'homogénéité. Durocher *et al.* (2015b) ont montré que la dispersion du coefficient de variations était en moyenne plus petit pour les voisinages délimités par

l'approche RVN que pour la méthode ROI et ACC. Ces résultats montrent qu'en choisissant les bonnes variables de référence, cela permet de regrouper des sites similaires selon les propriétés souhaitées et que ce choix se traduit par de meilleures prédictions.

L'interprétation de résultats de la méthode PPR sur les directions révélatrices est également plus riche. En effet, Durocher *et al.* (2015b) ont montré qu'en utilisant les transformations habituelles (logarithmes, racines carrées), la relation entre les caractéristiques physiométéorologiques et la moyenne des crues annuelles devient linéaire. Néamoins, des méthodes non paramétriques comme les PPR sont nécessaires afin de tenir compte de la non-linéarité des autres aspects, comme les L-coefficients de variation et L-coefficients d'asymétrie, pour lesquels aucune transformation habituelle n'est adéquate (non linéarisables).

3.3 MÉTHODES D'ESTIMATION DE LA DÉPENDANCE INTERSITE

3.3.1 Estimation de la dépendance entre les quantiles

La persistance à l'intérieur de données peut-être considérée comme une tendance déterministe ou comme une forme de dépendance entre les variables étudiées. Les méthodes non paramétriques décrivent cette persistance comme une tendance non linéaire, tandis que les méthodes géostatistiques adoptent l'approche opposée et représentent la persistance comme une dépendance. Toutefois en pratique, ces deux approches de modélisation peuvent être jumelées afin d'inclure une tendance et une structure de dépendance dans un même modèle. Par contre, ces deux parties doivent être clairement distinctes pour éviter tous problèmes d'identification (Opsomer *et al.*, 2001, Schabenberger *et al.*, 2004).

Pour les données du Québec, la dépendance entre les quantiles de crues a été modélisée par Chokmani *et al.* (2004) et par Nezhad *et al.* (2010). Chokmani *et al.* (2004) ont utilisé le krigeage ordinaire où la persistance est traitée uniquement comme une covariance déterminée par un semi-variogramme. Ce dernier décrit l'évolution de la covariance en fonction de la distance séparant deux

sites. L'effet pépite est une propriété du semi-variogramme qui introduit une erreur non négligeable pour les variables hydrologiques aux sites. En particulier, l'effet de pépites est nécessaire en AFR puisque les quantiles de crue ne sont pas mesurés, mais calculer à partir de données journalières. De plus, Chokmani *et al.* (2004) ont montré que l'espace construit à partir de l'ACC est plus adéquat que celui de l'ACP pour les données du Québec. Nezhad *et al.* (2010) ont fait l'analyse des mêmes données que Chokmani *et al.* (2004), mais ont utilisé la méthode du krigeage de résidues afin d'introduire une tendance déterministe. Plus précisément, une tendance quadratique a été ajustée aux variables hydrologiques, puis la méthode du krigeage ordinaire a été appliquée aux résidus. Ces deux parties sont estimées séparément en utilisant les moindres carrés pour ajuster la tendance et le semi-variogramme.

Les méthodes du krigeage calculent une prévision $\hat{y}$ à un site non jaugé comme un prédicteur linéaire:

$$\hat{y} = \sum_{i=1}^{m} \lambda_i y_i \qquad (2)$$

où y_i est une variable hydrologique observée au i-ème site jaugé et les λ_i sont des poids de krigeage. L'estimateur du krigeage est donc une combinaison linéaire des variables hydrologiques calculées aux sites jaugés. Cet estimateur est souhaitable lorsque les y_i proviennent d'un champ aléatoire gaussien puisque qu'il est le meilleur prédicteur sans biais (Schabenberger *et al.*, 2004). Par contre, l'estimateur du krigeage n'est pas le meilleur prédicteur lorsqu'une transformation logarithmique est utilisée, comme dans le cas des crues du Québec. À l'échelle logarithmique, la moyenne et la médiane de la loi des variables hydrologiques coïncident. Toutefois, la transformation inverse (exponentielle) préserve la médiane et non la moyenne. Par conséquent, les prédictions des quantiles de crues obtenues à partir d'une transformation sont les médianes des lois prédictives asymétriques et sont par conséquent biaisées. Ce biais est toutefois acceptable, puisque la correction

de ce biais, au profit de la moyenne, conduit à des prévisions ayant de plus grandes incertitudes (Girard *et al.*, 2004).

L'utilisation des copules spatiales proposée par Durocher *et al.* (2015c) a permis de déterminer séparément la loi des quantiles de crues et la structure de dépendance. La calibration du modèle des copules spatiales a considéré une loi marginale log-normale qui est cohérent avec celui des études antérieures utilisant une transformation logarithmique. Durocher *et al.* (2015c) ont eu recours également au test d'ajustement proposé par Bárdossy (2006) pour montrer que la dépendance entre les quantiles de crues pouvait être raisonnablement modélisé par une copule gaussienne. Ce choix est aussi cohérent avec les études antérieures sur ces mêmes données qui étaient basées sur des champs aléatoires gaussiens.

Le principal apport du cadre des copules spatiales dans l'étude des données du Québec est la considération d'un prédicteur non linéaire et l'ajout d'une tendance linéaire pour l'écart-type. Cette tendance montre que la variabilité tend à diminuer en direction du prédicteur fournit par la première paire canonique. Durocher *et al.* (2015c) ont aussi considéré l'utilisation de la vraisemblance par paires comme méthode d'estimation. Cette méthode est un cas particulier de la théorie de la vraisemblance composée adaptée au contexte de données spatiales (Heagerty *et al.*, 1998, Varin, 2008). Contrairement à la méthode des moindres carrées utilisée par Nezhad *et al.* (2010), cette méthode permet d'estimer conjointement la tendance et la dépendance en une seule étape.

Un avantage important de la méthode proposée par Durocher *et al.* (2015c) sur les techniques de krigeage est qu'elle a permis de calculer la loi prédictive complète des variables hydrologiques aux sites non jaugés. La connaissance de cette loi peut servir à calculer la moyenne des prévisions qui est le meilleur prédicteur sans biais (Bárdossy *et al.*, 2008). Par contre, Durocher *et al.* (2015c) ont également considéré la médiane de la loi prédictive qui a conduit à de meilleures performances en terme de RERM. En effet, Durocher *et al.* (2015c) ont montré qu'en utilisant la médiane, l'approche des copules spatiales a été la méthode géostatistique la plus performante avec un REQM de 41%

pour Q100, comparativement à 70% pour Chokmani *et al.* (2004) et 58% pour Nezhad *et al.* (2010). Dans les faits, l'approche des copules spatiales a surpassé les meilleures méthodes non paramétriques sur les mêmes données.

3.3.2 Estimation de la dépendance entre les maximums annuelles

Comme l'indique la Figure 2, les modèles hiérarchiques permettent de modéliser la structure de dépendance sur plusieurs paliers, ce qui permet de tenir compte d'une dépendance de type extrême entre les maximums annuels. Les précipitations extrêmes de la Californie ont été étudiées à l'aide de processus max-stables par Shang *et al.* (2011). L'objectif principal de cette étude était d'étudier l'influence de l'oscillation australe sur les précipitations extrêmes. La considération des processus max-stables permet d'étudier l'influence de l'oscillation australe sur l'ensemble des sites simultanément. Ceci a permis d'améliorer les résultats obtenus en analysant les sites individuellement (El Adlouni *et al.*, 2007, Zhang *et al.*, 2010). Une des limitations de l'étude de Shang *et al.* (2011) a été de se limiter à un modèle de Smith qui offre une représentation simpliste des phénomènes météorologiques, comme le démontre la Figure 3c.

Face à l'impossibilité de calculer explicitement les probabilités jointes d'un processus max-stable, l'estimation de ce dernier est généralement effectuée suivant la théorie de la vraisemblance composée (Padoan *et al.*, 2010). Notons que dans certaines situations, l'optimisation de la vraisemblance composée peut s'avérer difficile (Blanchet *et al.*, 2011). De plus, une mauvaise spécification de la structure de dépendance extrême peut entraîner d'important biais (Wang *et al.*, 2014). Ces difficultés ont motivé la recherche de méthodes d'estimation alternatives. E. L. Smith *et al.* (2009), M. Ribatet *et al.* (2009) et Erhardt *et al.* (2012) ont proposés des approches bayésiennes qui utilisent des algorithmes permettant l'échantillonnage de la loi a posteriori des paramètres. Toutefois, ces approches sont approximatives dans le sens où elles ne correspondent pas à la véritable loi a posteriori calculée par la théorie bayésienne. Durocher *et al.* (2015d) ont utilisé le calcul bayésien

approximatif, ou ABC pour *approximate bayesien computing*, comme proposé par Erhardt *et al.* (2012).

L'approche ABC est simple et vise à palier la connaissance de la formule analytique des probabilités jointes par des simulations. Un exemple typique de procédure ABC est l'algorithme de rejet. Pour cet algorithme, des statistiques sommaires $T(\mathbf{y})$ sont calculées à partir d'observations $\mathbf{y}$ afin de résumer l'information disponible. Idéalement T sont des statistiques exhaustives pour les paramètres θ du modèle estimé, c'est-à-dire que $P\left(\mathbf{y} \mid \hat{T}, \theta\right) = P\left(\mathbf{y} \mid \hat{T}\right)$. En d'autres mots, $T(\mathbf{y})$ contient toute l'information nécessaire sur $\mathbf{y}$ afin de réaliser l'inférence statistique du modèle. Malheureusement, de telles statistiques sont rarement disponibles et une importante partie de la mise en oeuvre d'une analyse par ABC consiste à mettre de l'avant des statistiques $T(\mathbf{y})$ appropriées. Dans l'algorithme de rejet, des paramètres $\hat{\theta}$ sont tirés d'une loi a priori et une simulation $S(\hat{\theta})$ est effectuée à partir de ces paramètres. Les statistiques $\hat{T} = T\left[S(\hat{\theta})\right]$ sont ensuite comparées aux statistisque observées $T(\mathbf{y})$. Si la distance d entre les deux groupes de statistiques est suffisamment faible $d\left(\hat{T}, T\left(\mathbf{y}\right)\right) < \varepsilon$, alors $\hat{\theta}$ est acceptée comme une réalisation de la loi a posteriori, sinon un nouveau candidat $\hat{\theta}$ est tiré de la loi a priori et le processus continue jusqu'à l'obtention d'un échantillon de taille désirée.

Cette approche simple conduit alors à une approximation de la loi a posteriori du cadre bayésien traditionnelle et la qualité de cette approximation est contrôlée par le seuil ε. De petites valeurs de ε mènent à de meilleures approximations, mais conduisent également à rejeter plusieurs candidats $\hat{\theta}$. Par conséquent, le choix de ε offre un compromis entre la qualité de l'approximation et le temps de calcul nécessaire pour piger le nombre de paramètres demandés. En pratiques, des algorithmes ABC plus rapides ont été proposés afin d'obtenir de meilleures approximations que

l'algorithme de rejet, en limitant le nombre de candidats $\hat{\theta}$ rejetés (Beaumont *et al.*, 2009, Beaumont *et al.*, 2002, Marjoram *et al.*, 2003).

Erhardt *et al.* (2012) ont montré que dans certaines conditions, l'estimateur ABC était plus efficace que celui de la vraisemblance composée. Notons que la procédure ABC dépend fortement du choix des statistiques T. Ainsi Erhardt *et al.* (2012) ont considéré plusieurs groupes de statistiques et ont montré que les statistiques basées sur des triplets de probabilités étaient plus efficaces. Par contre, Durocher *et al.* (2015d) ont choisi d'utiliser un madogram empirique qui est l'analogue d'un semi-variogramme dans le contexte des processus max-stables (Cooley *et al.*, 2006). Ce choix conduit à un groupe de statistiques sommaires de taille raisonnable qui est relativement rapide à calculer. Ce choix a été validé à l'aide d'une étude de simulations par Durocher *et al.* (2015d) qui a confirmé que les performances de l'approche ABC avec le madogram empirique étaient supérieures à celles obtenues avec la vraisemblance par paires.

Notons que les résultats de l'étude de simulation de Durocher *et al.* (2015d) vont dans une certaine mesure à l'encontre de ceux de Erhardt *et al.* (2012). Ces derniers ont constaté de meilleures performances de la part de l'estimateur de la vraisemblance par paires que l'approche ABC utilisant le mandogram empirique. Deux facteurs peuvent expliquer ces différences. Premièrement, le madogram théorique considéré par Durocher *et al.* (2015d) était basé sur une fonction de corrélation exponentielle avec un seul paramètre. Celui de Erhardt *et al.* (2012) était basé sur une fonction de corrélation de Whittle-Matérn avec deux paramètres. Ces résultats suggèrent que l'approche ABC peut perdre de son efficacité lorsque la structure de dépendance devient plus complexe et que le rôle de chaque paramètre devient plus difficile à identifier. Par ailleurs, Durocher *et al.* (2015d) ont utilisé de plus larges échantillons ABC en plus d'un post-traitement (Blum *et al.*, 2010). Ce post-traitement vise à corriger le biais et la variance de la distribution a posteriori issue d'un algorithme ABC en présence d'un seuil ε non négligeable. Ces améliorations techniques peuvent avoir conduit à une

amélioration suffisante de l'approximation de loi a posteriori afin de surpasser l'approche de la vraisemblance par paires.

Par ailleurs, l'étude de Erhardt *et al.* (2012) a estimé uniquement la structure de dépendance d'un processus max-stable. Une innovation de Durocher *et al.* (2015d) est de considérer conjointement l'estimation des lois marginales et la structure de dépendance. Durocher *et al.* (2015d) ont ajouté ainsi les L-moments au mandogram empirique dans le groupe de statistiques sommaires $T(\mathbf{y})$. Ce choix a également été validé par une seconde étude de simulations qui a toutefois conduit à des estimations moins efficaces que celles de la vraisemblance par paires. Les résultats obtenus sont par contre acceptables puisqu'ils sont comparables à ceux obtenus par un modèle à variables latentes, qui assume l'indépendance entre les sites.

Durocher *et al.* (2015d) ont comparé l'estimateur ABC avec l'estimateur de vraisemblance par paires sur les données de précipitations extrêmes de la Californie. La surface de réponse retenue a des lois marginales de la famille des valeurs extrêmes avec un paramètre de forme constant et un paramètre de dispersion proportionnelle aux paramètres de location. On remarquera que ces hypothèses sont similaires à ceux d'un modèle d'indice de crue. L'examen du madogram empirique a montré que l'estimateur du maximum de vraisemblance par paires n'est pas parvenu à estimer convenablement le paramètre de dépendance spatiale. De plus, les performances obtenues à partir d'un ensemble de validation ont indiqué des performances inférieures à celles de l'estimateur ABC. Ces résultats sont cohérents avec les conclusions de Wang *et al.* (2014), qui ont avisé de l'importance de bien choisir et de bien estimer la dépendance d'un processus max-stable.

Notons que Durocher *et al.* (2015d) et Erhardt *et al.* (2012) ont limité leurs études à des modèles de Schlater, en raison du temps nécessaire afin de réaliser une simulation, comparativement aux modèles de Brown-Resnick (Oesting *et al.*, 2012). Ces derniers pourraient par contre être plus appropriés dans le cas des précipitations extrêmes en Californie. En effet, le modèle de Schlater est un processus max-stable dont les processus sous-jacents sont des champs aléatoires gaussiens. Ce

modèle est plus réaliste que le modèle de Smith, mais possède une limitation importante puisqu'il ne permet pas l'indépendance entre deux sites. Ce comportement peut être justifié pour de petites régions comme dans le cas de Durocher *et al.* (2015d), mais n'est pas réaliste dans le cas de plus vastes territoires où deux sites éloignés ne sont pas toujours touchés par le même événement extrême. Le modèle de Brown-Resnick ne possède pas cette limitation et la considération de ces modèles pourrait améliorer les résultats pour les précipitations extrêmes en Californie. Néamoins ceci rendrait la procédure ABC plus difficilement utilisable. L'amélioration des méthodes de simulations pour les processus max-stables fait actuellement l'objet d'étude (Dombry *et al.*, 2013, Oesting *et al.*, 2014) et de futures améliorations pourraient rendre l'approche ABC plus attrayante et plus généralement applicable en AFR.

4 CONCLUSINS ET PERSPECTIVE DE RECHERCHE

4.1 CONCLUSIONS

En AFR, de nombreux outils statistiques sont utilisés afin de modéliser la relation entre les variables hydrologiques et les caractéristiques physiométéorologiques dans le but de prédire les risques d'occurrence d'événements extrêmes à des sites non jaugés. Cette thèse s'est attaquée à la validité des hypothèses de normalité et de linéarité qui découle de l'emploi de ces approches statistiques en AFR. Les principaux outils ciblés ont été les méthodes non paramétriques, l'ACC, les techniques de krigeages et les processus max-stables. Bien que les performances soient un aspect fondamental de l'AFR, cette thèse à chercher à mettre en valeur la qualité de la représentation des modèles proposés dans des contextes où les phénomènes hydrologiques extrêmes sont non linéaires et non normales.

Les méthodes non paramétriques, comme les RNA, offrent une interprétation limitée des résultats et demandent une quantité d'information importante afin d'être efficaces. La méthode PPR a été proposée afin de tenir compte de la non-linéarité dans les techniques de prédiction des quantiles de crues au Québec. Cette méthode a permis d'ajuster adéquatement les quantiles de crues estimés aux sites et de mettre en évidence d'importants prédicteurs (ou directions révélatrices). Ces prédicteurs permettent d'obtenir des équations de régression explicites et faciles d'interprétation. Par conséquent, la méthode proposée a conduit à des modèles plus parcimonieux que les modèles additifs et les RNA, sans pour autant sacrifier les performances prédictives.

L'utilisation de la méthode PPR a également été proposée afin de prédire des variables de référence à des sites non jaugés. Ces variables de référence ont servi à construire des voisinages permettant d'identifier la loi régionale de sites non jaugés et de prédire les quantiles de crues désirés. Cette méthode généralise l'approche de l'ACC, en considérant la non-linéarité dans la relation entre les variables hydrologiques et les caractéristiques physiométéorologiques. Dans les cas de l'AFR des

crues du Québec, cette thèse à montrer que l'utilisation de statistiques sommaires, comme les L-moments, conduit à la formation de voisinages plus homogènes. Cette amélioration s'est traduite par une réduction des incertitudes pour les modèles estimés à l'intérieur des voisinages, ce qui a conduit à des performances prédictives améliorées par rapport aux méthodes traditionnelles de l'ACC et ROI.

En AFR, les méthodes géostatistiques sont utilisées afin de prédire à l'intérieur d'espaces transformés des variables hydrologiques à des sites non jaugés. Cette thèse a proposé les copules spatiales comme cadre de travail afin de décrire la dépendance entre les quantiles de crues par des copules. La méthode des copules spatiale s'est montrée supérieure aux techniques de krigeage qui ne permettent pas de tenir compte adéquatement de la non-normalité de la distribution régionale des variables hydrologiques et d'inclure une structure tenant compte de l'hétéroscédasticité. Les résultats obtenus par l'utilisation des copules spatiales sur les données de crue du Québec ont montré d'importantes réductions du biais en plus d'offrir les prévisions les plus performantes de toutes les méthodes considérées dans cette thèse.

Les processus max-stables sont une généralisation de la théorie des valeurs extrêmes. Ces modèles conduisent à une représentation plus fidèle du véritable mécanisme de génération des extrêmes hydrologiques et donc de la dépendance entre les sites. Toutefois, plusieurs difficultés numériques résultent de l'absence d'une forme analytique des probabilités jointes. Ce problème peut être résolu par l'emploi d'une vraisemblance composée ou d'une procédure ABC. Cette thèse a montré à partir d'étude de simulation que l'approche ABC était dans certaines circonstances la meilleure méthode afin d'estimer la structure de dépendance des processus max-stables. Dans le cadre d'une application pratique sur les précipitations extrêmes en Californie, le calcul bayésien approximatif a conduit à des résultats de qualité supérieure à ceux de la vraisemblance composée.

4.2 PERSPECTIVES DE RECHERCHES

Cette thèse a proposé plusieurs méthodes qui permettent d'améliorer les méthodes actuelles rencontrées en AFR. Les résultats retrouvés dans cette thèse mènent à leurs tours à de nouvelles questions qui pourraient servir de pistes de recherche pour des futurs travaux.

L'utilisation des processus max-stables a été exclusivement considérée en AFR pour les précipitations extrêmes. Or les débits de crues peuvent être la conséquence de ces précipitations extrêmes. Par conséquent, il serait intéressant de vérifier si la forme de la dépendance entre les basins jaugées est celle de processus max-stables. Plusieurs questions sur la façon de spécifier cette structure demeurent puisque les débits de rivières ne sont pas continus dans l'espace géographique. Des réponses à ces questions pourraient venir des méthodes géostatistiques en AFR. En particulier, des versions max-stables des techniques de krigeage physiographique et topographique sont envisageables.

L'étude de Durocher *et al.* (2015c) a considéré la question de l'hétéroscédasticité et cette contribution a conduit à de bonnes performances. En général, les techniques de prédiction des paramètres et les modèles probabilistes vont prendre en compte cet aspect puisqu'un paramètre de dispersion est directement estimé. Dans les méthodes traditionnelles, l'utilisation de régions homogènes permet également de tenir compte de l'hétéroscédasticité. Toutefois, les méthodes non paramétriques considérées sur les données de crue au Québec assument une variance constante à l'échelle logarithmique. De travaux futurs devraient être considérés afin d'étudier la nécessité de modéliser spécifiquement la variance pour les méthodes non paramétriques en AFR. En particulier, des approches de modélisation telle que proposée par Fan *et al.* (1998) où la variance est estimée à partir des résidus devraient être validées dans ces situations.

L'étude de Durocher *et al.* (2015b) a montré que la proximité entre des variables de référence conduit à une notion de distance qui est bénéfique dans la formation de voisinages et surpasse la méthode ROI, basée sur la proximité entre les caractéristiques physiométéorologiques. Cette notion

de distance entre variables de référence pourrait également être intégrée aux méthodes géostatistiques en AFR. En effet, l'utilisation de cette approche pourrait conduire à de nouveaux espaces à l'intérieur desquels les méthodes géostatistiques seraient utilisées afin de prédire des variables hydrologiques. En générale, plusieurs autres concepts de proximité pourraient être combinés avec les méthodes géostatistiques en AFR. D'autres exemples seraient la dissimilarité issue des fonctions de profondeur ou à la notion de distance dans l'espace engendré par les directions révélatrices (PPR). Ces approches pourraient représenter des alternatives intéressantes à l'ACC et à l'ACP.

La validation des méthodes proposées dans cette thèse a été effectuée sur des cas d'études. Cette approche montre que les méthodes proposées peuvent être bénéfiques dans certaines situations. Dans de futures recherches, des études de simulations devraient être effectuées afin de déterminer plus précisément les conditions dans lesquelles les méthodes proposées dans cette thèse se démarquent.

REFERENCES

Acreman MC & Sinclair CD (1986) Classification of drainage basins according to their physical characteristics; an application for flood frequency analysis in Scotland. *Journal of Hydrology* 84(3):365-380.

Alila Y (1999) A hierarchical approach for the regionalization of precipitation annual maxima in Canada. *Journal of Geophysical Research D : Atmospheres* 104:31645-31655.

Antoniadis A, Grégoire G & McKeague IW (2004) Bayesian estimation in single-index models. *Statistica Sinica* 14(4):1147-1164.

Archfield SA, Pugliese A, Castellarin A, Skøien JO & Kiang JE (2013) Topological and canonical kriging for design flood prediction in ungauged catchments: an improvement over a traditional regional regression approach? *Hydrology & Earth System Sciences* 17(4).

Aziz K, Rahman A, Fang G & Shrestha S (2014) Application of artificial neural networks in regional flood frequency analysis: A case study for Australia. *Stochastic Environmental Research and Risk Assessment* 28(3):541-554.

Banerjee S, Carlin B & Gelfand A (2004) *Hierarchical modeling and analysis for spatial data.* Chapman & Hall/CRC,

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

Bárdossy A & Li J (2008) Geostatistical interpolation using copulas. *Water Resources Research* 44(7).

Basu B & Srinivas VV (2014) Regional flood frequency analysis using kernel-based fuzzy clustering approach. *Water Resources Research* 50(4):3295-3316.

Beaumont MA, Cornuet J-M, Marin J-M & Robert CP (2009) Adaptive approximate Bayesian computation. *Biometrika* 96:983-990.

Beaumont MA, Zhang W & Balding DJ (2002) Approximate Bayesian Computation in Population Genetics. *Genetics* 162:2025-2035.

Bhaskar N & O'Connor C (1989) Comparison of Method of Residuals and Cluster Analysis for Flood

Regionalization. *Journal of Water Resources Planning and Management* 115(6):793-808.

Bishop CM (1995) *Neural networks for pattern recognition.* Oxford university press,

Blanchet J & Davison A (2011) Spatial modeling of extreme snow depth. *The Annals of Applied Statistics* 5(3):1699-1725.

Blum M & Francois O (2010) Non-linear regression models for Approximate Bayesian Computation. *Statistics and Computing* 20(1):63-73.

Burn DH (1990) An appraisal of the "region of influence" approach to flood frequency analysis. *Hydrological Sciences Journal* 35(2):149-166.

Burn DH & Goel NK (2000) The formation of groups for regional flood frequency analysis. *Hydrological Sciences Journal* 45(1):97-112.

Carlin B & Louis T (2009) *Bayesian methods for data analysis.* Chapman & Hall/CRC,

Carroll RJ, Fan J, Gijbels I & Wand MP (1997) Generalized partially linear single-index models. *Journal of the American Statistical Association* 92:477-489.

Castiglioni S, Castellarin A & Montanari A (2009) Prediction of low-flow indices in ungauged basins through physiographical space-based interpolation. *Journal of Hydrology* 378(3–4):272-280.

Castiglioni S, Castellarin A, Montanari A, Skøien JO, Laaha G & Blöschl G (2011) Smooth regional estimation of low-flow indices: physiographical space based interpolation and top-kriging. *Hydrology and Earth System Sciences* 15(3):715-727.

Chebana F, Charron C, Ouarda TBMJ & Martel B (2014) Regional frequency analysis at ungauged sites with the generalized additive model. *Journal of hydrometeorology* (Accepted).

Chebana F & Ouarda TBMJ (2008) Depth and homogeneity in regional flood frequency analysis. *Water Resources Research* 44(11).

Chokmani K & Ouarda TBMJ (2004) Physiographical space-based kriging for regional flood frequency estimation at ungauged sites. *Water Resources Research* 40(12).

Coles S (1993) Regional Modelling of Extreme Storms via Max-Stable Processes. *Journal of the Royal Statistical Society* Series B (Methodological), 55:797-816.

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

Coles S & Casson E (1998) Extreme value modelling of hurricane wind speeds. *Structural Safety* 20(3):283-296.

Cooley D, Naveau P & Poncet P (2006) Variograms for spatial max-stable random fields. *Dependence in Probability and Statistics,* (Lecture Notes in Statistics: Springer New York, Vol 187. p 373-390.

Cooley D, Nychka D & Naveau P (2007) Bayesian Spatial Modeling of Extreme Precipitation Return Levels. *Journal of the American Statistical Association* 102(479):824-840.

Cunnane C (1988) Methods and merits of regional flood frequency analysis. *Journal of Hydrology* 100(1-3):269-290.

Dalrymple T (1960) Flood-frequency analysis. *Survey Water-Supply Paper* 1543.

Davie T (2008) *Fundamentals of hydrology.* Taylor & Francis,

Davis JM, Eder BK, Nychka D & Yang Q (1998) Modeling the effects of meteorology on ozone in Houston using cluster analysis and generalized additive models. *Atmospheric Environment* 32(14–15):2505-2520.

Davison AC, Padoan SA, Ribatet M & others (2012) Statistical modeling of spatial extremes. *Statistical Science* 27(2):161-186.

Dawson CW, Abrahart RJ, Shamseldin AY & Wilby RL (2006) Flood estimation at ungauged sites using artificial neural networks. *Journal of Hydrology* 319(1–4):391-409.

Demarta S & McNeil AJ (2005) The t Copula and Related Copulas. *International Statistical Review* 73(1):111-129.

Diggle PJ, Tawn JA & Moyeed RA (1998) Model-based geostatistics. *Journal of the Royal Statistical Society: Series C (Applied Statistics)* 47(3):299-350.

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

Dominici F, McDermott A, Zeger SL & Samet JM (2002) On the Use of Generalized Additive Models in Time-Series Studies of Air Pollution and Health. *American Journal of Epidemiology* 156(3):193-203.

El Adlouni S, Ouarda TBMJ, Zhang X, Roy R & Bobée B (2007) Generalized maximum likelihood estimators for the nonstationary generalized extreme value model. *Water Resources Research* 43(3):W03410.

Eng K, Tasker GD & Milly PCD (2005) An Analysis of Region-Of-Influence methods for flood regionalization in the Gulf-Atlantic rolling plain. *Journal of the American Water Resources Association* 41(1):135-143.

Erhardt RJ & Smith RL (2012) Approximate Bayesian computing for spatial extremes. *Computational Statistics & Data Analysis* 56:1468-1481.

Fan J & Yao Q (1998) Efficient estimation of conditional variance functions in stochastic regression. *Biometrika* 85(3):645-660.

Friedman JH & Tukey JW (1974) A projection pursuit algorithm for exploratory data analysis. *Computers, IEEE Transactions on* 100:881-890.

Gaál L, Kysely J & Szolgay J (2008) Region-of-influence approach to a frequency analysis of heavy precipitation in Slovakia. *Hydrology and Earth System Sciences* 12(3):825-839.

Gabriele S & Arnell N (1991) A hierarchical approach to regional flood frequency analysis. *Water Resources Research* 27(6):1281-1289.

Gelman A (2004) *Bayesian data analysis.* CRC press,

Girard C, Ouarda TBMJ & Bobée B (2004) Étude du biais dans le modèle log-linéaire d'estimation régionale. *Canadian Journal of Civil Engineering* 31(2):361-368.

Goovaerts P (2000) Geostatistical approaches for incorporating elevation into the spatial interpolation of rainfall. *Journal of Hydrology* 228(1–2):113-129.

GREHYS (1996a) Inter-comparison of regional flood frequency procedures for canadian rivers. *Journal of hydrology(Amsterdam)* 186:85-103.

GREHYS (1996b) Presentation and review of some methods for regional flood frequency analysis. *Journal of Hydrology* 186(1-4):63-84.

Griffis VW & Stedinger JR (2007) The use of GLS regression in regional hydrologic analyses. *Journal of Hydrology* 344(1):82-95.

Guillemette N, St-Hilaire A, Ouarda TBMJ, Bergeron N, Robichaud É & Bilodeau L (2009) Feasibility study of a geostatistical modelling of monthly maximum stream temperatures in a multivariate space. *Journal of hydrology* 364(1):1-12.

Guisan A, Edwards Jr TC & Hastie T (2002) Generalized linear and generalized additive models in studies of species distributions: setting the scene. *Ecological Modelling* 157(2–3):89-100.

Gupta VK, Mesa OJ & Dawdy DR (1994) Multiscaling theory of flood peaks: Regional quantile analysis. *Water Resources Research* 30(12):3405-3421.

Haan LD (1984) A Spectral Representation for Max-stable Processes. *The Annals of Probability* 12(4):1194-1204.

Haddad K & Rahman A (2012) Regional flood frequency analysis in eastern Australia: Bayesian GLS regression-based methods within fixed region and ROI framework – Quantile Regression vs. Parameter Regression Technique. *Journal of Hydrology* 430–431(0):142-161.

Hall MJ & Minns AW (1999) The classification of hydrologically homogeneous regions. *Hydrological Sciences Journal* 44(5):693-704.

Hamed K & Rao AR (1999) *Flood frequency analysis.* CRC press,

Hardle W, Hall P & Ichimura H (1993) Optimal smoothing in single-index models. *The annals of Statistics* 21(1):157-178.

Hastie T, Tibshirani R & Friedman JH (2009) *The elements of statistical learning: data mining, inference, and prediction.* Springer,

He Y, Bárdossy A & Zehe E (2011) A review of regionalisation for continuous streamflow simulation. *Hydrology and Earth System Sciences* 15(11):3539-3553.

Heagerty PJ & Lele SR (1998) A composite likelihood approach to binary spatial data. *Journal of the American Statistical Association* 93:1099-1111.

Hosking JRM (1990) L-moments: analysis and estimation of distributions using linear combinations of order statistics. *Journal of the Royal Statistical Society* Series B (Methodological), 52:105-124.

Hosking JRM & Wallis JR (1997) *Regional frequency analysis: an approach based on L-moments.* Cambridge Univ Pr,

Hosking JRM, Wallis JR & Wood EF (1985) Estimation of the generalized extreme-value distribution by the method of probability-weighted moments. *Technometrics*:251-261.

Hwang J-N, Lay S-R, Maechler M, Martin RD & Schimert J (1994) Regression modeling in back-propagation and projection pursuit learning. *Neural Networks, IEEE Transactions on* 5(3):342-353.

Jingyi Z & Hall MJ (2004) Regional flood frequency analysis for the Gan-Ming River basin in China. *Journal of Hydrology* 296(1–4):98-117.

Joe H (2014) *Dependence Modeling with Copulas.* CRC Press,

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

Katz RW, Parlange MB & Naveau P (2002) Statistics of extremes in hydrology. *Advances in Water Resources* 25(8–12):1287-1304.

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

Kjeldsen TR & Jones D (2007) Estimation of an index flood using data transfer in the UK. *Hydrological Sciences Journal* 52(1):86-98.

Kjeldsen TR & Jones DA (2009) An exploratory analysis of error components in hydrological regression modeling. *Water Resources Research* 45(2):n/a-n/a.

Kouider A, Gingras H, Ouarda TBMJ, Ristic-Rudolf Z & Bobee B (2002) Analyse fréquentielle locale et régionale et cartographie des crues au Québec. (INRS-ETE, Ste-Foy, Canada.).

Kroll CN & Stedinger JR (1998) Regional hydrologic analysis: Ordinary and generalized least squares revisited. *Water Resources Research* 34(1):121-128.

Kuczera G (1982) Combining site-specific and regional information: an empirical Bayes approach. 18.

Laaha G & Blöschl G (2006) A comparison of low flow regionalisation methods—catchment grouping. *Journal of Hydrology* 323(1–4):193-214.

Laio F, Ganora D, Claps P & Galeati G (2011) Spatially smooth regional estimation of the flood frequency curve (with uncertainty). *Journal of Hydrology* 408(1–2):67-77.

Landwehr JM, Matalas NC & Wallis JR (1979) Probability weighted moments compared with some traditional techniques in estimating Gumbel parameters and quantiles. *Water Resources Research* 15:1055-1064.

Lima CHR & Lall U (2010) Spatial scaling in a changing climate: A hierarchical bayesian model for non-stationary multi-site annual maximum and monthly streamflow. *Journal of Hydrology* 383(3–4):307-318.

Lin G-F & Chen L-H (2006) Identification of homogeneous regions for regional frequency analysis using the self-organizing map. *Journal of Hydrology* 324(1–4):1-9.

Madsen H, Rasmussen P & Rosbjerg D (1997) Comparison of annual maximum series and partial duration series methods for modeling extreme hydrologic events 1. At-site modeling. *Water Resources Research* 33(4):747-757.

Marjoram P, Molitor J, Plagnol V & Tavare S (2003) Markov chain Monte Carlo without likelihoods. *Proceedings of the National Academy of Sciences of the United States of America* 100:15324-15328.

Merz R & Blöschl G (2005) Flood frequency regionalisation—spatial proximity vs. catchment attributes. *Journal of Hydrology* 302(1):283-306.

Micevski T & Kuczera G (2009) Combining site and regional flood information using a Bayesian Monte Carlo approach. *Water Resources Research* 45(4).

Mosley MP (1981) Delimitation of New Zealand hydrologic regions. *Journal of Hydrology* 49(1–2):173-192.

Nathan RJ & McMahon TA (1990) Identification of homogeneous regions for the purposes of regionalisation. *Journal of Hydrology* 121(1–4):217-238.

Neves M & Gomes D (2011) Geostatistics for spatial extremes. A case study of maximum annual rainfall in Portugal**. *Procedia Environmental Sciences* 7(0):246-251.

Nezhad MK, Chokmani K, Ouarda TBMJ, Barbet M & Bruneau P (2010) Regional flood frequency analysis using residual kriging in physiographical space. *Hydrological Processes* 24(15):2045-2055.

Nikoloulopoulos A, Joe H & Li H (2009) Extreme value properties of multivariate t copulas. *Extremes* 12(2):129-148.

Oesting M, Kabluchko Z & Schlather M (2012) Simulation of Brown–Resnick processes. *Extremes* 15(1):89-107.

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

Opitz T (2013) Extremal processes: Elliptical domain of attraction and a spectral representation. *Journal of Multivariate Analysis* 122(0):409-413.

Opsomer J, Wang Y & Yang Y (2001) Nonparametric Regressin with Correlated Errors. *Statist. Sci.* 10.1214/ss/1009213287(2):134-153.

Ouali D, Chebana F & Ouarda TBMJ (2015) Non-linear canonical correlation analysis in regional frequency analysis. *Stochastic Environmental Research and Risk Assessment* 10.1007/s00477-015-1092-7:1-14.

Ouarda TBMJ, Ba KM, Diaz-Delgado C, Carsteanu A, Chokmani K, Gingras H, Quentin E, Trujillo E & Bobee B (2008a) Intercomparison of regional flood frequency estimation methods at ungauged sites for a Mexican case study. *Journal of Hydrology* 348(1-2):40-58.

Ouarda TBMJ, Girard C, Cavadias GS & Bobée B (2001) Regional flood frequency estimation with canonical correlation analysis. *Journal of Hydrology* 254(1-4):157-173.

Ouarda TBMJ & Shu C (2009) Regional low-flow frequency analysis using single and ensemble artificial neural networks. *Water Resources Research* 45(11).

Ouarda TBMJ, St-Hilaire A & Bobée B (2008b) Synthèse des développements récents en analyse régionale des extrêmes hydrologiques. *Revue des sciences de l'eau:/Journal of Water Science* 21(2):219-232.

Oudin L, Kay A, Andréassian V & Perrin C (2010) Are seemingly physically similar catchments truly hydrologically similar? *Water Resources Research* 46(11):n/a-n/a.

Padoan SA, Ribatet M & Sisson SA (2010) Likelihood-based inference for max-stable processes. *Journal of the American Statistical Association* 105:263-277.

Pandey GR & Nguyen VTV (1999) A comparative study of regression based methods in regional flood frequency analysis. *Journal of Hydrology* 225(1–2):92-101.

Rao AR & Srinivas VV (2006) Regionalization of watersheds by fuzzy cluster analysis. *Journal of Hydrology* 318(1–4):57-79.

Reis DS, Stedinger JR & Martins ES (2005) Bayesian generalized least squares regression with application to log Pearson type 3 regional skew estimation. *Water Resources Research* 41(10).

Renard B (2011) A Bayesian hierarchical approach to regional frequency analysis. *Water Resources Research* 47(11):n/a-n/a.

Ribatet M, Cooley D & Davison AC (2009) Bayesian Inference from Composite Likelihoods, with an Application to Spatial Extremes. *ArXiv e-prints*.

Ribatet M, Sauquet E, Grésillon J-M & Ouarda TBMJ (2007) A regional Bayesian POT model for flood frequency analysis. *Stochastic Environmental Research and Risk Assessment* 21(4):327-339.

Robinson JS & Sivapalan M (1997) An investigation into the physical causes of scaling and heterogeneity of regional flood frequency. *Water Resources Research* 33(5):1045-1059.

Salvadori G, De Michele C, Kottegoda N & Rosso R (2007) *Extremes in nature: an approach using copulas.* Springer Verlag,

Sang H & Gelfand A (2009) Hierarchical modeling for extreme values observed over space and time. *Environmental and Ecological Statistics* 16(3):407-426.

Sang H & Gelfand A (2010) Continuous Spatial Process Models for Spatial Extreme Values. *Journal of Agricultural, Biological, and Environmental Statistics* 15(1):49-65.

Sauquet E, Gottschalk L & Leblois E (2000) Mapping average annual runoff: a hierarchical approach applying a stochastic interpolation scheme. *Hydrological Sciences Journal* 45(6):799-815.

Schabenberger O & Gotway CA (2004) *Statistical methods for spatial data analysis.* CRC Press. 512 p

Schlather M (2002) Models for stationary max-stable random fields. *Extremes* 5:33-44.

Schnier S & Cai X (2014) Prediction of regional streamflow frequency using model tree ensembles. *Journal of Hydrology* 517(0):298-309.

Seidou O, Ouarda TBMJ, Barbet M, Bruneau P & Bobée B (2006) A parametric Bayesian combination of local and regional information in flood frequency analysis. *Water Resources Research* 42(11).

Shang H, Yan J & Zhang X (2011) El Niño–Southern Oscillation influence on winter maximum daily precipitation in California in a spatial model. *Water Resources Research* 47(11).

Shaw EM, Beven KJ, Chappell NA & Lamb R (2010) *Hydrology in practice.* CRC Press,

Shu C & Burn D (2004) Artificial neural network ensembles and their application in pooled flood frequency analysis. *Water Resources Research* 40.

Shu C & Ouarda TBMJ (2007) Flood frequency analysis at ungauged sites using artificial neural networks in canonical correlation analysis physiographic space. *Water Resources Research* 43.

Shu C & Ouarda TBMJ (2008) Regional flood frequency analysis at ungauged sites using the adaptive neuro-fuzzy inference system. *Journal of Hydrology* 349(1):31-43.

Skøien JO & Blöschl G (2007) Spatiotemporal topological kriging of runoff time series. *Water Resources Research* 43(9):n/a-n/a.

Skøien JO, Merz R & Blöschl G (2006) Top-kriging - geostatistics on stream networks. *Hydrology and Earth System Sciences* 10(2):277-287.

Smith EL & Stephenson AG (2009) An extended Gaussian max-stable process model for spatial extremes. *Journal of Statistical Planning and Inference* 139(4):1266-1275.

Smith JA (1989) Regional flood frequency analysis using extreme order statistics of the annual peak record. *Water Resources Research* 25(2):311-317.

Smith RL (1990) Max-stable processes and spatial extremes. *Unpublished manuscript*.

Solomatine DP & Dulal KN (2003) Model trees as an alternative to neural networks in rainfall—runoff modelling. *Hydrological Sciences Journal* 48(3):399-411.

Stedinger J & Tasker G (1985) Regional hydrologic analysis: 1. Ordinary, weighted, and generalized least squares compared. *Water Resources Research* 21(9):1421-1432.

Tasker G, Hodge S & Bark S (1996) Region of Influence regression for estimating the 50-years flood

at ungaged sites. *Water Resources Bulletin* 10.1111/j.1752-1688.1996.tb03444.x.

Tasker G & Stedinger J (1989) An operational GLS model for hydrologic regression. *Journal of Hydrology* 111(1):361-375.

Tasker GD (1980) Hydrologic regression with weighted least squares. *Water Resources Research* 16(6):1107-1113.

Thibaud E, Mutzner R & Davison AC (2013) Threshold modeling of extreme spatial rainfall. *Water Resources Research* 49(8):4633-4644.

Tukey J (1975) Mathematics and the picturing of data.1975), p 523-531.

Varin C (2008) On composite marginal likelihoods. *AStA Advances in Statistical Analysis* 92(1):1-28.

Vida S, Durocher M, Ouarda TBMJ & Gosselin P (2012) Relationship Between Ambient Temperature and Humidity and Visits to Mental Health Emergency Departments in Québec. *Psychiatric Services* 63(11):1150-1153.

Wallis JR, Schaefer MG, Barker BL & Taylor GH (2007) Regional precipitation-frequency analysis and spatial mapping for 24-hour and 2-hour durations for Washington State. *Hydrology and Earth System Sciences* 11(1):415-442.

Wang Z, Yan J & Zhang X (2014) Incorporating spatial dependence in regional frequency analysis. *Water Resources Research* 50(12):9570-9585.

Wazneh H, Chebana F & Ouarda TBMJ (2013) Optimal depth-based regional frequency analysis. *Hydrol. Earth Syst. Sci.* 10.5194/hess-17-2281-2013(17):2281-2296.

Weisberg S & Welsh AH (1994) Adapting for the missing link. *The Annals of Statistics* 10.1214/aos/1176325749:1674-1700.

Westra S & Sisson SA (2011) Detection of non-stationarity in precipitation extremes using a max-stable process model. *Journal of Hydrology* 406(1-2):119-128.

Wiltshire SE (1985) Grouping basins for regional flood frequency analysis. *Hydrological Sciences Journal* 30(1):151-159.

Wittenberg H (1999) Baseflow recession and recharge as nonlinear storage processes. *Hydrological Processes* 13(5):715-726.

Zhang X, Wang J, Zwiers FW & Groisman PY (2010) The Influence of Large-Scale Climate Variability on Winter Maximum Daily Precipitation over North America. *Journal of Climate* 23(11):2902-2915.

CHAPITRE 2 :

A NONLINEAR APPROACH TO REGIONAL FLOOD FREQUENCY ANALYSIS USING PROJECTION PURSUIT REGRESSION

A Nonlinear Approach to Regional Flood Frequency Analysis Using Projection Pursuit Regression

Martin Durocher [*,1], Fateh Chebana [1], and Taha B. M. J. Ouarda[2]

[1]Institut National de Recherche Scientifique (INRS-ETE),
University of Québec
490 de la Couronne, Québec G1K 9A9, Canada

[2] Institute Center for Water Advanced Technology and Environnemental Research (iWater),
Masdar Institue of Science and Technology
P.O. Box 54224, Abu Dhabi, UAE

(Accepted Journal of Hydrometeorology)

ABSTRACT

This paper presents an approach for Regional Flood Frequency Analysis (RFFA) in the presence of nonlinearity and problematic stations, which requires adapted methodologies. To this end, we propose the Projection Pursuit Regression (PPR). The latter is a family of regression models that applies smooth functions on intermediate predictors to fit complex patterns. The PPR approach can be seen as a hybrid method between the Generalized Additive Model (GAM) and the Artificial Neural Network (ANN), which combines the advantages of both methods. On one hand, the PPR approach has the structure of a GAM to describe nonlinear relations between hydrological variables and other basin characteristics. On the other hand, PPR can consider interactions between basin characteristics to improve the predictive capabilities in a similar, but simpler way to ANN. The methodology developed in the present study is applied to a case study represented by hydrometric stations from Southern Quebec, Canada. It is shown that flood quantiles are mostly associated to a dominant intermediate predictor, which provides a parsimonious representation of the nonlinearity in the flood generating processes. The model performance is compared to eight other methods available in the literature for the same dataset, including GAM and ANN. When using the same basin characteristics, the results indicate that the simpler structure of PPR does not affect the global performance and that PPR is competitive with the best existing methods in RFFA. Particular attention is also given to the performance resulting from the choice of the basin characteristics and the presence of problematic stations.

Keywords: Flood, Projection Pursuit Regression, Québec, GAM, ANN, Regional Frequency Analysis, Ungauged Sites, Non-linear Process.

1. INTRODUCTION

A T-year return period is an essential measure for the design and the management of water resources. In practice, the necessary information is not always available at the desired sites. To this end, adapted methodologies are required to transfer information from gauged stations to ungauged locations. At ungauged locations, river discharge data is not available whereas basin characteristics are usually available. Consequently, various approaches have been developed for predicting the behaviour of hydrological variables at ungauged locations on the basis of the relation they share with their basin characteristics. Regional Flood Frequency Analysis (RFFA) is the appropriate framework for such applications.

In traditional methods (Burn, 1990; Eng et al., 2007; Hosking and Wallis, 1997; Ouarda et al., 2001; Reis et al., 2005), RFFA is decomposed into two main steps. The initial step consists in pooling stations into homogenous regions. Afterwards, regional estimation of flood quantiles is usually performed by multiple regression techniques. This approach assumes a linear association between the hydrological variables and the basin characteristics. However, nonlinear models may be more justified as hydrological processes are naturally nonlinear (Wittenberg, 1999). In RFFA, Chebana et al. (2014) investigated Generalized Additive Models (GAM), where basin characteristics are related to hydrological variables by individual smooth functions. This approach provides a more realistic way of taking into account the effect of basin characteristics on the hydrological variables with high performances.

Other approaches that can account for nonlinearity are based on machine learning, which designates a family of data-driven algorithms that are usually considered for large datasets in which complex patterns may be learned. Artificial Neural Networks (ANN) is a machine learning method that has been already used in RFFA. However, since ANN calibration requires large datasets, it is employed in RFFA for regions with a large number of stations (Dawson et al., 2006; Shu and Burn,

2004). Similarly to traditional methods, ANN consists in a prediction model that estimates desired quantiles at ungauged locations from available basin characteristics (Ouarda and Shu, 2009). ANN has the appealing property of being able to approximate any continuous surface with desired accuracy (Bishop, 1995). However, ANN is a non-parametric model that does not lead to explicit regression equations and its calibration is not an easy task for non-initiate users as some problems may occur if proper guidelines are not followed (Hastie et al., 2009; Khalil et al., 2011). Part of the difficulties arises from the overparametrized nature of ANN, which implies the existence of several locally optimal estimators. Therefore, the convergence of standard numerical algorithms to a global solution is not possible. Hence, ANN must be fitted several times with different starting values. Afterwards, ANN requires usually an ensemble strategy for merging the individual ANNs. Comparative studies, such as Shu and Burn, (2004), provided guidelines for better fitting and assembling ANN in RFFA. However, as far as we know, the complexity of the ANN structure leads to models which do not provide simple and direct understanding of the nonlinearity in the flood generating processes.

Several methods apply predictive models not directly to the observed variables but to intermediate predictors that are linear combinations of basin characteristics. For instance, principal component regression corresponds to multiple regression that is performed on the outputs of a principal component analysis (Hastie et al., 2009). In this method, the outputs of principal component analysis are the intermediate predictors and the purpose of this substitution is to overcome multicollinearity problems. Other examples that consider intermediate predictors in RFFA are spatial methods (Archfield et al., 2013; Castiglioni et al., 2009; Chokmani and Ouarda, 2004). In these approaches, the intermediate predictors form the basis of a subspace of the basin characteristics in which hydrological variables can be considered as continuous. Hence, in these spaces usual interpolation methods represent natural ways of prediction at ungauged locations. Moreover, comparative studies show that spatial methods have a competitive performance with the traditional RFFA methods and are robust to the quality of the data (Archfield et al., 2013; Ouarda et al., 2008). A

last example is ANN, for which the neurons (the basic elements) correspond to the transformation of an intermediate predictor by specific activation functions (Bishop, 1995).

The last examples show how intermediate predictors can be useful for dealing with multicollinearity or improving prediction power. However, except for ANN, these intermediate predictors are obtained and fit by separate procedures that do not share a common goal (e.g. principal component analysis is performed on inputs without considering the output of the regression model). On the other hand, ANN does jointly estimate the intermediate predictors, but the resulting structure may be overparametrized and does not provide a parsimonious representation that helps to understand the studied phenomenon. In this study, the rationale for using intermediate predictors, instead of the basin characteristics, is investigated from a class of regression models, called projection pursuit regression (PPR) (Friedman et al., 1983; Friedman and Stuetzle, 1981; Hwang et al., 1994; Roosen and Hastie, 1994). The structure of a PPR model consists in fitting smooth functions on intermediate predictors, which is similar to fitting a GAM model on the coordinates of a physiographical space. A particularity of PPR is that both smooth functions and intermediate predictors are jointly estimated within the model. Moreover, in special cases PPR can be sufficiently parsimonious to provide a meaningful structure.

The first objective of this research is to investigate the predictive performance of PPR in comparison to other methods, like GAM and ANN. A second objective is to verify if PPR can bring further understanding of the nonlinearity in the association between flood quantiles and basin characteristics. A third objective of this study is to assess the robustness of PPR in the presence of problematic stations. The present paper is organized as follows: The description of the PPR methodology is developed in section 2. Section 3 illustrates a practical application of PPR on hydrometric stations from Southern Quebec, Canada. Finally, section 4 provides concluding remarks on the present work.

2. METHODOLOGY

Regression models are usually applied in RFFA to predict flood quantiles (output variables) from relevant basin characteristics (input variables). One of the most commonly used regression models for this purpose is the log linear model (Ouarda et al., 2001; Pandey and Nguyen, 1999). In this model, the logarithm transformation has the effect of stabilizing the variance, which otherwise increases with the return level of flood quantiles. Hence, a PPR methodology adapted to flood quantiles should be performed on the logarithm scale. Alternatively, instead of quantiles, other hydrological variables such as empirical moments or parameters of at-site distributions, may be predicted and then combined to predict flood quantiles (Haddad et al., 2014; Haddad and Rahman, 2012). Nevertheless, the comparison between regressions of the at-site parameters versus regression of the flood quantile is not part of the present objectives. The direct modelling of the at-site flood quantiles is investigated.

2.1 PROJECTION PURSUIT REGRESSION

The simplest form of the PPR model is similar to a linear regression model and can be applied to characterize the link between a flood quantile $Y \in \mathbb{R}$ and basin characteristics $X = \left(X_1, \ldots \right.$ $\mathbb{R}$. First, denote $\alpha = \left(a_1, \ldots \right.$ $\mathbb{R}$ a vector called direction representing a set of coefficients such that:

$$\sum_{j=1}^{p} a_j^2 = 1 \qquad (1)$$

The purpose of the direction α is to define a new (intermediate) predictor $\alpha' X$ that can be fitted by a smooth function g where the simplest PPR model for RFFA has the form:

$$\log(Y) = g\left(\alpha' X\right) + \varepsilon \qquad (2)$$

where ε is an error term with zero mean and constant variance. The absolute value $|a_j|$ is thus associated to the variable X_j and indicates the relative importance of the basin characteristic X_j. The smooth function g can have different formulations, such as polynomials or spline polynomials among others. More details on the choice of the smooth function g is provided in Section 2.3.

Joint estimation of α and g distinguishes PPR from regression methods applied on a known intermediate predictor $\alpha' X$. The smooth function g in (2) can be optimized with respect to α. However, there is only one direction for which the fitting criterion is optimized with respect to both α and g (Yu and Ruppert, 2002). Estimation corresponds to a nonlinear least squares problem that can be solved by standard algorithms found in most numerical software, such as R and MATLAB. The quality of the model depends on the smoothness of g. In one hand, if the shape of g is too wiggly, the data will be overfitted and the model will have an unnecessary large variance. On the other hand, putting too many restrictions on g will decrease the variance, but increase the bias. This is a classical situation of trade-off between bias and variance for which objective criteria are proposed for guiding the regularization of the model. Popular choices are the cross-validation score, the generalized cross-validation score and the Akaike information criterion (Hastie et al., 2009).

Model (2) can be considered as a regression model where the conditional mean is $h(X) = g(\alpha' X)$. For more flexibility, the general PPR model includes the sum of $k = 1,...,q$ similar functions called terms (based on the same vector *X*):

$$h_k(X) = \beta_k g_k(\alpha_k' X) \tag{3}$$

Accordingly, the PPR model is written as:

$$\log(Y) = \mu + \sum_{k=1}^{q} h_k(X) + \varepsilon \tag{4}$$

where μ represents the overall mean. Each term includes a direction vector α_k, a smooth function g_k and a scale factor β_k. For the PPR model in (4), each direction α_k must respect condition (1). Additionally, two constrains are imposed to the smooth functions:

$$\int_{-\infty}^{\infty} g_k(u)\,du = 0 \tag{5}$$

$$\int_{-\infty}^{\infty} g_k^2(u)\,du = 1 \tag{6}$$

where u is in the domain of g_k for each k. Conditions (5) and (6) insure respectively that each smooth function g_k has zero mean and a unit scale. These constraints are necessary for the uniqueness of the components of the model. Therefore, the scale factor β_k represents the importance of the k-th terms. In general, the intermediate predictors $\alpha_k{}'X$ are correlated, which implies that PPR models can account for the interactions between predictors (Hastie et al., 2009), e.g. basin characteristics. Such interactions may occur when the combining effect of basin characteristics has different effect on the flood quantile in comparison with their individual sums.

The structure of PPR models is polyvalent and can be related to other regression methods. The PPR model with a single term is closely related to the generalized linear model (GLM) (Nelder and Wedderburn, 1972). As shown by (2) the smooth function g is the inverse of the link function of a GLM, which relates the response variable to a linear combination of explanatory variables. Consequently, a PPR model with a single term may be seen as a GLM for which a smooth link function is specified. However, notice that g is estimated in PPR whereas in GLM g is given. This model may be useful to produce visual diagnostics of existing link functions (like logarithm) or to provide alternatives if no link functions are appropriate (Weisberg and Welsh, 1994).

Figure 1 presents a diagram illustrating the structure of the PPR model as a network. It shows the close connection between PPR and single hidden layer ANN as they are both represented by the same network (Bishop, 1995). Like ANN, PPR with multiple terms can take advantage of the

combination of several intermediate predictors to fit complex patterns. Indeed, PPR can approximate, with the desired precision, any regression models having a continuous mean (Hastie et al., 2009). However, an advantage of PPR is that the smooth functions are not imposed/selected by the user and the estimation procedure uses a specific algorithm.

2.2 ESTIMATION

The least squares method is the usual way of fitting a PPR model. Let y_i be a flood quantile observed at a site i with basin characteristics $\mathbf{x}_i$. Accordingly, the residuals of a PPR model are $\mathbf{e} = (e_1, \ldots$ where:

$$e_i = \log(y_i) - \mu - \sum_{k=1}^{q} g_k(\alpha^T \mathbf{x}_i) = \log(y_i) - \mu - \sum_{k=1}^{q} h_k(\mathbf{x}_i) \qquad (7)$$

In matrix notation, the ordinary least squares criterion, the residual sum of squares, is written as $RSS = \mathbf{e}'\mathbf{e}$ and must be optimized over μ and h_k. Recall that the latter includes the directions α_k, the smooth functions g_k and the scale factors β_k. Consequently, except for a single term model, PPR models are generally overparametrized with multiple local solutions. To find an adequate solution, PPR relies on the following estimation algorithm:

For $k = 1,...,q$,

1. Fit h_k on the residuals of partial model $\mu + \sum_{j=1}^{k-1} h_j(X)$
2. Optimize RSS for h_k, with $\{h_j\}_{j \neq k}$ fixed until RSS converges

The first step aims to find a proper starting solution. This is done by fitting a series of single term models on the residuals of the previous fit. Afterwards, the second step updates separately each one of the terms h_k. Therefore, the updating process is repeated until a solution is found. It is important to remember that the estimation algorithm does not ensure a global optimum, but aims at converging to a relatively good solution. A property of this estimation algorithm is that it tends to provide terms h_k of

decreasing importance ($\beta_k \geq \beta_{k+1}$) (Hwang et al., 1994). This ordering makes it more likely to keep terms h_k with more meaningful interpretation. This property may be useful in the RFFA context to extract important components of the relation between flood quantiles and basin characteristics. Notice that some variants of the estimation algorithm are also proposed to reduce possibilities of converging to poor local estimate. In this regard, more details can be found in Hwang et al. (1994) as well as Roosen and Hastie (1994).

2.3 CALIBRATION

As indicated in (4), the calibration of a PPR model requires the prior selection of the number q of terms. Since PPR can easily overfit the data, this choice should be guided by an objective criterion. Except for cross-validation, popular criteria require a measure of complexity that is not generally available for PPR (Hastie et al., 2009). For instance, the Akaike Information Criterion (AIC) penalized the fitting of a model according to the number of parameters. Consequently, the leave-one-out cross-validation is adopted in the present study. In turn, each gauged station is considered as ungauged and a PPR model is fitted on the remaining stations to obtain the predicted values $\hat{y}_i$ of y_i . The collection of all predicted values is used to evaluate the Nash-Sutcliffe at the logarithm scale:

$$NASH = 1 - \frac{\sum_{i=1}^{n}\left(y_i^* - \hat{y}_i^*\right)^2}{\sum_{i=1}^{n}\left(y_i^* - \overline{y}^*\right)^2} \qquad (8)$$

where $y_i^*, \hat{y}_i^*$ are the logarithms of $y_i, \hat{y}_i$ and $\overline{y}_i^*$ is the empirical mean of y_i^* . Note that the Nash-Sutcliffe criterion is a standardized version of the mean square (additive) errors (Schaefli and Gupta, 2007) and that the errors of the PPR model in (4) are additive at the logarithm scale. This motivates the calculation of the $NASH$ on the y_i^* instead of the y_i . In the following, the $NASH$ criterion is adopted as guideline for calibrating the PPR model. To evaluate the predictive performance at the original scale, the relative root mean square error is used:

$$RMSEr = \sqrt{\frac{1}{n}\sum_{i=1}^{n}\left(\frac{y_i - \hat{y}_i}{y_i}\right)^2} \quad (9)$$

Additionally, the relative bias is examined for detecting systematic errors:

$$BIASr = \frac{1}{n}\sum_{i=1}^{n}\left(\frac{y_i - \hat{y}_i}{y_i}\right) \quad (10)$$

The nature of the smooth functions g_k can take different forms. Friedman and Stuetzle (1981) used what they called a “super-smoother”, which consists of a local polynomial with specific rules to regularize the smoothness of the g_k. Other choices of smooth functions are polynomials (Hwang et al., 1994) and spline functions of a limited degree of freedom (Friedman et al., 1983). In this study, cubic B-spline functions (Hastie et al., 2009) with equally space knots are considered. The latter choice, as well as the degree of freedom value, are based on cross-validation.

The inclusion of basin characteristics that are not significant adds unnecessary complexity. This effect may be particularly important for PPR models as their estimation is an overparametrized least squares problem. Indeed, notice that if each smooth function is determined by m parameters, or degrees of freedom, the PPR model in (4) has $q(m+p)+1$ parameters where p is the number of basin characteristics and q the number of terms. Consequently, the unnecessary basin characteristics may reduce the quality of the estimation even though the model is more general. For this reason, either close prior examination of the basin characteristics should be done or PPR should be combined to stepwise selection of basin characteristics (Hastie et al., 2009). In this study, forward-stepwise is preferred to backward-stepwise for limiting the complexity throughout the procedure. The strategy consists in testing individually the impact of each basin characteristic on the cross-validation criteria. Afterwards, the basin characteristic that best improves the model is added permanently and the procedure is repeated until no basin characteristic can improve the model.

2.4 EXTENSIONS

In RFFA the hydrological data are not direct observations, but at-site flood quantiles. They are hence the output of at-site flood frequency analysis, and may possess different levels of variability resulting from at-site modelling errors and unequal record lengths. In this context, the ordinary least squares estimator remains unbiased but does not have minimal variance (Pandey and Nguyen, 1999). Moreover, the cross-correlations between gauged stations and the skewness of the at-site distributions represent some aspects that may also affect the quality of the least squares estimator (Griffis and Stedinger, 2007). For these reasons, several researches adopt instead the Generalized Least Squares (GLS) framework (Eng et al., 2007; Haddad et al., 2013; Reis et al., 2005).

In the GLS framework the logarithms of flood quantiles $\log(y_i)$ are not direct observations, rather estimates of hydrological variables with sampling errors η_i. Hence, the total error in (4) is the sum $\varepsilon_i = \eta_i + \delta_i$ where δ_i are the model errors attributed to the fitting of the PPR model. Assuming that the model errors are normally and independently distributed with variance σ_δ^2, the covariance matrix of the total errors is:

$$\mathbf{V} = E\left(\varepsilon'\varepsilon\right) = \sigma_\delta^2 I + \Sigma_\eta \qquad (11)$$

where I is the identity matrix and Σ_η is the covariance matrix of the sampling errors (Stedinger and Tasker, 1985). The form of Σ_η reflects the results of the at-site frequency analysis and must be specified before the estimation of the model. Practical guidelines to incorporate cross-correlation and unequal sampling errors in Σ_η are provided for instance by Griffis and Stedinger (2007) and Kjeldsen and Jones (2009).

Section 2.2 indicates that the estimation of a PPR model is conducted by a specific algorithm using ordinary least squares, and is based on $RSS = \mathbf{e}'\mathbf{e}$. If desired, the estimation algorithm of section 2.2 can be adapted directly where $RSS = \mathbf{e}'\mathbf{e}$ is replaced by $RSS = \mathbf{e}'\mathbf{V}^{-1}\mathbf{e}$. However, including

correlation in nonparametric regression techniques such as PPR and GAM, may result in a number of fundamental problems. The principal issue is that persistence due to either trend or correlation may not be identifiable (Opsomer et al., 2001). Consequently, iterative procedures as proposed by (G. Tasker and Stedinger, 1989), designed to estimate simultaneously a regression model with its correlation structure, should be performed with care.

3. CASE STUDY

3.1 DATA

The adaptation of PPR to the prediction of flood quantiles is applied to a case study including 151 hydrometric stations managed by the ministry of the environment of Quebec, Canada. These hydrometric stations were part of an at-site frequency analysis described in Chokmani and Ouarda (2004). Precisely, different estimation methods or distributions can be selected for each gauged station. The final choices are guided by visual diagnostics and the Akaike information criterion. In most cases maximum likelihood theory is used to estimate the parameters of the best distribution among common distributions, including the generalized extreme values, Log-normal and Pearson III distributions. The results of the at-site frequency analysis have since been used in several other studies that serve as reference for this study (e.g. Chebana and Ouarda, 2008; Nezhad et al., 2010; Shu and Ouarda, 2007).

Figure 2 presents a map of the location of the hydrometric stations located between the 45th and the 55th parallel North in the Southern part of the province of Quebec, Canada. The selected stations meet the conditions of having a natural flow regime with at least 15 years of recorded data. Moreover, the stations respect the usual conditions of stationarity, homogeneity and independence. After further examination, Chokmani and Ouarda (2004) indicated that 6 gauged stations are problematic and lead to important discrepancies in the predictions. For 4 of them, this was explained

by an undervaluation of the true drainage area, while for the other 2 gauged stations, the difficulties were attributed to the overvaluation of the true percentage of drainage area covered by lakes. These problematic stations are maintained in the dataset, but are examined closely.

In the present case study, we consider prediction using PPR for the flood quantiles, especially for return periods 10 and 100 years. In Canada, it is shown that river discharges have a strong log-log relationship with the drainage areas (Eaton et al., 2002; Ribeiro and Rousselle, 1996). Therefore, flood quantiles are standardized by their drainage areas. The obtained specific quantiles QS10 and QS100, allow reducing the scaling effect and provide a better understanding of the impact for the other basin characteristics. At each gauged station, twelve basin characteristics are available and defined in Table 1.

3.2 COMPARED MODELS

Comparisons are performed first between variants of the PPR model and then with other available models in RFFA. According to the choices of basin characteristics and included hydrometric stations, three PPR models are considered. In addition, the PPR models are compared to eight other models reported in previous literature using the same dataset. These eight models are part of four families of regional models that are discussed in the introduction: traditional methods, spatial methods, GAM and ANN. To ensure comparability with the previous studies, ordinary least squares theory is also employed for all PPR regional models as described in section 2.2.

The three PPR regional models are denoted PPR_ALL, PPR_PBL and PPR_STW. For comparison with previous works, the PPR_ALL regional model is fitted using the same basin characteristics and the same prior transformations (i.e. Shu and Ouarda, 2007). The latter being the logarithm function except for the fraction of basin occupied by lake (PLAC), for which the square root is preferred. Evaluation of the quality of the PPR_ALL model is the main interest of this study. To measure the impact of problematic stations on the cross-validation criteria, PPR_PBL designates the PPR regional model with the same basin characteristics as PPR_ALL, but calibrated excluding the

problematic stations identified by Chokmani and Ouarda (2004). Hence, PPR_PBL is performed on 145 sites. On the other hand, PPR_STW designate the regional model (including problematic stations) calibrated using a stepwise procedure to select the basin characteristics. The purpose of PPR_STW is to verify if the selections of the basin characteristics is appropriate for PPR_ALL. Moreover, the PPR_STW includes the geographical coordinates: LON and LAT, which is motivated by the fact that combining similarity between the physiographical characteristics and geographical proximity can improve the prediction of regional models (Chebana et al., 2014; Eng et al., 2007).

Among the eight models selected for comparison with the PPR regional models, two traditional methods are considered and employed with two different methods to delineate the homogenous regions. A well-known method for delineation is Canonical Correlation Analysis (CCA), which pools together the closest gauged stations to the ungauged location in the canonical space (Ouarda et al., 2001). Alternatively, depth function is a statistical notion that provides a center-outward ordering in multidimensional space (Tukey, 1975) and can be employed to identify the nearest gauged stations to an ungauged location (Chebana and Ouarda, 2008). More recently, depth function is employed as an important ingredient to define new methods in RFFA (e.g. Wazneh et al., 2013a, 2013b). In the analysis of the Quebec dataset by spatial methods, two methods of interpolation are considered. They are the ordinary kriging and the residuals kriging in the canonical space as proposed by Chokmani and Ouarda (2004) and Nezhad et al. (2010). Chebana et al. (2014) used the GAM model to carry out the RFFA of the Quebec gauged stations. In the latter, the analysis is provided with the same basin characteristics as PPR_ALL, but also in combination with a stepwise procedure. Shu and Ouarda (2007) performed a RFFA using predictions from a single ANN and by aggregating several ANNs

3.3 RESULTS

The regional models PPR_ALL and PPR_STW are described to provide a better illustration of the components of a PPR model. Due to space limitation and because its estimation is closely similar to PPR_ALL, further description of the PPR_PBL model is not reported. Table 2 presents the direction

vectors α_k along with the scale factors β_k. The smooth functions for PPR_ALL are calibrated with five degrees of freedom as described in section 2.3. For QS10 obtained from PPR_ALL, only one single term is selected by cross-validation. The coefficients of the direction are $\alpha_1 = \left(a_{1,1}, a_{1,2}, a_{1,3}, a_{1,4}, a_{1,5}\right)$ and the coefficient associated to the basin characteristics PLAC is $a_{1,3} = -0.86$, which is the largest one in absolute value. Accordingly, the model with a single term as in (2) can be written as:

$$\log(QS10) = 0.61\, g\Big[-0.21\log(BV) + 0.09\log(PMBV) - 0.86\sqrt{PLAC} + 0.42\log(PTMA) + 0.18\log(DJBZ)\Big] \quad (12)$$

where g is the smooth function. In absolute value, PLAC represents the highest coefficients, which indicates the dominant effect of PLAC in the prediction of flood quantiles. On the other hand, the intermediate predictor $\alpha_1' X$ is different for the two return periods. Indeed, for QS100 the coefficient of PTMA ($a_{1,4} = 0.63$) is equivalent in absolute value to the one of PLAC ($a_{1,3} = -0.64$). Moreover, the difference of sign indicates that both basin characteristics have opposite effects on flood quantiles. With scale factors $\beta_1 > \beta_2$ for QS100, it shows that the first term represents a more important component of PPR_ALL.

The PPR_STW regional model selects the basin characteristics by a stepwise procedure and results in a different selection than PPR_ALL as shown in Table 2. This new selection of basin characteristics is interesting as it leads to a PPR_STW regional model with only a single term for both QS10 and QS100. The smooth functions of PPR_STW are calibrated with five degrees of freedom. In comparison with PPR_ALL, it is seen that the basin characteristics are more coherent in both return periods as suggested by the greater similarity between the direction vectors α_1.

In previous studies, the geographical coordinates were not considered and the selection of significant basin characteristics were based on a correlation analysis (Chokmani and Ouarda, 2004), which does not account for nonlinearity. The basin characteristics selected by PPR_STW are more coherent with the results of a GAM model using stepwise selection of the basin characteristics

(Chebana et al., 2014). Both models include the following basin characteristics: BV, PLAC, PLMA, LON. Among them, only BV and PLAC are also considered by Chokmani and Ouarda (2004) and the other studies on the same database. The longitude (LON) may be indirectly related to different factors that are not actually measured by the other basin characteristics. For instance, in Southern Quebec the longitude is an indicator of the proximity to the Atlantic Ocean and thus to its influence on the local climate. Also, for QS100, both GAM and PPR_STW models prefer to consider the mean annual liquid precipitation (PLMA) instead of the mean annual total precipitation (PTMA), which may suggest the importance of distinguishing between floods generated by intense rainfall versus spring snowmelt.

In this study, PPR aims at giving another point of view to the nonlinearity in the process generating floods. The smooth functions $b_k g_k$ are presented in Figure 3 as scatterplots of the centered flood quantiles with respect to the predictors $\alpha_k' X$. Regarding Term 1, sub-Figures 3A,B,D,E show that the first smooth function $b_1 g_1$ approximately has a S-shape, which testifies to the nonlinearity in the data. Moreover, a new fitting of the smooth function without the problematic stations (dashed lines in Figure 3) shows that these stations have a considerable influence on the shape of the smooth functions $b_1 g_1$ for the lowest values of the intermediate predictor $\alpha_1' X$.

For QS100, the smooth function of the second term is presented in sub-Figure 3C. It is seen that in the middle section, the smooth function has no trend and high variability, which suggests that the second term, mostly influenced by PTMA and DJBZ (see associated coefficients in Table 2) brings no additional information to the first term in sub-Figure 3B. On the other hand, examination the left part of sub-Figure 3C shows lower variability, which for few sites indicates that some information is missed by the first term.

Visual diagnostics for the residuals of PPR_ALL are presented in Figure 4 to assess their quality and the impact of the problematic stations. The examination is performed on the relative residuals (sub-Figures 4A,B) as well as the standardized residuals on the logarithm scale (sub-Figures 4C,D). These plots show that most of the problematic stations correspond to atypically large residuals.

Notice that the relative residuals are calculated at the logarithm scale. Hence, their comparison with the standard residuals reveals that the logarithm transformation has an amplified effect on the lowest residuals. Indeed, for the same sites, sub-Figures 4A,B exhibit more atypical values than sub-Figures 4C,D. This implies that the criterion $RMSEr$ (based on the relative residuals) is more influenced by the problematic stations than the $NASH$ criterion (based on the residuals at the logarithm scale).

The performances of the 3 regional models are compared and the results of these comparisons are reported in Table 3. With respect to PPR_ALL, removing the problematic stations (PPR_PBL) leads to *NASH* improvements of 5% (QS10) and 9% (QS100). The same pattern remains true when $RMSEr$ is considered. These differences of performance between PPR_ALL and PPR_PBL quantify the impact of the problematic stations in the cross-validation performance as anticipated by the examination of the residuals in Figure 4. Similarly, the stepwise selection (PPR_STW) is associated to improvements of 6% (QS10) and 8% (QS100) in comparison with PPR_ALL, even though the problematic stations are included. Notice that the regional model PPR_STW uses a single term for both flood quantiles conversely to PPR_ALL that requires 2 terms for QS100. This indicates that with a proper selection of basin characteristics, the return level of the flood quantiles in Southern Quebec can be successfully associated to a single intermediate predictor.

The results of previous studies can serve as reference to evaluate the relative quality of the PPR approach. The corresponding cross-validation criteria values are presented in Table 4. Of all methods considered, PPR_STW is the model that has the best performance in terms of $RMSEr$. However, notice that PPR_STW and GAM-stepwise use automatic selection of the basin characteristics. Consequently, for these two regional models the difference of performance cannot be uniquely attributed to the method of prediction, but also to the selection of more adequate basin characteristics. When considering the same basin characteristics and the same dataset, PPR_ALL leads to competitive predictive performance with the best other methods. Only two methods have slightly better $RMSEr$: the ensemble ANN in the CCA-space and the traditional method with depth

functions. This represents differences in the $RMSEr$ of respectively 3 % and 4% for QS100. PPR_ALL also has similar $RMSEr$ to GAM, but, when a stepwise procedure is used (PPR_STW), PPR performs slightly better than GAM-stepwise. In terms of bias, it can be seen that overall PPR performs similarly to the GAM and the ANN approaches.

The previous studies listed in Table 4 have not provided details on the impact of the problematic stations. Consequently, it is not possible to know if the differences of $RMSEr$ are caused by a general improvement over all gauged stations or if it is mostly influenced by the individual fit of some problematic stations. For instance, Chokmani and Ouarda (2004) have reported that for QS100, the $RMSEr$ obtained for ordinary kriging drops from 70% to 41% when the problematic stations are removed. Consequently, when all hydrometric stations are included, the difference of $RMSEr$ between PPR_ALL and ordinary kriging is 22%. However, without the problematic stations, the difference with PPR_PBL drops to 7%. This shows that without consideration of the problematic stations, the variation of the $RMSEr$ can be misleading on the true gain of performance and more details are necessary.

GAM was very recently introduced in RFFA to provide a better understanding of the nonlinearity in the association between flood quantiles and basin characteristics (Chebana et al., 2014). Because both GAM and PPR use smooth functions to account for nonlinearity, the comparison between them allows to quantify the advantage of considering intermediate predictors (i.e. $\alpha_k X$). For PPR_STW, only a single smooth function is used, while GAM-stepwise requires one smooth function f_k for each basin characteristic. Accordingly, the formulation of PPR_ALL model for QS10 in (12) can be compared the GAM model:

$$\begin{aligned}\log(QS10) &= f_1\left[\log(BV)\right] + f_2\left[\log(PMBV)\right] + f_3\left[\sqrt{PLAC}\right] \\ &\quad + f_4\left[\log(PTMA)\right] + f_5\left[\log(DJBZ)\right]\end{aligned} \tag{13}$$

Consequently, PPR can be seen as a more parsimonious model than GAM, without loss of predictive power (see Table 4).

Figure 3 shows that the relation between flood quantiles and basin characteristics corresponds to a smooth nonlinear curve, or surface, for which PPR provides a clear analytic formulation (12). GAM provides a similar representation (13) and both aims to obtain the smoothest possible surface during the calibration, without sacrificing performance. When a linear regression is applied on a subset of gauged stations, it achieves a linear approximation of this nonlinear surface for a restricted zone. The comparison in Table 4 indicates that different approaches for the delineation of homogeneous regions lead to different results. Indeed, for QS100 the traditional methods achieved a $RMSEr$ of 51% with CCA delineation and 44% with depth function. On the opposite, PPR and GAM have respectively a $RMSEr$ of 48% and 49%, which suggests that searching for the smoothest nonlinear surface is a more rigorous approach, because it does not depend on the type of homogenous regions and thus leads to less subjective results.

The estimations of PPR as well as ANN models are overparametrized least squares problems. The single ANN method in Table 4 corresponds to the case where only the best of 15 ANNs is kept. In terms of $RMSEr$, it is seen that PPR_ALL has better accuracy. PPR_ALL can also be seen as more parsimonious model than the single ANN model, because it uses a single nonlinear function in comparison to 5 neurones. Hence it shows that the estimation algorithm is able to identify directly a better model without several trials. In the present case study, regional models using PPR mostly have a single term. By analogy, the single ANN regional model used by Shu and Ouarda (2007) has 5 neurons, which implies that 5 intermediate predictors are involved. For the ensemble ANN approaches, it is 75 intermediate predictors that are involved due to the bagging of 15 ANNs. This study does not provide a thorough comparison of PPR and ANN in terms of performance, but shows that PPR can reach a favorable balance between parsimony and performance.

4. CONCLUSIONS

The present work investigates the adaptation of the PPR approach to RFFA to deal with the nonlinearity of the hydrologic processes involved and to properly handle problematic stations. Namely, PPR models are used to approximate a regional model that predicts flood quantiles at ungauged locations from observed basin characteristics. The calibration of the PPR model is made by an algorithm that jointly finds and smoothly fits relevant intermediate predictors. The present approach aims to provide answers to drawbacks of actual methods in terms of parsimony and interpretability. A case study using hydrometric stations from Southern Quebec, Canada, is carried out to illustrate the investigated methodology. Leave-one-out cross-validation is used to evaluate and compare the predictive performance of the PPR methods with a variety of other available and recent methods where the same dataset is considered.

In the present study, when the same basin characteristics are chosen, PPR provides predictions of similar quality to other sophisticated and optimized methods. Nevertheless, the comparison between PPR_STW and PPR_ALL indicates that with proper selection of the basin characteristics, flood quantiles in Southern Quebec can be successfully associated with a single intermediate predictor and lead to the best predictive performance of all considered methods. The present study considers a specific choice of smooth functions (i.e. spline polynomial with equally space knots), which is developed to accommodate PPR models with a various number of terms (Friedman et al., 1983). However, the present study led to PPR models with a single term. In these situations, more rigorous options for smoothing are available, such as penalized splines (Yu and Ruppert, 2002). Further efforts should focus on these particular cases of PPR in RFFA, for which a better calibration can be achieved.

Data screening is an essential step in RFFA to assure the integrity of the data (Hosking and Wallis, 1997). It allows detecting discordant sites, which should be examined closely and eventually

removed, or kept in the estimation step to evaluate their effects. These stations can also be detected after the estimation step as outliers by the PPR model and should be re-examined more closely for gross errors and structural changes. Nevertheless, apparent abnormal behaviors may be legitimate and should not be discarded lightly. For the Quebec dataset, a previous study identified few problematic stations. The difference of performance between the PPR model with (PPR_ALL) and without (PPR_PBL) problematic stations quantifies the impact of the problematic stations as anticipated by the examination of the residuals and shows their non-negligible impact on the performance criteria.

As GAM, PPR uses smooth functions to describe the association between hydrological variables and basin characteristics. In the present case study, the combination of intermediate predictors and smooth functions are able to account adequately for the nonlinearity in this relation. Moreover, this approach is shown to be more parsimonious than ANN and GAM, while having competitive predictive power.

The present work focused on the adaptation of PPR for the regression of the at-site flood quantiles. However, further research is necessary to provide a more complete picture of the usefulness of PPR in RFFA. More efforts may also be put in better understanding the meaning of the intermediate predictors and the smooth functions when multiple terms are found. In the present case study, the data are explained by very few terms, which is convenient, but may not be the general case for RFFA. The adaptation of the PPR method for the multivariate case (Chebana and Ouarda, 2009; Sadri and Burn, 2011) needs also to be investigated in future research activities.

ACKNOWLEDGMENTS

Financial support for this study was graciously provided by the Natural Sciences and Engineering Research Council (NSERC) of Canada. The authors are graceful to the Editor, Dr. L. Ruby Leung and two anonymous reviewers whose comments and suggestions contributed to the improvement of the manuscript.

REFERENCES

Archfield, S., Pugliese, A., Castellarin, A., Skøien, J., Kiang, J., 2013. Topological and canonical kriging for design flood prediction in ungauged catchments: an improvement over a traditional regional regression approach? Hydrol. Earth Syst. Sci. 17. doi:10.5194/hess-17-1575-2013

Bishop, C.M., 1995. Neural networks for pattern recognition. Oxford university press.

Burn, D.H., 1990. An appraisal of the “region of influence” approach to flood frequency analysis. Hydrol. Sci. J. 35, 149–166. doi:10.1080/02626669009492415

Castiglioni, S., Castellarin, A., Montanari, A., 2009. Prediction of low-flow indices in ungauged basins through physiographical space-based interpolation. J. Hydrol. 378, 272 – 280. doi:10.1016/j.jhydrol.2009.09.032

Chebana, F., Charron, C., Ouarda, T.B.M.J., Martel, B., 2014. Regional frequency analysis at ungauged sites with the generalized additive model. J. Hydrometeorol. (Accepted).

Chebana, F., Ouarda, T.B.M.J., 2009. Index flood–based multivariate regional frequency analysis. Water Resour. Res. 45. doi:10.1029/2008WR007490

Chebana, F., Ouarda, T.B.M.J., 2008. Depth and homogeneity in regional flood frequency analysis. Water Resour. Res. 44. doi:10.1029/2007WR006771

Chokmani, K., Ouarda, T.B.M.J., 2004. Physiographical space-based kriging for regional flood frequency estimation at ungauged sites. Water Resour. Res. 40. doi:10.1029/2003WR002983

Dawson, C.W., Abrahart, R.J., Shamseldin, A.Y., Wilby, R.L., 2006. Flood estimation at ungauged sites using artificial neural networks. J. Hydrol. 319, 391 – 409. doi:10.1016/j.jhydrol.2005.07.032

Eaton, B., Church, M., Ham, D., 2002. Scaling and regionalization of flood flows in British Columbia, Canada. Hydrol. Process. 16, 3245–3263. doi:10.1002/hyp.1100

Eng, K., Milly, P., Tasker, G., 2007. Flood regionalization: q hybrid geographic and predictor-variable region-of-influence regression method. J. Hydrol. Eng. 12, 585–591. doi:10.1061/(ASCE)1084-

0699(2007)12:6(585)

Friedman, J., Grosse, E., Stuetzle, W., 1983. Multidimensional Additive Spline Approximation. SIAM J. Sci. Stat. Comput. 4, 291–301. doi:10.1137/0904023

Friedman, J.H., Stuetzle, W., 1981. Projection Pursuit Regression. J. Am. Stat. Assoc. 76, 817–823. doi:10.1080/01621459.1981.10477729

Griffis, V., Stedinger, J., 2007. The use of GLS regression in regional hydrologic analyses. J. Hydrol. 344, 82–95.

G. Tasker, Stedinger, J., 1989. An operational GLS model for hydrologic regression. J. Hydrol. 111, 361–375. doi:10.1016/0022-1694(89)90268-0

Haddad, K., Rahman, A., 2012. Regional flood frequency analysis in eastern Australia: Bayesian GLS regression-based methods within fixed region and ROI framework – Quantile Regression vs. Parameter Regression Technique. J. Hydrol. 430–431, 142 – 161. doi:10.1016/j.jhydrol.2012.02.012

Haddad, K., Rahman, A., Ling, F., 2014. Regional flood frequency analysis method for Tasmania, Australia: A case study on the comparison of fixed region and region-of-influence approaches. Hydrol. Sci. J. null–null. doi:10.1080/02626667.2014.950583

Haddad, K., Rahman, A., Zaman, M.A., Shrestha, S., 2013. Applicability of Monte Carlo cross validation technique for model development and validation using generalised least squares regression. J. Hydrol. 482, 119 – 128. doi:http://dx.doi.org/10.1016/j.jhydrol.2012.12.041

Hastie, T., Tibshirani, R., Friedman, J.H., 2009. The elements of statistical learning: data mining, inference, and prediction, Springer series in statistics. Springer.

Hosking, J.R.M., Wallis, J.R., 1997. Regional frequency analysis: an approach based on L-moments. Cambridge Univ Pr.

Hwang, J.-N., Lay, S.-R., Maechler, M., Martin, R.D., Schimert, J., 1994. Regression modeling in back-propagation and projection pursuit learning. Neural Netw. IEEE Trans. On 5, 342–353. doi:10.1109/72.286906

Khalil, B., Ouarda, T.B.M.J., St-Hilaire, A., 2011. Estimation of water quality characteristics at

ungauged sites using artificial neural networks and canonical correlation analysis. J. Hydrol. 405, 277–287. doi:10.1016/j.jhydrol.2011.05.024

Kjeldsen, T.R., Jones, D.A., 2009. An exploratory analysis of error components in hydrological regression modeling. Water Resour. Res. 45, n/a–n/a. doi:10.1029/2007WR006283

Nelder, J.A., Wedderburn, R.W., 1972. Generalized linear models. J. R. Stat. Soc. Ser. Gen. 370–384.

Nezhad, M.K., Chokmani, K., Ouarda, T.B.M.J., Barbet, M., Bruneau, P., 2010. Regional flood frequency analysis using residual kriging in physiographical space. Hydrol. Process. 24, 2045–2055. doi:10.1002/hyp.7631

Opsomer, J., Wang, Y., Yang, Y., 2001. Nonparametric Regressin with Correlated Errors. Stat. Sci 134–153. doi:10.1214/ss/1009213287

Ouarda, T.B.M.J., Ba, K.M., Diaz-Delgado, C., Carsteanu, A., Chokmani, K., Gingras, H., Quentin, E., Trujillo, E., Bobee, B., 2008. Intercomparison of regional flood frequency estimation methods at ungauged sites for a Mexican case study. J. Hydrol. 348, 40–58. doi:10.1016/j.jhydrol.2007.09.031

Ouarda, T.B.M.J., Girard, C., Cavadias, G.S., Bobée, B., 2001. Regional flood frequency estimation with canonical correlation analysis. J. Hydrol. 254, 157 – 173. doi:10.1016/S0022-1694(01)00488-7

Ouarda, T.B.M.J., Shu, C., 2009. Regional low-flow frequency analysis using single and ensemble artificial neural networks. Water Resour. Res. 45. doi:10.1029/2008WR007196

Pandey, G., Nguyen, V., 1999. A comparative study of regression based methods in regional flood frequency analysis. J. Hydrol. 225, 92 – 101. doi:10.1016/S0022-1694(99)00135-3

Reis, D., Stedinger, J., Martins, E., 2005. Bayesian generalized least squares regression with application to log Pearson type 3 regional skew estimation. Water Resour. Res. 41.

Ribeiro, J., Rousselle, J., 1996. Robust simple scaling analysis of flood peaks series. Can. J. Civ. Eng. 23, 1139–1145. doi:10.1139/l96-923

Roosen, C.B., Hastie, T.J., 1994. Automatic smoothing spline projection pursuit. J. Comput. Graph. Stat. 3, 235–248.

Sadri, S., Burn, D.H., 2011. A Fuzzy C-Means approach for regionalization using a bivariate homogeneity and discordancy approach. J. Hydrol. 401, 231 – 239. doi:10.1016/j.jhydrol.2011.02.027

Schaefli, B., Gupta, H.V., 2007. Do Nash values have value? Hydrol. Process. 21, 2075–2080. doi:10.1002/hyp.6825

Shu, C., Burn, D.H., 2004. Artificial neural network ensembles and their application in pooled flood frequency analysis. Water Resour. Res. 40. doi:10.1029/2003WR002816

Shu, C., Ouarda, T.B.M.J., 2007. Flood frequency analysis at ungauged sites using artificial neural networks in canonical correlation analysis physiographic space. Water Resour. Res. 43. doi:10.1029/2006WR005142

Stedinger, J., Tasker, G., 1985. Regional hydrologic analysis: 1. Ordinary, weighted, and generalized least squares compared. Water Resour. Res. 21, 1421–1432. doi:10.1029/WR021i009p01421

Tukey, J., 1975. Mathematics and the picturing of data, in: Proceedings of the International Congress of Mathematicians. pp. 523–531.

Wazneh, H., Chebana, F., Ouarda, T.B.M.J., 2013a. Optimal depth-based regional frequency analysis. Hydrol Earth Syst Sci 2281–2296. doi:10.5194/hess-17-2281-2013

Wazneh, H., Chebana, F., Ouarda, T.B.M.J., 2013b. Depth-based regional index-flood model. Water Resour. Res. 49, 7957–7972. doi:10.1002/2013WR013523

Weisberg, S., Welsh, A., 1994. Adapting for the missing link. Ann. Stat. 1674–1700. doi:10.1214/aos/1176325749

Wittenberg, H., 1999. Baseflow recession and recharge as nonlinear storage processes. Hydrol. Process. 13, 715–726. doi:10.1002/(SICI)1099-1085(19990415)13:5<715::AID-HYP775>3.0.CO;2-N

Yu, Y., Ruppert, D., 2002. Penalized spline estimation for partially linear single-index models. J. Am. Stat. Assoc. 97, 1042–1054. doi:10.1198/016214502388618861

Table 1: Descriptive statistics for hydrological variables and basin characteristics.

Variable	Notation	min	mean	max	sd
Flood quantile of 10 years (m^3/s)	QS10	53	698	5649	828
Flood quantile of 100 years (m^3/s)	QS100	64	913	7013	1048
Drainage area (km^2)	BV	208	6 265	96 600	11 713
Length of the main channel (km)	LCP	17	157	855	142
Slope of the main channel (m/km)	PCP	0.20	3.23	23.60	3.22
Mean slope of the basin (°)	PMBV	0.96	2.43	6.81	0.99
Fraction of basin occupied by forest (%)	PFOR	18.0	83.1	99.8	16.6
Fraction of basin occupied by lakes (%)	PLAC	0.03	7.72	47	7.99
Mean total annual precipitation (mm)	PTMA	646	988	1 534	154
Mean liquid annual precipitation (mm)	PLMA	423	717	1 625	176
Mean solid annual precipitation (cm)	PSMA	166	302	720	86
Mean liquid precipitation during Jul-Dec (mm)	PLME	306	455	664	72
Mean level of snow on 30th of March (cm)	MNS30	4.9	50.7	98.8	22.6
Degree-day below 0 Celsius (dgr-day)	DJBZ	8 589	16 346	29 631	5 385

Table 2: Direction and scale factors of the regional model.

Regional Model	Flood quantiles	Basin characteristics	Terms 1	Terms 2
PPR_ALL	QS10	BV	$a_{1,1}$ = -0.21	
		PMBV	$a_{1,2}$ = 0.09	
		PLAC	$a_{1,3}$ = -0.86	
		PTMA	$a_{1,4}$ = 0.42	
		DJBZ	$a_{1,5}$ = 0.18	
		scale factors	β_1 = 0.61	
	QS100	BV	$a_{1,1}$ = -0.09	$a_{2,1}$ = -0.27
		PMBV	$a_{1,2}$ = -0.11	$a_{2,2}$ = 0.22
		PLAC	$a_{1,3}$ = -0.64	$a_{2,3}$ = 0.15
		PTMA	$a_{1,4}$ = 0.63	$a_{2,4}$ = -0.73
		DJBZ	$a_{1,5}$ = 0.43	$a_{2,5}$ = -0.57
		scale factors	β_1 = 0.63	β_2 = 0.28
PPR_STW	QS10	BV	$a_{1,1}$ = 0.06	
		LCP	$a_{1,2}$ = -0.28	
		PLAC	$a_{1,3}$ = -0.85	
		PLMA	$a_{1,4}$ = 0.14	
		LON	$a_{1,5}$ = 0.42	
		scale factors	β_1 = 0.63	
	QS100	BV	$a_{1,1}$ = -0.09	
		LCP	$a_{1,2}$ = -0.26	
		PCP	$a_{1,3}$ = -0.10	
		PLAC	$a_{1,4}$ = -0.82	
		PLMA	$a_{1,5}$ = 0.17	
		LON	$a_{1,6}$ = 0.46	
		scale factors	β_1 = 0.63	

Table 3: Cross-validation criteria of the regional PPR models.

Regional model	Quantile	Num. Terms	*BIASr* (%)	*RMSEr* (%)	*NASH* (%)
PPR_ALL [1,3]	QS10	1	-6	40	76
	QS100	2	-7	48	71
PPR_PBL [2,3]	QS10	1	-3	31	81
	QS100	2	-5	34	80
PPR_STW [1,4]	QS10	1	-4	34	82
	QS100	1	-6	40	79

[1] Use all available stations
[2] Remove problematic stations
[3] Same basin characteristics as in previous work on the same dataset (i.e. Shu and Ouarda, 2007).
[4] Basin characteristics selected from stepwise procedure.

Table 4: Comparison of the PPR method with other methods found in previous studies using the same dataset.

	Reference	Q10		Q100	
		BIASr (%)	*RMSEr* (%)	*BIASr* (%)	*RMSEr* (%)
PPR methods					
PPR_ALL[1]	Table 3	-6	40	-7	48
PPR_STW[2]	Table 3	-4	**34**	-6	**40**
Traditional methods					
with CCA delineation[1]	Chokmani & Ouarda, 2004	-9	43	-11	51
with depth functions[1]	(Wazneh et al., 2013a)	**-3**	38	**-2**	44
GAM methods					
GAM[1]	Chebana et al., 2014	-5	41	-8	49
GAM-stepwise[2]	Chebana et al., 2014	-5	38	-7	42
Spatial methods					
Ordinary kriging[3]	Chokmani & Ouarda, 2004	-16	51	-23	70
Residual kriging[3]	Nezhad et al., 2010	-7	39	-14	58
ANN methods					
Single ANN[3]	Shu & Ouarda, 2007	-7	47	-7	64
Ensemble ANN[3]	Shu & Ouarda, 2007	-5	37	-6	45

Bold character indicates the best results

[1] Same basin characteristics as in previous work on the same dataset (Shu et Ouarda 2007).

[2] Basin characteristics selected from stepwise procedure.

[3] Method applied on the physiographical space constructed by CCA[1]

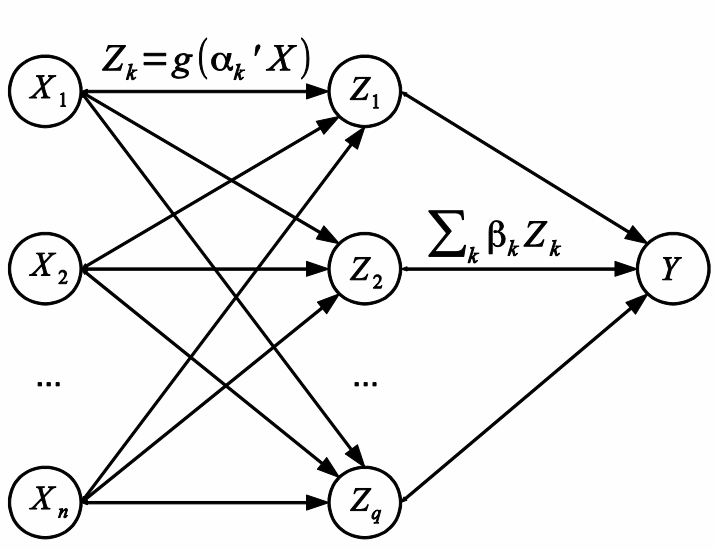

Figure 1: Representation of the PPR model as a network.

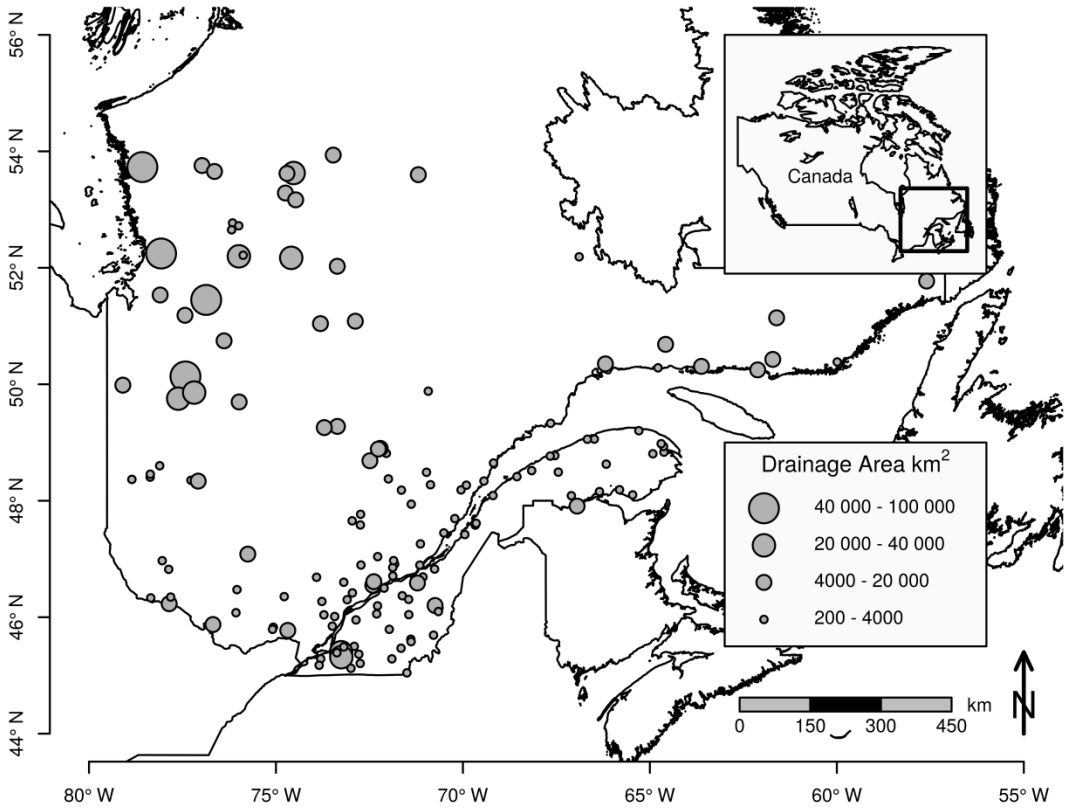

Figure 2: Location of the 151 hydrometric stations in Southern Quebec, Canada

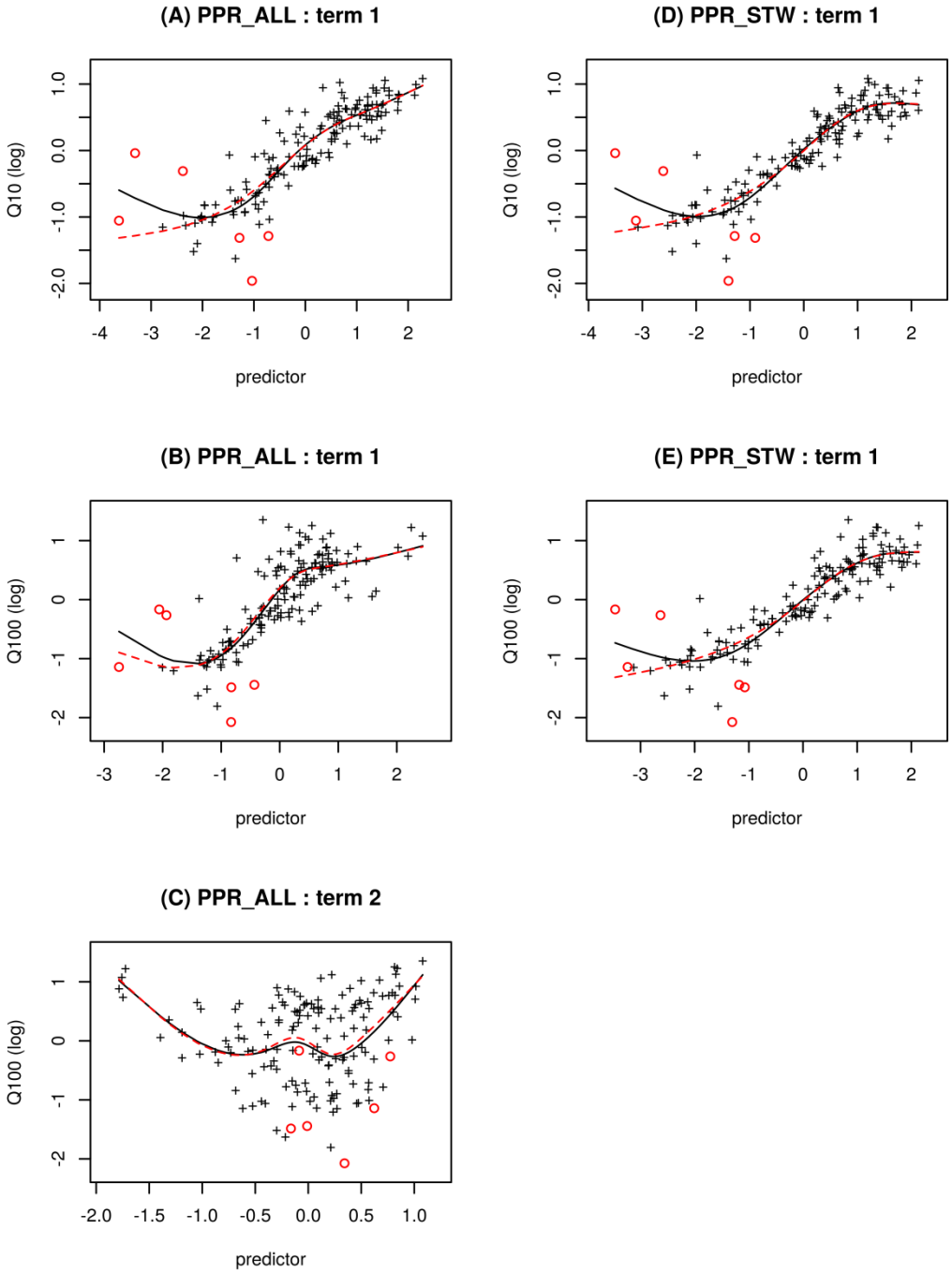

Figure 3: Centred flood quantiles as a function of predictors $\alpha_k{}'X$. The solid lines represent the smooth functions $b_k g_k$. The red circles represent the problematic gauged stations and the dashed line represents the smooth functions without the problematic stations.

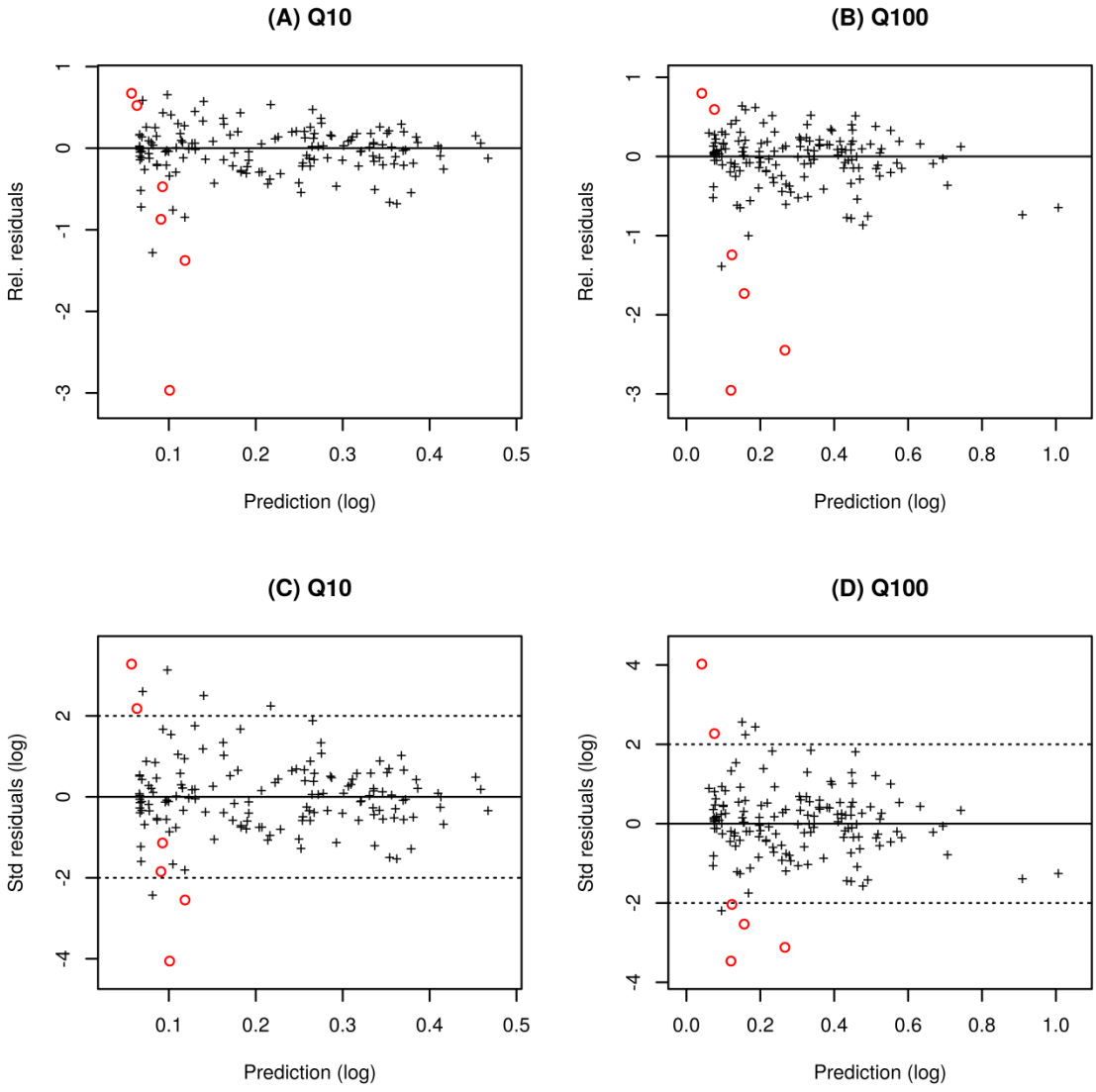

Figure 4: Visuals diagnostics of the residuals for the PPR_ALL regional model. The dashed lines correspond to intervals of two standard deviations. The red circles represent the problematic gauged stations.

CHAPITRE 3:

ON THE IMPORTANCE OF HYDROLOGICAL INFORMATION FOR THE DELINEATION OF NEIGHBORHOODS IN REGIONAL FREQUENCY ANALYSIS

On the importance of hydrological information for the delineation of neighborhoods in regional frequency analysis

Martin Durocher [*,1], Fateh Chebana [*,1], Taha B. M. J. Ouarda[1,2]

[1]Institut National de Recherche Scientifique (INRS-ETE),

University of Quebec

490 de la Couronne, Québec G1K 9A9, Canada

[2] Institute Center for Water Advanced Technology and Environnemental Research (iWater),

Masdar Institue of Science and Technology

P.O. Box 54224, Abu Dhabi, UAE

Submitted for publication

ABSTRACT

This study investigates the utilization of hydrological information in Regional Frequency Analysis (RFA) to enforce properties for a group of gauged stations. Neighborhoods are a particular type of regions that are centered on target locations. A difficulty using neighborhoods in RFA is that hydrological information is not available at target locations. Instead of using known site characteristics (not hydrological) to define the center of a target location, this study proposes to introduce proper estimates of (hydrological) reference variables to ensure better properties. These reference variables represent nonlinear relations with the site characteristics obtained by projection pursuit regression; a nonparametric regression method. The resulting neighborhoods are investigated in combination of common regional models: the index-flood model and the regression-based models. The complete approach is illustrated on a real case study with gauged sites located in Southern Quebec (Canada) and is compared with traditional approaches: region of influence and canonical correlation analysis. The result shows improvements of the neighborhood properties as well as the predicting performances, with special attention on problematic stations.

Keywords: Flood, Index-flood model, L-moments, Regional frequency analysis, Ungauged site, Region of influence, Projection pursuit regression, CCA.

1. INTRODUCTION

Accurate estimates of the risk of occurrence of extreme hydrological events are necessary for optimal management of water resource systems and for the minimization of the impacts of these events. However, necessary information is not always available at the sites of interest. Hence, it is necessary to develop procedures to transfer, or to regionalize, the information available at existing gauged sites to ungauged ones. Regional Frequency Analysis (RFA) represents a large class of techniques commonly used in water sciences to evaluate the risk of occurrence of extreme hydrological phenomena of rare magnitude at ungauged locations (Haddad and Rahman, 2012; Hosking and Wallis, 1997; Laio et al., 2011; Pandey, 1998; Reis et al., 2005).

RFA methods are usually composed of two main steps. The first step is the formation of homogenous regions, which aims at pooling together sites that are approximately similar according to homogenous criteria. Inside these homogenous regions, it is assumed that hydrological information can be reasonably transferred from gauged sites to ungauged locations (Cunnane, 1988). The second step, the estimation of flood quantiles, consists in the calibration of a regional model that characterizes the interrelation between hydrological variables of interest and explanatory physio-meteorological variables that correspond to known site characteristics. Consequently, RFA is used to study unobserved hydrological behaviours from available hydrological and physio-meteorological information.

Neighborhoods are specific forms of regions inside which gauged sites are not classified into fixed regions, but are composed of gauged sites that are the most similar to the target location (Acreman and Sinclair, 1986; Burn, 1990). Hence, two distinct target locations have their own neighborhoods that may overlap. Comparative studies showed that neighborhoods lead to better regional estimates than fixed regions (Burn, 1990; Ouarda et al., 2008; Tasker et al., 1996). To identify the most similar gauged sites, a notion of distance is needed to evaluate the proximity, or relevance, of each gauged

site to the target location and identify the most similar gauged sites. However, when the target location is ungauged, the distance between hydrological variables cannot be directly calculated due to the missing hydrological information. Physio-meteorological information is hence used for similarity evaluation. The traditional approach, based on the distance between site characteristics, is commonly referred to as Region of Influence (ROI) model (Burn, 1990).

Alternatively, (Ouarda et al., 2001) used Canonical Correlation Analysis (CCA) to build neighborhoods from a canonical distance that accounts for the interrelation between flood quantiles and site characteristics. For this method, neighborhoods are formed by gauged sites that are the most similar to the target location, according to the distance between vectors of flood quantiles of different return periods. Due to the missing hydrological information, the CCA method in RFA estimates the unavailable hydrological variables as linear combinations of site characteristics. Consequently, the available site characteristics are transformed into more meaningful “hydrological” quantities for the purpose of delineating neighborhoods. However, the CCA method suffers from some limitations, such as linearity and normality assumptions (He et al., 2011). Subsequent studies aimed to improve the CCA method by improving the CCA technique itself (Chebana and Ouarda, 2008; Ouali et al., 2015). However, little attention has been paid to the importance of properly choosing the hydrological quantities in the delineation step whereas much effort has been done in the modeling step. Indeed, Chebana and Ouarda (2008) employed an iterative linear procedure to estimate neighborhood centers and they showed that the estimation quality of these centers is the crucial element to improve the final model performance.

This study aims to provide a general framework with more flexibility regarding the linearity and normality assumptions by replacing CCA in the prior analysis of hydrological variables by Projection Pursuit Regression (PPR), a nonparametric regression method recently considered in RFA as an estimation model (Durocher et al. 2015). This study is also interested in validating the advantage of employing other hydrological variables than the at-site flood quantiles in prior modeling as well as considering a combination of these hydrological variables with site characteristics. L-moments have

already been used in RFA to test the homogeneity of fixed regions when the target site is gauged (Chebana et Ouarda 2007; Hosking and Wallis 1997). In this study, the prediction of the L-moments at ungauged sites is considered to improve the delineation of the neighborhoods by reducing uncertainties. Moreover, a conceptual advantage of using L-moments conversely to at-site flood quantiles is that the L-moments do not depend on the subjective selection of at-site distributions.

The present paper is organized as follows. Section 2 presents the background for the techniques commonly used in RFA. Section 3 elaborates on the prior analysis of hydrological variables and their integration with the techniques presented in Section 2 to form a complete procedure. Section 3 suggests also criteria for the evaluation of the predictive performances and the neighborhood properties. Section 4 illustrates the application of the method on a case study. Traditional ROI and CCA methods serve as references in order to evaluate the relative performance of the investigated method. Finally, concluding remarks are provided in the last section.

2. BACKGROUND

2.1 DELINEATION OF NEIGHBORHOODS

In RFA, neighborhoods are used to identify gauged sites from which it is beneficial to transfer information to a target location. A neighborhood is characterized by a center and a radius that delimits an area (not necessary in the geographical sense). Gauged sites inside the area delineate a region that includes relevant sites to the target location. At each site $i = 1, \ldots$, p characteristics $\mathbf{x}_i = \left(x_{i,1}, \ldots \right.$ are available. Typically, the ROI method forms neighborhoods according to a radius based on a metric d :

$$d\left(\mathbf{x}_i, \mathbf{x}_j\right) = \sqrt{\sum_{k=1}^{p} \frac{\left(x_{i,k} - x_{j,k}\right)^2}{\sigma_k^{\,2}}} \qquad (1)$$

where σ_k is the standard deviation of $\{x_{i,k}\}_{i=1}^{n}$ the kth site characteristic (Eng et al., 2005).

Alternatively, CCA is a multivariate technique used to unveil the interrelation between two groups of variables. Let Y and X be normally distributed random vectors with zero means. The CCA method defines canonical pairs (U_k, V_k) as linear combinations of the original random variables:

$$U_k = a_k X \tag{2}$$

$$V_k = b_k Y \tag{3}$$

where the correlations $\rho_k = corr(U_k, V_k)$ are sequentially maximal for $k = 1,\ldots$ under the conditions $corr(U_k, U_l) = corr(V_k, V_l) = 0$ for $k \neq l$. Moreover, only the canonical pairs (U_k, V_k) with unit variances are considered.

To delineate neighborhoods, the CCA approach considers the canonical scores $\mathbf{u}_i = (a_1,\ldots$ and $\mathbf{v}_i = (b_1,\ldots$ that are respectively linear combinations of site characteristics $\mathbf{x}_i$ and flood quantiles of different return periods $\mathbf{y}_i$ for site *i*. Due to the missing hydrological information at the ungauged location denoted $i = 0$, the flood quantiles $\mathbf{y}_0$ and the corresponding linear combination $\mathbf{v}_0$ are unknown. Nevertheless, CCA provides a linear estimate $\mathbf{v}_0 \approx \Lambda \mathbf{u}_0$, where $\Lambda = \mathrm{diag}(\rho_1,\ldots$. Accordingly, a neighborhood is delineated in the canonical space according to the distance:

$$d(\mathbf{v}_i, \Lambda \mathbf{u}_0) = (\mathbf{v}_i - \Lambda \mathbf{u}_0)' \left(I - \Lambda^2\right)^{-1} (\mathbf{v}_i - \Lambda \mathbf{u}_0) \tag{4}$$

More details on the CCA approach in RFA can be obtained in Ouarda et al. (2001).

2.2 MULTIPLE REGRESSION

In RFA, two types of regional models are often considered to predict flood quantiles corresponding to given return periods: the index-flood model and the regression-based model (Ouarda et al., 2008).

The index-flood model predicts a target distribution by assuming that all distributions inside a region are proportional to a regional distribution, up to a scale factor called index-flood. The flood quantile of interest at a target location is then calculated from the regional distribution based on the predicted index-flood (e.g. Chebana and Ouarda, 2009; Dalrymple, 1960; Stedinger and Lu, 1995). Conversely, the regression-based model considers directly the at-site estimates of the desired flood quantiles for prediction. The flood quantiles are then predicted at their target locations by the regression equations estimated within the neighborhoods (Pandey and Nguyen, 1999).

Even though they proceed differently, both the index-flood model and the regression-based model may use the same multiple regression techniques to transfer information to an ungauged location. For the sake of simplicity, the term hydrological variables is used to designate the corresponding output variables z_i of these models at location $i = 1, \ldots$. Consequently, for the index-flood model, z_i is the index flood, while for a regression-based model the hydrological variable z_i is the flood quantile of interest.

Multiple regression models assume linear interrelation between the hydrological variable z_i and the site characteristics $\mathbf{x}_i$. Consequently, in several cases, transformations are necessary to fulfill this hypothesis. For instance, the power law form is frequently used to model flood quantiles:

$$z_i = e^{\beta_0} \times x_{i,1}{}^{\beta_1} \times \ldots \qquad \times \varepsilon_i \qquad (5)$$

where $\beta' = \left(\beta_0, \beta_1, \ldots\right.$ are parameters and ε_i is an error term. Applying a logarithmic transformation is sufficient to cast (5) into a linear model. In general, a proper transformation is assumed for the hydrological variables $y_i = g\left(z_i\right)$ being linearly related to its sites characteristics.

According to previous notations, let $\mathbf{y} = (y_1, \ldots$ be the hydrological variables, $\mathbf{X}$ be the design matrix of the site characteristics $x_{i,j}$ with intercept, and $\varepsilon = (\varepsilon_1, \ldots$ be the error term with zero mean and constant variance. Hence in matrix notation, a multiple regression model has the form:

$$\mathbf{y} = \mathbf{X}\beta + \varepsilon \tag{6}$$

and according to least-squares theory, the estimates of the parameters are:

$$\hat{\beta} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y} \tag{7}$$

2.3 PROJECTION PURSUIT REGRESSION

Some methods predict hydrological variables without the formation of regions, such as physiographical kriging (Castiglioni et al., 2009; Chokmani and Ouarda, 2004), generalized additive models (Chebana et al., 2014) and artificial neural networks (Dawson et al., 2006; Ouarda and Shu, 2009). More recently, Projection Pursuit Regression (PPR) was introduced to provide a flexible nonparametric regression approach to describe the nonlinearity that is present in the relationship between hydrological variables and site characteristics. PPR was used in the RFA context by Durocher et al. (2015) to predict flood quantiles

The basic elements of a PPR model are $k = 1, \ldots$ functions f_k called terms and defined as:

$$f_k(\mathbf{X}) = g_k(\alpha_k'\mathbf{X}) \tag{8}$$

where directions $\alpha_k \in \mathbb{R}$ are vectors of coefficients and g_k are smooth functions such as polynomials or spline polynomials among others.. The directions α_k are coefficients that respect $|\alpha| = 1$ and determine a predictor $\alpha_k'\mathbf{X}$ as relevant linear combinations of the site characteristics $\mathbf{X}$. The terms are then combined into a regression model:

$$\mathbf{y} = \mu + \sum_{k=1}^{m} f_k(\mathbf{X}) + \varepsilon \qquad (9)$$

where μ is the global mean and ε is a term of error with zero mean and constant variance. Notice that the orthogonality between directions α_k is not imposed, hence the predictors $\alpha_k'\mathbf{X}$ and $\alpha_l'\mathbf{X}$ for $k \neq l$ may be correlated. Consequently, PPR allows for interaction between site characteristics, which leads to a large variety of regression models (Hastie et al., 2009).

The components α_k and g_k of model (9) are estimated by the least-squares approach (Friedman et al., 1983). For a unique direction ($m = 1$), PPR can be estimated by standard nonlinear algorithms (Yu and Ruppert, 2002), but in general a stagewise algorithm is adopted to find a proper solution (Friedman and Tukey, 1974). Comparative studies show that PPR has a similar predictive performance to artificial neural networks (Bishop, 1995; Hwang et al., 1994). However, Durocher et al. (2015) indicated that in RFA, PPR reduces to more parsimonious models than artificial neural networks, which provides an explicit expression of the regression equations.

3 METHODOLOGY

This study deals with neighborhood delineation and more precisely it focus on the identification of a reliable estimates of the hydrological centers of these neighborhoods. For simplicity, the variables forming these centers will be referred to as reference variables, because they represent the reference to evaluate the similarity between a target location and the gauged sites. Reference variables can take different forms, such as site characteristics, hydrological variables or a combination of both. Their nature is important, because it determines the properties that will be similar between close sites. The particularity of the present method is that PPR can be used to predict these neighborhood centers (prior to the RFA modeling step) when some of the reference variables are unknown hydrological

variables. Accordingly, the proposed method will be referred to as RVN for Reference Variable Neighborhoods.

3.1 ESTIMATION OF THE REFERENCE VARIABLES

The general procedure of the RVN method can be described by the main steps below:

(i) Estimation of the hydrological reference variables at the target centers;

(ii) Delineation of the neighborhoods;

(iii) Estimation of the flood quantiles at the target locations.

Step (i) is the particularity of the RVN. If a target location is designated by $i = 0$, the radius of the neighborhood can be computed as $h_i = d\left(\mathbf{t}_i, \mathbf{t}_0\right)$ where d is a metric and $\mathbf{t}_i' = \left(t_{i,1}, \ldots\right.$ are the reference variables of the *i*th site. For simplicity, the Euclidian metric d is considered throughout the present study, but other metrics or dissimilarity measures could be employed as well. In particular, the Mahalanobis distance, the weighted distance and the depth function could be considered (Chebana and Ouarda, 2008; Cunderlik and Burn, 2006; Ouarda et al., 2000).

As hydrological information is unavailable at the target location, the estimation of the hydrological reference variables is necessary to produce an estimate $\mathbf{t}_0 = f(\mathbf{x}_0)$ from site characteristics $\mathbf{x}_0$ at the target location. This substitution leads to the distance $h_{(i)} = d\left[\mathbf{t}_i, f(\mathbf{x}_0)\right]$, which may be seen as an approximation of the true distance h_i. This study considers PPR models in order to fit every hydrological reference variable as described in section 2.3. The motivations for adopting PPR are that it does not require a prior delineation of regions, it accounts for nonlinear relationships, it has good predictive performances and it leads to a straightforward interpretation of the reference variables when a few directions α_k are necessary (Durocher et al., 2015).

Figure 1 illustrates two neighborhoods resulting from the RVN method. It shows the importance of correctly predicting the reference variables in order to be representative of the true center of the target location and hence the appropriate sites to be included in the neighborhood. Indeed, in green, the true neighborhood designates the neighborhood that would be delineated if all hydrological variables were known at the target location. Alternatively, the red and the blue neighborhoods are identified from different estimates of the reference variables. Figure 1 indicates that if the reference variables are well predicted, then the corresponding RVN neighborhoods will most likely include the same gauged sites as the true neighborhood.

Steps (ii-iii) are common in RFA and are explained in sections 2.1 and 2.2. In the remainder of this study, step (ii) uses a specific type of neighborhoods that is composed of a fixed number of the nearest sites (Eng et al., 2005; Tasker et al., 1996), but could also be constrained to the degree of the homogeneity of the neighborhoods (Ouarda et al., 2001). Consequently, the selected gauged sites can be obtained by sorting $h_{(i)}$ and keeping the desired number of sites. Notice that even though $h_{(i)}$ does not approximate exactly h_i, both distances will lead to the same neighborhoods if they preserve the ranks. Finally, step (iii) consists in the estimation of the flood quantiles using either the index-flood or the regression-based model.

Notice that the RVN method may be seen as a generalization of the ROI and the CCA methods in RFA. Indeed, the ROI method corresponds to the RVN method, for which all the reference variables are site characteristics. In that case, $\mathbf{t}_0 = f(\mathbf{x}_0)$ is known and PPR is not necessary in step (i). Similarly, the CCA approach may be seen as the special case for which the reference variables are the canonical pairs (4) and CCA is used, instead of PPR, to predict them in step (i).

3.2 EVALUATION CRITERIA

For the RVN method presented above, the neighborhood sizes must be calibrated according to an objective criterion. In this regards, the leave-one-out cross-validation approach is a general strategy to

assess the performance of the predicted hydrological variables z_i at site $i = 1,\ldots$. In turn, each gauged site i is considered as an ungauged target location. From the remaining gauged sites, predicted values $z_{(i)}$ can be obtained without using the hydrological information at the target location. Discrepancies between the sampled and the predicted values are used to define evaluation criteria. Notice that the hydrological variables are transformed $y_i = g(z_i)$. Hence, if $\bar{y}$ is the sample mean of the y_i, then an appropriate global performance measure is the Nash-Sutcliffe criterion:

$$\mathrm{NHS} = 1 - \frac{\sum_{i=1}^{n}\left[y_i - y_{(i)}\right]^2}{\sum_{i=1}^{n}\left[y_i - \bar{y}\right]^2} \tag{10}$$

Additionally, the predictive performance is examined at the original scale by the relative root mean square error:

$$\mathrm{RRMSE} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}\left(1 - \frac{z_{(i)}}{z_i}\right)^2} \tag{11}$$

The choice of the reference variables is an important aspect and a set of reference variables should be chosen in order to enforce the desired properties. For instance with the index-flood model the assumption of a regional distribution suggests that, apart from the index-flood, the at-site distributions must be proportional to a regional distribution. Heterogeneity measure based on the dispersion of the L-coefficient of variation (LCV) is shown to be a proper way to ensure that the LCV is relatively constant (Viglione et al., 2007). Accordingly, let I_j be the set of index for the N nearest gauged sites on target location j during the cross-validation process. The regional LCV $\hat{\theta}_{(j)}$ is calculated as the average:

$$\hat{\theta}_{(j)} = \frac{1}{N}\sum_{i \in I_j}\theta_i \tag{12}$$

of the at-site LCV θ_i inside the jth region. The heterogeneity measure is defined as:

$$H_{(j)} = \sum_{i \in I_j} \left(\theta_i - \hat{\theta}_{(j)} \right)^2 \quad (13)$$

In their procedure, Hosking and Wallis (1997) used this heterogeneity measure to test the homogeneity of a region, which implies that the regional LCV can be considered constant. Hence, the result of this test allows to decide if a region must be divided into smaller and more homogenous sub-regions. In the present study the size of the neighborhoods is the same for every neighborhood delineated during the cross-validation procedure. Hence, if a homogeneity test is performed with a given neighborhood size, some of the neighborhoods will be considered homogenous, while the others will be considered heterogeneous (Das and Cunnane, 2010) However, the heterogeneity measure (13) remains a useful indicator of dispersion for the regional LCV $\hat{\theta}_{(j)}$ inside a neighborhood. Consequently, a smaller $H_{(j)}$ suggests that the regional LCV $\hat{\theta}_{(j)}$ is measured with less uncertainty.

To facilitate the interpretation of the results and to ensure the comparability between neighborhoods, the heterogeneity measure $H_{(j)}/N$ is considered instead. The measure represents the sample variance of the LCV for the jth target location. This heterogeneity measure is standardized by H/n, where H is the heterogeneity measure (13) calculated on all n available gauged sites. The resulting ratio corresponds to a scale-free heterogeneity measure, where a value under one provides evidence of a less heterogeneous neighborhood in comparison to the whole dataset. Therefore, the Average Heterogeneity Measure (AHM) criterion below is defined as the average of every neighborhood considered in the cross-validation process:

$$\mathrm{AHM} = \frac{1}{NH} \sum_{j=1}^{n} H_{(j)} \quad (14)$$

This criterion is not specific to a given target location, but represents the global level of heterogeneity resulting from a given delineation method, such as ROI, CCA or RVN. In particular, a delineation

method with a smaller AHM suggests that on average a more precise regional LCV is used to predict the flood quantiles.

Another desired property for a neighborhood is to lead to estimation models with less uncertainty. For the index-flood models, this implies in particular less uncertainty in the prediction of the index-flood, while for the regression-based models, it implies less uncertainty in the prediction of flood quantiles. For a multiple regression model, the uncertainty can be quantified by the residual variance:

$$s_{(j)}^2 = \frac{1}{N}\sum_{i \in I_j}\left(e_{i,(j)}\right)^2 \qquad (15)$$

where $e_{i,(j)}$ is the residual at the ith gauged site, when predicting the jth target location in the cross-validation process. Notice that a regression model fitted on two different neighborhoods (for the same target location) can obtain identical values, but lead to different levels of uncertainty. In this study, a neighborhood with a smaller residual variance than another one is said to be relatively more efficient.

During the cross-validation process, the sample variance of the regression models can be calculated for every site, which leads to the Average Relative Efficiency (ARE) criterion defined by :

$$\mathrm{ARE} = \frac{1}{ns^2}\sum_{j=1}^{n} s_{(j)}^2 \qquad (16)$$

where the residual variance s^2 is calculated from the multiple regression model on the whole dataset. This criterion is similar to the AHM criterion as it is standardized to a scale-free measure. This criterion can be used to identify the delineation method which achieved on average the smallest residual variances for each neighborhood. The ARE and the AHM criteria are used in the present study, along the NHS and RRMSE to access the performances of the various models.

4. APPLICATIONS

4.1 DATA

To validate the RVN method on a practical situation, RFA is carried out in a real-world case study using both the index-flood model and the regression-based model. The hydrological variables of interest are the flood quantiles corresponding to a return period of 100 years, denoted Q100. The analysis is performed on 151 sites located in Southern Quebec, Canada, for which at least 15 years of data are available and the usual hypotheses of stationarity, homogeneity and independence are verified. Only a brief description of the data and the at-site frequency analysis is provided since the elements were already presented in detail in previous studies (e.g Chokmani and Ouarda, 2004).

The at-site distributions are selected among several families including: generalized extreme values (GEV), Pearson type III (P3), generalized logistic (GLO) and log-normal with 3 parameters (LN3). In general, the estimation of the at-site distribution was achieved by maximum likelihood and the final choices of distributions are based on the Akaike information criterion. Recent studies on the same dataset have identified 4 relevant site characteristics (Chebana et al., 2014; Durocher et al., 2015), which are used in the present analysis: the drainage area or BV (km^2), the fraction of the basin area occupied by lakes or PLAC (%), the annual mean liquid precipitation or PLMA (mm) and the longitude or LON. Proper transformations are applied on these site characteristics in order to obtain approximately standard normal distributions (Chokmani and Ouarda, 2004).

4.2 DETERMINATION OF THE NEIGHBORHOOD CENTERS

The first step of the RVN method is the estimation of the hydrological reference variables at the target locations. Two groups of reference variables are considered. The first group is based on L-moments only and the second is based on the combination of L-moments and site-characteristics. More precisely, the L-moments considered for both groups are the sample average (L1), the LCV, the L-

coefficient of skewness (LSK) and the L-coefficient of kurtosis (LKT). These reference variables are transformed logarithmically and standardized to obtain zero mean and unit variance. For LSK and LKT, an additional translation is necessary to avoid numerical difficulties due to negative values. Moreover, a specific implementation of PPR is assumed, which considers the smooth functions g_k in (8) as cubic spline polynomials with 5 equally spaced knots. The number of knots is validated by cross-validation using the NHS criterion. Notice that for the fitting of LSK, one site has a very low standardized residual of approximately -6. Consequently, this site is considered as an outlier and removed from the estimation of the reference variables. In previous studies (Chokmani and Ouarda, 2004), this site was identified as one of a few problematic sites that are difficult to predict due to an underestimated drainage area or overevaluated percentage of area covered by lakes. Nevertheless, in the present study, this site is only removed only during the prediction of the reference variables and all sites are included in the rest of the analysis.

Figure 2 shows the fitting of the four reference variables by the PPR models. Cross-validation has selected PPR models with a unique direction α for all reference variables. Figure 2a shows a strong linear relationship between L1 and the predictor $\alpha'\mathbf{X}$. Conversely, Figures 2b,c,d show nonlinearity and hence indicate the need for a nonlinear model such as PPR. The PPR equations that describe the relation between the reference variables and the site characteristics are explicit, for instance, the regression equation for the LCV has the form:

$$\log(\mathrm{LCV}) = -1.80 + 0.26 \times f\left[-0.67 \times \log(\mathrm{BV}) - 0.09 \times \sqrt{PLAC} + 1.27 \times \log(\mathrm{PLMA}) + 0.06 \times \mathrm{LON} - 1.32\right] \quad (17)$$

Notice the constant term -1.32 and the norm of direction $|\alpha| \neq 1$ inside the function f in (17). The difference in (17) in comparison to the general form of the PPR model in (9) is the consequence of transformations on the explanatory variables. Indeed, during the optimization procedure of a PPR model, it is suggested to scale the explanatory variables in order to avoid the scale effect in the

coefficients of the direction α (Hastie et al., 2009). Nevertheless, notice that the formula inside the function f corresponds to a linear model.

The predictive performances of the reference variables are evaluated by the NHS criterion with values 91%, 33%, 7% and 56% respectively for L1, LCV, LSK and LKT. These results show that L1 is accurately predicted by the site characteristics, while a poor fit is associated to LSK. Indeed, Figure 2c suggests that apart from a few sites at the right of the curve, LSK appears not highly related to the predictor $\alpha'\mathbf{X}$. Due to its poor fit, LSK may not be a proper reference variable for the delineation step. To validate this assumption, the neighborhoods are formed with and without using LSK and the rest of the analysis is carried out for both scenarios. Based on the RRMSE criterion, LSK must be maintained as it is associated to better predictive performances. Similarly, the same procedure is applied to validate the usefulness of each reference variable, which leads to discarding LKT and to maintaining L1, LCV and LSK.

The second group of reference variables contains both the L-moments and the site characteristics. As with the first group, the complete analysis is performed with and without each of the reference variables. The final reference variables that are kept are: BV, PLAC, LCV and LSK. In order to distinguish the two groups of reference variables, RVN-LM will designate the first group with the L-moments only and RVN-HYB will designate the second group with both the L-moments and the site characteristics.

4.3 RESULTS OF THE INDEX-FLOOD MODEL

One of the objectives of RFA is to identify a proper family of distributions from regional information, which is achieved here by analysing the distribution of the gauged sites inside a neighborhood. The index-flood model and the L-moments algorithm were proven to lead to a reliable procedure to identify a regional distribution and to estimate its parameter (Hosking and Wallis, 1997). In this model, the

quantile $Q_i(r)$ corresponding to a return period r at a target location i is of the form $Q_i(r) = \mu_i Q(r)$, where μ_i is the index-flood. In the present study, the index-flood is taken to be the means of the at-site distributions and is predicted at the target location by multiple regression.

The index-flood model is fitted inside the neighborhoods obtained by each one of the four methods: ROI, CCA, RVN-LM and RVN-HYB. For CCA, two canonical pairs are calculated as described in section 2.1 using flood quantiles corresponding to 10 and 100 year return periods as hydrological variables. The choice of the regional distribution is made between the four common families of distributions that were mentioned earlier: GEV, GLO, LN3 and P3. The parameters of the regional quantile function $Q(r)$ are calculated from the regional LCV and the regional LSK as the respective averages, see (12). Figure 3a shows the L-moment ratio diagram for the regional LSK and LKT with RVN-LM. For each neighborhood, the distribution family is selected as the one having the nearest regional LKT to the theoretical value, given the regional LSK. RVN-HYB is omitted in Figure 3 for the clarity of the graphics, but has similar behaviour to RVN_LM.

Figures 3b,c,d present the L-moment ratio diagram of the at-site LCV and LSK for three given target locations as an illustration of the gauged sites found in the respective neighborhoods. In these diagrams, the nearest gauged sites selected for RVN-LM, CCA and ROI are highlighted. Figure 3b shows that RVN_LM has a denser cluster of gauged sites in terms of LCV and is approximately centered on the true target. Conversely, Figures 3c and 3d show situations where the true targets do not correspond to the predicted target. Although, all the reference variables are known at target location for the ROI method, Figures 3b and 3c show that the selected sites are also not located around the true target. This finding is coherent with the results of (GREHYS, 1996a, 1996b) which indicates that delineation according to physiographical similarity can lead to substantially different regions than according to hydrological similarity.

Results of cross-validation are presented in Figure 4. The evaluation criteria are calculated for every neighborhood with size superior to 15 in order to calibrate the model. The tendency illustrated in this

figure helps to visualize the evolution of these criteria with better perspective. The comparison of Figures 4a and 4b indicates that the optimal neighborhood sizes for RRMSE and NHS are not always in agreement. In particular, the best RRMSE for the RVN-HYB method is with 24 sites, while the best NHS is with near 80 sites. Nevertheless, the optimal values for the three other methods are obtained with approximately 30 sites for both criteria. Figure 4b indicates that all methods have relatively stable NHS between 86% and 87%, but the best NHS is obtained by RVN-LM. Conversely, Figure 4a shows clearer improvements of the calibration in terms of the RRMSE criterion. Hence, the calibrated models are set according to the RRMSE criterion and are represented by circles in Figure 4. RVN-HYB, with a RRMSE of 40% outperforms the methods RVN-LM and CCA, with a RRMSE of 45% and ROI, with a RRMSE of 46%.

Figures 4c,d present respectively the AHM and the ARE criteria. The AHM criterion indicates that the ROI and the CCA methods have in general lower heterogeneity than the whole dataset, but are outperformed by RVN-LM and RVN-HYB methods especially for smaller neighborhoods. This validates and quantifies the intuitive assumption that the regional LCV is calculated with less uncertainty when the L-moments are directly considered instead of other reference variables. Moreover, the ARE criterion reveals that RVN-LM leads to neighborhoods with the most relatively efficient regression models for predicting the index-flood. In particular, the calibrated model of the RVN-LM method has an ARE of 36% and the ROI method has an ARE of 67%, which is more than the double. Overall, this indicates that a strategic pooling of the gauged sites using hydrological reference variables reduces the uncertainty of both the regional LCV and the index-flood.

As mentioned in section 4.2, previous studies have identified few problematic stations in the considered dataset. Figure 5 presents the relative residuals between different methods. In general, the points associated to the largest discrepancies are close to the $y = x$ line, which indicates that the sites that are difficult to predict are essentially the same for all methods. However, Figures 5a,b show that the RVN-HYB specifically improves the prediction of the sites with the largest discrepancies as

their points are mostly located under the $y = x$ lines, which explains that this method leads to the best RRMSE. On the other hand, Figures 5c,d demonstrate that at the logarithmic scale, the RVN-LM method achieved mostly similar predicted values as ROI and CCA methods, which explains the similarity of the NHS criteria for all the compared methods.

4.4 RESULTS OF THE REGRESSION-BASED MODEL

Prediction of Q100 at the target location is also computed by the regression-based model using the same delineation methods as with the index-flood model, but with different calibration values for the neighborhood sizes. Cross-validation results for the regression-based model are presented in Figures 6. As with the index-flood model, Figure 6a reveals that the RVN-HYB method leads to the best performance in terms of the RRMSE. Although all methods differ by less than 2% in terms of NHS, results indicate that NHS values corresponding to CCA and RVN-HYB are inferior to those corresponding to the regression model applied on all gauged sites, which corresponds to $n = 150$ in Figure 6b. However, CCA leads to the best relative efficiency as indicated by the ARE criterion in Figure 6d. Hence, CCA corresponds to the regression models with, on average, the lowest uncertainties. This indicates that flood quantiles may be better reference variables for the regression-based model than for the index-flood model and suggests that in general different reference variables may be more appropriate for different situations. Nevertheless, the two close lines in Figure 6d reveal that for the same neighborhood size the RVN-LM has similar ARE values than with CCA. In terms of AHM, Figure 6c is identical to Figure 4c except that new neighborhood sizes are indicated in circles.

5. CONCLUSIONS

A general methodology was investigated to improve homogenous properties of neighborhoods in RFA. A procedure to calculate relevant reference variables at a target location prior to the RFA was proposed to improve neighborhood properties and to reduce uncertainties. The predicted values of

reference variables represent the unknown centers of neighborhoods delineated according to a distance of gauged sites with respect to the centers. The proposed method represents a generalization of both ROI and CCA methods in RFA. The proposed RVN method has the advantages of accepting various groups of reference variables, of considering nonlinear interrelations and of being more objective since L-moments are used instead of estimated flood quantiles from at-site analysis.

In this study, the reference variables correspond to transformed L-moments. The resulting RVN-LM and RVN-HYB methods were applied on sites located in Southern Quebec, Canada, to predict flood quantiles corresponding to the 100 years by both index-flood and regression-based models. The prediction of the reference variables at target locations showed that after proper transformations, L1 can be linearly related to the site characteristics, but no proper transformations are found for the other L-moments. This justifies the consideration of the PPR method to account for the nonlinearity in the prediction of the reference variables. In general, any non-parametric regression model, such as generalized additive models or artificial neural networks, could be considered instead of PPR to account for the nonlinearity. Nevertheless, the PPR approach unveils direction vectors that provide explicit, parsimonious and meaningful regression equations.

Although none of the methods performed best for all criteria, cross-validation showed that the proposed RVN method perform well in comparison to the traditional ROI and CCA methods. In both the index-flood and the regression-based model the best RRMSE is obtained by RVN_HYB and the best NHS is obtained by RVN_LM. In particular, the favorable RRMSE obtained by RVN-HYB are due to more robust estimation of problematic sites. However, RVN_LM has the best balance, because it achieves the best or the second best values for all criteria. Most importantly, the utilization of hydrological reference variables with the CCA and RVN methods has reduced the uncertainty on the regional LCV, the index-flood and the predicted flood quantiles, in comparison to ROI. Consequently, prior modeling of hydrological reference variables was shown to be advantageous to the delineation of neighborhoods in RFA.

The present study has made specific assumptions in order to investigate the RVN method in a certain well defined condition. Nevertheless, the rational of predicting hydrological reference variables in *a priori* analysis remains a valid approach when other choices of regression models, neighborhood forms and metrics are considered. Hence, more comparative studies should be carried out to evaluate alternatives to fixed size neighborhoods and Euclidian distances in the specific context of the RVN framework.

The L-coefficient of skewness is commonly used in RFA to describe the shape of a distribution. Consequently, to improve the result of the RVN method, further research efforts could focus on improving the prediction of this crucial reference variable. One way to improve the prior analysis of the hydrological reference variables is the consideration of the unequal sampling error. This aspect is often considered in the estimation of flood quantiles in RFA, but may also play an important role in the prior analysis of the RVN method.

ACKNOWLEDGEMENT

Financial support for this study was graciously provided by the Natural Sciences and Engineering Research Council (NSERC) of Canada.

REFERENCES

Acreman, M., Sinclair, C., 1986. Classification of drainage basins according to their physical characteristics; an application for flood frequency analysis in Scotland. J. Hydrol. 84, 365–380.

Bishop, C.M., 1995. Neural networks for pattern recognition. Oxford university press.

Burn, D.H., 1990. An appraisal of the “region of influence” approach to flood frequency analysis. Hydrol. Sci. J. 35, 149–166. doi:10.1080/02626669009492415

Castiglioni, S., Castellarin, A., Montanari, A., 2009. Prediction of low-flow indices in ungauged basins through physiographical space-based interpolation. J. Hydrol. 378, 272 – 280. doi:10.1016/j.jhydrol.2009.09.032

Chebana, F., Charron, C., Ouarda, T.B.M.J., Martel, B., 2014. Regional frequency analysis at ungauged sites with the generalized additive model. J. Hydrometeorol. 15, 2418–2428. doi:10.1175/JHM-D-14-0060.1

Chebana, F., Ouarda, T.B.M.J., 2009. Index flood–based multivariate regional frequency analysis. Water Resour. Res. 45. doi:10.1029/2008WR007490

Chebana, F., Ouarda, T.B.M.J., 2008. Depth and homogeneity in regional flood frequency analysis. Water Resour. Res. 44. doi:10.1029/2007WR006771

Chebana, F., Ouarda, T.B.M.J., 2007. Multivariate L-moment homogeneity test. Water Resour. Res. 43. doi:10.1029/2006WR005639

Chokmani, K., Ouarda, T.B.M.J., 2004. Physiographical space-based kriging for regional flood frequency estimation at ungauged sites. Water Resour. Res. 40. doi:10.1029/2003WR002983

Cunderlik, J.M., Burn, D.H., 2006. Switching the pooling similarity distances: Mahalanobis for Euclidean. Water Resour. Res. 42. doi:10.1029/2005WR004245

Cunnane, C., 1988. Methods and merits of regional flood frequency analysis. J. Hydrol. 100, 269–290.

Dalrymple, T., 1960. Flood-frequency analysis. Surv. Water-Supply Pap. 1543.

Das, S., Cunnane, C., 2010. Examination of homogeneity of selected Irish pooling groups. Hydrol.

Earth Syst. Sci. Discuss. 7, 5099–5130. doi:10.5194/hessd-7-5099-2010

Dawson, C.W., Abrahart, R.J., Shamseldin, A.Y., Wilby, R.L., 2006. Flood estimation at ungauged sites using artificial neural networks. J. Hydrol. 319, 391 – 409. doi:10.1016/j.jhydrol.2005.07.032

Durocher, M., Chebana, F., Ouarda, T.B.M.J., 2015. A Nonlinear Approach to Regional Flood Frequency Analysis Using Projection Pursuit Regression. J. Hydrometeorol. doi:10.1175/JHM-D-14-0227.1

Eng, K., Tasker, G.D., Milly, P., 2005. An Analysis of Region-Of-Influence methods for flood regionalization in the Gulf-Atlantic rolling plain. J. Am. Water Resour. Assoc. 41, 135–143. doi:10.1111/j.1752-1688.2005.tb03723.x

Friedman, J., Grosse, E., Stuetzle, W., 1983. Multidimensional Additive Spline Approximation. SIAM J. Sci. Stat. Comput. 4, 291–301. doi:10.1137/0904023

Friedman, J.H., Tukey, J.W., 1974. A projection pursuit algorithm for exploratory data analysis. Comput. IEEE Trans. On 100, 881–890.

GREHYS, 1996a. Presentation and review of some methods for regional flood frequency analysis. J. Hydrol. 186, 63–84.

GREHYS, 1996b. Inter-comparison of regional flood frequency procedures for canadian rivers. J. Hydrol. 186, 85–103.

Haddad, K., Rahman, A., 2012. Regional flood frequency analysis in eastern Australia: Bayesian GLS regression-based methods within fixed region and ROI framework – Quantile Regression vs. Parameter Regression Technique. J. Hydrol. 430–431, 142 – 161. doi:10.1016/j.jhydrol.2012.02.012

Hastie, T., Tibshirani, R., Friedman, J.H., 2009. The elements of statistical learning: data mining, inference, and prediction, Springer series in statistics. Springer.

He, Y., Bárdossy, A., Zehe, E., 2011. A review of regionalisation for continuous streamflow simulation. Hydrol. Earth Syst. Sci. 15, 3539–3553. doi:10.5194/hess-15-3539-2011

Hosking, J.R.M., Wallis, J.R., 1997. Regional frequency analysis: an approach based on L-moments.

Cambridge Univ Pr.

Hwang, J.-N., Lay, S.-R., Maechler, M., Martin, R.D., Schimert, J., 1994. Regression modeling in back-propagation and projection pursuit learning. Neural Netw. IEEE Trans. On 5, 342–353. doi:10.1109/72.286906

Laio, F., Ganora, D., Claps, P., Galeati, G., 2011. Spatially smooth regional estimation of the flood frequency curve (with uncertainty). J. Hydrol. 408, 67 – 77. doi:http://dx.doi.org/10.1016/j.jhydrol.2011.07.022

Ouali, D., Chebana, F., Ouarda, T.B.M.J., 2015. Non-linear canonical correlation analysis in regional frequency analysis. Stoch. Environ. Res. Risk Assess. 1–14. doi:10.1007/s00477-015-1092-7

Ouarda, T.B.M.J., Ba, K.M., Diaz-Delgado, C., Carsteanu, A., Chokmani, K., Gingras, H., Quentin, E., Trujillo, E., Bobee, B., 2008. Intercomparison of regional flood frequency estimation methods at ungauged sites for a Mexican case study. J. Hydrol. 348, 40–58. doi:10.1016/j.jhydrol.2007.09.031

Ouarda, T.B.M.J., Girard, C., Cavadias, G.S., Bobée, B., 2001. Regional flood frequency estimation with canonical correlation analysis. J. Hydrol. 254, 157 – 173. doi:10.1016/S0022-1694(01)00488-7

Ouarda, T.B.M.J., Shu, C., 2009. Regional low-flow frequency analysis using single and ensemble artificial neural networks. Water Resour. Res. 45. doi:10.1029/2008WR007196

Ouarda, T., Haché, M., Bruneau, P., Bobée, B., 2000. Regional Flood Peak and Volume Estimation in Northern Canadian Basin. J. Cold Reg. Eng. 14, 176–191. doi:10.1061/(ASCE)0887-381X(2000)14:4(176)

Pandey, G.R., 1998. Assessment of scaling behavior of regional floods. J. Hydrol. Eng. 3, 169–173. doi:10.1061/(ASCE)1084-0699(1998)3:3(169)

Pandey, G.R., Nguyen, V.-T.-V., 1999. A comparative study of regression based methods in regional flood frequency analysis. J. Hydrol. 225, 92–101. doi:10.1016/S0022-1694(99)00135-3

Reis, D., Stedinger, J., Martins, E., 2005. Bayesian generalized least squares regression with application to log Pearson type 3 regional skew estimation. Water Resour. Res. 41.

Stedinger, J., Lu, L.-H., 1995. Appraisal of regional and index flood quantile estimators. Stoch. Hydrol. Hydraul. 9, 49–75.

Tasker, G., Hodge, S., Bark, S., 1996. Region of Influence regression for estimating the 50-years flood at ungaged sites. Water Resour. Bull. doi:10.1111/j.1752-1688.1996.tb03444.x

Viglione, A., Laio, F., Claps, P., 2007. A comparison of homogeneity tests for regional frequency analysis. Water Resour. Res. 43, n/a–n/a. doi:10.1029/2006WR005095

Yu, Y., Ruppert, D., 2002. Penalized spline estimation for partially linear single-index models. J. Am. Stat. Assoc. 97, 1042–1054. doi:10.1198/016214502388618861

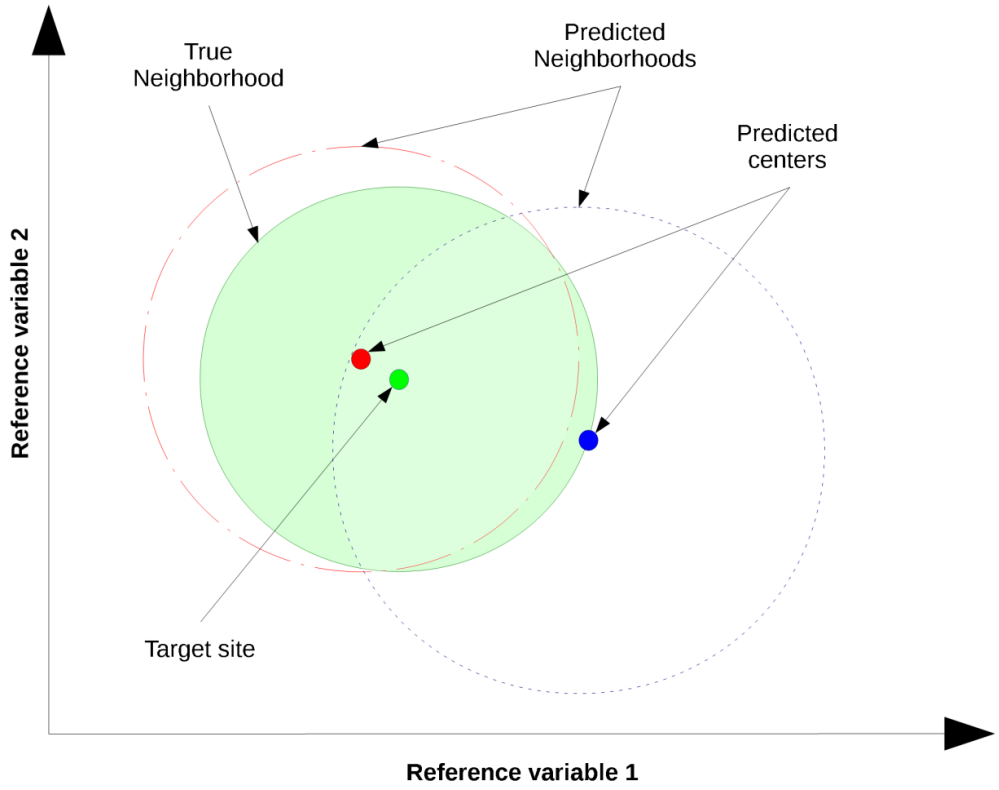

Figure 1: Illustration of the neighborhoods obtained by the RVN method.

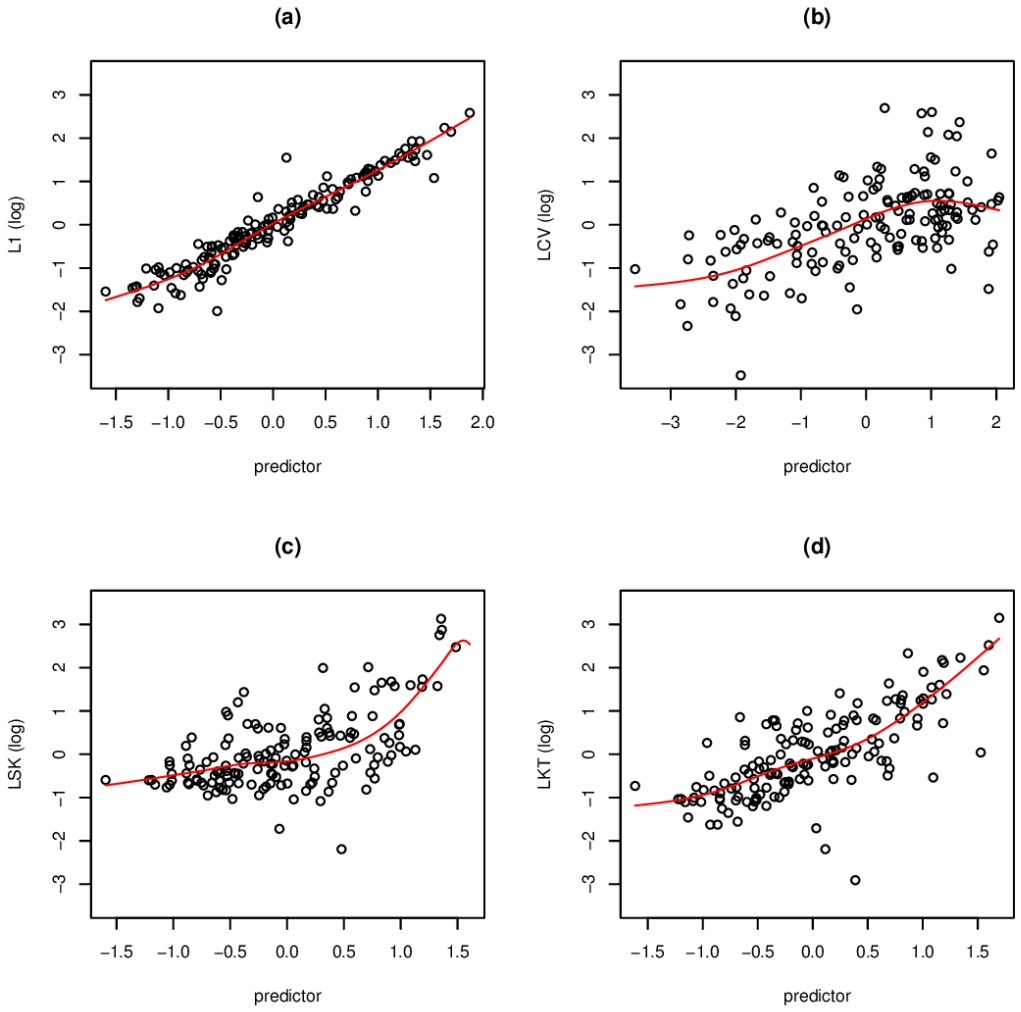

Figure 2: Residuals of the reference variables by PPR methods.

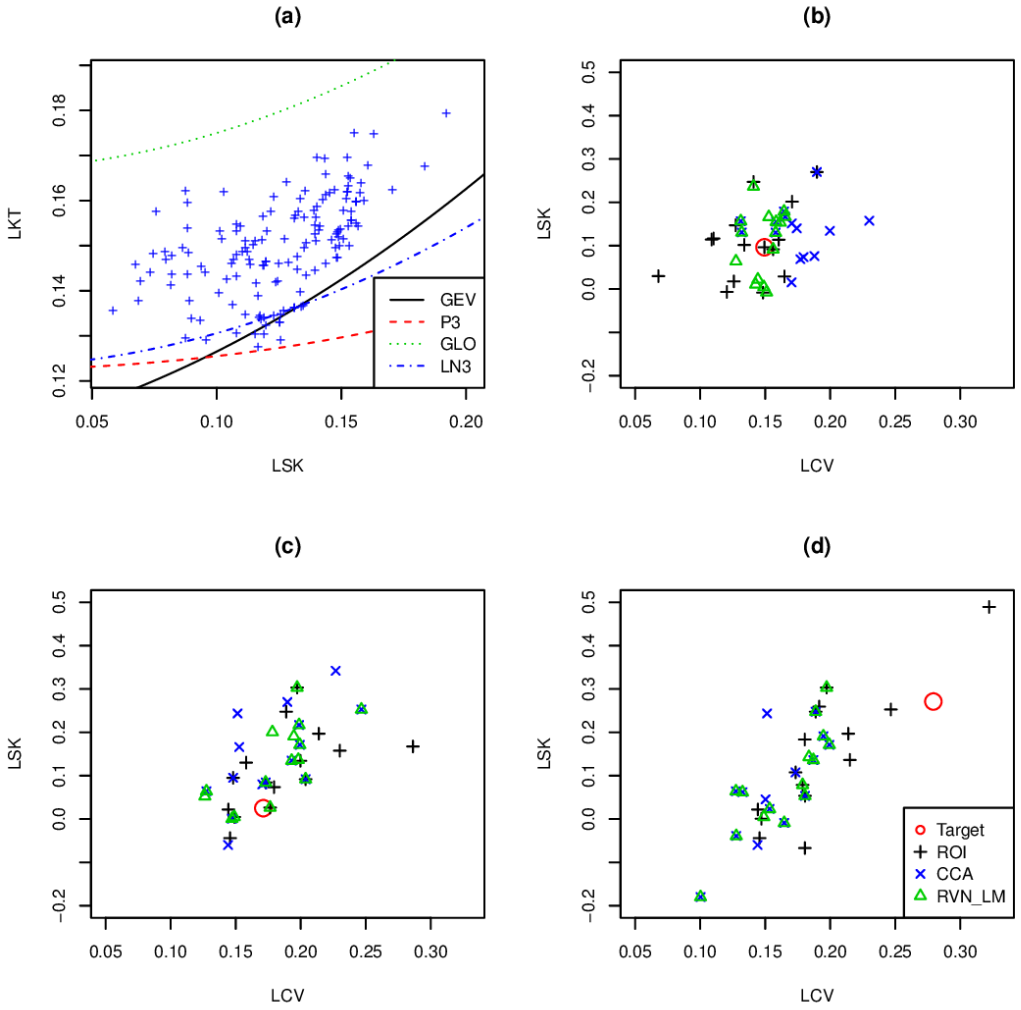

Figure 3: L-moments ratio diagram for index-flood model. (a) Regional L-moments for RVN_LM with 29 gauged sites. (b) Regional L-moments based on the 15 nearest gauged sites for 3 selected target locations.

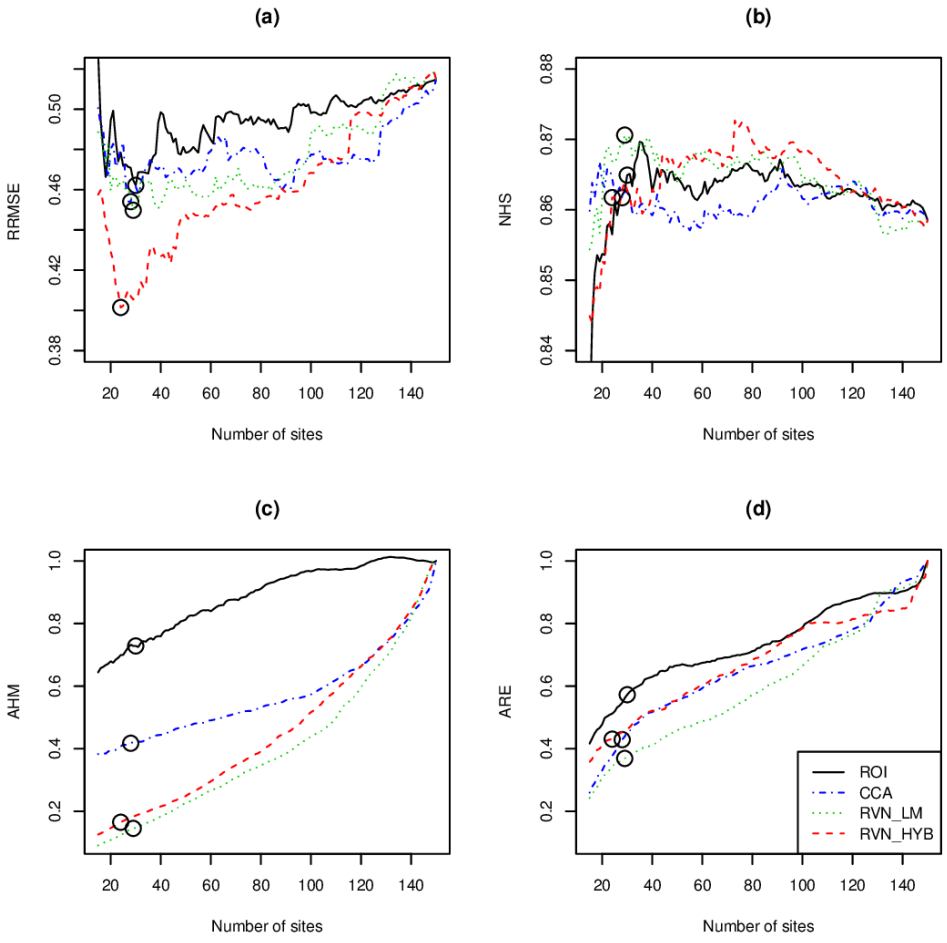

Figure 4: Evaluation criteria for the index-flood model. Calibrated models are represented by circles.

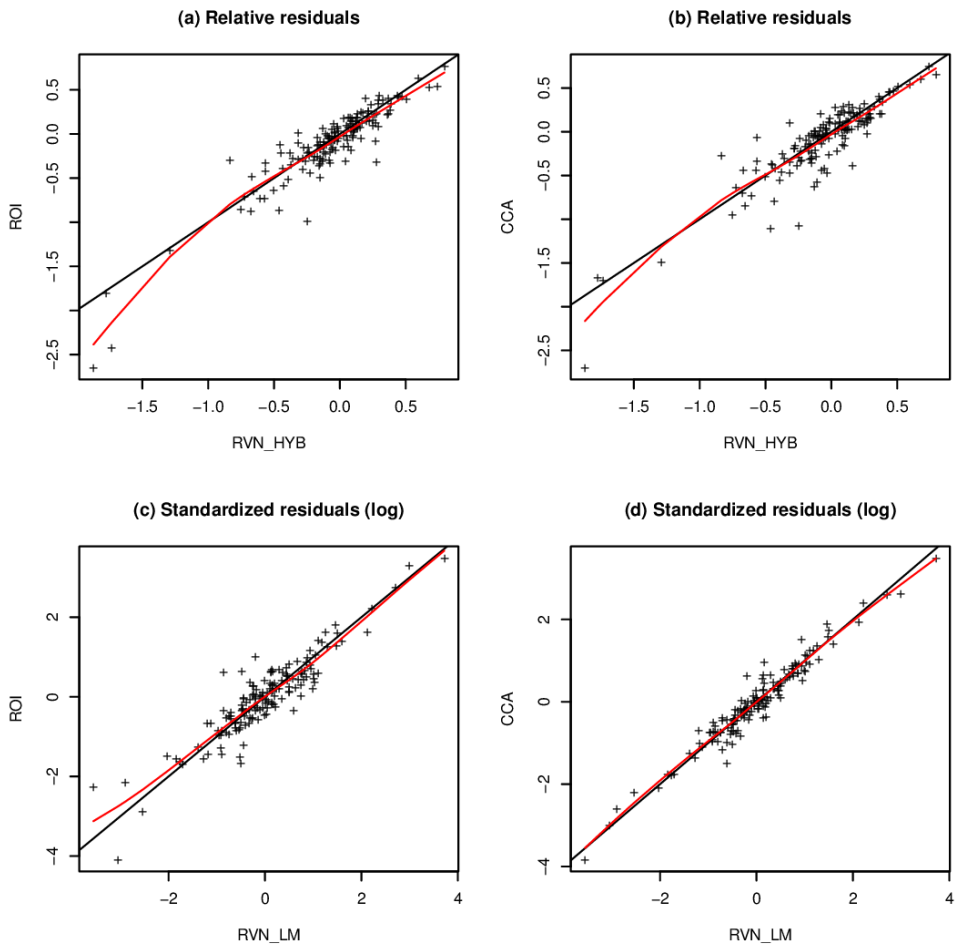

Figure 5: Comparison of the cross-validation residuals for Q100 between different methods. The black line is the unitary slope and the red line is a smooth fitting of the residuals.

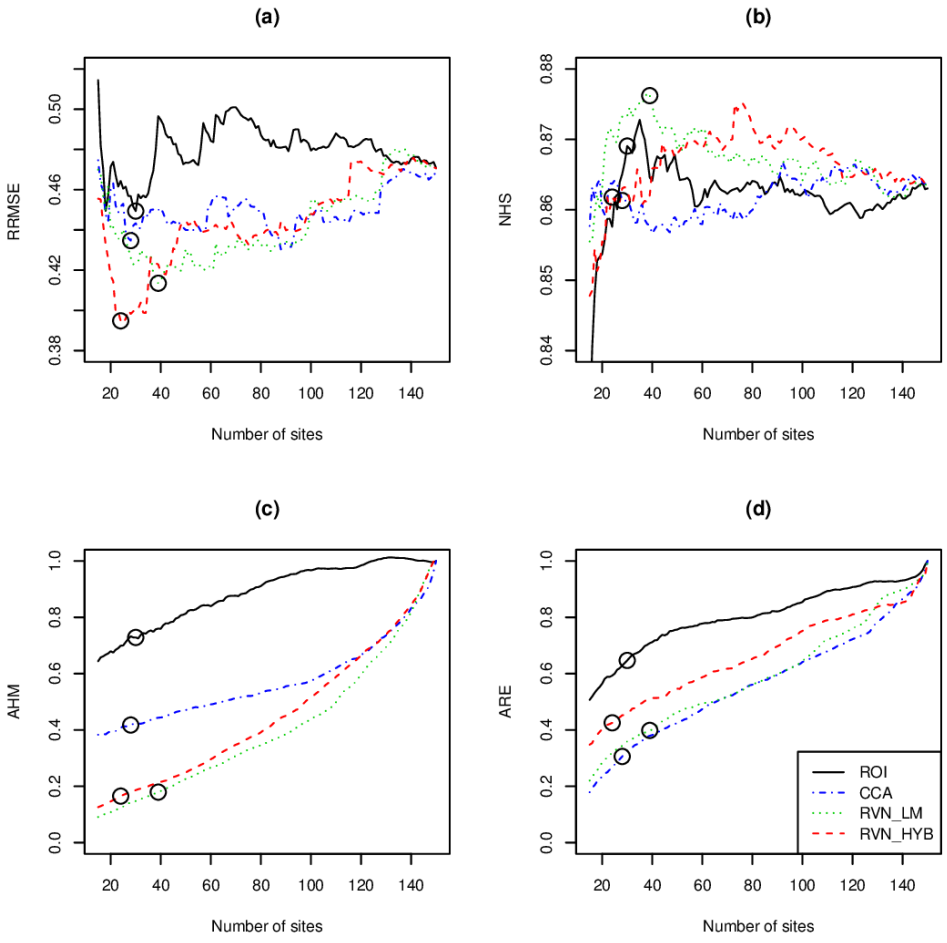

Figure 6: Evaluation criteria for the quantile-based technique. Calibrated models are represented by circles.

CHAPITRE 4:

ON THE PREDICTION OF EXTREME FLOOD QUANTILES AT UNGAUGED LOCATIONS WITH SPATIAL GAUSSIAN COPULA

On the prediction of extreme flood quantiles at ungauged locations with spatial gaussian copula

Martin Durocher [*,1], Fateh Chebana [1] and Taha B. M. J. Ouarda[1,2]

[1]Institut National de Recherche Scientifique (INRS-ETE), University of Québec
490 de la Couronne, Québec G1K 9A9, Canada

[2] Institute Center for Water Advanced Technology and Environnemental Research (iWater),
Masdar Institute of Science and Technology
P.O. Box 54224, Abu Dhabi, UAE

Submitted for publication

ABSTRACT

The present study investigates the use of the spatial copula approach for predicting flood quantiles at ungauged basins. Spatial copulas are the formalization of traditional geostatistics by copulas. In regional flood frequency analysis (RFFA), the regression of flood quantiles is often carried out at the logarithmic scale. Consequently, traditional interpolation methods introduce a bias and provide suboptimal predictions. In this study, the copula framework is examined for offering proper corrections in this framework. Moreover, copula techniques separate the regional distribution of flood quantiles from spatial dependence. This provides a full probabilistic model that represents a more flexible framework where proper combinations of regional distribution and dependence can be adapted to various situations that are encountered in RFFA. The adequacy of the investigated methodology is evaluated on a real world case study involving hydrometric stations from southern Quebec, Canada. Results show that the spatial copula framework is able to deal with the problem of bias, is robust to the presence of problematic stations and may improve the quality of quantile predictions while reducing the level of complexity of the models used in RFFA.

Keywords: Flood analysis, physiographical-based kriging, regional frequency analysis, spatial copula, ungauged basin, interpolation.

1. INTRODUCTION

Knowing the likelihood of threatening events is of primary interest for water resource managers. In practice, a large number of years of record is required in order to collect adequate information for carrying out a reliable at-site frequency analysis. Regional flood frequency analysis relies on at-site quantile estimates to transfer hydrological information to an ungauged site. There are two main approaches in regional flood frequency analysis (RFFA) for predicting flood quantiles at ungauged locations: The quantile based regression which predicts directly at-site flood quantiles corresponding to a given return period (Chebana and Ouarda, 2008; Ouarda et al., 2001; Pandey and Nguyen, 1999; Tasker et al., 1996) and the parameter based regression that predicts the at-site distribution before calculating the desired flood quantiles (Chebana and Ouarda, 2009; Eng et al., 2007; Haddad and Rahman, 2012; Hosking and Wallis, 1997). Notice that the latter approach includes as a special case the popular index flood model (Hosking and Wallis, 1997). In the following, the quantile based regression approach is investigated.

Spatial modeling was shown to be useful in RFFA in a large number of publications (Archfield et al., 2013; Castiglioni et al., 2009; Chokmani and Ouarda, 2004; Hundecha et al., 2008; Nezhad et al., 2010; Skøien et al., 2006). The flood generating processes of a given basin may depend on several factors including the physiographical characteristics of the basin and the meteorological conditions in the region of study. Two rivers representing different basin characteristics can produce river discharges possessing very different characteristics even if they are very close geographically. Direct interpolation from the geographical coordinates is therefore inappropriate. This motivated the development of suitable spaces, called physiographical spaces, where hydrological variables can be treated as spatial data. On the other hand, it was found that hybrid techniques that combine both geographical distance and similarity between the basin characteristics can improve the quality of regional models (Eng et al., 2007; Haddad and Rahman, 2012). Geographical distance is often

implicitly included in RFFA models through the use of the latitude and longitude as physiographical characteristics. Therefore, the term physiographical space is used in the present work to designate general spaces, in which spatial methods can be applied.

Kriging techniques are developed for predicting statistical variables, such as precipitation, at unknown locations as linear combinations of observed values. One property that may explain the popularity of kriging, is that it minimizes the predictive mean square errors (Schabenberger and Gotway, 2004). However, this property assumes that spatial data can be characterized by a multivariate normal distribution (MVN), which is usually not valid for flood quantiles. Conversely, the logarithm transformation is commonly used for describing the exponential relationship between flood quantiles and basin characteristics (Pandey and Nguyen, 1999). At the original scale, this strategy creates bias and produces suboptimal predictions because of the nonlinear nature of the data (Schabenberger and Gotway, 2004).

Copulas have been introduced in the field of hydrology for providing a more realistic evaluation of risks associated to flood events by characterizing multiple aspects of the hydrographs (e.g. Chebana and Ouarda, 2009, 2007; Requena et al., 2013; Salvadori and De Michele, 2004; Shiau et al., 2006). In this framework, copulas offer a simple and efficient way to account for the dependence between different aspects of the hydrograph, such as peak, duration and volume. The usefulness of the copula framework emanates from the simplicity by which the model separates the marginal distribution from the dependence structure. This decomposition of the model allows combinations of proper components for fitting data that would otherwise be difficult to describe.

The motivation for using copulas to describe the dependence structure of spatial data was discussed by Bárdossy (2006). The objective was to develop a more general approach for characterizing spatial patterns that traditional methods fail to model. Building a copula that allows easy formulation of the spatial structure according to the separating distance can be a difficult task and not all copulas are suited for spatial analysis. The Gaussian copula is a particular case that ensures the

continuity with the existing methodologies, while allowing marginals to belong to various classes of distributions. For predictive purposes, Bárdossy and Li (2008) showed that copulas provide a practical way of writing a predictive distribution at a new location. Adopting a copula framework for spatial modelling, therefore called spatial copula, is a way of generalizing traditional kriging techniques such as indicator kriging, trans-Gaussian kriging and normal rank kriging (Kazianka and Pilz, 2010), as well as providing additional flexibility.

This study investigates the spatial copula approach for predicting flood quantiles at ungauged locations. One of the interests of considering spatial copulas is to provide a correction for the bias and to account properly for the nonnormal distribution of the at-site flood quantiles in order to define optimal predictors. To this end, spatial copulas provide a simple expression of the mean of the predictive distribution, which in terms of square errors corresponds to the best predictors at the original scale (Bárdossy and Li, 2008). Accordingly, the present work aims at evaluating the gain of accuracy made by adopting spatial copulas in RFFA. A second objective is to examine how the more general framework of spatial copulas may be used to provide predictions that are more robust and to adapt to problematic situations.

The present study is organised as follows. Section 2 provides a review of the methodology of spatial copulas and proposes some specific adaptations to RFFA. Section 3 illustrates a real world application to hydrometric stations in the southern region of the province of Quebec, Canada. Finally, discussions are provided in Section 4 and concluding remarks are drawn in Section 5.

2. METHODOLOGY

We only provide a short introduction on copula theory in this section for assuring the completeness of the document. For a more detailed treatment on the topic of copulas, the reader is referred to Nelsen (2006) and Salvadori et al., 2007. The present section also follows closely the

general methodology developed in Bárdossy (2006) and Bárdossy and Li (2008), where further discussion on the technical aspect of spatial copulas can be found.

The methodology section is divided in several subsections that develop the steps to follow for performing RFFA in a spatial copula framework. Section 2.1 reviews the basics of the copula theory. These are necessary to develop the methodology of the present work. Section 2.2 discusses the particularity of the copulas adapted to spatial analysis and provides an adaptation of the general structure to RFFA. Section 2.3, presents common methods used to build the physiographical space and shows how to integrate this information in a regional model. Sections 2.4 and 2.5 describe respectively the dependence and the marginal components that characterize the regional model. Finally, Section 2.6 proposes estimation procedures and Section 2.7 explains how to predict the flood quantiles of the spatial copula model.

2.1 BASIC CONCEPTS

A d -dimensional copula

$$C:[0,1]^d \rightarrow [0,1] \qquad (1)$$

is a multivariate distribution defined on the unit hypercube, which respects some regularity conditions (Nelsen, 2006). The main result concerning copulas is known as the Sklar's theorem that states that every multivariate distribution *G* can be expressed in function of a copula C and marginals $\{F_i\}_{i=1}^{d}$:

$$G(\mathbf{x}) = C\left[F_1(x_1),\ldots \qquad)\right] \qquad (2)$$

where $\mathbf{x}' = (x_1,\ldots \qquad \mathbb{R}$. Moreover, if each marginal F_i is continuous, then C is unique. Conversely, copulas can be constructed from existing distributions:

$$C(\mathbf{u}) = G\left[F_1^{-1}(u_1),\ldots \qquad {}_{l})\right] \qquad (3)$$

where $\mathbf{u}' = (u_1, \ldots$ $0,1]^d$. The values u_i correspond to observations on the unit hypercube. An important copula built from (3) is the Gaussian copula for which $G = \Phi_\Sigma$ is a MVN with correlation matrix Σ and $F_i = \phi$ are standard univariate normal distributions. Both Φ_Σ and ϕ have a zero mean and a unit variance.

Copulas aim to provide a framework where dependence is treated separately from marginal distributions. As suggested by (2), the modelling strategy consists of finding the best combination of a copula *C* and marginals F_i, even if the resulting distribution G is not a member of a well-known distribution family. Another appealing property of copulas is their invariance to strictly monotonic transformations of the variables. This property removes the subjectivity of choosing the best transformation, which otherwise affects the shape of the dependence (Nelsen, 2006).

2.2 SPATIAL COPULAS

Spatial data may be seen as a single realization of a random field $Z(\mathbf{s}) \in \mathbb{R}$ defined as a set of random variables indexed by a location $\mathbf{s} \in \mathbb{R}$. Accordingly, let $\mathbf{z} = (z_1, \ldots$ be a set of observations of the random field Z where every observation is simultaneously collected at respective location $\mathbf{s}_i$. For RFFA, $\mathbf{z}$ represents at-site flood quantiles estimated from gauged stations located at $\mathbf{s}_i$ in the physiographical space. The modeling strategy is to represent the multivariate distribution of the random field Z with marginal distributions $F_Z = F_1 = \ldots$ and a dependence structure characterized by a spatial copula C_Z.

A spatial copula is a n-dimensional copula that has the same dimension as the number of locations. Hence, most copulas developed uniquely for restricted dimensions, usually 2 or 3, are inappropriate. Moreover, the structure of a spatial copula C_Z must allow for the evolution of the dependence with respect to the separation $\mathbf{h}$ between two locations $\mathbf{s}$ and $\mathbf{s} + \mathbf{h}$. Close locations

must allow for complete dependence and conversely, the strength of the dependence must gradually fade out as the distance $h = \|\mathbf{h}\|$ increases. This can be written in a copula notation as:

$$\text{if } h \to 0 \text{ then } C_Z \to M_n \tag{4}$$

where M_n is the Frechet upper bound, as well as:

$$\text{if } h \to \infty \text{ then } C_Z \to \Pi_n \tag{5}$$

where Π_n is the independent copula (Nelsen, 2006). The properties (4) and (5) rule out some of the best known n-dimensionnal copulas like the Fairlie-Gumbel-Morgenstern family (Bárdossy, 2006). Although an interesting family of copulas is the elliptical family, however only the Gaussian copulas have a practical structure that can properly relate the spatial dependence to the separating distance (Kazianka and Pilz, 2010). These conditions imposed to the spatial copulas, highlight the attractiveness of the Gaussian copula for spatial modelling and illustrate the difficulty of finding suitable alternatives.

2.3 PHYSIOGRAPHICAL SPACE

The first step in RFFA for spatial modelling is to define a physiographical space in which hydrological variables can be considered as continuous and for which interpolation methods can be applied. Chokmani and Ouarda (2004) indicated procedures for building physiographical spaces from statistical techniques, such as principal component analysis (PCA) and canonical correlation analysis (CCA). In particular, they recommend using the first two canonical pairs obtained from CCA as new axis of the physiographical space. The CCA approach has the property of optimizing the correlation between the axis of the physiographical space and the hydrological variables. Comparison with PCA has showed that CCA tends to lead to better prediction performances (Chokmani and Ouarda, 2004; Guillemette et al., 2009).

Both PCA and the CCA approaches provide a projection matrix A allowing passing from the original basin characteristics $\{\mathbf{x}_i\}_{i=1}^{n}$ to the coordinates of the physiographical space:

$$\mathbf{s}_i = A\mathbf{x}_i \tag{6}$$

For spatial methods in RFFA, the coordinates $\mathbf{s}_i$ are the locations of the gauged stations for which flood quantiles are observed as a random field Z. Every $\mathbf{s}_i$ summarizes then the information necessary to evaluate the proximity between sites. In practice, the dimension of the physiographical space is generally inferior to the number of basin characteristics. Consequently, two rivers with different basin characteristics may have the same coordinates $\mathbf{s}_i$ in the physiographical space. This aspect creates a measurement error that needs to be accounted for in the dependence structure of the spatial copula model.

2.4 DEPENDENCE COMPONENT

Traditional geostatistics prefer describing the spatial dependence by the mean of a variogram, which expresses the variance between pairs of locations with respect to the separating distance. However, the variogram is associated to both the marginals and the copula. Consequently, the variogram is an unnatural way of describing the dependence in the spatial copula framework.

In spatial analysis, the utilization of the Gaussian copula C_θ, of parameter θ is motivated by the practical form of the correlation matrix $\Sigma = \{\rho_{i,j}\}_{i=1,\mathrm{j}=1}^{n,n}$. If $h_{i,j}$ represent the distance in the physiographical space between locations $\mathbf{s}_i$ and $\mathbf{s}_j$, then each coefficient can be directly determined by a correlation function $\rho_{i,j} = \rho(h_{i,j} \mid \theta)$. For instance, the correlation function may be written as:

$$\rho(h \mid \lambda, \tau) = \begin{cases} 1 & h = 0 \\ (1-\tau)\,\gamma(h \mid \lambda) & h > 0 \end{cases} \tag{7}$$

where $\lambda > 0$ and $0 \le \tau \le 1$. Notice that in (7) the whole correlation matrix Σ is efficiently characterized by two parameters $\theta = (\lambda, \tau)$. The function γ represents any basic correlation functions that are frequently used in spatial analysis (see e.g. Schabenberger and Gotway, 2004). A discontinuity at $h = 0$ is created by the parameter τ and plays a similar role to the nugget effect in the traditional variogram. In RFFA, one justification for the consideration of $\tau \neq 0$ is to account for the measurement error created by the reduction of the dimension of the physiographical space that may create identical coordinates $\mathbf{s}_i$. Additionally, the parameter λ represents the range of the correlation function γ, which corresponds to the point where the correlation almost vanishes.

2.5 MARGINAL COMPONENT

Additionally to the dependence component of the regional model that describes the spatial structures, a marginal component that characterizes the regional distribution of the at-site flood quantiles is required. In the following, the vector of parameters defining the form of this marginal is denoted η. Traditionally, adding a spatial trend is subject to debate as it is usually impossible to discriminate between persistence due to either spatial correlation or deterministic trend. In RFFA, it is shown that the addition of a trend may improve the prediction capabilities of the interpolating methods (Castiglioni et al., 2009; Nezhad et al., 2010). These results should not be surprising because of the way the physiographical space is built. For instance, CCA explicitly finds canonical vectors that optimize the correlation with the flood quantiles. Hence, the axis of the physiographical space generated by these canonical vectors should have a strong correlation with the flood quantiles, which creates the spatial trend. Consequently, it is advised to consider η as a more general response surface.

In general, there is no systematic way for defining a response surface. However, some guidelines may be followed. As mentioned earlier, the logarithm transformation is frequently applied to the flood quantiles. In continuity with this traditional approach, a lognormal distribution with location

parameter μ^* and a scale parameter σ^* may be used. These parameters correspond respectively to the mean and the standard deviation of the distribution at the logarithm scale. In the spatial copula framework, the traditional geostatistics models with trend are equivalent to choosing the response surface:

$$\begin{aligned} \mu^*(\mathbf{s}_i) &= \mathbf{s}_i \beta_\mu \\ \sigma^*(\mathbf{s}_i) &= \beta_\sigma \end{aligned} \tag{8}$$

where β_μ and $\beta_\sigma > 0$ are parameters in $\mathbb{R}$. The linear trend in (8) for the location parameter μ^* corresponds to the deterministic pattern expected for usual physiographical spaces. Moreover, the scale parameter σ^* is assumed constant and reflects the belief that the variability does not change in the physiographical space.

Nevertheless, the spatial copula framework accepts more sophisticated response surfaces. This allows for a generalisation of the methods commonly used in RFFA. For instance, the variability may evolve differently in the physiographical space and hence, the scale parameter σ^* should be allowed to change with respect to $\mathbf{s}_i$. More flexible probability distributions with 3 or more parameters can also be considered. In a non RFFA context, Kazianka and Pilz (2010) used the generalized extreme value distribution that accounts for various forms of skewness in the response surface.

2.6 ESTIMATION

Joint estimation of parameters (η, θ) can be obtained with the maximum likelihood method. To account for the spatial trend, distinct marginals are assumed for each location $\mathbf{s}_i$, where the density of the distribution $F_{\eta,i}$ at location $\mathbf{s}_i$ is denoted $f_{\eta,i}$. In addition, let c_θ be the density of the spatial copula C_θ. Specific formulas for the Gaussian copulas are provided in Appendix 1. Following this notation, the full log-likelihood can be expressed as:

$$l(\mathbf{z} \mid \eta, \theta) = l_{ind}(\mathbf{z} \mid \eta) + \log\left\{c_\theta\left[F_{\eta,1}(z_1), \ldots, F_{\eta,n}(z_n)\right]\right\} \quad (9)$$

where $l_{ind}\left(\mathbf{z} \mid \eta\right) = \sum_{i=1}^{n} \log\left[f_{\eta,i}\left(z_i\right)\right]$ is the independent log-likelihood (Bárdossy and Li, 2008). It is seen in (9) that the log-likelihood is the combination of the usual log-likelihood function l_{ind} (as if the dependence structure is ignored) and a dependence factor associated to the copula density c_θ.

Alternative estimation procedures exist for situations where the maximum likelihood is difficult to compute or is leading to unsatisfactory estimations. This includes the least squares fitting of the empirical rank correlations (Bárdossy, 2006) and the optimization of composite likelihood functions (Kazianka and Pilz, 2010). For the latter, one composite likelihood that is useful for spatial data analysis is the pairwise likelihood (Heagerty and Lele, 1998; Varin, 2008) which corresponds to the product of all possible bivariate density functions. If $\mathbf{z}_{i,j} = \left(z_i, z_j\right)$ denotes a pair of locations, then the pairwise log-likelihood takes the form:

$$l_p(\mathbf{z} \mid \eta, \theta) = \sum_{i<j}\left(l_{ind}(\mathbf{z}_{i,j} \mid \eta) + \log\left\{c_\theta\left[F_\eta(z_i), F_\eta(z_j)\right]\right\}\right) \quad (10)$$

Inference from the composite likelihood estimator is similar to the traditional likelihood approach. Under a number of assumptions generally met in practice, the estimates converge to an asymptotic normal distribution. The resulting estimator is unbiased and its covariance structure is given by the robust covariance matrix (Varin, 2008).

2.7 PREDICTION

In traditional applications of spatial methods in RFFA, flood quantile prediction is performed by kriging at the logarithmic scale. In this situation, the regional distribution of the at-site flood quantiles is almost normal and the median agrees with the mean. Conversely, at the original scale the median and the mean disagree as the inverse transformation preserves the median, but not the mean. This creates a bias in the prediction process. Moreover, traditional kriging predictors are linear as they

correspond to a weighted sum of the observations, which is not generally optimal (Schabenberger and Gotway, 2004).

In the construction of a spatial model that is primordial for interpolation that a spatial copula can pass from dimension n to $n+1$. One possible way of predicting spatial data in the copula framework consists in calculating the plugin predictive distribution (PPD) of a new location $\mathbf{s}_0$, which is the predictive distribution of $z_0 = Z(\mathbf{s}_0)$ given the estimated parameters $(\hat{\eta}, \hat{\theta})$ and the observations $\mathbf{z}$. For simplicity, let $\mathbf{w}' = (w_1, \ldots$ be the pseudo-observations calculated as $w_i = F_{\hat{\eta},i}(z_i)$. Accordingly, the conditional density of z_0 knowing $\mathbf{z}$ is given by (Bárdossy and Li, 2008):

$$p(z_0 \mid \mathbf{z}) = f_{\hat{\eta},0}(z_0) \times c_{\hat{\theta}}\left[F_{\hat{\eta},0}^{-1}(z_0) \mid \mathbf{w}\right] \tag{11}$$

where the conditional copula density is:

$$c_{\hat{\theta}}(w \mid \mathbf{w}) = \frac{c_{\hat{\theta}}(w, \mathbf{w})}{c_{\hat{\theta}}(\mathbf{w})} \tag{12}$$

Formulas for evaluating the conditional density $c_{\hat{\theta}}$ of the Gaussian copula are provided in Appendix 1.

The expression of the PPD in (11) allows the calculation of any quantities of the distribution of z_0 given $\mathbf{z}$, and not only the means as it is the case with traditional kriging predictors. For instance, the quantile $Q_t = F_{\hat{\eta},0}^{-1}(w^*)$ corresponding to a probability t for the PPD may be obtained by solving numerically for w^* in the equation:

$$t = \int_0^{w^*} c_{\hat{\theta}}(u \mid \mathbf{w})\, du \tag{13}$$

For the spatial copula model, a first predictor of flood quantiles at an ungauged location is obtained by evaluating the median of the PPD $\tilde{}$. Also, solving (13) for $t = (\alpha/2, 1-\alpha/2)$ can be used to provide confidence intervals of level α.

A second predictor of z_0 is the mean of the PPD, which results in a predictor that minimizes the mean square error (Bárdossy and Li, 2008). There is no general formula for computing the mean of the PPD, but its value may be evaluated by computing the finite integral:

$$\bar{z}_0 = \int_0^1 F_{\hat{\eta},0}^{-1}(u) \times c_{\hat{\theta}}(u \mid \mathbf{w})\, du \tag{14}$$

To quantify the variability associated to the predictor $\bar{z}_0$, the variance of the PPD is given by:

$$\hat{\mathrm{V}}(z_0) = \int_0^1 \left[F_{\hat{\eta},0}^{-1}(u) - \bar{z}_0 \right]^2 \times c_{\hat{\theta}}(u \mid \mathbf{w})\, du \tag{15}$$

Notice that conversely to the traditional kriging predictor, the predictions obtained as the median or the mean of the PPD are constructed nonlinearly and are thus more general.

3. CASE STUDY

3.1 DATA

In this section, the methodology of spatial copula is applied to 151 hydrometric stations located in Southern Quebec (Canada) between the 45th and 55th parallel North. Figure 1 illustrates a map of the available stations. The selected hydrometric stations are located in the inhabited region of the province and include rivers with natural flow regimes. An historical record period of at least 15 years is imposed for the selected stations. The drainage areas of the selected stations range between 200 km^2 and 100 000 km^2. The data series of all stations pass the tests of stationarity, homogeneity and independence. For more information concerning the at-site frequency analysis of these rivers the reader is referred to Chokmani and Ouarda (2004). Several references have used these results for investigating new RFFA methodologies (Chebana and Ouarda, 2008; Chebana et al., 2014; Chokmani and Ouarda, 2004; Nezhad et al., 2010; Shu and Ouarda, 2007; Wazneh et al., 2013).

In the present study, the hydrological variables of interest are the flood quantiles corresponding to return periods of 10 and 100 years standardized by the drainage area. These variables are respectively denoted Q10 and Q100. As shown in Table 1, to insure the comparability with previous studies the same 5 basin characteristics are used. Accordingly, for the construction of the physiographical space, the basin characteristics are transformed by a logarithmic function and standardized to obtain approximately normal distributions. One exception is PLAC for which a square root transformation is preferred. Then, CCA is used for extracting 2 new axis, which constitute the desired physiographical space. The coordinates in the resulting physiographical space are finally standardized to provide comparable scales between them.

3.2 MODEL CALIBRATION

Leave-one-out cross-validation is used for guiding the configuration of the regional model. In turn, each gauged station $\mathbf{s}_i$ with flood quantiles z_i is considered as ungauged and the spatial model is fitted on the remaining stations for obtaining the predicted flood quantiles $\hat{z}_i$. The first performance criterion for the evaluation of the predictive model is the relative root mean square error:

$$RMSEr = 100\sqrt{\frac{1}{n}\sum_{i=1}^{n}\left(\frac{z_i - \hat{z}_i}{z_i}\right)^2} \qquad (16)$$

This performance criterion is preferred in the present situation for evaluating of the global accuracy of the model relatively to the levels of the hydrological variables. Additionally to the $RMSEr$, the relative bias:

$$BIASr = 100\frac{1}{n}\sum_{i=1}^{n}\left(\frac{y_i - \hat{y}_i}{y_i}\right) \qquad (17)$$

is also examined for detecting systematic errors.

Bárdossy (2006) showed that examining bivariate spatial copulas with respect to the distance are useful tools. The advantage of using a copula in that context is that not only the correlation can be

examined, but the adequacy of the copula itself. In the specific case of a Gaussian copula the parameter of the copula is the coefficient of correlation. Accordingly, the pairs of locations are grouped into l classes of distances $h_1,\ldots$ for which empirical copulas can be fitted individually. This way, each empirical copula is fitted separately and offers an approximation of the strength of the spatial dependence for the classes h_k. The classes are separated to have equivalent size. Preliminary analysis shows that empirical correlations became low or negative for distances $h > 1.0$. Consequently, these classes are ignored in the calibration of the model. Figure 2 presents the empirical correlation coefficients obtained by fitting the classes h_k. It is observed that the strength of the spatial dependence is not very high and that important nugget effects are present.

In a preliminary study not reported here, the maximum likelihood procedure was found to lead to an unsatisfactory estimation of the dependence parameters θ, because all pairs of sites are used, including those at great distances. Alternatively, the pairwise likelihood approach is preferred, where the evaluation of the pairwise likelihood in (10) is modified to include only the pairs of sites with pairwise distances inferior to the threshold $h \leq 1.0$. Accordingly, the correlation function of the best spatial copula model is of the form (8) where the basic correlation function is:

$$\gamma(h \mid \lambda) = \exp\left[-3\left(\frac{h}{\lambda} \right)^2 \right] \qquad (18)$$

The theoretical correlation coefficients are represented in Figure 2 as solid lines and show acceptable agreements.

In addition to the visual diagnostic provided by Figure 2, the goodness-of-fit test of Bárdossy, (2006) is carried out to validate the Gaussian Copula. Note that the test does not validate the n-dimensional copula, but instead it validates the bivariate copulas for each class h_k. A failure of the test for the classes h_k indicates the inadequacy of the bivariate copula and hence the Gaussian Copula. In

the present study, for each class h_k inside the threshold $h \leq 1.0$ the p-values found validate the hypothesis of a Gaussian copula with significant levels of at least 10%.

Based on the cross-validation study, it is found that the marginal component of the best regional model uses a lognormal distribution with a linear trend for both location μ^* and scale σ^* parameters as a function of the first coordinates of the physiographical space. The validation of the response surface can be carried out by comparing the sample quantiles with the theoretical quantiles of normal distribution. To simplify the comparison, the quantiles are compared at the scale of a standard normal distribution. To this end, the pseudo-observations $w_i = F_{\hat{\eta},i}(z_i)$ are calculated and transformed to a standard normal distribution function as $\phi^{-1}(w_i)$. The results are illustrated in Figure 3 and show that sample quantiles are inside the 95% confidence interval bounds, which confirms the choice of the response surface.

3.3 RESULTS

A grid of locations in the physiographical space is used to produce maps of the flood quantiles. For being numerically more efficient, only a neighbourhood of the closest gauged stations is used to compute the PPD (13). This simplification is preferred to a distance threshold for assuring a minimum of gauged stations during each computation of the PPD (11). Neighborhoods of size 15 to 80 are tested by steps of 5 stations. Based on the cross-validation, the optimal size is 20. The results are presented in Figure 4 and indicate the evolution of the mean of the PPD (14) as well as the standard deviation (15). In the four maps, the deterministic trends may be distinguished by the global persistence from left to right.

In a previous study on the same dataset, Chokmani and Ouarda (2004) have identified 6 problematic gauged stations, for which the drainage area (BV) is underestimated or the fraction of basin covered by lakes (PLAC) is overevaluated. Visual diagnostics are presented in Figure 5 to examine the residuals of the cross-validation procedures and to verify the impact of these problematic

stations on the estimation of the spatial copula model. In panels (A) and (C) of Figure 5, the relative residuals are presented with respect to the predicted values. It is seen that the problematic stations correspond to the most important discrepancies. The spatial copula model included a linear trend for the scale parameter σ^* with respect of the first coordinates of the physiographical space. Because the first coordinate is constructed by CCA, it is strongly correlated with flood quantiles. Consequently, non-constant variability must be observed in the residuals. In panels (B) and (D) of Figure 5, it is seen in the lower flood quantiles that the problematic stations create higher variability. This shows that the full probabilistic model provided by the spatial copula approach is able to adapt to the particularity of the data, which makes the present methodology more robust to the influence of problematic stations.

To evaluate the relative performance of the spatial copula model, the performance criteria are compared to those obtained in previous studies carried out on the same set of hydrometric stations. The results of the comparison are presented in Table 2. It is observed that the predictions given by the median of the PPD are superior to the mean of the PPD for both bias and accuracy. Their difference between the two predictors is the consequence of the skewness of the lognormal distribution that composes the response surface, for which the two central tendency measures disagree.

Comparatively to classical ordinary and residual kriging approaches, the spatial copula improved the performance criteria. In general, the $RMSEr$ indicates that the spatial copula using the median of the PPD has an accuracy that is better to the best other methods, including the Ensemble ANN in the CCA-space and the traditional method with depth functions. Moreover, the important measures of bias for traditional kriging methods show their inadequacy in the present situation. This may be explained by the difficulty of the method to properly deal with the problematic stations and the non-constant variability. Conversely, the regional model with spatial copulas reaches the best levels of relative bias of all spatial methods for this dataset.

4. DISCUSSION

The strategy of using the PPD for predicting flood quantiles makes the hypothesis that (η, θ) is known. Consequently, the variability measured by the plugin variance (15) or the confidence intervals evaluated by the quantiles of the PPD (13) does not account for the uncertainty due to the estimation of the model. One answer to this issue is provided by Kazianka and Pilz, (2011) who developed a Bayesian framework for obtaining the a posteriori distribution of the parameters. Therefore, the true predictive distribution at a new location can be approximated by computing the plugin estimator for each element of the posterior distribution.

The present study focuses uniquely on the Gaussian copula. However, every step of the present methodology may be adapted to other spatial copulas. An important property of the Gaussian copula is radial symmetry, which implies that the dependence structure is identical for lower and upper quantiles. In mathematical terms, radial symmetry means

$$c(u_1, \ldots \qquad \ldots \qquad (19)$$

where c is the density of the copula C. This property is not always realistic (Bárdossy, 2006), which leads to the adoption of asymmetrical families like the υ-transformed copulas (Bárdossy and Li, 2008). Another approach for building non-Gaussian copulas is by using vine copulas, in which multidimensional copulas are constructed by assembling copulas of lower dimension (Gräler and Pebesma, 2011). The adaptation of the present methodology to general spatial copulas is straightforward and consists in using the proper copula densities for evaluating the likelihood function and the conditional density.

5. CONCLUSIONS

The present study investigates the benefit of using the spatial copula approach in RFFA. A methodology based on the spatial copula framework is investigated for providing a more general

framework for the prediction of flood quantiles at ungauged basins. The spatial copula methodology leads to a simple expression of the PPD, which allows straightforward calculations of nonlinear predictors at ungauged locations.

The case study of the hydrometric stations located in southern Quebec provides evidence that proper interpolation methods can improve the quality of the predictions. The flexibility of the spatial copula framework allowed specifying a response surface that accounts for the difference of variability observed between the lower and the higher flood quantiles, which in the present case study was caused by problematic stations. This shows that with the proper settings, the spatial copula approach is a robust prediction method. The important relative bias observed with traditional kriging approaches is reduced considerably with the spatial copula approach. Actually, the overall predictive performance of the spatial copula model reaches similar levels to the best methods applied on the same dataset. These excellent performances are obtained despite the fact that the spatial copula model is simpler and less computationally intensive than these RFFA methods. The results of cross-validation suggest also that, in terms of relative accuracy, the median of the PPD represents a superior predictor than the mean.

A limited number of non-Gaussian copulas are known to be appropriate for spatial analysis. Nevertheless, choosing from different marginal distributions and copulas gives the possibility of adapting spatial copula models to various situations where traditional approaches may fail. Future research efforts may lead to new spatial copulas that possess interesting properties for RFFA. Their adaptation to the present methodology may contribute to a better modelling of the dependence between hydrometrics stations.

ACKNOWLEDGMENTS

Financial support for this study was graciously provided by the Natural Sciences and Engineering Research Council (NSERC) of Canada.

APPENDIX 1

Here the procedure for calculating the density of a Gaussian copula is presented. The density is required for evaluating the likelihood and the pairwise likelihood function in (9) and (10). Namely, a Gaussian copula is constructed as:

$$C(\mathbf{u}) = \Phi\ \left[\phi^{-1}(u_1),\dots\qquad\right] \tag{A.1}$$

where $\mathbf{u}' = (u_1,\dots\qquad 0,1]^d$. The univariate distribution ϕ designates a standard normal distribution and the MVN Φ have zero mean and unit variance. The Gaussian copula C is absolutely continuous and hence, the derivative of (A.1) leads to the density

$$c(u) = \frac{g\left[\phi^{-1}(u_1),\dots\qquad\right]}{\prod_{i=1}^{n} f\left(\phi_i^{-1}(u_i)\right)} \tag{A.2}$$

where g and f_i are respectively the density functions of Φ and ϕ (Kazianka and Pilz, 2010).

The evaluation of the conditional density is necessary for computing the characteristics of the PPD including the mean (13) and the standard deviation (14). Let $\mathbf{w} = (w_1,\dots\qquad$ be the pseudo-observations in the unit hypercube $[0,1]^n$ and w_0 the pseudo-observation of an unknown location. Accordingly, denote:

$$\Sigma = \begin{pmatrix} \Sigma_{11} & \Sigma_{12} \\ \Sigma_{21} & \Sigma_{22} \end{pmatrix} \tag{A.3}$$

a block decomposition of the covariance matrice of $(w_0, \mathbf{w})$ where the vector Σ_{12} is the vector of correlation between the value w_0 and $\mathbf{w}$. Σ_{22} is the correlation matrix of $\mathbf{w}$. Assuming $\mathbf{a} = \left[F^{-1}(w_1),\dots\qquad_n)\right]$ as well as $h \sim N(\mu, \sigma^2)$ with mean $\mu = \Sigma_{12}\Sigma_{22}^{-1}\mathbf{a}$ and variance $\sigma^2 = 1 - \Sigma_{12}\Sigma_{22}^{-1}\Sigma_{21}$, the desired conditional density for the Gaussian copula is (Kazianka et Pilz 2010):

$$c(u \mid \mathbf{w}) = \frac{h\left[\phi^{-1}(u) \mid \mathbf{w}\right]}{f\left[\phi^{-1}(u)\right]} \tag{A.4}$$

REFERENCES

Archfield, S., Pugliese, A., Castellarin, A., Skøien, J., Kiang, J., 2013. Topological and canonical kriging for design flood prediction in ungauged catchments: an improvement over a traditional regional regression approach? Hydrol. Earth Syst. Sci. 17. doi:10.5194/hess-17-1575-2013

Bárdossy, A., 2006. Copula-based geostatistical models for groundwater quality parameters. Water Resour. Res. 42. doi:10.1029/2005WR004754

Bárdossy, A., Li, J., 2008. Geostatistical interpolation using copulas. Water Resour. Res. 44.

Castiglioni, S., Castellarin, A., Montanari, A., 2009. Prediction of low-flow indices in ungauged basins through physiographical space-based interpolation. J. Hydrol. 378, 272 – 280. doi:10.1016/j.jhydrol.2009.09.032

Chebana, F., Charron, C., Ouarda, T.B.M.J., Martel, B., 2014. Regional frequency analysis at ungauged sites with the generalized additive model. J. Hydrometeorol. (Accepted).

Chebana, F., Ouarda, T.B.M.J., 2007. Multivariate L-moment homogeneity test. Water Resour. Res. 43.

Chebana, F., Ouarda, T.B.M.J., 2008. Depth and homogeneity in regional flood frequency analysis. Water Resour. Res. 44. doi:10.1029/2007WR006771

Chebana, F., Ouarda, T.B.M.J., 2009. Index flood–based multivariate regional frequency analysis. Water Resour. Res. 45. doi:10.1029/2008WR007490

Chokmani, K., Ouarda, T.B.M.J., 2004. Physiographical space-based kriging for regional flood frequency estimation at ungauged sites. Water Resour. Res. 40. doi:10.1029/2003WR002983

Eaton, B., Church, M., Ham, D., 2002. Scaling and regionalization of flood flows in British Columbia, Canada. Hydrol. Process. 16, 3245–3263. doi:10.1002/hyp.1100

Eng, K., Milly, P., Tasker, G., 2007. Flood regionalization: q hybrid geographic and predictor-variable region-of-influence regression method. J. Hydrol. Eng. 12, 585–591. doi:10.1061/(ASCE)1084-0699(2007)12:6(585)

Gräler, B., Pebesma, E., 2011. The pair-copula construction for spatial data: a new approach to model spatial dependency. Procedia Environ. Sci. 7, 206–211. doi:10.1016/j.proenv.2011.07.036

Guillemette, N., St-Hilaire, A., Ouarda, T.B.M.J., Bergeron, N., Robichaud, É., Bilodeau, L., 2009. Feasibility study of a geostatistical modelling of monthly maximum stream temperatures in a multivariate space. J. Hydrol. 364, 1–12. doi:10.1016/j.jhydrol.2008.10.002

Haddad, K., Rahman, A., 2012. Regional flood frequency analysis in eastern Australia: Bayesian GLS regression-based methods within fixed region and ROI framework – Quantile Regression vs. Parameter Regression Technique. J. Hydrol. 430–431, 142 – 161. doi:10.1016/j.jhydrol.2012.02.012

Heagerty, P.J., Lele, S.R., 1998. A composite likelihood approach to binary spatial data. J. Am. Stat. Assoc. 93, 1099–1111. doi:10.1080/01621459.1998.10473771

Hosking, J.R.M., Wallis, J.R., 1997. Regional frequency analysis: an approach based on L-moments. Cambridge Univ Pr.

Hundecha, Y., Ouarda, T.B.M.J., Bárdossy, A., 2008. Regional estimation of parameters of a rainfall-runoff model at ungauged watersheds using the “spatial” structures of the parameters within a canonical physiographic-climatic space. Water Resour. Res. 44. doi:10.1029/2006WR005439

Kazianka, H., Pilz, J., 2010. Copula-based geostatistical modeling of continuous and discrete data including covariates. Stoch. Environ. Res. Risk Assess. 24, 661–673. doi:10.1007/s00477-009-0353-8

Kazianka, H., Pilz, J., 2011. Bayesian spatial modeling and interpolation using copulas. Comput. Geosci. 37, 310–319. doi:10.1016/j.cageo.2010.06.005

Nelsen, R.B., 2006. An introduction to copulas. Springer.

Nezhad, M.K., Chokmani, K., Ouarda, T.B.M.J., Barbet, M., Bruneau, P., 2010. Regional flood frequency analysis using residual kriging in physiographical space. Hydrol. Process. 24, 2045–2055. doi:10.1002/hyp.7631

Ouarda, T.B.M.J., Girard, C., Cavadias, G.S., Bobée, B., 2001. Regional flood frequency estimation with canonical correlation analysis. J. Hydrol. 254, 157 – 173. doi:10.1016/S0022-

1694(01)00488-7

Pandey, G., Nguyen, V., 1999. A comparative study of regression based methods in regional flood frequency analysis. J. Hydrol. 225, 92 – 101. doi:10.1016/S0022-1694(99)00135-3

Requena, A.I., Mediero, L., Garrote, L., 2013. A bivariate return period based on copulas for hydrologic dam design: accounting for reservoir routing in risk estimation. Hydrol. Earth Syst. Sci. 17, 3023–3038. doi:10.5194/hess-17-3023-2013

Salvadori, G., De Michele, C., 2004. Frequency analysis via copulas: Theoretical aspects and applications to hydrological events. Water Resour. Res. 40. doi:10.1029/2004WR003133

Salvadori, G., De Michele, C., Kottegoda, N., Rosso, R., 2007. Extremes in nature: an approach using copulas. Springer Verlag.

Schabenberger, O., Gotway, C.A., 2004. Statistical methods for spatial data analysis. CRC Press.

Shiau, J.-T., Wang, H.-Y., Tsai, C.-T., 2006. Bivariate frequency analysis of floods using copulas. J. Am. Water Resour. Assoc. 42, 1549–1564. doi:10.1111/j.1752-1688.2006.tb06020.x

Shu, C., Ouarda, T.B.M.J., 2007. Flood frequency analysis at ungauged sites using artificial neural networks in canonical correlation analysis physiographic space. Water Resour. Res. 43. doi:10.1029/2006WR005142

Skøien, J.O., Merz, R., Blöschl, G., 2006. Top-kriging - geostatistics on stream networks. Hydrol. Earth Syst. Sci. 10, 277–287. doi:10.5194/hess-10-277-2006

Tasker, G., Hodge, S., Bark, S., 1996. Region of Influence regression for estimating the 50-years flood at ungaged sites. Water Resour. Bull. doi:10.1111/j.1752-1688.1996.tb03444.x

Varin, C., 2008. On composite marginal likelihoods. AStA Adv. Stat. Anal. 92, 1–28. doi:10.1007/s10182-008-0060-7

Wazneh, H., Chebana, F., Ouarda, T.B.M.J., 2013. Optimal depth-based regional frequency analysis. Hydrol Earth Syst Sci 2281–2296. doi:10.5194/hess-17-2281-2013

Table 1: Descriptive statistics for hydrological variables and basin characteristics

Variable	Notation	min	mean	max	sd
Flood quantile of 10 years (m^3/s)	Q10	53	698	5649	828
Flood quantile of 100 years (m^3/s)	Q100	64	913	7013	1048
Drainage area (km^2)	BV	208	6 265	96 600	11 713
Mean slope of the basin (°)	PMBV	0.96	2.43	6.81	0.99
Fraction of basin occupied by lakes (%)	PLAC	0.03	7.72	47	7.99
Mean total annual precipitation (mm)	PTMA	646	988	1 534	154
Degree-day below 0 Celsius (dgr-day)	DJBZ	8 589	16 346	29 631	5 385

Table 2: Leave-one-out cross-validation for spatial copulas and previous research using the same dataset.

		Q10		Q100	
Method	**Reference**	*BIASr* **(%)**	*RMSEr* **(%)**	*BIASr* **(%)**	*RMSEr* **(%)**
Spatial Methods					
Spatial Copula – mean	This study	-8	38	-11	45
Spatial Copula – median	This study	**-3**	**35**	**-4**	**41**
Residual kriging in CCA-space	Nezhad et al., 2010	-7	39	-14	58
Ordinary kriging in CCA-space	Chokmani and Ouarda, 2004	-16	51	-23	70
Ordinary kriging in PCA-space	Chokmani and Ouarda, 2004	-20	66	-27	86
Other methods					
Traditional CCA method	Chokmani and Ouarda, 2004	-9	43	-11	51
Traditional with depth function	Wazneh et al., 2013	**-3**	38	**-2**	**44**
GAM	Chebana et al., 2014	-5	41	-8	49
Single ANN	Shu and Ouarda, 2007	-7	47	-7	64
Ensemble ANN	Shu and Ouarda, 2007	-11	44	10	60
Single ANN in CCA-space	Shu and Ouarda, 2007	-5	38	-7	46
Ensemble ANN in CCA-space	Shu and Ouarda, 2007	-5	**37**	-6	45

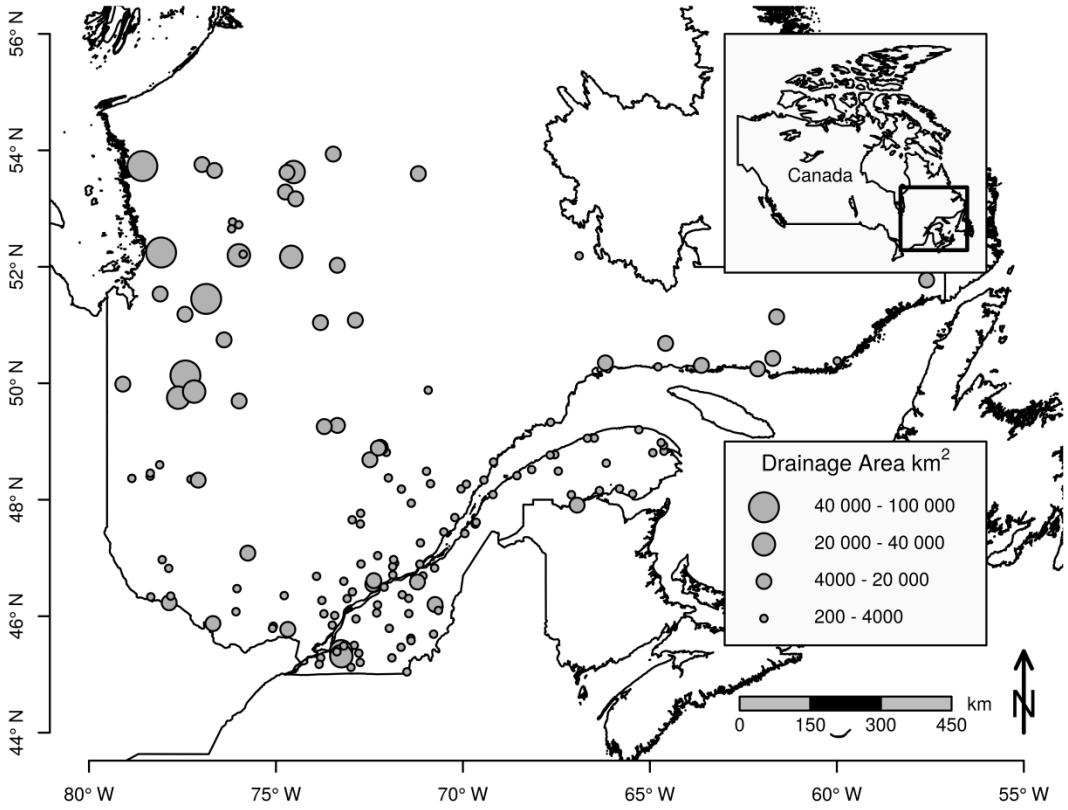

Figure 1: Location of the 151 hydrometric stations in southern Quebec, Canada

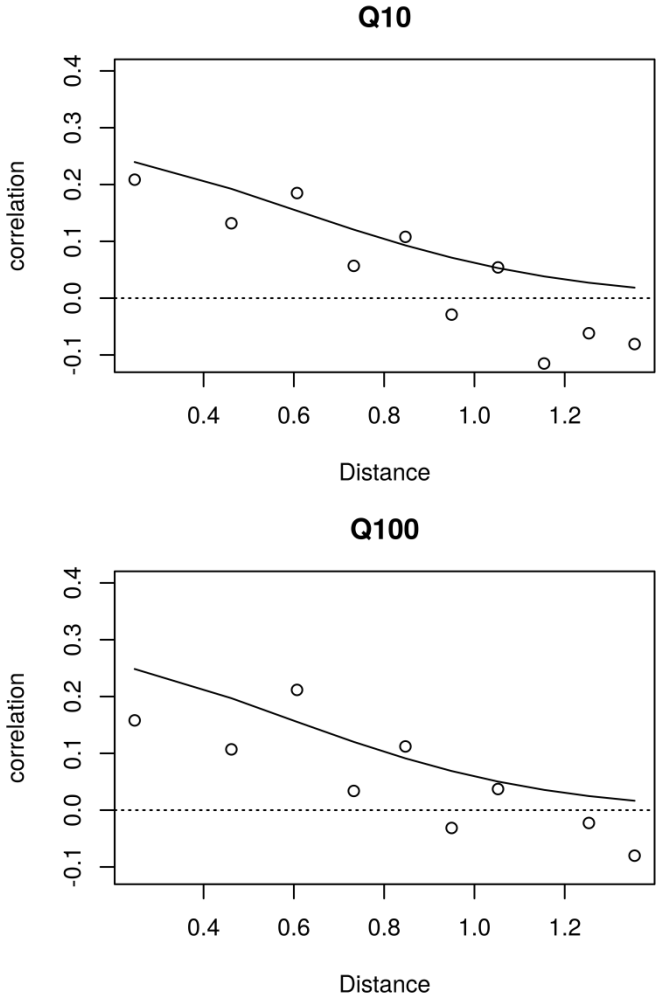

Figure 2: Correlogram with respect of the separating distance. The circles represent empirical correlations and the solid line designates the fitted correlation function.

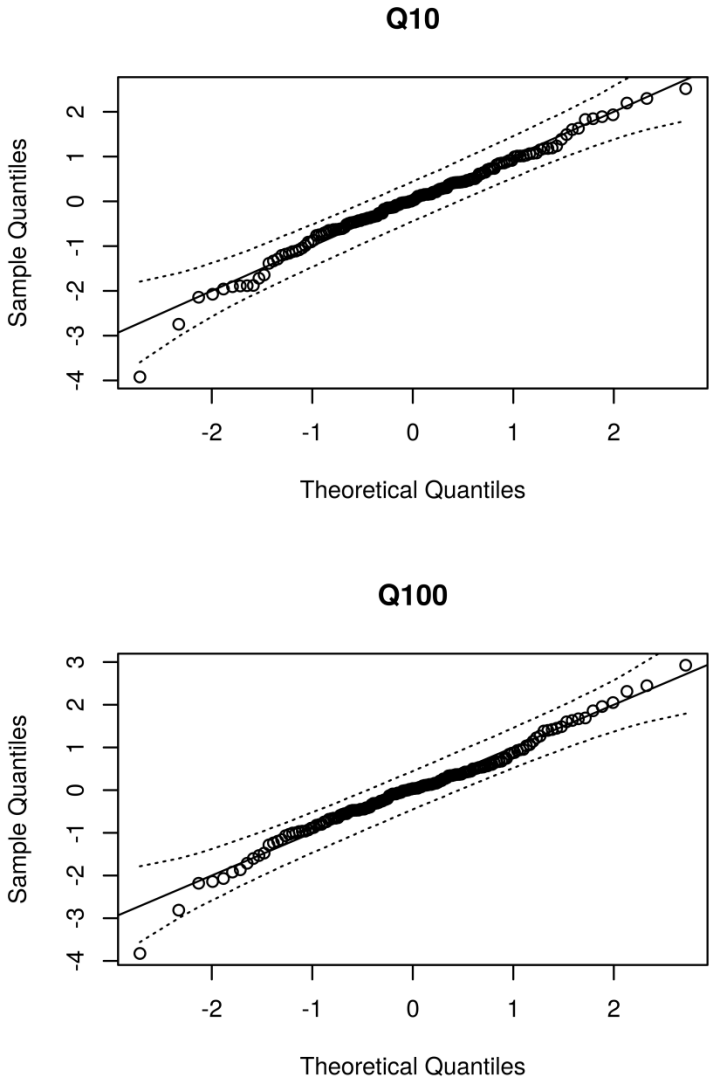

Figure 3: Theoretical quantiles versus sample quantiles of the response surface. The dashed lines are the bounds of the 95% confidence interval.

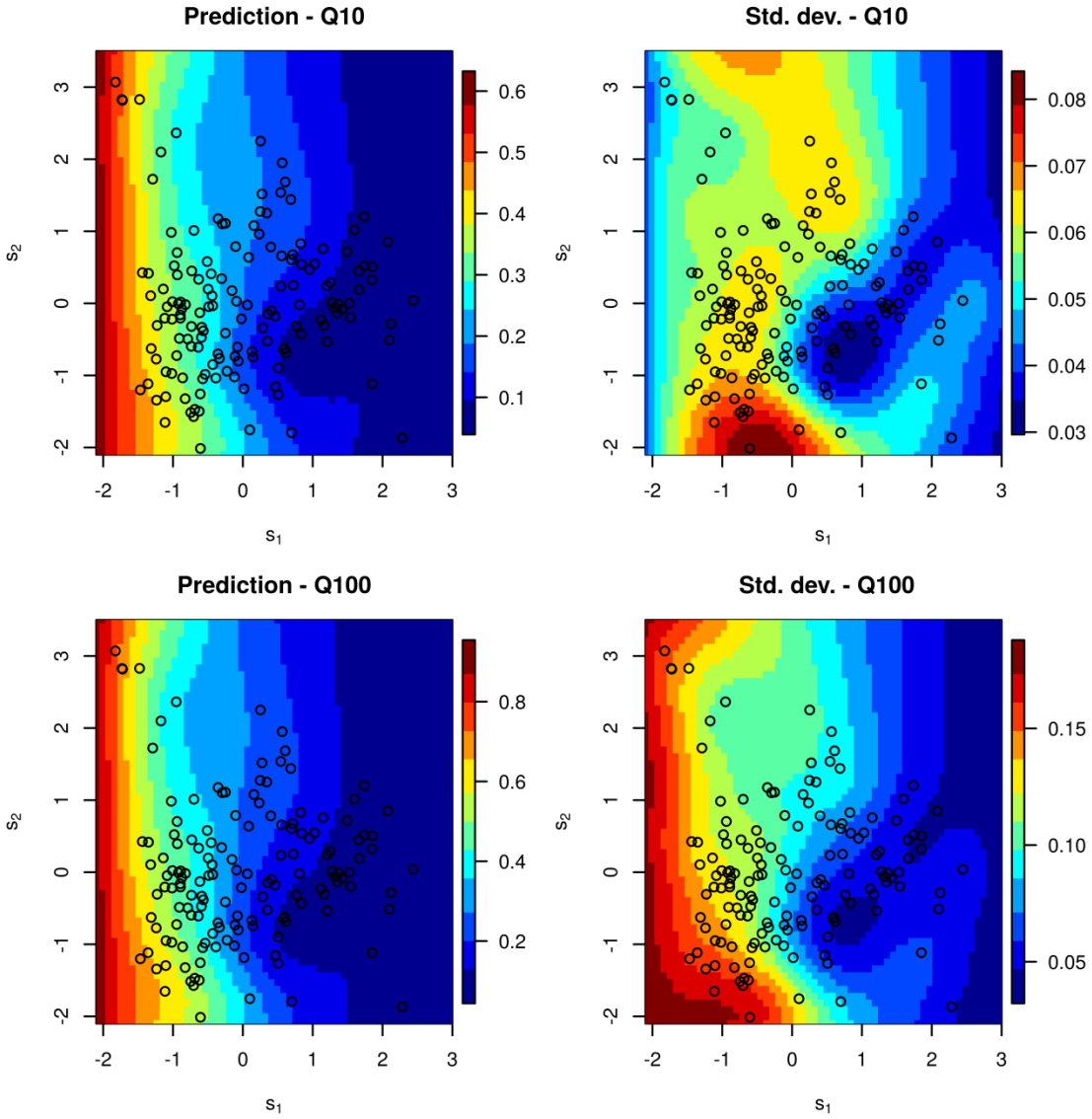

Figure 4: Maps of the mean of the PPD and standard deviation of the PPD. The circles represent the gauged stations in the physiographical space with coordinates (s_1, s_2) .

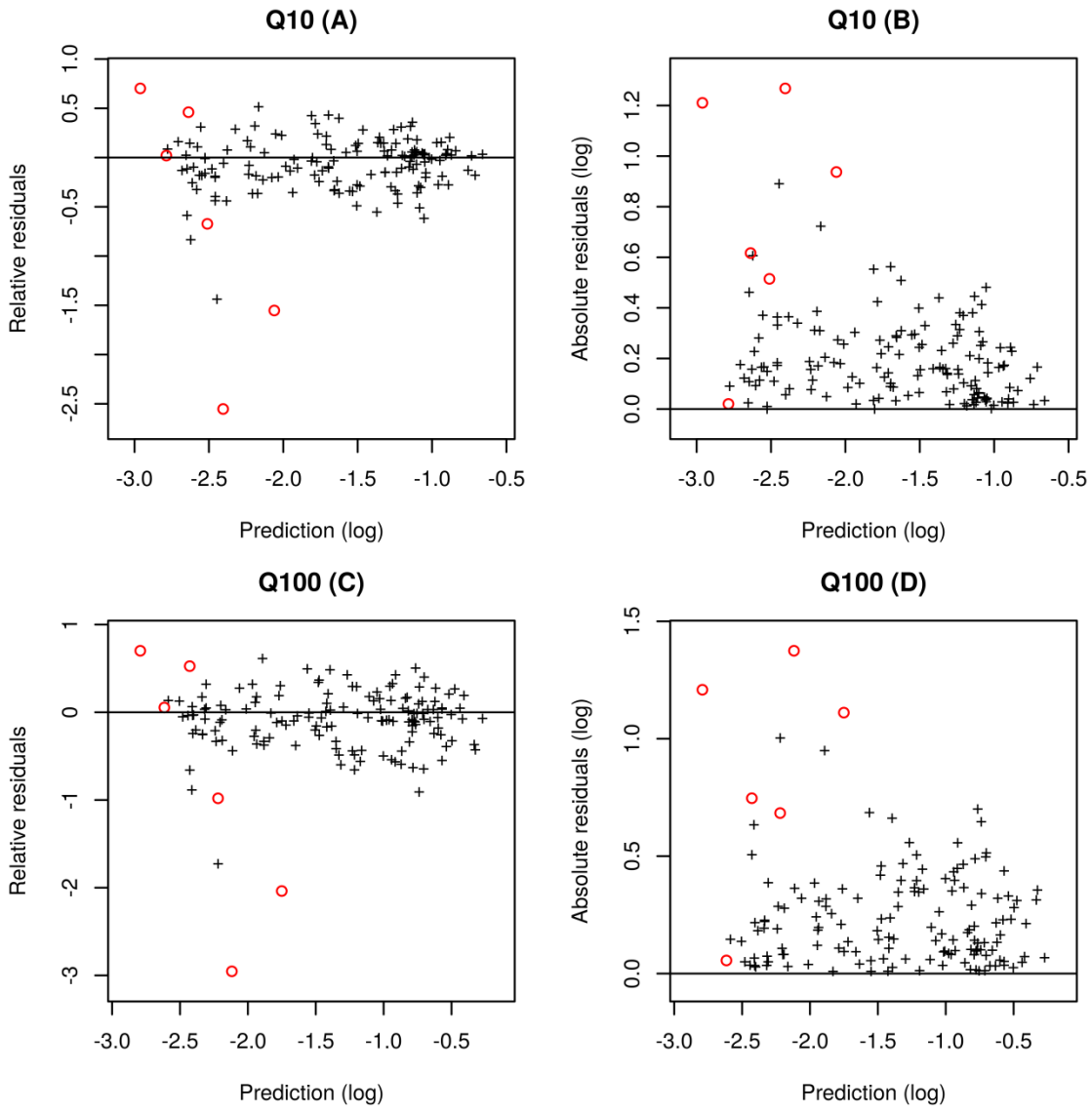

Figure 5: Visual diagnostics of the leave-one-out cross-validation. The predictions are obtained as the mean of the PPD.

CHAPITRE 5:

REGIONAL FREQUENCY ANALYSIS FROM APPROXIMATE BAYESIAN COMPUTING OF MAX-STABLE PROCESSES

Regional Frequency Analysis From Approximate Bayesian Computing of Max-Stable Processes

Martin Durocher[1], Mathieu Ribatet [2,3], Taha B.M.J. Ouarda[4,1]

[1]Institut National de Recherche Scientifique (INRS-ETE), University of Québec
490 de la Couronne, Québec G1K 9A9, Canada

[2]Université de Lyon, Université Lyon 1, Institut de Science Financière et d'Assurances,
50 Avenue Tony Garnier, F-69007 Lyon, France

[3]Department of Mathematics, University of Montpellier 2,
4 Place Eugène Bataillon, 34095 Montpellier Cedex 5, France

[4]Institute Center for Water Advanced Technology and Environnemental Research (iWater), Masdar Institue of Science and Technology
P.O. Box 54224, Abu Dhabi, UAE

Submitted for publication

ABSTRACT

As generalized extreme value distributions became standard distributions in the study of univariate extreme values, mathematical arguments justify the use of max-stable processes for modelling spatial extremes. However, since no closed form exists for their joint distribution, standard likelihood-based techniques are impossible to use in this context. The present work investigates the use of approximate Bayesian computing as an alternate estimation method for max-stable processes. Some guidelines for using approximate Bayesian computing with max-stable models are given and the performance is analyzed through simulation studies. A max-stable model similar to the index flood model is developed and applied to a small region in California, USA. The results of approximate Bayesian computing are compared to those of composite likelihood and show that approximate Bayesian computing can offer good fitting of spatial dependency and provide reliable return levels.

Some key words: Approximate Bayesian Computing, Composite likelihood, F-madogram, Index Flood Model, L-moments, Max-stable process, Regional Frequency Analysis.

1 INTRODUCTION

Intense precipitation represents a natural hazard that can be responsible for sudden flood events. The application of extreme value theory in environmental sciences arises from the necessity of estimating the risk level associated with extreme events corresponding to large return periods that exceed the length of the available record. The concept of a return period is well established as a standard risk measure in water resources planning and management (El Adlouni et al., 2007, Shu and Ouarda, 2007). However, extrapolation is always a delicate task as it requires strong confidence in the selected model. Consequently, it is important to rely on methodologies that have solid justification. In the univariate case, well established methods have demonstrated their adequacy in real world applications (e.g. Coles, 2001; Ahmadi-Nedushan et al., 2007). However, to address the actual needs in environmental sciences, it is necessary to consider more complex scenarios where multivariate tools are used. In these situations, extrapolation of the risk should be subject to the same justification as in the univariate case.

A number of natural phenomena, like precipitation, may be spatially distributed over relatively large areas. Consequently, they are likely to record extreme events simultaneously at several close sites. Meteorological stations are usually part of a large network that systematically collects information where the spatial distribution of the phenomena can be unveiled (St-Hilaire et al., 2003). Max-stable processes (de Haan, 1984) are the infinite-dimensional analogue of the classical extreme value theory and have been considered recently in the field of hydrology (Westra and Sisson, 2011; Shang, Yan, Zhang, 2011). A practical characterization for spatial extremes was proposed by Haan (1984) and Schlather (2002). This representation allows max-stable processes to be built as the superposition of an infinity of underlying processes. This characterization leads to the intuitive interpretation of the observations as the maxima taken over several storm cells. Max-stable processes do not have a

closed form for the cumulative distribution, but the later representation borrows the parametric structure of the underlying process to characterize the spatial dependency at the max-stable level.

The main difficulty when dealing with a max-stable model is that the full likelihood is missing and hence the usual estimators and statistical tests based on the likelihood principles cannot be used. One estimation technique for max-stable processes consists of optimizing a composite likelihood function composed by lower marginal densities (Padoan et al., 2010). The resulting objective function is known as the marginal likelihood (Cox and Reid, 2004) and is part of the larger family of composite likelihood functions (Lindsay, 1988). Composite likelihood is now widely used in several domains of statistics (Le Cessie and Van Houwelingen, 1994; Heagerty and Lele, 1998; Zhao and Joe, 2005), where one convenient form for spatial observations is the pairwise likelihood. This fitting criterion involves the optimisation of the product of all pairs of bivariate densities. A general discussion on the application of the pairwise likelihood to the problem of fitting max-stable processes was presented by Padoan et al. (2010). A practical challenge with the pairwise likelihood for spatial extremes is that numerical optimization is sometimes difficult due to the slow convergence rate of standard numerical algorithms.

Composite likelihood theory is not the only way to deal with models where full likelihood is unavailable. Another possible solution is Approximate Bayesian Computing (ABC) that aims to sample from the posterior distribution by replacing the likelihood with numerous simulations. Approximate Bayesian methods fall into the class of intensive computing techniques, but are relatively simple to implement (Pritchard et al., 1999; Tavare et al., 1997). The basic steps of the method can be described as picking a parameter from a prior distribution, simulating the model according to the proposed parameter and keeping it if the simulation is “close enough” to the observation. Recently, more sophisticated algorithms have been proposed and make Approximate Bayesian methods relevant to more situations. In particular, Marjoram et al. (2003) adapted the Metropolis-Hastings

algorithm to ABC and Beaumont et al. (2009) provided an adapted version of the Population Monte-Carlo algorithm (Cappé et al., 2004).

For max-stable processes, Erhardt and Smith (2012) provided an approximate Bayesian framework to realize statistical analysis of the spatial dependency. An approximate Bayesian analysis of spatial extremes requires several choices from the user and in particular, it requires a proxy measure to evaluate the closeness between the simulations and the observations. Erhardt and Smith (2012) compared different metrics that can be used for this purpose. They also compared their results with the more established method of composite likelihood. The application of their method to U.S. temperature data in Texas shows that approximate Bayesian methods can lead to smaller root mean square errors than composite likelihood. However, it has to be noted that the methodology of Erhardt and Smith (2012) does not consider the joint estimation of the marginal distribution as does the composite likelihood method.

Methods involving max-stable models are relatively new, but as max-stable processes represent one of the few mathematically justified ways of describing extreme spatial structures, their use in the hydrological field should increase in the future. The present work investigates the use of the approximate Bayesian method to carry out the joint estimation of all the parameters of a max-stable process in the context of regional frequency analysis (RFA). The present work aims to check if the ABC methodology can be a good competitor to the widely used pairwise likelihood estimator. A popular model in hydrologic frequency analysis is the index flood model (Dalrymple, 1960; Chebana and Ouarda, 2009). In this work, a max-stable model closely related to the index flood model is proposed.

Given the computational burden associated with ABC methods, this study restricts its focus to the Schlather max-stable model. Other important max-stable models include the storm model of Smith (1990) and the Brown-Resnick model of Kabluchko et al. (2009). The Schlather model is preferred in this study because its underlying structure is more realistic then the storm model (Schlather 2002) and

it is also faster to simulate than the Brown-Resnick model (Oesting et al. 2012). Nevertheless, the methodology developed here is not restricted to the specific Schlather model and could be applied to other max-stable models.

This document is organized as follows: Section 2 provides the theoretical background for max-stable processes and composite likelihood estimation. In section 3, the methodological aspects of approximate Bayesian computing with adaptation to RFA are described. In section 4, simulation studies are used to assess the general performance of the approximate Bayesian estimation for a max-stable process. In section 5, the proposed methodology is applied to precipitation data from a small region in California, USA. Finally concluding remarks are provided in section 6.

2. THEORETICAL BACKGROUND

2.1 MAX-STABLE PROCESSES

All max-stable models have the representation proposed by Schlather (2002). Without loss of generality, the study of max-stable models can be limited to the case of unit Fréchet margins, if proper transformations are assumed. Suppose that Y_i are independent copies of a non negative stationary process Y defined on a region $\Omega \subset \mathbb{R}$ such that $E\left[Y(x)\right]=1$ for all $x \in \Omega$. Let u_i with $i=1,2,\ldots$ be the points of a Poisson process on $\left(0,\infty\right)$ with intensity $u^{-2}du$. Thus,

$$Z\left(x\right)=\max_{i\geq 1} u_i Y_i\left(x\right), \qquad x \in \Omega \tag{4}$$

is a stationary max-stable process with unit Fréchet margins (i.e. $\Pr\left[Z(x)<z\right]=\exp\left(-1/z\right)$ for all *x*). A distinction must be made between the above representation (4) and the Schlather model that takes *Y* as stationary Gaussian random field with correlation function ρ.

A useful way to describe the spatial structure of a max-stable process *Z* is by its extremal coefficient function θ (Smith, 1990) which satisfies:

$$\Pr\left[Z(x) < z, Z(x+h) < z\right] = \exp\left(-\frac{\theta(h)}{z}\right), \qquad z > 0 \tag{5}$$

The function $1 \le \theta(h) \le 2$ links the strength of the spatial dependency to the distance h between two points. A value $\theta(h) = 1$ corresponds to complete dependence, while $\theta(h) = 2$ indicates perfect independence. The extremal coefficients $\theta(h)$ only consider pairs of observations and hence do not fully characterize the spatial structure of Z. However, $\theta(h)$ provides a practical tool to assess the pairwise dependence of a max-stable process. Its utility is then similar to the variogram of a Gaussian random field (Ancona-Navarrete and Tawn, 2002). It was shown by Schlather (2002) that the bivariate cumulative distribution function of the Schlather model is given by:

$$\Pr\left[Z(x_1) < z_1, Z(x_2) < z_2\right] = \exp\left[-\frac{1}{2}\left(\frac{1}{z_1}+\frac{1}{z_2}\right)\left(1+\sqrt{1-2\left[1+\rho(h)\right]\frac{z_1 z_2}{(z_1+z_2)^2}}\right)\right] \tag{6}$$

where *h* is the Euclidean distance between x_1 and x_2. It is not difficult to show that $\theta(h) = 1+\sqrt{\left[1-\rho(h)\right]/2}$. If $\rho(h)$ is a monotone positive definite function, necessarily as $h \to \infty$, $\rho(h) \to 0$ and the Schlather model does not allow for complete pairwise independence, i.e., $\theta(h) \to 1+\sqrt{1/2}$ as $h \to \infty$.

A equivalent tool for the description of the pairwise dependence is the F-madogram

$$2\nu(h) = \mathbb{E} \qquad \left] - F_2\left[Z(x+h)\right]\right| \tag{7}$$

where F_i designates the cumulative distribution function of $Z(x_i)$. A natural estimator for the F-madogram consists in using the sample mean in (7) and replacing F_1, F_2 by their empirical cumulative distribution functions.More precisely, if B_h is the set of pair of locations (or bin) whose pairwise distances belong to the interval $[h-\delta, h+\delta[$ and the z_i are realizations of $Z(x_i)$, then a binned version of the F-madogram is:

$$\hat{\nu}(h) = \frac{1}{2|B_h|} \sum_{(i,j)\in B_h} \left| \hat{F}_i(z_i) - \hat{F}_j(z_j) \right| \tag{8}$$

Cooley et al. (2006) showed that the F-madogram is related to the extremal coefficient function by

$$2\nu(h) = \frac{\theta(h)-1}{\theta(h)+1} \tag{9}$$

and indirectly to the correlation function ρ. The resulting empirical F-madogram provides a simple and effective way to estimate the extremal coefficients (Cooley et al., 2006).

2.2 COMPOSITE LIKELIHOOD

The estimation of the max-stable model by the maximum of a composite likelihood function is not much different than the traditional optimisation of a full likelihood. The primary distinction concerns variance estimation. Let $\mathbf{z}$ be a matrix of observations with elements $z_{i,k}$. These elements are obtained from independent copies $\{Z_i\}_{i=1,\ldots,n}$ of a max-stable process Z evaluated at points $\{x_k\}_{k=1,\ldots,K}$. Let also $f(z_{i,j}, z_{i,k} \mid \psi)$ denote their bivariate densities at parameter $\psi \in \mathbb{R}$. Under simple regular conditions (Lindsay, 1988), the optimum $\hat{\psi}$ of the log pairwise likelihood

$$l_{pl}(\psi) = \sum_{i=1}^{n} \sum_{k>j} \log f(z_{i,j}, z_{i,k} \mid \psi) \tag{10}$$

is the solution of the score equation:

$$U(\psi \mid \mathbf{z}) = \sum_{i=1}^{n} \sum_{k>j} \nabla \log f(z_{i,j}, z_{i,k} \mid \psi) = 0 \tag{11}$$

The maximum pairwise likelihood $\hat{\psi}$ is a consistent and unbiased estimator that asymptotically follows a Gaussian distribution with covariance matrix

$$V(\hat{\psi}) = H(\hat{\psi})^{-1} J(\hat{\psi}) H(\hat{\psi})^{-1} \tag{12}$$

where $H(\hat{\psi}) = \mathbb{E}$ $\mid \mathbf{z}))$ is the Hessian and $J(\hat{\psi}) = \mathrm{Var}(U(\hat{\psi} \mid \mathbf{z}))$ is the variance of the score equations. A more detailed review of the application of composite likelihood in statistics is given by Varin (2008).

The variance-covariance matrix can be evaluated analytically at $\hat{\psi}$. Padoan et al. (2010) provided these formulas when the parameters of the marginal distributions depend on a set of covariates. For more general situations, substantial efforts might be required to obtain the necessary derivations. Furthermore, the matrix *J* can also be difficult to estimate (Varin, 2008). To avoid non robust estimation of the covariance matrix (12), an alternate solution is the jackknife estimator (Zhao and Joe, 2005)

$$\hat{V} = \sum_{i=1}^{n} (\hat{\psi}_{-i} - \hat{\psi})(\hat{\psi}_{-i} - \hat{\psi})^{T} \tag{13}$$

where $\hat{\psi}_{-1}$ is the estimator of ψ obtained when the ith year of observations is deleted.

3. METHODOLOGY

3.1 APPROXIMATE BAYESIAN COMPUTING

The simplest sampling strategy for ABC is the rejection scheme (ABC-REJ) (Appendix 1). Although a naive application of ABC-REJ might be effective for discrete random variables, the algorithm is unrealistic when dealing with continuous data. For a practical implementation of ABC-REJ, the equality $z^* = z$ must be relaxed using a proxy measure $d(\mathrm{z}, z^*) < \epsilon$ with tolerance value ϵ (Pritchard et al., 1999; Tavare et al., 1997). Such substitution adds a degree of approximation. The resulting sample is more formally generated from the approximate posterior distribution:

$$\pi\{\psi, z^* \mid d(z, z^*) < \epsilon \qquad \mid z). \tag{14}$$

The acceptance rate for this sampling scheme can be very low. A better approach is to work with a posterior distribution that is conditional to a set of summary statistics T. The sufficiency principle says that if T is sufficient, the posterior distribution $\pi(\psi \mid z)$ conditional on the observation is equivalent to the posterior distribution $\pi\left(\psi \mid T(z)\right)$ conditional on the sufficient statistics. Using a proxy measure $d\left(T(z), T(z^*)\right)$ with sufficient summary statistics can improve the acceptance rate of ABC-REJ without loss of information. However, in most cases, the available statistics are not sufficient and the general hypothesis is that *T* is simply informative about ψ. This adds a second degree of approximation that depends on the relevance of T. Finding an appropriate metric and a set of summary statistics is an essential step in the development of the approximate Bayesian methodology and must be adapted to every problem.

3.2 POST-PROCESSING

Another way to improve the quality of ABC-REJ is to apply a post-processing treatment (Beaumont et al., 2002). At some points reducing ε becomes an inefficient way to improve the quality of the approximate posterior sample, especially when several parameters are involved. The purpose of the post-processing strategy is to improve the acceptance rate of ABC-REJ by accepting a larger tolerance value before the post-treatment. Let $\{\psi_i\}_{i=1,\ldots,p}$ be an approximate posterior sample, say obtained by ABC-REJ, with associated summary statistics $s_i = T(z_i^*)$ where z_i^* is the data generated when ψ_i was accepted. Denote $m(s_i) = \mathbb{E}$ | the conditional mean, $v^2(s_i) = \mathrm{Var}(\psi \mid s_i)$ the conditional variance and e_i a standard residual term. The nonparametric regression model

$$\psi_i = m(s_i) + v(s_i)e_i, \qquad i = 1, \ldots, p \tag{15}$$

relates the approximate posterior sample to its summary statistics. The post-processing model (15) contains information about the departure from the target summary statistics $s = T(z)$. In the best scenario, we have ϵ and $s_i = s$ for all i. Consequently, the mean $m(s)$ and variance $v(s)$ represent the first two moments of the true approximate posterior distribution when $s* = s$. The linear transformation

$$\psi_i^* = m(s) + v(s)\frac{\psi_i - m(s_i)}{v(s_i)} \tag{16}$$

corrects the initial approximate posterior sample to respect the mean and the variance of the true approximate posterior distribution. Blum and Francois (2010) provided two approaches to fit the post-processing model: a local polynomial and an artificial neural network estimator. Although they showed that the neural network strategy is superior when the tolerance is large, the two approaches give

similar results when the tolerance becomes smaller. In the present paper we prefer the local polynomial approach for its simplicity.

3.3 MONTE CARLO SAMPLER

An alternative approach to ABC-REJ is the Markov Chain Monte Carlo sampler (ABC-MCMC) (Appendix 2). This algorithm is inspired from the Metropolis algorithm. The ABC-MCMC algorithm improves over ABC-REJ as it requires fewer simulations to obtain the same number of posterior parameters with the same ϵ. Another important advantage is that ABC-MCMC can work with an arbitrary vague prior because the proposal vector is picked from a distribution that depends only on the location of the last accepted vector. Unlike ABC-REJ, the algorithm does not sample candidate values directly from the prior distribution where most of its elements might be unlikely to generate an accepted simulation. Although Marjoram et al. (2003) showed that the chain obtained from ABC-MCMC has the right stationary distribution, the chain usually suffers from very high autocorrelations. A potential side effect from this strong autocorrelation is the creation of artificial modes where the chain gets stuck in a small region with a low probability to accept the next proposal. Consequently, it becomes difficult with ABC-MCMC to select tolerance values that are as small as possible.

Population Monte Carlo (ABC-PMC) (Appendix 3) is another sampling scheme that is adapted to ABC. The algorithm shows a similar gain of speed than ABC-MCMC, but benefits from better mixing properties (Beaumont et al., 2009). The main idea of ABC-PMC is to serially pass through a sample starting from the prior distribution and gradually evolving toward the target distribution by importance sampling. At each iteration, a decreasing sequence of tolerance values $\{\epsilon$ is set until it reaches the desired value. Notice that ABC-REJ samples candidates parameters the prior distribution, which in practice prevents the utilization of too vague priors that would lead to reject almost every simulation. The first iteration of ABC-PMC is a ABC-REJ and consequently, this algorithm also suffers from the same difficulty.

3.4 SETTINGS FOR REGIONAL FREQUENCY ANALYSIS

To develop an approximate Bayesian strategy for max-stable processes, a set of practical summary statistics and a proximity measure are required. The complete set of summary statistics must contain information about the spatial dependency parameters as well as the marginal distribution parameters. Concerning the spatial structure, the obvious choice is the F-madogram, which is easy to compute and contains indirectly the same information as the pairwise extremal coefficients. Inside an approximate Bayesian algorithm, the empirical F-madogram should be aggregated into a set of bins to maintain a manageable set of summary statistics.

For the marginal parameters, a trend surface must be specified for the parameters of the generalized extreme value (GEV) distribution (Padoan et al., 2010) with cumulative distribution $F(\mathrm{z}) = \mathrm{e}^{-t(\mathrm{z})}$ with

$$t(\mathrm{z}) = \begin{cases} \left[1 + \xi(\mathrm{z} - \mu)/\sigma\right]^{-1/\xi} & \xi \neq 0 \\ e^{-(\mathrm{z}-\mu)/\sigma} & \xi = 0 \end{cases} \tag{17}$$

The parameters $\mu \in \mathbb{R}$ and $\sigma > 0$ control respectively the position and the scale of the GEV, while the third parameter $\xi \in \mathbb{R}$ defines the shape of the distribution. Here, the L-moments (Hosking, 1990) are proposed as summary statistics. This choice is motivated by their robustness to the presence of outliers and their good performance for small to moderate sample sizes (Hosking et al., 1985). An important argument in favour of L-moments versus traditional moments is that, for the GEV distribution, the traditional moments of order q are finite if only if $\xi < 1/q$. Consequently, traditional moments can only be used when $\xi < 1/3$. On the other hands, the L-moments estimator requires that $\xi < 1$. In the approximate Bayesian context, this means that L-moments allow for prior distributions with more realistic supports.

There are a number of approaches to use the at-site L-moments to create an informative set of summary statistics. The working hypothesis of the index flood model is that all the marginal distributions are the same apart from a scale factor. The empirical mean is usually chosen as the scale factor for the at-site distributions and hence a constant coefficient of variation and a constant shape parameter are used. For this work, the objective is to reduce the number of summary statistics involved in the post-processing models. For instance, let μ, σ and ξ be respectively the location, scale and shape parameters of a GEV. Therefore, assume a constant "coefficient of variation" $\sigma / |\mu|$ and a constant shape parameter ξ for the marginal distributions. If $l_j^{(i)}$ are the jth L-moments of sites x_i, then the average L-CV

$$\overline{l}_{cv} = \frac{1}{n}\sum_{i=1}^{n} \frac{l_2^{(i)}}{l_1^{(i)}} \tag{18}$$

and the average L-coefficients

$$\overline{\tau}_j = \frac{1}{n}\sum_{i=1}^{n} \frac{l_j^{(i)}}{l_2^{(i)}} \tag{19}$$

of order $j = 3,4$ are central tendency measures that respectively inform about the constant parameters. For a given coefficient of variation and shape parameter, the response surface $\mu(x_i)$ of the location parameters is described by the first L-moments. Overall, the complete set of summary statistics includes the F-madogram, the first L-moment of each site and the L-statistics (18) and (19).

Practical proxy measures are the quadratic $\| (s^* - s) \|^2$ or the relative quadratic $\| (s^* - s)/s \|^2$ distances between the vector of summary statistics $s^* = T(z^*)$ and its target $s = T(z)$. Because of the relative complexity of the max-stable model, a good approach is to use different proxy measures for separate groups of comparable statistics. Hence, to accept candidate parameters, all proxy measures

must respect the conditions imposed by their respective tolerence values. This makes the algorithm easier to set up as more meaningful tolerance values are associated to comparable statistics. Also, by proceeding by iterative testing of the several proxy measures, it is possible to reject a proposal vector in earlier steps and thus avoiding unnecessary operations.

We propose to use three conditions and to reject the proposal vector as soon as one of the conditions fails. First, the relative quadratic distance between the observed and the simulated first L-moments is verified. Second, the L-statistics (18) and (19) are tested by quadratic distance. Finally, the proposal vector is accepted if the quadratic distance between the observed and simulated F-madogram respects its tolerance value.

4. SIMULATION STUDIES

The approximate Bayesian methodology presented in the last section implies several choices that aim to balance between the feasibility and the quality of the approximation. To assess the global performance of the method in the context of a max-stable process and to investigate the impact of different choices, two separate simulation studies are conducted. In order to compare the approximate Bayesian computing and the likelihood based methods on a fair basis, uniform prior distributions are assumed since, in that case, the posterior distribution is proportional to the likelihood. For both simulation studies, a Schlather model is considered with random locations inside the unit square. The spatial structure is built from Gaussian random fields with an exponential correlation function

$$\rho(h) = \exp(-h/\lambda) \tag{20}$$

and range parameter $\lambda > 0$. The simulation studies are investigated from 500 Monte-Carlo experiments where the tolerance values are set in order to make the Monte-Carlo experiments take approximately one CPU-hour on a Intel Core i5 processor of 2.4GHz.

4.1 SPATIAL DEPENDENCE

The first simulation study examines the accuracy of the empirical *F*-madogram with 20 bins as an effective set of summary statistics for the dependence parameters. To this aim, unit Fréchet margins are assumed and the range parameter in (17) takes values λ=0.15, 0.3, 0.5. The size of the data set is considered to be k=15, 25, 50 random points and respectively n=15, 25, 50 independent realizations. Overall, there are 9 configurations for which an adjusted posterior sample of size 2000 is obtained by ABC-PMC.

Table 1 presents the performance statistics for this simulation study. For both estimation methods, we observe a small systematic error as described by the emprirical bias. Based on standard error, the approximate Bayesian estimator outperforms the pairwise likelihood except for the case when λ=0.50 and n=15. This special case can be explained by the observed trend in the results that shows that the approximate Bayesian estimator has better relative efficiency with smaller spatial dependencies and larger sample sizes.

4.2 RESPONSE SURFACE

The second simulation study evaluates the performance of our estimation method for the prediction of the 20 and 100-year return levels. The selected configuration aims to isolate the specific effect of choosing the L-moments as summary statistics. For this reason, the parameters controling the dependence structure are assumed to be known. Here, the Monte-Carlo experiments have k=30 random points and n=50 independent observations. The marginal shape parameter ξ and the coefficient of variation $\sigma/|\mu|$ are assumed to be constant with respective values of 0.1 and 0.3. If (x, y) designates the system of coordinates, the linear response surface for the location parameters is chosen to be

$$\mu(x,y) = \beta_{\mu,0} + \beta_{\mu,1}x + \beta_{\mu,2}y \tag{21}$$

with $\beta_{\mu,0} = 270$, $\beta_{\mu,1} = 350$ and $\beta_{\mu,2} = -60$. This specific response surface is chosen because of its similarity to that appearing in the case study section. Under the constraint of approximately one CPU-hour for the Monte-Carlo experiments, previous experiments show that the quality of the adjusted approximate posterior sample is better when imposing a larger size for the posterior sample. We adopt an adjusted posterior sample of size 15 000. This is justified by the fact that 5 parameters are involved here and improving the quality of the posterior sample by decreasing the tolerance value becomes rapidly inefficient. To deal with this phenomenon, increasing the size of the posterior sample reduces the sampling error of the post-treatment model and consequently provides a better correction of the initial posterior sample.

The empirical mode of the approximate posterior sample is compared to the maximum of the pairwise likelihood. The comparison also includes the estimation of the response surface by an independent likelihood function. Following the same notation as in (10), the independent log likelihood function

$$l_{ind}(\psi') = \sum_{i=1}^{n}\sum_{k=1}^{K} \log\left[f(\mathrm{z}_{i,\mathrm{k}} \mid \psi') \right] \tag{22}$$

is equivalent to ignoring the spatial structure and optimising the sum of all marginal log densities. The independent likelihood is another form of the composite likelihood and hence has the same asymptotic properties as the pairwise likelihood. Table 2 presents the performance statistics for the parameters of this second simulation study and Table 3 shows the performance statistics for the prediction of the return levels. All three methods have a relatively low bias and pairwise likelihood proves to lead to the best predictions. The independent likelihood has a slightly better relative efficiency concerning the parameters of the model, while on the opposite, the approximate Bayesian estimator has a slight advantage for the prediction of the return levels when λ>0.15.

5. CASE STUDY

California is characterized by rainy winters and dry summers. The southern region of California is known to be more arid than the northern region. The analysis of the maximum annual precipitation includes 39 daily gauged stations located in the middle region of the state. They are part of the Cooperative Observer Program from the National Weather Services (COPNWS, 2012). The location of the stations is presented in Figure 1. From West to East, the region covered by the gauged stations varies significantly in elevation. The stations start at the Pacific shore, including the Central Valley, and end in front of the Sierra Nevada Mountains. Each site has a minimum of 55 years of observations covering the period from 1949 to 2004.

In the present analysis, the gauged stations were divided into a training and a validation set. The validation set includes 8 randomly chosen sites, which are represented by circles in Figure 1. This choice of the cross-validation method is made to avoid multiple refitting of the model, because of the computational cost of the approximate Bayesian method. Examination of the distances between gauged stations for both sets reveals that the distances range from 11 km to 376 km for the training set and from 11 km to 348 km for the validation set.

Approximate Bayesian methods involve several choices to adapt the sampling algorithm to a specific problem. A premilinary study led to the adoption of similar assumptions to the index flood model as presented in section 2.4. If we denote the longitude, latitude and elevation as the respective coordinates (x, y, z), the model for the response surface takes the form

$$\begin{aligned} \mu(x,y,z) &= \beta_{0,\mu} + \beta_{1,\mu}x + \beta_{2,\mu}x^2 + \beta_{3,\mu}y + \beta_{4,\mu}y^2 + \beta_{5,\mu}z \\ \sigma(x,y,z) &= \beta_{\sigma}\ \mu(x,y,z) \\ \xi(x,y,z) &= \beta_{\xi} \end{aligned} \tag{23}$$

where the $\beta_i \in \mathbb{R}$ and $\mu,\sigma>0$ are the GEV parameters. The estimation of the Schlather model with an exponential correlation function is performed jointly with the identified response surface. The ABC-MCMC algorithm is preferred for its capacity to deal with vague uniform priors and several parameters. The tolerance values are set to accepted about 15% of the chain proposals. Only one proposal vector out of 100 is systematically subsampled to deal with the strong autocorrelations. The algorithm stops when an approximate posterior sample of size 20 000 is reached. The first 500 proposal vectors are burned even if the chain appears to start at its stationary distribution. A post-processing treatment is then applied. Figure 2 illustrates the Markov chains for the parameter $\beta_{\mu,0}$ and its autocorrelation function after the correction. Similar conclusions are drawn from the verification of all the β_i.

Visual analysis of the spatial structure is carried out by the extremal coefficients as illustrated in Figure 3. Their examination gives evidence of a stronger dependency between close stations and the presence of the upper bound near the limit of the Schlather model. The dashed line of Figure 3 presents the theoretical extremal coefficients for the approximate Bayesian model. The estimate is given by the empirical mode of the approximate posterior sample. The maximum pairwise likelihood estimator is also illustrated in the same figure by the solid line. Graphical validation of the F-madogram suggests that the maximum pairwise likelihood estimator tends to underestimate the spatial dependency. A potential explanation for this poor fit is that the true parametric structure of the data might not be exactly a Schlather model. On the other hand, the approximate Bayesian estimator shows a relatively good fit. This can be explained by the robustness of ABC to model misspecification as discussed by Wilkinson (2008).

Visual examination of the marginal distributions is performed by quantile-quantile plots for the validation sites. Figure 4 illustrates these diagnostics for 4 selected stations. Notice that the tails of some marginals are not as good as they could have been if the distributions were fitted separately. These discrepencies are attributed to a too restrictive hypothesis concerning the constant shape

parameter ξ. The primary distinction between the graphs is the different asymptotic slope that is due to an approximate Bayesian shape parameter of 0.09 and a pairwise likelihood shape parameter of 0.12. To assess the predictive performance of the model, let $q_{i,\mathrm{k}}$ be a kth sample quantile of the ith of the $n'=8$ validation sites. For a Schlather model with parameter ψ, the predicted quantiles $\tilde{\iota}$ can be calculated, which gives rise to relative error terms

$$e_{i,\mathrm{k}}(\psi) = \frac{\tilde{\iota}}{q_{\mathrm{i,k}}} - \frac{'_{\mathrm{i,k}}}{} . \qquad (24)$$

From the simulation of every parameter $\{\psi_j\}_{j=1,\ldots}$ in the posterior sample, we compute the performance statistics

$$RMSE_k = \sqrt{\frac{1}{n'm}\sum_{i=1}^{n'}\sum_{j=1}^{m} e_{i,k}^2\left(\psi_j\right)} \qquad (25)$$

for every sample quantile of the validation sites. The same performance statistics are also evaluated for the pairwise likelihood using a random sample of size m coming from the asymptotic distribution of the parameters. Figure 5 presents the plot of $RMSE_k$ versus the sample quantile. According to this criterion, the approximate Bayesian estimation has clearly better predictive results for all sample quantiles.

The 20-year return level maps for the approximate Bayesian estimation are presented in Figure 6. The similarity between the return levels and the standard deviation maps is a consequence of the relation between the location and the scale parameter of the marginal distribution. The effect of the altitude is clearly seen as the Central Valley shows the lower return levels. On the other hand, the higher return levels are found in the Sierra Nevada Mountains. The maps are relatively smooth as a consequence of the polynomial surface. In Figure 7, four unconditional simulations from the max-

stable model with ABC parameters are presented. They illustrate the yearly effect of the spatial structure characterized by the dependence of a Schlather model. In particular, the upper right figure represents a simulated year where important precipitation occurs on the coast.

6. DISCUSSIONS AND CONCLUSIONS

In this paper we have presented an adaptation of the approximate Bayesian method to the estimation of a max-stable model. Simulation studies are used to asses the performance of the method and to compare it to the reference method of composite likelihood. The results of these simulation studies show that the approximate Bayesian approach is generally more adequate for spatial dependency estimation than composite likelihood, while the predictive performance of the return levels is similar to one obtained by independent likelihood. The RFA of precipitation data in the State of California is carried out. For this specific dataset, cross-validation demonstrates that the approximate Bayesian method is superior because of its predictive performance and its robustness to model misspecification.

A weakness of the model used in the present case study is the hypothesis of a constant shape parameter for the whole studied region. This hypothesis led to poor fit of the tails of the marginal distributions for some sites. In future work, more effort should be focused on the improvement of the prediction of the shape parameter. For instance, a hierarchical classification may be considered to delineate homogenous regions inside which the hypothesis of a constant shape parameter can be tested (i.e. Hosking and Wallis, 1997).

The principal limitation of the approximate Bayesian approach comes from its difficulty to deal with high dimensional models that may result from more complex response surfaces. The Monte-Carlo algorithm and post-processing treatment are useful tools in these circumstances. However, they have limitations and unreasonable acceptance rates might sometimes be unavoidable.

As mentioned earlier, the methodology presented here is not limited to the Schlather model, and other max-stable models can be applied directly. It is also important to notice that the full power of the Bayesian framework has not been explored, given that non-informative priors were selected in this work in order to carry out the comparison with the composite likelihood approach on the same ground. A more judicious specification of the prior distribution could contribute to improving parameter estimation and enhancing the efficiency of the ABC sampler.

This study illustrates the feasibility of using ABC approaches with max-stable models in a pratical application. In particular, the present results go in the same direction as Erhardt and Smith (2012) who suggested that ABC is actually one of the best methods for estimating the spatial dependance of a max-stable process. For researchers interested in using the ABC framework for the max-stable process, this study shows that a separate treatement of the marginal distribution and the spatial dependance is not the only option. Good results can be obtained from a full Bayesian model where these two parts are jointly considered in one step. This provides a more coherent framework with reliable return levels.

ACKNOWLEDGMENTS

Financial support for this study was graciously provided by the Natural Sciences and Engineering Research Council (NSERC) of Canada.

APPENDIX 1

Algorithm ABC-REJ : Approximate rejection scheme

Repeat

Propose ψ^* from prior distribution $\pi(\psi)$.

Simulate z^* from ψ^*.

Accept ψ^* if $z^* = z$ the observation

Until p vectors are sampled

APPENDIX 2

Algorithm ABC-MCMC: Approximate Markov Chain Monte-Carlo algorithm

Compute target summary statistics $s = T(z)$

Start the chain at ψ_0

For $i = 1, \ldots, p$ **do**

Propose ψ^* from transition kernel $K(\psi^* \mid \psi)$

Simulate z^* form ψ^*

Compute $s^* = T(z^*)$

If $d(s, s^*) > \epsilon$ **then**

Stay at ψ_{i-1}

Else

Calculate

$$h = \min\left\{1, \frac{\pi(\psi^*)K(\psi^* \mid \psi)}{\pi(\psi)K(\psi \mid \psi^*)}\right\}$$

Accept $\psi_i = \psi^*$ with propability h, otherwise $\psi_i = \psi_{i-1}$

End if

End for

APPENDIX 3

Algorithm ABC-PMC: Approximate Population Monte-Carlo

At $j = 1$, pick initial set $\{\psi_i^{(1)}\}_{i=1,\ldots,p}$ from ABC-REJ

Set weights $w_i^{(1)} = 1/p$ and η_1^2 the variance of transition kernel K_j as twice the empirical variance of $\psi_i^{(1)}$

For $j = 2,\ldots,m$ **do**

For $i = 1,\ldots,p$ **do**

Repeat

Pick ψ^{**} from $\{\psi_i^{j-1}\}_i$ with probability $w_i^{(j-1)}$

Proposed ψ^* from $K_j(\psi^* \mid \psi^{**})$

Simulate data z^* from ψ^*

Compute summary statistics $s^* = T(z^*)$

Until $d(s, s^*) \leq \epsilon$

Accept $\psi_i^{(j)} = \psi^*$

Calculate

$$w_i^{(j)} \propto \sum_{k=1}^{n} w_k^{(j-1)} K_j \left\{ \psi_i^{(j)} \mid \psi_k^{(j-1)} \right\}$$

End for

Set η_j^2 as twice the weighted empirical variance of $\psi_i^{(j)}$

End for

REFERENCES

Ahmadi-Nedushan, B., A. St-Hilaire, T. B.M.J. Ouarda, L. Bilodeau, É. Robichaud, N. Thiémonge, and B. Bobée. (2007). Predicting river water temperatures using stochastic models: Case study of Moisie River (Québec, Canada). Hydrological Processes, 21 : 21-34, DOI: 10.1002/hyp.6353.

Ancona-Navarrete, M., Tawn, J., 2002. Diagnostics for Pairwise Extremal Dependence in Spatial Processes. Extremes 5, 271–285.

Beaumont, M.A., Cornuet, J.-M., Marin, J.-M., Robert, C.P., 2009. Adaptive approximate Bayesian computation. Biometrika 96, 983–990.

Beaumont, M.A., Zhang, W., Balding, D.J., 2002. Approximate Bayesian Computation in Population Genetics. Genetics 162, 2025–2035.

Blum, M., Francois, O., 2010. Non-linear regression models for Approximate Bayesian Computation. Statistics and Computing 20 (1), 63–73.

Cappé, O., Guillin, A., Marin, J., Robert, C., 2004. Population monte carlo. Journal of Computational and Graphical Statistics 13 (4), 907–929.

El Adlouni, S., Ouarda, T.B.M.J., Zhang, X., Roy, R. and B. Bobée. (2007). Generalized maximum likelihood estimators for the nonstationary generalized extreme value model. Water Resources Research. 43. W03410, doi: 10.1029/2005WR004545.

Le Cessie, S., Van Houwelingen, J., 1994. Logistic regression for correlated binary data. Applied Statistics 43, 95–108.

Chebana, F., and Ouarda, T.B.M.J., 2009. Index flood-based multivariate regional frequency analysis, Water Resources Research., 45 (10).

Coles, Stuart, 2001. An introduction to statistical modeling of extreme values. Springer.

Cooley, D., Naveau, P., Poncet, P., 2006. Variograms for spatial max-stable random fields. In: Dependence in Probability and Statistics, Lecture Notes in Statistics. Springer New York.

COPNWS, 2012. Cooperative Observer Program from National Weather Services,

http://www.nws.noaa.gov/om/coop/.

Cox, D., Reid, N., 2004. A note on pseudolikelihood constructed from marginal densities. Biometrika 91 (3), 729.

Dalrymple, T., 1960. Flood-frequency analyses. Survey Water-Supply Paper 1543.

Erhardt, R.J., Smith, R.L., 2012. Approximate Bayesian computing for spatial extremes. Computational Statistics & Data Analysis 56, 1468 – 1481.

Haan, L.D., 1984. A Spectral Representation for Max-stable Processes. The Annals of Probability 12 (4), 1194–1204.

Heagerty, P.J., Lele, S.R., 1998. A composite likelihood approach to binary spatial data. Journal of the American Statistical Association 93, 1099–1111.

Hosking, J., Wallis, J.R., Wood, E., 1985. Estimation of the generalized extreme-value distribution by the method of probability-weighted moments. Technometrics 251–261.

Hosking, J., 1990. L-moments: analysis and estimation of distributions using linear combinations of order statistics. Journal of the Royal Statistical Society, Series B (Methodological), 52, 105–124.

Hosking, J., & Wallis, J.R., 1997. Regional frequency analysis: an approach based on L-moments. Cambridge Univ Pr.

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

Lindsay, B.G., 1988. Composite likelihood methods. Contemporary Mathematics 80 (1), 22139.

Marjoram, P., Molitor, J., Plagnol, V., Tavare, S., 2003. Markov chain Monte Carlo without likelihoods. Proceedings of the National Academy of Sciences of the United States of America 100, 15324–15328.

Oesting, M., Kabluchko, Z., & Schlather, M. (2012). Simulation of Brown–Resnick processes. Extremes, 15, 89-107.

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

Pritchard, J.K., Seielstad, M.T., Perez-Lezaun, A., Feldman, M.W., 1999. Population growth of human Y chromosomes: a study of Y chromosome microsatellites. Molecular Biology and Evolution 16 (12), 1791–1798.

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

Shang, H., Yan, J., Zhang, X. (2011). El Niño–Southern Oscillation influence on winter maximum daily precipitation in California in a spatial model. Water Resources Research, 47(11).

Shu, C., and T.B.M.J. Ouarda (2007). Flood frequency analysis at ungauged sites using artificial neural networks in canonical correlation analysis physiographic space, Water Resources Research., 43, W07438, doi:10.1029/2006WR005142, 1-12.

Smith, R.L., 1990. Max-stable processes and spatial extremes. Unpublished manuscript.

St-Hilaire, A., Ouarda, T.B.M.J., Lachance, M., Bobée, B., Gaudet, J. and C. Gignac (2003). Assessment of the impact of meteorological network density on the estimation of precipitation and runoff. Hydrological Processes. 17(18) : 3561-3580.

Tavare, S., Balding, D.J., Griffiths, R.C., Donnelly, P., 1997. Inferring Coalescence Times From DNA Sequence Data. Genetics 145 (2), 505–518.

Varin, C., 2008. On composite marginal likelihoods. AStA Advances in Statistical Analysis 92 (1), 1–28.

Westra, S. and Sisson, S. A., 2011. Detection of non-stationarity in precipitation extremes using a max-stable process model, Journal of Hydrology, 406, 119-128.

Wilkinson, R., 2008. Approximate Bayesian computation (ABC) gives exact results under the assumption of model error. ArXiv e-prints.

Zhao, Y., Joe, H., 2005. Composite likelihood estimation in multivariate data analysis. Canadian Journal of Statistics 33, 335–356.

Table 1 : Performance statistics for the first simulation study. The ABC estimator is compared to the maximum pairwise likelihood estimator (MPL). The bias and standard error (SE) are presented. The relative efficiency (RE) is defined as the ratio of the variance with ABC at the denominator.

		ABC		MPL		
n	λ	**Bias**	**SE**	**Bias**	**SE**	**RE (%)**
15	0.15	-0.001	0.057	0.016	0.088	2.38
	0.30	0.013	0.106	0.019	0.126	1.41
	0.50	0.030	0.182	0.017	0.180	0.98
25	0.15	0.013	0.055	0.024	0.072	1.71
	0.30	0.005	0.064	0.011	0.092	2.07
	0.50	0.012	0.107	0.020	0.140	1.71
50	0.15	-0.001	0.014	0.005	0.038	7.37
	0.30	0.000	0.034	0.006	0.060	3.11
	0.50	-0.002	0.068	-0.004	0.092	1.83

Table 2 : Performance statistics for the parameters of the second simulation study. The ABC estimator is compared to the maximum independent likelihood (MIL) and the maximum pairwise likelihood (MPL). The bias and standard error (SE) are presented. The relative efficiency (RE) is defined as the ratio of the variance with ABC at the denominator.

		ABC		MIL			MPL		
Parameter*	λ	**Bias**	**SE**	**Bias**	**SE**	**RE**	**Bias**	**SE**	**RE**
$\beta_{\mu,0}$	15	-0.081	13.133	-0.230	11.670	0.79	-0.048	11.231	0.73
	30	0.335	14.112	-0.003	13.638	0.93	0.100	12.579	0.79
	50	1.887	13.932	1.388	13.588	0.95	1.384	12.397	0.79
$\beta_{\mu,1}$	15	1.022	18.045	0.663	16.086	0.79	0.316	15.413	0.73
	30	0.567	19.234	1.320	19.103	0.99	0.767	17.666	0.84
	50	0.315	21.011	1.293	20.337	0.94	1.022	18.649	0.79
$\beta_{\mu,2}$	15	0.772	16.465	1.127	14.034	0.73	0.830	12.924	0.62
	30	-0.037	16.399	0.017	15.935	0.94	-0.054	14.194	0.75
	50	-1.525	16.419	-1.475	15.446	0.88	-1.714	13.642	0.69
β_{σ}	15	0.002	0.016	-0.001	0.015	0.88	-0.001	0.014	0.77
	30	0.001	0.017	-0.002	0.017	1.00	-0.002	0.015	0.78
	50	0.002	0.018	-0.002	0.018	1.00	-0.002	0.015	0.69
β_{ξ}	15	0.005	0.066	-0.008	0.061	0.85	-0.006	0.053	0.64
	30	0.002	0.071	-0.012	0.066	0.86	-0.008	0.053	0.56
	50	-0.003	0.079	-0.017	0.072	0.83	-0.012	0.057	0.52

*See (19) for explanation of the model parameters.

Table 3 : Performance statistics for the return levels of the second simulation study. The ABC estimator is compared to the maximum independent likelihood (MIL) and the maximum pairwise likelihood (MPL). The relative bias and relative variance are computed for individual sites and the overall performance statistic is taken as the mean. The relative efficiency (RE) is the ratio of the variance, with ABC estimator at the denumerator

		ABC		MIL			MPL		
years	λ	**Bias**	**SE**	**Bias**	**SE**	**RE**	**Bias**	**SE**	**RE**
20	0.15	-0.002	0.071	0.003	0.069	0.97	0.002	0.061	0.86
	0.30	0.003	0.075	0.007	0.076	1.01	0.006	0.062	0.83
	0.50	0.005	0.079	0.009	0.079	1.00	0.007	0.064	0.81
100	0.15	0.002	0.118	0.002	0.115	0.97	0.001	0.100	0.85
	0.30	0.014	0.125	0.009	0.127	1.02	0.007	0.100	0.80
	0.50	0.021	0.130	0.014	0.131	1.01	0.011	0.102	0.78

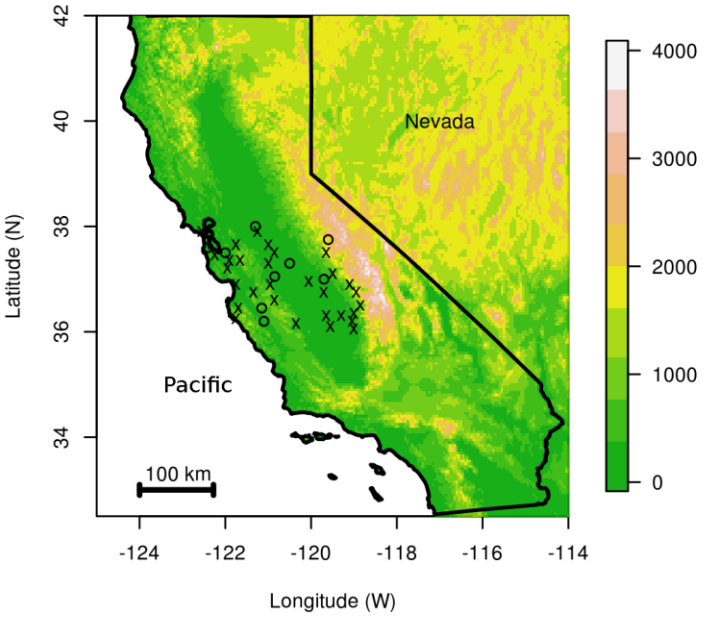

Figure 1 : Digital elevation map for the state of California. The crosses indicate the locations of the 30 meteorological stations used to estimate the return levels of extreme precipitations. The circles represent the 8 additional stations retained for cross-validation.

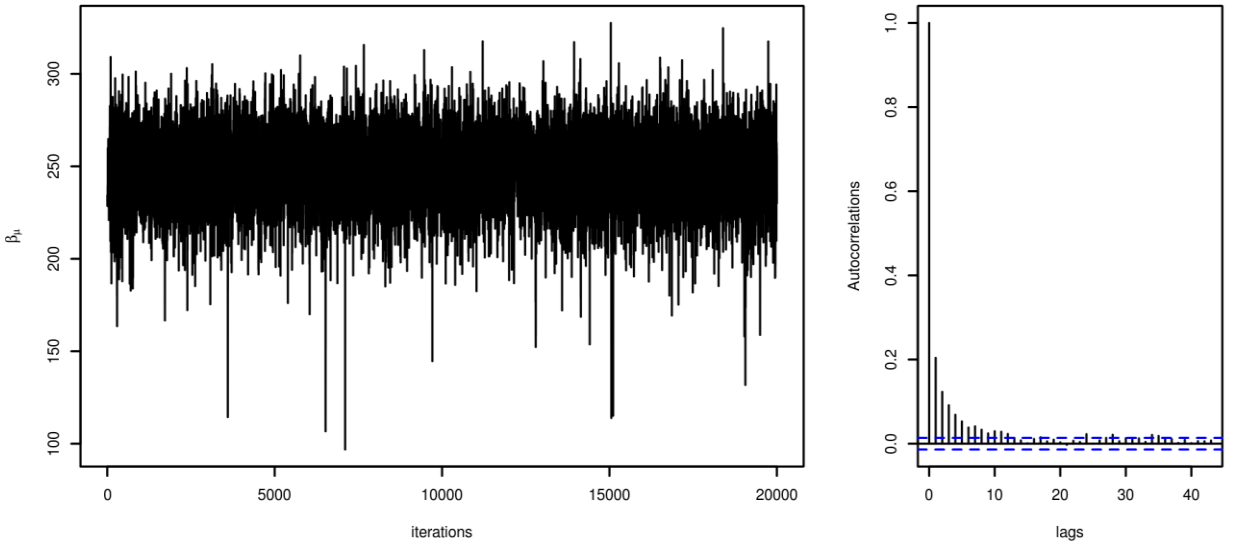

Figure 2 : Chain from ABC-MCMC for the parameters $\beta_{\mu,0}$ for the adjustment of the Californian precipitation. On the right, the autocorrelation function (ACF) of the chain.

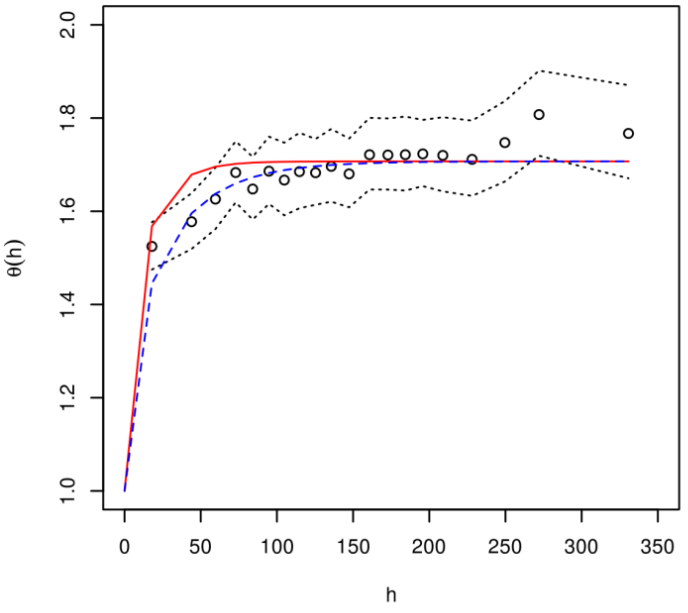

Figure 3: Extremal coefficients for the meteorological stations in California computed from the empirical F-madogram. The dotted lines represent the two-standard error interval. The dashed line represents the approximate Bayesian estimator and the solid line the maximum pairwise likelihood estimator.

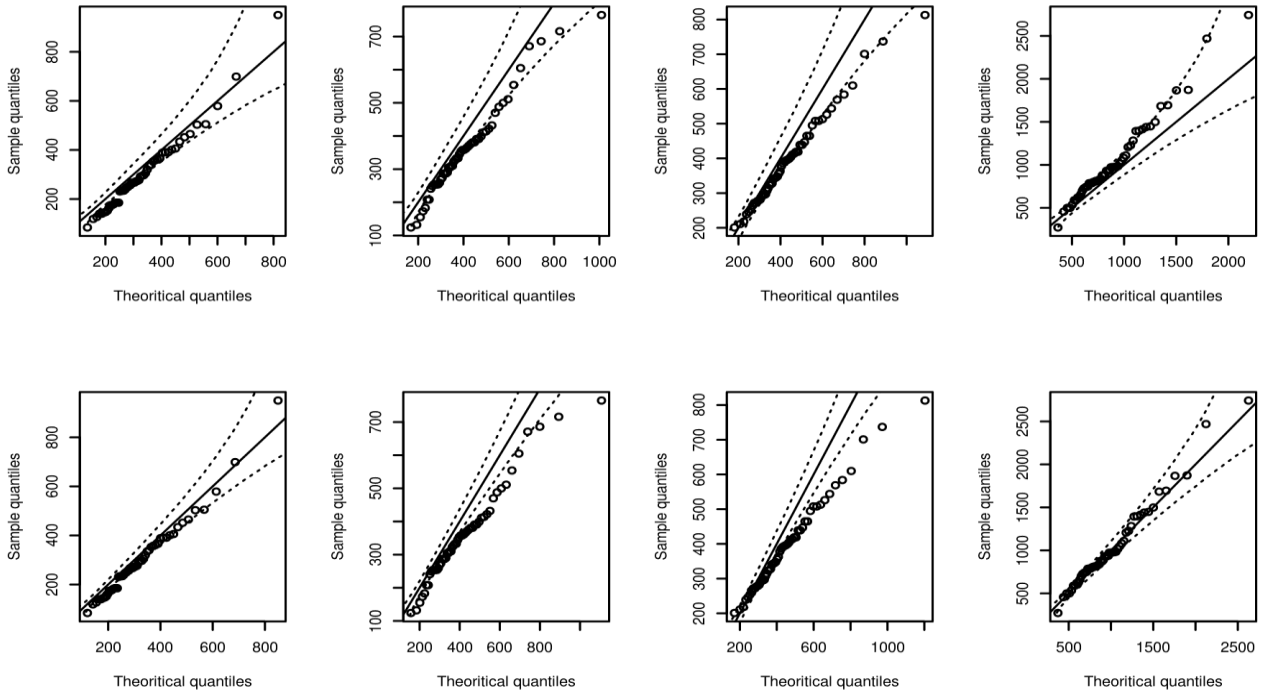

Figure 4: Quantile-Quantlie plots of the 4 validation stations (each column). The approximate Bayesian method (first row) is compared to the maximum pairwise likelihood method (second row). The solid lines represent the 95% confidence interval bound. The dashed line is the unitary slope that crosses the origin.

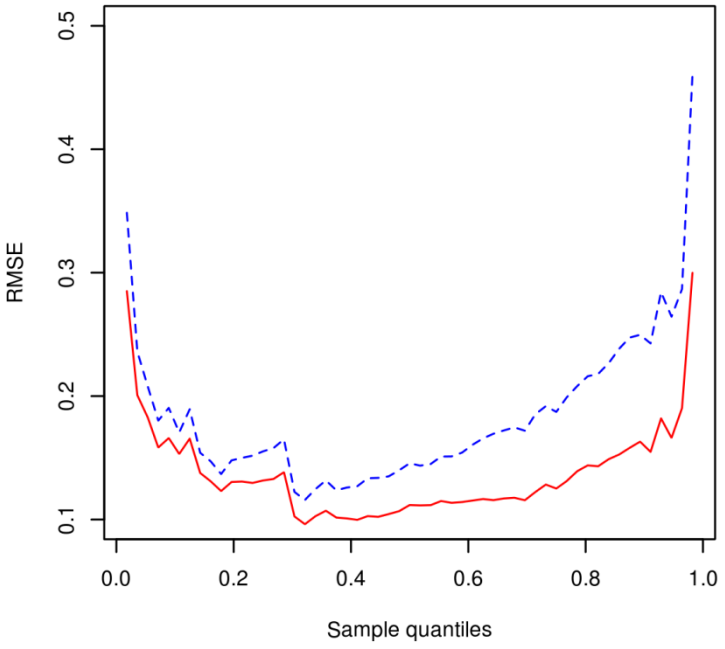

Figure 5 : Performance statistics for the sample quantiles of the validation stations. The satistics are computed from the approximate Bayesian method (solid) and the maximum pairwise likelihood (dashed).

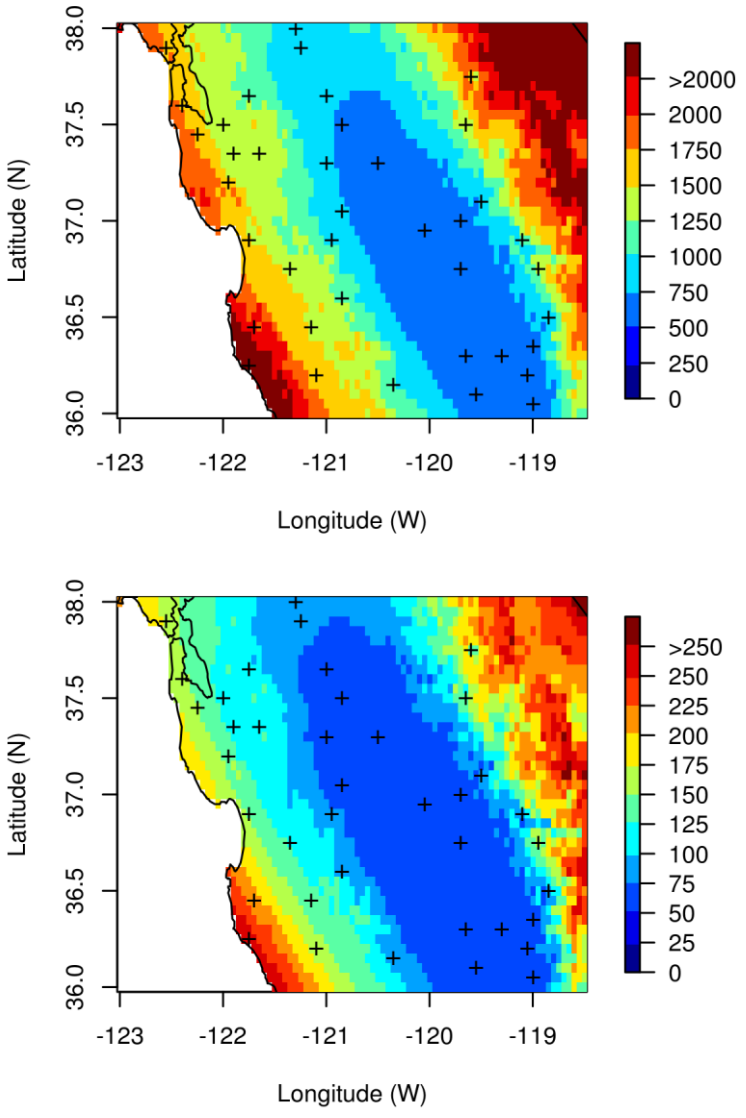

Figure 6 : Top, 20-year return level map for extreme precipitations (mm) and bottom, the predicted standard deviation map. The crosses indicate the 39 stations in the case study.

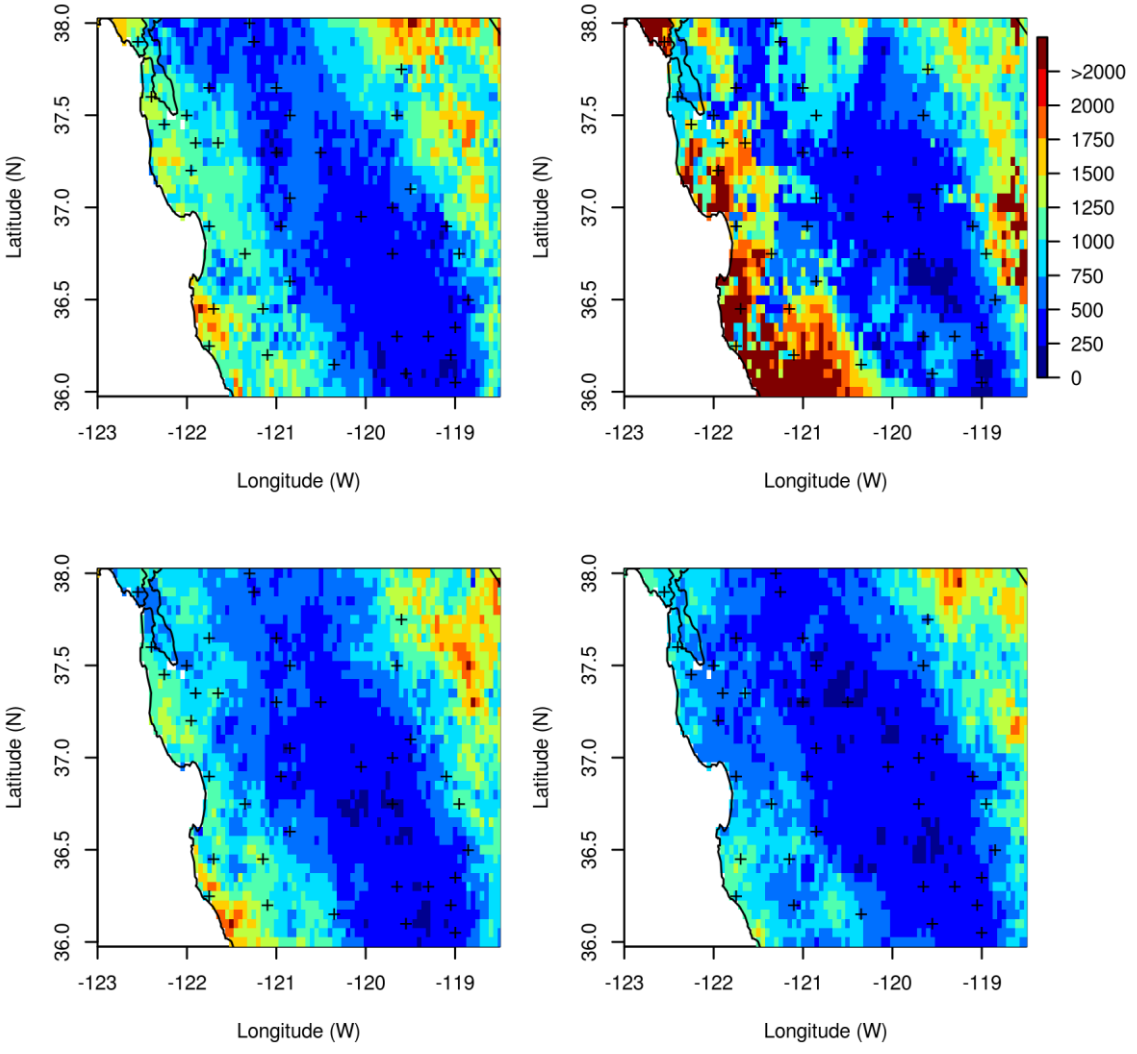

Figure 7: Four unconditional simulations of the max-stable model with parameters estimated by ABC. The crosses indicate the 39 stations in the case study.