Université du Québec
Institut Nationale de la Recherche Scientifique
Eau, Terre et Environnement

Aspects non standards en analyse fréquentielle régionale des variables hydrologiques

Par

Dhouha Ouali

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

Jury d'évaluation

Examinateur externe	Boualem Khouider University of Victoria
Examinateur externe	Marie Amélie Boucher Université de Québec à Chicoutimi
Examinateur interne	Sophie Duchesne INRS-ETE
Co-directeur de recherche	Taha B.M.J. Ouarda INRS-ETE - MASDAR Institute
Directeur de recherche	Fateh Chebana INRS-ETE

Thèse présentée le 29 Septembre 2016

Remerciements

Je tiens à adresser mes plus sincères remerciements à de nombreuses personnes qui, sans leur soutien, cette thèse n'aurait pu voir le jour. Un grand merci doit tout d'abord être remis à mes encadrants Fateh Chebana et Taha B.M.J Ouarda pour toutes les connaissances qu'ils m'ont transmises, pour la confiance et la liberté qu'ils m'ont accordées tout au long de cette aventure. C'est grâce à vos encouragements incessants et vos précieux conseils que j'arrive aujourd'hui à écrire ces mots. Recevez ici l'expression de mes sincères gratitudes.

J'aimerais également remercier les membres du comité : Boualem Khouider, Marie Amélie Boucher et Sophie Duchesne, pour avoir accepté et pris le temps d'évaluer ce travail.

J'adresse mes remerciements à toute l'équipe de recherche en hydrologie statistique à l'INRS pour leur coopération et leur aide. Je n'aurais jamais passé sans avoir remercier mes chers amis de l'INRS qui ont toujours été présents pour me soutenir.

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

Finalement je désire exprimer mes chaleureux remerciements aux personnes les plus proches, qui ont toujours été là, et le seront, dans toutes les circonstances : Ma mère, mon père, mes sœurs, mon frère, mon mari et ma petite. Je ne vous remercierai jamais assez.

Préface

Cette thèse présente les travaux de recherche menés au cours de mes études doctorales. La structure de la présente thèse suit la structure standard des thèses par articles de l'INRS-ETE. La première partie de la thèse comporte une synthèse générale des travaux effectués. Cette synthèse a pour objectif de survoler la méthodologie adoptée et les principaux résultats obtenus au cours de la thèse. La deuxième partie de la thèse contient quatre articles publiés, soumis ou sur le point d'être soumis à des revues internationales avec comité de lecture.

Articles et contribution des auteurs

1. D. Ouali, F. Chebana et T.B.M.J. Ouarda (2015). "Non-linear canonical correlation analysis in regional frequency analysis". Stoch Environ Res Risk Assess. DOI 10.1007/s00477-015-1092-7.

2. D. Ouali, F. Chebana et T.B.M.J. Ouarda (2016a). "Fully nonlinear regional hydrological frequency analysis". Soumis.

3. D. Ouali, F. Chebana et T.B.M.J. Ouarda (2016b). "Quantile regression in regional frequency analysis: a better exploitation of the available information". Journal of Hydrometeorology. DOI: 10.1175/JHM-D-15-0187.1

4. D. Ouali, F. Chebana et T.B.M.J. Ouarda (2016c). " Hydro-climatic frequency analysis in a changing climate using additive quantile regression: an exploratory analysis ". À soumettre.

Dans le premier article, D. Ouali a intégré une nouvelle technique dans l'analyse fréquentielle régionale, particulièrement pour la délimitation des régions hydrologiques homogènes, basée sur les réseaux de neurones artificiels. Les co-auteurs F. Chebana et T. B. M. J. Ouarda ont commenté et révisé la version finale du manuscrit.

Dans le deuxième article, D. Ouali a mené une étude comparative complète entre plusieurs modèles d'analyse fréquentielle régionale pour évaluer l'utilité de considérer des outils non linéaires et identifier la meilleure combinaison possible. F. Chebana et T. B. M. J. Ouarda ont fourni leurs commentaires durant l'exécution du travail et ont révisé la version finale du manuscrit.

Dans le troisième article, D. Ouali a proposé une nouvelle approche d'estimation régionale des crues en se basant sur la régression des quantiles. L'évaluation de la performance de cette approche

ainsi que des approches classiques s'est basée sur le développement d'un nouveau critère d'évaluation objectif. Tout au long de ce travail, F. Chebana et T. B. M. J. Ouarda ont discuté l'aspect méthodologique et les résultats obtenus, et ont révisé la version finale du manuscrit.

Dans le quatrième article, D. Ouali a mené une analyse fréquentielle locale pour l'estimation des quantiles de crues en utilisant la régression des quantiles. Plusieurs aspects physiques ont été considérés dans cette étude notamment la non-linéarité et la non-stationnarité des processus hydrologiques et les variables météorologiques associées. F. Chebana et T. B. M. J. Ouarda ont donné de précieux conseils et suggestions durant l'élaboration de ce travail.

Résumé de la thèse

La protection et la gestion des ressources en eau reposent dans une large mesure sur la maitrise et la compréhension des phénomènes hydrologiques extrêmes et la capacité à estimer adéquatement les risques hydrologiques que ce soit dans les conditions actuelles ou futures. Dans ce cadre, les outils statistiques trouvent une large application, allant des méthodes linéaires simples pour déterminer l'incertitude d'une moyenne hydrologique à des techniques sophistiquées non linéaires qui révèlent la dynamique et la complexité des évènements hydrologiques extrêmes. Dans le cas de l'analyse fréquentielle (AF), le but est de prédire adéquatement la fréquence de l'occurrence de ces évènements dans un site jaugé. Toutefois, il arrive souvent qu'on se trouve amené à produire des estimations dans des sites non jaugés. Dans de telles circonstances, les hydrologues et les praticiens font appel à des procédures de régionalisation ou ce qu'on appelle également une Analyse Fréquentielle Régionale (AFR).

L'AFR consiste à estimer les quantiles (de dépassement et/ou de non-dépassement) des évènements extrêmes (les crues et/ou les étiages) dans un site cible non jaugé à partir des données émanant des sites jaugés. Pour une meilleure estimation, ces derniers doivent être hydrologiquement similaires au site cible. Ainsi, l'AFR comporte deux étapes, la délimitation des régions hydrologiquement homogènes (DRH) en utilisant des méthodes de classification, et l'estimation régionale (ER) pour transférer l'information au site cible non jaugé, en se basant sur des méthodes de régression.

Malgré l'existence dans la littérature de diverses approches en AF locale et régionale, ces approches présentent des contraintes et limitations. En réalité, différentes conditions non standards telles que la complexité topographique des bassins versants, le manque et/ou la non-disponibilité des données de débits, les perturbations par les aménagements urbains et/ou les changements

climatiques peuvent influencer la réponse hydrologique des bassins versants. De telles conditions rendent la prédétermination des crues par les méthodes classiques d'AF et AFR un exercice difficile, non efficace et mal adapté à de tels contextes non standards.

L'objectif de cette étude consiste à proposer de nouvelles approches et de nouveaux modèles en AF locale et régionale des crues. Ces approches et modèles visent à contourner les limites de ceux utilisés dans la littérature et à considérer des aspects non standards en AF. En AFR, l'accent est mis sur le problème de la non-linéarité dans l'étape de la DRH et le problème de la mauvaise exploitation des données disponibles dans l'étape de l'estimation. D'autre part, en AF locale l'accent est mis sur le problème de la non-stationnarité et l'inclusion de plus d'information dans le modèle. Ces nouveaux modèles sont basés sur des outils statistiques en plein essor dans la littérature statistique au cours des dernières années, y compris des outils de régionalisation non linéaire et des outils de régression récents.

Précisément, on s'intéresse dans une première partie à intégrer la notion de la non-linéarité en AFR dans l'étape de la DRH. La méthode adoptée est l'analyse canonique des corrélations non linéaire (ACCNL), présentée dans le Chapitre 2 de ce manuscrit. Elle permet de considérer la complexité du processus hydrologique en considérant une variante non linéaire de l'analyse canonique des corrélations (ACC) pour identifier un voisinage homogène d'un site non jaugé.

Par la suite, dans le but d'identifier les combinaisons de méthodes de DRH et d'ER les plus prometteuses permettant une meilleure estimation des risques des extrêmes hydrologiques, une étude comparative a été mise au point incluant différentes approches d'AFR. Les techniques considérées au niveau des deux étapes de la procédure d'AFR sont des techniques linéaires (telles que l'ACC et la régression multiple) et non linéaires (telles que l'ACCNL, les réseaux de neurones artificiels et les modèles additifs généralisés). Les résultats d'une étude comparative, détaillés dans

le Chapitre 3, sont en faveur de l'introduction d'une composante non linéaire au niveau de chacune des deux étapes de l'analyse régionale.

Un modèle d'AFR basé sur la notion de la régression quantile (RQ) a été conçu afin d'améliorer l'exploitation des données hydrologiques disponibles. En développant un critère d'évaluation objectif, nous montrons l'intérêt de considérer un tel outil assez puissant dans l'AFR des crues. Le développement de ce modèle et les résultats obtenus sont présentés dans le Chapitre 4 de ce rapport.

Parallèlement à l'aspect non linéaire du processus hydrologique, la non-stationnarité des extrêmes hydrologiques est également l'un des facteurs déterminants lors de la modélisation statistique des extrêmes hydrologiques. À cet égard, un modèle d'AF non linéaire non stationnaire basé sur la notion de la RQ a été développé à l'échelle locale. Ce modèle servira comme base pour intégrer cet aspect de non-stationnarité dans l'AFR. Nous montrons dans le Chapitre 5 que, comparée aux approches classiques, cette approche peut être prometteuse non seulement en termes de performances mais également au niveau conceptuel.

Table des matières

Remerciements iii

Préface iv

Articles et contribution des auteurs v

Résumé de la thèse vii

Table des matières x

Liste des tableaux xii

Liste des figures xiii

CHAPITRE 1 : SYNTHÈSE 1

1. Contexte et revue de littérature 2

1.1. Analyse fréquentielle locale 2

1.2. Analyse fréquentielle régionale 3

1.2.1. Délinéation des régions homogènes 3

1.2.2. L'estimation régionale 5

1.3. Organisation de la synthèse 9

2. Problématiques et objectifs de la recherche 9

2.1. Problématiques 9

2.2. Objectifs de la thèse 12

3. Méthodologie 14

3.1. Un modèle d'AFR introduisant la non-linéarité dans la DRH 14

3.2. Combinaisons des approches et étude comparative 17

3.3. Approche régionale par RQ 20

3.4. Modèle d'AF non-linéaire non-stationnaire en utilisant la RQ 23

4. Applications et résultats 26

4.1. Zones d'études et données 26
4.2. Principaux résultats et discussions 28
4.2.1. Résultats des approches non linéaires 28
4.2.2. Résultats de l'approche régionale par RQ 35
4.2.3. Résultats de l'AF locale non linéaire non stationnaire 39
5. Conclusions et perspectives de la recherche 46
5.2. Conclusions générales 46
5.3. Perspectives de la recherche 48
6. Références bibliographiques 50

CHAPITRE 2 : NON-LINEAR CANONICAL CORRELATION ANALYSIS IN REGIONAL FREQUENCY ANALYSIS……...57

CHAPITRE 3 : FULLY NONLINEAR REGIONAL HYDROLOGICAL FREQUENCY ANALYSIS …….……....99

CHAPITRE 4 : QUANTILE REGRESSION IN REGIONAL FREQUENCY ANALYSIS: A BETTER EXPLOITATION OF THE AVAILABLE INFORMATION …..…….……....137

CHAPITRE 5 : HYDRO-CLIMATIC FREQUENCY ANALYSIS IN A CHANGING CLIMATE USING ADDITIVE QUANTILE REGRESSION: AN EXPLORATORY ANALYSIS …...…183

LISTE DES TABLEAUX

Tableau 1. Corrélations entre les variables météo-physiographiques et hydrologiques (Québec) 30

Tableau 2. Modèles régionaux semi-linéaires et non linéaires adoptés 31

Tableau 3. Résultats de la validation croisée des estimations des quantiles par les différents modèles adoptés.....32

Tableau 4. Résultats des simulations Monte-Carlo : RBIAS et RRMSE des quantiles estimés, conditionnellement à la co-variable, par le modèle GEV_{10} et GEV_{20}.....41

LISTE DES FIGURES

Figure 1. Schéma illustratif du principe de l'ACCNL 17

Figure 2. Différentes combinaisons et modèles adoptés 19

Figure 3. Localisation géographique des sites étudiés dans la partie sud de la province de Québec, Canada 27

Figure 4. Diagramme de dispersion des caractéristiques physiographiques des bassins versants et des quantiles de crue (Québec).....29

Figure 5. Résultats de la DRH en utilisant l'ACC (a) et l'ACCNL (b) pour la station Gatineau (ID: 040830), Québec. Le site cible est présenté par une étoile verte, les stations de tout le réseau hydrographique sont présentées en points noirs et les stations formant la région homogène du site cible sont présentées en points rouges.....34

Figure 6. Diagrammes de dispersion des quantiles régionaux en fonction des quantiles estimés localement en utilisant le modèle RLM (première colonne) et le modèle RQ (deuxième colonne) pour les quantiles Q_{S10}, Q_{S50} et Q_{S100}. Les deux modèles sont calibrés et évalués en utilisant tous les sites. Les points foncés désignent les sites avec de longues séries de données. 36

Figure 7. RMSE des estimations régionales de Q_{S50} (a) et Q_{S100} (b) ainsi que la MPLF des estimations régionales de Q_{S50} (c) et Q_{S100} (d) en fonction de la longueur des séries de données. Les deux modèles sont calibrés en utilisant des sites avec une longueur d'enregistrement dépassant l années, à l'exception de (c) et (d) où le modèle RQ a été calibré en utilisant toutes les données; la validation de la RQ et de la RLM se fait en utilisant tous les sites. 38

Figure 8. Estimations des quantiles associées aux probabilités de non-dépassement 0.90 (a) et 0.99 (b) par les modèles RQMA, RQL, GEV_{20}, GEV_{01} et GEV_{00}, conditionnellement aux valeurs du SOI, Arroyo Seco. 43

Figure 9. Estimations des quantiles associées aux probabilités de non-dépassement 0.90 par les modèles RQMA, RQL, GEV_{20}, GEV_{01} et GEV_{00}, conditionnellement aux valeurs du SOI, durant la période de calibration (a) et de validation (b), Bear Creek 44

Figufre 10. Hydrogramme de crues de la Station Dartmouth, Gaspésie, superposé aux formes médianes (quantile 0.50) résultantes des modèles RQL (a), RQMA et GEV_{20} (b).....45

Figure 11. Approches classiques déjà existantes en AFR ainsi que les approches proposées dans cette recherche.....47

LISTE DES ABRÉVIATIONS

ACC	Analyse Canonique des Corrélations
ACCNL	Analyse Canonique des Corrélations non linéaire
ACC-RLM	Modèle de RLM couplé à l'ACC dans l'étape de la DRH
ACC-GAM	Modèle de GAM couplé à l'ACC dans l'étape de la DRH
ACC-RNA	Modèle de RNA couplé à l'ACC dans l'étape de la DRH
ACC-RNE	Modèle de RNE couplé à l'ACC dans l'étape de la DRH
ACCNL-RLM	Modèle de RLM couplé à l'ACCNL dans l'étape de la DRH
ACCNL-GAM	Modèle de GAM couplé à l'ACCNL dans l'étape de la DRH
ACCNL-RNA	Modèle de RNA couplé à l'ACCNL dans l'étape de la DRH
ACCNL-RNE	Modèle de RNE couplé à l'ACCNL dans l'étape de la DRH
AF	Analyse fréquentielle
AFR	Analyse fréquentielle régionale
AMP	Moyennes des précipitations totales annuelles
AMD	Moyenne annuelle des degrés-jours supérieurs à 0°C
BV	Superficie du bassin versant
DMA	Débit maximum annuel
DRH	Délinéation des régions homogènes
ER	Estimation régionale
FAL	Fraction de la superficie couverte par des lacs
GAM	Modèles additifs généralisés

MBS	Pente moyenne du bassin versant
MPLF	Mean piecewise loss function
NAO	Indice de l'oscillation Atlantique du Nord
NASH	Critère de l'efficacité de Nash-Sutcliffe
PMC	Perceptron multicouches
Q_{ST}	Quantile spécifique associé au période de retour T
RLM	Régression linéaire multiple
RNA	Réseaux de neurones artificiels
RNE	Réseaux de neurones ensemble
RQ	Régression quantile
RQMA	Régression quantile par modèle additif
RRMSE	Racine carrée de l'erreur quadratique moyenne relative
RBIAS	Biais relatif
SOI	Indice de l'oscillation australe
TMA	Température moyenne annuelle

CHAPITRE 1 : SYNTHÈSE

1. Contexte et revue de littérature

Cette section a pour objectif de faire le lien entre les travaux existants dans la littérature de l'AFR et les modèles proposés dans le cadre de cette thèse. Les principales méthodes classiques considérées pour des fins de comparaison y sont brièvement présentées et discutées.

1.1. Analyse fréquentielle locale

L'estimation du risque associé aux événements hydrologiques extrêmes, tels que les crues, constitue toute une branche de la modélisation hydrologique. En effet, la prédétermination et l'évaluation des risques associés aux crues formaient, depuis des décennies, le premier souci tant pour les décideurs que pour les hydrologues. Lorsque nous disposons de suffisamment de données dans un site, l'AF locale de séries du débit observé constitue un outil statistique privilégié par les hydrologues et les ingénieurs facilitant la prise de décision. La finalité de l'AF est d'estimer les probabilités d'occurrence de certains événements, souvent extrêmes, dans des sites jaugés [Hamed et Rao, 1999]. Les principales étapes d'une AF consistent à : i) vérifier les hypothèses de base; ii) ajuster une distribution statistique à l'échantillon, et finalement iii) estimer le(s) quantile(s) ainsi que la(es) période(s) de retour associée(s).

Généralement, deux approches d'extraction des extrêmes sont utilisées lors de l'élaboration d'une AF, à savoir l'approche par blocs et l'approche de dépassement de seuil (POT). L'approche par blocs, qui consiste à identifier le maximum des données sur une période de temps, souvent une année, est couramment utilisée en hydrologie [Martins et Stedinger, 2000]. Toutefois, la période de mesure des séries de débits maximaux annuels (DMA) est souvent limitée et généralement ne dépassent pas une trentaine d'années, limitant ainsi les informations sur les distributions des débits extrêmes. Un tel effet peut réduire significativement la précision des estimations obtenues. En effet,

la qualité de cette estimation dépend essentiellement de la longueur de la série utilisée pour identifier la loi de probabilité et estimer ses paramètres [Castellarin et al., 2001].

1.2. Analyse fréquentielle régionale

Lorsqu'on ne dispose pas d'assez de données dans un site, ou que les données y sont rares, on fait appel à l'analyse fréquentielle régionale (AFR). Cette méthode est une approche statistique très utile qui vise à prédire l'occurrence des événements hydrologiques extrêmes dans des sites non jaugés. Ainsi, l'estimation régionale des quantiles de crue est effectuée via le transfert d'information à partir d'autres sites jaugés hydrologiquement similaires au site cible [Dalrymple, 1960; Burn, 1990].

Les approches menées dans le cadre de l'AFR comportent deux principales étapes, à savoir la détermination des régions homogènes (DRH) et l'estimation régionale (ER). Une description détaillée des méthodes de DRH et d'ER adoptées dans cette thèse est présentée dans les sections suivantes.

1.2.1. Délinéation des régions homogènes

La DRH consiste à regrouper des sites ayant un comportement hydrologique similaire au site cible non jaugé par l'intermédiaire de diverses méthodes statistiques [e.g. Burn, 1990; Cavadias, 1990; Ouarda et al., 2001]. Cette délinéation est basée essentiellement sur les caractéristiques physiographiques, météorologiques et hydrologiques des bassins versants. Diverses méthodes de DRH ont été proposées dans la littérature [e.g. Ouarda, 2013]. Le choix de la meilleure méthode de regroupement a fait également l'objet de plusieurs études. En fonction de l'approche adoptée, on identifie trois façons pour former les régions homogènes qui peuvent être géographiquement contiguës (par exemple la méthode des L-moments), géographiquement non contiguës (l'Analyse

en composantes principales et la Classification Ascendante Hiérarchique) ou encore de type voisinage (tels que l'Analyse Canonique des Corrélations, ACC, et la région d'influence, ROI). Les résultats de deux études d'inter-comparaison réalisées par un groupe d'hydrologues [GREHYS, 1996a, 1996b] montrent que les méthodes de type voisinage se distinguent des autres [Ouarda et al., 2008b]. Dans cette même direction, les approches de type voisinage, particulièrement l'ACC, sont choisies dans la présente étude.

L'ACC est une méthode statistique d'analyse multivariée utilisée pour explorer et décrire les relations de dépendance qui peuvent exister entre deux groupes de variables aléatoires. Cette approche a été utilisée avec succès dans plusieurs domaines tels que la prévision climatique saisonnière [Barnett et Preisendorfer, 1987], la gestion financière [Tishler et Lipovetsky, 1996] et la prévision des risques d'accidents [Michael et Raymond, 2003]. Dans l'AFR des crues, cette approche a été initialement introduite par Cavadias [1990] pour identifier les voisinages hydrologiques. En revanche, cette technique a été exploitée dans de nombreuses études de régionalisation pour estimer aussi bien les débits de crues [e.g. Ouarda et al., 2001; Chokmani et Ouarda, 2004] que les débits d'étiages [e.g.Ouarda et al., 2008a; Tsakiris et al., 2011].

Formellement, le principe de l'ACC consiste à créer, à partir des variables hydrologiques et physiographiques des bassins versants : $X = (X_1, X_2, ..., X_q)$ et $Y = (Y_1, Y_2, ..., Y_r)$ respectivement, de nouvelles variables appelées variables canoniques U et V, de sorte que la corrélation, dite canonique, $\lambda_i = corr\left(U_i, V_i\right)$ soit maximale en imposant une variance unitaire. Il importe de mentionner qu'il s'agit bien des transformations linéaires des variables originales X et Y:

$$U_i = a_{i1}X_1 + a_{i2}X_2 + ... + a_{iq}X_q \qquad (1)$$

$$V_i = b_{i1}Y_1 + b_{i2}Y_2 + ... + b_{ir}Y_r \qquad (2)$$

où $i = 1,..., p$ et $p = \min(r, q)$.

Pour un site non jaugé dans lequel l'information hydrologique est indisponible, l'information météo-physiographique canonique est généralement connue, permettant une estimation de l'information hydrologique. Ainsi, une région homogène à un niveau de confiance 100 (1-α) % avec $\alpha \in [0,1]$ et par exemple, alpha = 0,2 est la valeur prise dans le cas d'étude du chapitre 1. La région de confiance est identifiée par la distance de Mahalanobis entre l'estimation de la variable hydrologique du site cible et les sites voisins. Plus de détails techniques sur cette approche sont donnés dans Ouarda et al. [2001].

1.2.2. L'estimation régionale

L'estimation régionale (ER), la deuxième étape de l'AFR, consiste à transférer l'information hydrologique des sites jaugés vers un site non jaugé ou partiellement jaugé, au sein de la même région hydrologique homogène. On reconnait deux principales catégories de méthodes pour cette étape à savoir la méthode de l'indice de crue [Dalrymple, 1960] et les approches régressives [e.g. Pandey et Nguyen, 1999]. La première catégorie, l'indice de crue, communément utilisée en AFR fait l'hypothèse que toutes les données des sites appartenant à une même région homogène ont la même distribution statistique à un facteur d'échelle près. La deuxième catégorie inclut toute une panoplie de méthodes régressives permettant d'établir des relations directes entre les variables explicatives et la variable réponse en utilisant différentes fonctions de transfert.

Un fait important à noter est que ces deux approches, des deux catégories précédentes, se basent sur les quantiles de crue estimés dans des sites jaugés pour calibrer la fonction de transfert. En règle générale, uniquement les quantiles estimés avec des séries de données suffisamment longues

(dépassant généralement les trentaines d'années) sont retenus pour la calibration et l'évaluation du modèle RFA, tandis que l'information associée à des sites avec de courtes séries de données est souvent ignorée [Marco et al., 2012].

Dans les sous-sections suivantes les principaux outils régressifs adoptés dans nos travaux de thèse sont résumés.

- **Régression linéaire multiple**

La réponse hydrologique d'un bassin versant dépend principalement de ses facteurs physiographiques et météorologiques. Pour décrire adéquatement la relation entre une variable caractéristique du régime hydrologique (telle que le quantile de crue Q_T associé à une période de retour T) d'un bassin versant et ses $\boldsymbol{q}$ caractéristiques physiographiques, le modèle de forme puissance est généralement le plus utilisé [Pandey et Nguyen, 1999]:

$$Q_T = \alpha_o X_1^{\alpha_1} X_2^{\alpha_2} .... X_q^{\alpha_q} \varepsilon \qquad (3)$$

où $\alpha_o, \alpha_1, ..., \alpha_q$ sont des paramètres à estimer et ε est l'erreur du modèle.

En pratique, cette relation peut être linéarisée en prenant le logarithme de l'équation (3):

$$\log(Q_T) = \beta_0 + \beta_1 \log(X_1) + ... + \beta_q \log(X_q) + \varepsilon_1 \qquad (4)$$

On se retrouve ainsi face à un modèle classique de régression linéaire multiple (RLM) dont les paramètres sont à estimer en utilisant des méthodes d'estimation qui consistent, à titre d'exemple, à minimiser une somme des carrés des résidus (méthode des moindres carrés) [Pandey et Nguyen, 1999].

- **Modèles additifs généralisés**

Initialement développés par Hastie et Tibshirani [1986], les modèles additifs généralisés (GAM pour Generalized Additive Models) constituent une extension des modèles linéaires généralisés (GLM pour Generalized Linear Models) [McCullagh et Nelder, 1989]. Ces derniers sont également une extension flexible du modèle de la régression linéaire permettant de modéliser des variables réponse non Gaussiennes.

D'une manière similaire au GLM, le GAM permet de modéliser une variable réponse Y non forcément normale (dont la distribution appartient à la famille exponentielle) avec des relations de dépendance flexibles. À la différence du GLM, le GAM ne se restreint pas à décrire des relations linéaires. De plus, il intègre des fonctions qui ne sont pas nécessairement paramétriques. La formulation du modèle de base est explicitement donnée par [Wood, 2006]:

$$g(Y) = \alpha + \sum_{i=1}^{m} f_i\left(X_i\right) + \varepsilon \qquad (5)$$

où g est une fonction de lien monotone et différentiable, f_i sont des fonctions de lissage. Cette approche a été largement utilisée pour des applications en médicine [Austin, 2007], en pollution atmosphérique [Davis et al., 1998], en épidémiologie environnementale [Bayentin et al., 2010] et en hydrologie [López-Moreno et Nogués-Bravo, 2005]. Dans le contexte de l'AFR, le modèle GAM a été introduit par Chebana et al. [2014] pour l'estimation des quantiles de crues dans des sites non jaugés dans la province du Québec.

- **Réseaux de neurones artificiels**

Un réseau de neurones artificiel (RNA) est un modèle mathématique dont la conception est inspirée du fonctionnement des neurones biologiques. À ce jour, plusieurs modèles de RNA ont été

développés et mis en place permettant la résolution d'un grand nombre de problèmes complexes tels que ceux liés aux assurances, la finance, la médecine, l'environnement [e.g. Ashtiani et al., 2014; Coad et al., 2014; Benzer et Benzer, 2015; Wang et al., 2015] et également l'hydrologie [e.g. Dawson et Wilby, 2001; Nohair et al., 2008; Huo et al., 2012]. Les RNA ont été également utilisés avec succès dans l'AFR [e.g. Ouarda et Shu, 2009; Aziz et al., 2014; Alobaidi et al., 2015].

Les différences entre plusieurs classes de RNA peuvent résider, par exemple, dans la topologie du modèle adopté, dans l'algorithme d'apprentissage et/ou la fonction de transfert utilisée. Parmi les différents types de RNA, le perceptron multicouches (PMC) est, jusqu'à présent, le modèle le plus couramment utilisé pour les applications hydrologiques [e.g. Chokmani et al., 2008; Chen et al., 2013]. Une architecture typique d'un PMC est caractérisée par une couche d'entrée, une ou plusieurs couches cachées et une couche de sortie. Chaque couche contient des unités de calcul directement interconnectées dans un seul sens (RNA non-bouclés ou "feed-forward"), en d'autres termes, les neurones de la couche de sortie correspondent toujours aux sorties du modèle. Les connexions entre les neurones de deux couches successives sont assurées par des fonctions de transfert conçues pour l'estimation des paramètres appropriés. Ces derniers sont estimés, au cours du processus d'apprentissage, en utilisant une procédure d'optimisation. En effet, contrairement aux modèles statistiques habituels, les RNA ne fournissent pas une solution analytique aux problèmes d'optimisation [e.g. Bekey et Goldberg, 2012]. Ainsi, la fonction objectif (par exemple la somme des erreurs au carré) doit être minimisée numériquement durant la période d'apprentissage.

Différents algorithmes d'apprentissage pour le PMC sont proposés dans la littérature parmi lesquels l'algorithme de rétro-propagation est le plus répandu [Shu et Burn, 2004]. Plus de détails sur cet algorithme sont fournis dans Haykin et Lippmann [1994]. Notons que la phase

d’apprentissage est une étape primordiale lors de la modélisation par RNA dans le sens où un bon apprentissage nous évite le problème de sur-apprentissage. Ce dernier se manifeste quand le modèle performe bien durant l’étape de l’apprentissage (de calibration) mais perd ses pouvoirs prédictifs sur les échantillons de la validation.

1.3. Organisation de la synthèse

La synthèse de cette thèse est organisée comme suit : la section 2 présente la problématique, les objectifs et l’originalité du projet de recherche. La section 3 résume la méthodologie adoptée ainsi que les outils statistiques utilisés pour atteindre les objectifs de la thèse. Les principaux résultats obtenus sont présentés dans la section 4. Enfin, la conclusion et des perspectives de recherche sont présentées dans la section 5.

2. Problématiques et objectifs de la recherche

Dans cette section, les principaux problématiques et objectifs de recherche sont présentés.

2.1. Problématiques

L'objectif de la présente sous-section est d’évoquer les principales questions qui seront traitées dans cette thèse. La problématique générale de la thèse repose sur le constat que les approches disponibles dans la littérature et couramment utilisées en AFR soient inappropriées pour diverses situations réalistes telles que la non-linéarité, l’incompatibilité entre les étapes de la DRH et de l’ER, l’ignorance d’une partie de l’information, la négligence de courtes séries de données et l’agrégation de l’information. Cette problématique générale se décompose en les problématiques spécifiques suivantes :

A. Non-linéarité du processus hydrologique dans l'étape de la délinéation

La complexité naturelle du processus hydrologique, dérivant par exemple de la topographie des bassins versants, leurs formations géologiques ou également de la variation météorologique, a été largement reconnue et documentée dans la littérature hydrologique [Riad et Mania, 2004]. Cette propriété, contraignante pour la modélisation hydrologique, doit être prise en compte dans une démarche de modélisation pour aboutir à des modèles non seulement précis en termes de critères de performance mais également capable de reproduire la dynamique des processus hydrologiques. À cet égard, une panoplie de méthodes non linéaires a été proposée dans diverses études portant sur l'AFR des crues [e.g. Shu et Burn, 2004]. Plus précisément, cet aspect non linaire a été considéré uniquement dans l'étape de l'ER via l'utilisation des RNA [e.g. Shu et Ouarda, 2007], d'un modèle de régression linéaire basé sur une méthode d'estimation non linéaire [e.g. Pandey et Nguyen, 1999] ou récemment à travers les GAM [Chebana et al., 2014]. Cependant, l'aspect non linéaire n'a pas été considéré dans l'étape de la DRH. Par conséquent, les approches communément utilisées en AFR sont partiellement inappropriées puisqu'elles utilisent un modèle non linéaire dans l'étape de l'estimation combiné avec une approche linéaire dans l'étape de la délinéation. Une question qui se pose à ce niveau est la suivante : est-il utile de considérer la non-linéarité des processus hydrologiques dans la première étape de l'AFR à savoir celle de l'identification des voisinages hydrologiquement homogènes ?

B. Non-linéarité dans les deux étapes de l'AFR

En dépit des efforts déployés dans des études antérieures [e.g. Shu et Ouarda, 2007; Chebana et al., 2014] qui visent une meilleure estimation du risque lié aux évènements hydrologiques extrêmes dans des sites non jaugés, il importe de noter que ces méthodes peuvent être relativement inadaptées et incompatibles dans la mesure où elles intégrèrent l'aspect non linéaire uniquement dans l'étape de l'ER. En effet, mis à part le fait que la non-linéarité n'a pas été convenablement intégrée dans

l'étape de la DRH, les méthodes non linéaires n'ont pas encore été prises en compte simultanément dans les deux étapes de l'AFR.

C. Exploitation insuffisante de l'information hydrologique disponible

La revue de littérature de l'AFR des variables hydrologiques montre que toutes les études menées dans ce cadre implémentent et évaluent leurs modèles régressifs en utilisant des séries hydrologiques estimées (quantiles aux sites jaugés). En effet, en disposant des données de DMA observés, une estimation locale dans un site jaugé est fournie moyennant une AF locale dont la performance dépend, entre autres, de la taille des observations, de la qualité d'ajustement des distributions statistiques et des méthodes d'estimation des paramètres associés. Cette estimation des quantiles constitue l'entrant des modèles régionaux. Par conséquent, les incertitudes associées à chaque estimation locale vont être additionnées à celles de l'AFR. En outre, effectuer une AF locale dans chaque site jaugé est un processus long. Enfin, l'évaluation de la performance des modèles régionaux est généralement basée sur cette estimation locale des quantiles considérée comme référence. Subséquemment, une estimation de moindre qualité pourra influencer négativement la qualité de l'estimation régionale.

D. Limitations des méthodes classiques d'analyse non stationnaire locale

Au cours des dernières décennies, le sujet de la non-stationnarité des processus hydrologiques, entre autres les crues, a reçu une attention considérable chez la communauté scientifique en général et les hydrologues en particulier. Outre cet aspect, la non-linéarité est également l'un des aspects incontestables qui caractérisent un processus hydrologique.

Différents points critiques peuvent être révélés lors de l'élaboration d'une AF non stationnaire classique à savoir:

- L'inclusion de plusieurs étapes, particulièrement une première analyse exploratrice des données, l'ajustement des lois statistiques et l'estimation des paramètres;

- Le choix d’une distribution de probabilité. En fait, la distribution de probabilité réelle des données observées est inconnue. Lors de l'identification de la distribution "adéquate", différents candidats se présentent et des analyses relativement complexes seront effectuées touchant aussi bien la partie centrale que les queues supérieure et inférieure de la loi de probabilité.
- La complexité et la difficulté de choisir une distribution non stationnaire au sein de la même famille de loi de probabilité (ex. la GEV pour Generalized Extreme Value). Ceci comporte l’identification des variables (des indices climatiques et/ou anthropiques) qui incluent une information supplémentaire liée à la non stationnarité, la forme (ex. linéaire, quadratique) de cette dernière, ainsi que les paramètres de la distribution (position, échelle) qui sont affectés ;
- En termes de développement, chaque distribution utilisée dans le cadre stationnaire nécessite des adaptations spécifiques propres à elle (ex. les méthodes d’estimation des paramètres). C’est le cas par exemple de la GEV et de la loi Log-Normale;
- Le problème de manque de données est une question traditionnelle communément reconnue dans le cas d’une analyse stationnaire des variables hydrologiques. En effet, les courtes séries d’observations sont généralement inappropriées pour obtenir des estimations fiables des quantiles. Ce problème persiste et s’aggrave lors de l’élaboration d’une analyse non stationnaire. En effet, dans le cadre non stationnaire, plus de paramètres sont à estimer. En outre, une tendance (une des formes de la non-stationnarité) n’est détectable que sur une assez longue période d’enregistrement;
- La difficulté d’interpréter les résultats vue l’absence d’un lien direct entre les co-variables et la variable réponse.

2.2. Objectifs de la thèse

Le principal objectif de la thèse consiste à développer des approches qui doivent être non seulement précises en termes de performances mais également capables de tenir compte de la complexité du

processus hydrologique. Ainsi, on se sert de certaines approches statistiques prometteuses pour pallier aux limitations des méthodes existantes et couramment utilisées en AF locale et régionale.

Afin d'atteindre l'objectif principal, ce dernier se décompose en quatre objectifs primaires associés aux problématiques décrites ci-dessus :

A. Développer une nouvelle approche d'AFR permettant de considérer la non-linéarité du processus hydrologique dans l'étape de la DRH;
B. Proposer et évaluer différentes combinaisons de méthodes linéaires et non linéaires au niveau de chacune des deux étapes de l'AFR. Ceci revient à effectuer une étude comparative entre les combinaisons proposées et des combinaisons classiques afin d'identifier laquelle des deux étapes est plus affectée par l'aspect non linéaire ;
C. Développer un modèle moins restrictif en termes de données utilisées permettant :

- D'intégrer directement des données observées plutôt que des séries estimées localement
- De maximiser l'exploitation de l'information hydrologique disponible sans imposer de contraintes sur la taille de la série d'observations ;

D. Proposer un modèle d'AF des crues dans un cadre local non linéaire et non stationnaire, permettant :

- D'intégrer plus d'information hydro-climatique,
- De réduire et simplifier les étapes techniques associées aux approches classiques.

En outre, ce modèle constitue une étape préliminaire pour fonder un modèle régional qui intègre simultanément l'aspect non linéaire et l'aspect non stationnaire des processus hydrologiques.

Avant d'élaborer sur chacun de ces éléments, il importe de noter que les problématiques et les objectifs énoncés sont valables pour toutes les variables hydrologiques, aussi bien les débits de crues que les débits d'étiages, entre autres. Dans le cadre de la présente thèse, on se concentre sur l'étude des caractéristiques des crues.

3. Méthodologie

Dans cette section, on présente les nouvelles approches développées dans ce projet de thèse qui visent à répondre à chacune des problématiques discutées ci-dessus. La stratégie de modélisation développée combine :

A. Un modèle d'AFR introduisant la non-linéarité dans l'étape de la DRH ;
B. Une étude comparative entre une panoplie de modèles et méthodes d'AFR visant l'aspect non linaire ;
C. Un modèle d'AFR permettant une meilleure exploitation des données;
D. Un modèle d'AF locale non linéaire et non stationnaire.

Soulignons à ce niveau que les travaux de recherche élaborés dans le cadre de cette thèse s'appuient principalement sur des outils statistiques bien fondés afin de résoudre les problématiques énoncées préalablement qui touchent essentiellement à l'estimation du risque lié aux extrêmes hydrologiques.

3.1. Un modèle d'AFR introduisant la non-linéarité dans la DRH

Plusieurs méthodes de régionalisation ont été développées dans la littérature récente pour améliorer l'estimation des périodes de retour des extrêmes hydrologiques dans des sites non jaugés [Ouarda, 2013]. Parmi les méthodes dédiées à l'identification des régions homogènes, l'ACC présente un

outil théorique et pratique important [e.g. Ribeiro-Corréa et al., 1995]. Toutefois, il s'agit d'une approche *linéaire* ne permettant pas de décrire les éventuelles relations non linéaires entre les variables. Par conséquent, l'ACC pourrait ne pas être adaptée pour la DRH ou ne pas conduire nécessairement aux meilleurs résultats notamment en traitant les processus hydrologiques.

Dans le but de tenir compte de la complexité du processus hydrologique, et en vue d'une meilleure évaluation du risque, une méthode non linéaire de DRH est considérée dans une approche de régionalisation des crues à savoir l'ACC non linéaire (ACCNL). Initialement introduite par Hsieh [2000] pour des applications en climatologie, l'ACCNL est une extension non linéaire de l'ACC, basée sur les RNA. Cette approche statistique, tout comme l'ACC, sert à réduire les dimensions d'un espace en tenant compte des relations entre les variables d'intérêt. Dans notre étude, l'utilisation de cette approche a été motivée par la complexité des relations entre les variables hydrologiques et les variables météo-physiographiques.

Sur le plan pratique, cette approche a été adoptée avec succès dans divers domaines tels que l'analyse de la conversion de la voix [e.g.Zhihua et Zhen, 2010], la biomédecine [e.g. Campi et al., 2013], la médecine [e.g. Wang et al., 2005], la sociologie [e.g. Frie et Janssen, 2009] et notamment en météorologie et en climatologie [e.g. Hsieh, 2001; Wu et Hsieh, 2003]. En hydrologie en général et en AFR en particulier, le potentiel de cette approche n'a pas été encore exploité.

Conçue avec le même principe que l'ACC, l'idée de l'ACCNL consiste également à identifier des variables canoniques (U, V), à la seule différence que les combinaisons linéaires entre les variables canoniques et les variables originales sont remplacées par des combinaisons non linéaires en utilisant les RNA:

$$U = w^{(x)} h^{(x)} + \bar{b}^{(x)} \tag{6}$$

$$V = w^{(y)} h^{(y)} + \bar{b}^{(y)} \tag{7}$$

avec $\bar{b}^{(x)} = -\left\langle w^{(x)} h^{(x)} \right\rangle$ et $\bar{b}^{(y)} = -\left\langle w^{(y)} h^{(y)} \right\rangle$.

Les fonctions $h^{(x)}$ et $h^{(y)}$ sont des couches cachées définies comme suit:

$$h_k^{(x)} = f\left(\left(W^{(x)} x + b^{(x)}\right)_k\right) \quad ; \quad k \text{ et } n = 1 \ldots l \tag{8}$$

$$h_n^{(y)} = f\left(\left(W^{(y)} y + b^{(y)}\right)_n\right) \tag{9}$$

où $W^{(x)}$ et $W^{(y)}$ sont des matrices de poids, $b^{(x)}$ et $b^{(y)}$ sont des vecteurs de paramètres, l indique le nombre de neurones cachés et *f* est une fonction de transfert, généralement choisie comme la tangente hyperbolique. L'estimation de ces paramètres revient à maximiser la corrélation canonique ou encore à minimiser la fonction objective *J* = - *corr* (*u*, *v*) en utilisant une procédure d'optimisation. Une illustration graphique du principe du fonctionnement de l'ACCNL est présentée dans la Figure 1. Dans celle-ci S_o désigne la position du site non jaugé dans l'espace physiographique canonique. Notons que les valeurs des variables météo-physiographiques de ce site sont connues. Le but de l'ACCNL étant de donner une estimation $\hat{S}_o$ de la position du site non jaugé dans l'espace hydrologique canonique. Pour plus de détails sur l'ACCNL, le lecteur peut se référer à Hsieh [2000] et à Ouali et al. [2015] (Chapitre 2 de cette thèse).

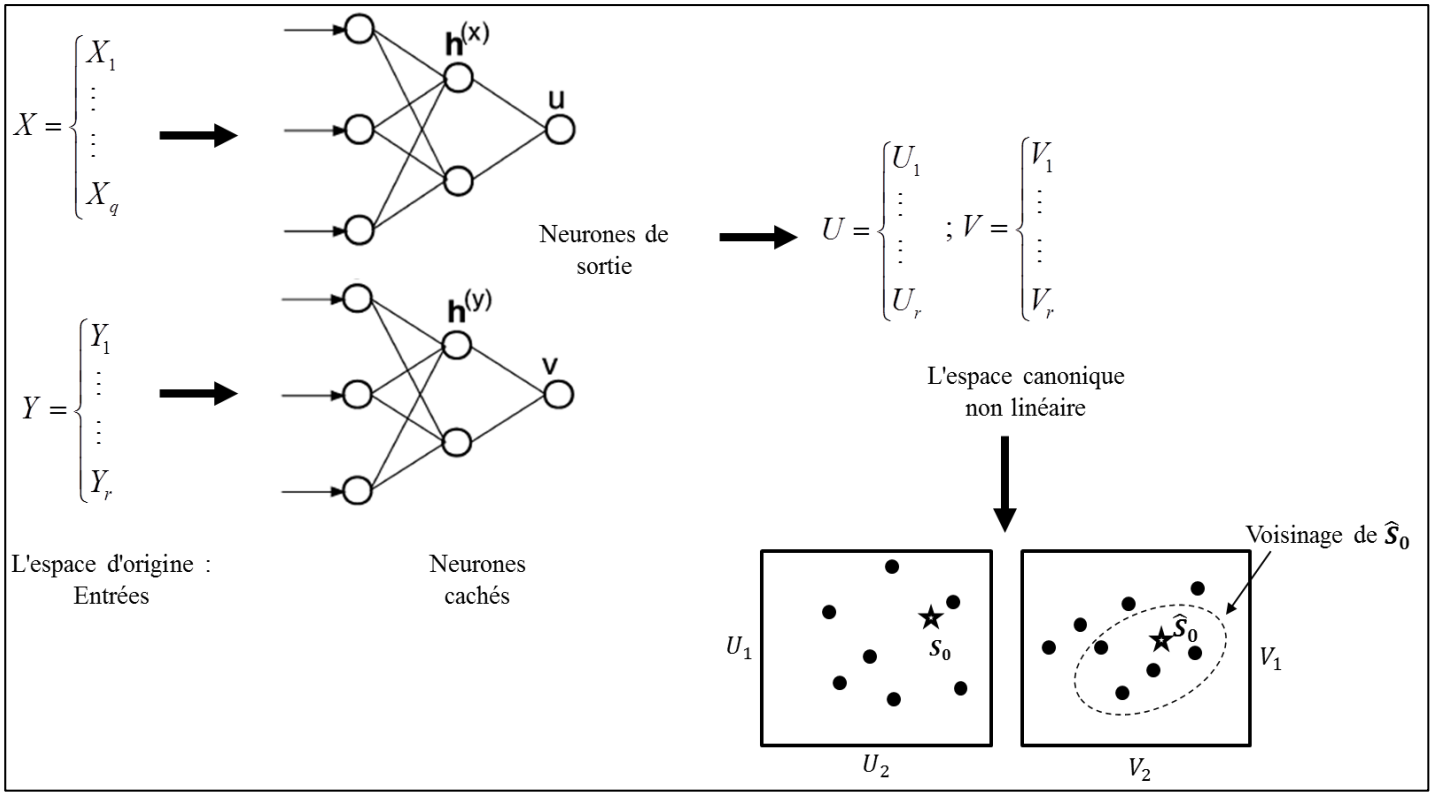


Figure 1. Schéma illustratif du principe de l'ACCNL

Dans le cadre de cette étude, l'ACCNL a été couplée à un modèle de régression log-linéaire dans l'étape de l'ER (ACCNL & RLM). Pour évaluer la performance d'une telle approche, la combinaison ACCNL & RLM est comparée à une combinaison purement linéaire très utilisée dans la littérature de l'AFR, à savoir l'ACC & RLM.

3.2. Combinaisons des approches et étude comparative

Comme mentionné ci-dessus, les processus hydrologiques sont des processus assez complexes incorporant une forte non-linéarité. À cet effet, des progrès importants ont été accomplis dans des outils statistiques afin de tenir compte de cette réalité. Dans le cadre de l'AFR, ces progrès ont touchés particulièrement aux méthodes de l'ER en intégrant les RNA et les GAM pour l'estimation des quantiles de crue [Shu et Burn, 2004; Chebana et al., 2014]. Dans ce travail de thèse, toutes ces

techniques, entre autres, sont regroupées constituant de nouvelles combinaisons (DRH-ER) semi-linéaires et non linéaires. Les approches proposées à ce niveau sont des approches plus compatibles dans la mesure où des techniques non linéaires sont considérées au niveau des deux étapes de l'AFR. Une étude comparative des différentes approches est ainsi mise en œuvre pour pouvoir identifier la meilleure combinaison [Ouali et al., 2016a], (Chapitre 3 de ce manuscrit). Ceci revient à adopter les nouvelles combinaisons non linéaires suivantes:

- *L'ACCNL dans l'étape de la DRH couplée à deux modèles de RNA dans l'étape de l'ER :* Afin d'assurer une plus grande compatibilité, un modèle d'estimation basé sur les RNA est adopté dans l'étape de l'ER en combinaison avec la méthode de l'ACCNL basée également sur les RNA. En réalité, on distingue deux types de modèles d'estimation à savoir un modèle de RNA simple et un modèle de RNA ensemble. Éventuellement, le plus grand souci lors de l'utilisation d'un modèle de RNA est bien sa capacité de prédire en intégrant des données différentes de celles utilisées pour la calibration. À cet égard, des études pertinentes montrent que ce point peut être amélioré significativement en combinant un ensemble de RNAs dans un seul modèle de façon à ce qu'ils fournissent différentes solutions. Bien que ceci semble être redondant, cette approche de généralisation, formellement connue comme réseau de neurones d'ensemble (RNE), offre une meilleure performance que celle associée à un seul RNA. Le principe de base consiste à considérer à chaque fois un ensemble de données différent pour la calibration du modèle en utilisant des techniques de ré-échantillonnages [Shu et Burn, 2004]. Les sorties de tous les modèles RNA sont ensuite combinées pour fournir la sortie du modèle global. Ainsi, les deux approches, simple et ensemble, ont été implémentées en combinaison avec l'ACCNL dans la DRH.

- *L'ACCNL dans l'étape de la DRH couplée à un GAM dans l'étape de l'ER* : Parallèlement à l'approche par RNA, le GAM a récemment été introduit dans l'AFR par Chebana et al. [2014]. Ce modèle est considéré dans cette analyse pour trois raisons : i) le GAM fournit un outil

particulièrement bien adapté pour l'étude des relations complexes, ii) comme indiqué dans Chebana et al. [2014], il aboutit à des estimations régionales nettement plus précises que celles associées aux approches classiques et finalement iii) le GAM permet une interprétation facile des résultats. En effet, une illustration graphique des courbes de lissages permet une compréhension plus réaliste de la véritable relation entre la variable réponse et les variables explicatives et, en conséquence, des phénomènes sous-jacents.

Dans le but d'alimenter et compléter l'étude comparative menée dans ce travail, ces approches non linéaires ont été comparées à d'autres approches linéaires et semi-linéaires. Une illustration graphique de toutes les approches de régionalisation considérées dans ce travail est présentée dans la Figure 2.

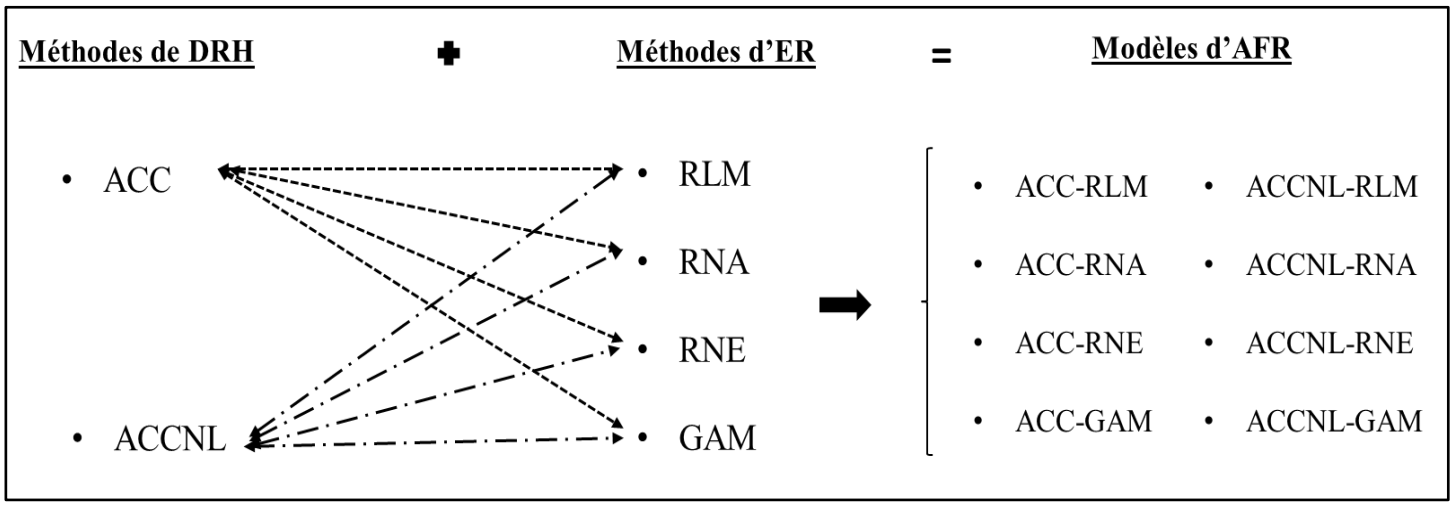


Figure 2. Différentes combinaisons et modèles adoptés

Pour quantifier l'erreur relative à chaque approche, différents critères de performance sont utilisés tels que l'efficacité de Nash-Sutcliffe (NASH), la racine carrée de l'erreur quadratique moyenne relative (*RRMSE*) et le biais relatif (*RBIAS*). Ces critères sont utilisés au sein d'une procédure de

validation croisée (*jackknife* ou *leave-one-out*) qui consiste à considérer, à tour de rôle, chaque site de la région comme un site non jaugé.

3.3. Approche régionale par RQ

Malgré le potentiel de chacune des méthodes d'estimation adoptées dans les études de régionalisation précédentes, une limitation majeure et commune pour toutes ces méthodes est qu'elles soient basées sur des méthodes régressives qui donnent des estimations de la moyenne conditionnelle de la variable réponse. Ainsi, pour fournir des estimations des quantiles de crues, ces modèles sont calibrés par des quantiles estimés localement. Cette procédure permet de trainer les erreurs et les incertitudes associées à chaque AF locale effectuée dans chaque site de la région d'étude. Rappelons que l'une des grandes faiblesses de l'AF consiste à ignorer les sites avec de courtes séries de données sous prétexte qu'elles ne produisent pas une bonne estimation locale. Le but du présent travail étant de proposer un modèle d'AFR qui fournit directement le quantile conditionnel de crue dans un site non jaugé, tout en retenant toute source de données. C'est dans cette optique que la régression des quantiles (RQ) a été introduite dans le cadre de cette étude [Ouali et al., 2016b]. Cette approche est présentée plus en détails dans le Chapitre 4 de ce manuscrit de thèse.

La RQ est une approche statistique permettant une description complète de la variable d'intérêt. En fait, elle repose sur un principe similaire à celui de la régression classique. Celle-ci, fondée sur l'estimateur du moindre carré, fournit la moyenne conditionnelle de la variable réponse comme solution au problème de minimisation de la somme des carrés des écarts. D'une manière analogue, la régression médiane, ou régression des moindres écarts absolus, fournit une estimation de la médiane de la réponse définie comme étant la solution au problème de minimisation d'une somme des résidus absolus. Cette approche est reconnue plus robuste aux valeurs aberrantes que la

régression classique des moindres carrés et évite l'imposition d'une distribution paramétrique au processus des erreurs. Dans cette même direction, qui adopte la moyenne et la médiane d'un échantillon comme des solutions à des problèmes de minimisation bien spécifiques, il s'est avéré utile d'étendre ces approches vers d'autres quantiles de l'échantillon. C'est dans ce sens que Koenker et Bassett [1978] ont introduit l'approche de la RQ qui fournit une description plus riche de la variable réponse du fait qu'elle s'intéresse à l'ensemble de la distribution conditionnelle plutôt qu'à sa moyenne conditionnelle ou à sa médiane.

Bien que le principe soit relativement ancien, la RQ a connu récemment un gain d'intérêt notamment dans les domaines qui touche à la modélisation environnementale et l'évaluation de l'impact des changements climatiques [e.g. Friederichs et Hense, 2007; Elsner et al., 2008; Jagger et Elsner, 2009; Cannon, 2011; Ben Alaya et al., 2015]. Cette approche a été également introduite dans le cadre de l'AF locale des crues par Sankarasubramanian et Lall [2003] mais n'a jamais été exploitée dans le cadre régional.

La procédure d'estimation des coefficients de régression consiste à minimiser la somme pondérée des valeurs des termes d'erreurs positifs et négatifs respectivement par le quantile d'intérêt τ et son complémentaire (1-τ) :

$$\arg\min_{\mathbf{b}} \sum_{i=1}^{n} \rho_\tau \left(y_i - \mathbf{x}_i^T \mathbf{b} \right) \tag{10}$$

avec $\rho_\tau(.)$ est le "check function" défini par:

$$\rho_\tau(u) = \begin{cases} u(\tau - 1) & if\ u < 0 \\ u\tau & if\ u \geq 0 \end{cases} \qquad ;0 < \tau < 1 \tag{11}$$

L'approche proposée permet d'établir une relation directe entre les variables météo-physiographiques et les quantiles de crues.

En revanche, l'un des objectifs de la présente recherche est d'évaluer la performance de l'approche par RQ en la comparant aux approches classiques. Toutefois, les critères d'évaluation couramment utilisés pour cette fin sont établis en se basant sur des quantiles estimés localement. Ces critères considèrent les estimations locales des quantiles comme des estimations parfaites. Or, en réalité, l'erreur totale de l'estimation régionale en utilisant la RLM émane de deux sources principales: i) l'erreur de l'estimation locale qui n'est souvent pas prise en compte dans l'évaluation de la modélisation régionale et qui dépend principalement de la longueur des séries de données observées [Tasker et Moss, 1979], et ii) l'erreur régionale qui est évaluée en utilisant les critères classiques.

A cet égard, un critère d'évaluation de la performance des modèles régionaux, basé sur les données observées, est également proposé dans le cadre de ce travail. Le principe derrière ce critère consiste à utiliser la fonction-objectif de la RQ, plutôt que la somme de l'erreur quadratique utilisée dans le cas classique, en intégrant les séries de DMA dans chaque site. Le critère d'évaluation proposé dans la présente étude, la moyenne de la fonction définie par morceau (MPLF pour Mean Piecewise Loss Function), en utilisant la fonction-objectif de Konker est exprimé comme suit :

$$MPLF\left(p\right)=\frac{10^3}{n}\sum_{i=1}^{N}\sum_{j=1}^{n_i}\rho_p\left(y_{ij}-\hat{q}_{ip}^{R}\right)\quad ;\quad p\in(0,1) \tag{12}$$

où n désigne le nombre total d'observations dans toutes les stations, $n=\sum_{i=1}^{N}n_i$, et $\hat{q}_{ip}^{R}$ est le quantile régional d'ordre p estimé dans un site i.

En somme, la présente partie de la thèse vise à combler des lacunes des RFA classiques en : i) effectuant une estimation directe des quantiles sans effectuer une AF locale et ii) proposant un

critère d'évaluation objectif. Ceci est effectué en considérant différents cas de figures selon les données utilisées pour calibrer le modèle:

- La calibration et l'application des deux modèles considérés (RQ et RLM) sont réalisées en utilisant tous les sites;
- Uniquement les sites ayant des séries de données de longueurs supérieures à 30 ans ont été considérés pour la calibration des deux modèles;
- Le modèle RLM est construit en utilisant uniquement les sites dont les longueurs sont supérieures à 30 ans, et évalué en utilisant tous les sites.

Pour plus de détails sur le modèle et le critère d'évaluation proposés, le lecteur est référé à Ouali et al. [2016b] correspondant au chapitre 4 de ce manuscrit.

3.4. Modèle d'AF non linéaire non stationnaire en utilisant la RQ

L'évolution du climat peut être détectée à partir d'un changement des grandeurs statistiques qui le décrit. Cette évolution peut impliquer des changements dans l'occurrence des crues. Dans cette optique, de nombreuses études ont investigué l'impact des changements climatiques sur l'estimation des extrêmes hydrologiques [e.g. He et al., 2006; El Adlouni et al., 2007; Aissaoui-Fqayeh et al., 2009; López et Francés, 2013]. La non-stationnarité des événements hydrologiques est souvent traitée en introduisant des co-variables comme le temps ou des indices climatiques dans les paramètres de la fonction de distribution des DMA. Dans cette direction, les modèles non stationnaires classiques suivants ont été considérés :

- GEV_{00} : modèle stationnaire classique avec des paramètres constants;
- GEV_{10} : modèle dont le paramètre de location dépend linéairement d'une co-variable;
- GEV_{01} : modèle dont le paramètre d'échelle dépend linéairement d'une co-variable;

- GEV_{11} : modèle dont les paramètres de location et d'échelle dépendent linéairement d'une co-variable;
- GEV_{20} : modèle dont le paramètre de location est une fonction quadratique d'une co-variable;
- GEV_{21} : modèle dont le paramètre de location est une fonction quadratique d'une co-variable et dont le paramètre d'échelle dépend linéairement d'une co-variable.

Une attention particulière est portée dans cette étude sur les limitations de ces approches aussi bien sur les plans pratique que théorique. En effet, tel qu'indiqué ci-dessus (dans les problématiques), les approches non stationnaires existantes présentent plusieurs limitations liées, entre autres, à la quantité de l'information introduite, la lourdeur de la procédure, le choix de la loi de probabilité et l'interprétation de l'effet de chaque co-variable. En plus, en adoptant ces approches on se trouve dès le début dans l'obligation de supposer une hypothèse de stationnarité ou de non-stationnarité.

À ce stade, on vise à développer des techniques et des outils qui sont en mesure de prendre en compte l'aspect non stationnaire et/ou non linéaire des crues et de les intégrer dans les processus de modélisation afin de fournir une meilleure estimation du risque hydrologique. Ainsi, un modèle d'AF locale basé sur la RQ est proposé et discuté en détails dans le chapitre 5 de cette thèse [Ouali et al., 2016c].

L'aspect non stationnaire est examiné et intégré naturellement via l'inclusion des variables météorologiques comme des variables explicatives dans le modèle de régression. En revanche, l'aspect non linéaire est considéré en adoptant une extension non linéaire du modèle de la RQ via l'introduction des modèles additifs (MA) [Buja et al., 1989]. Ce dernier fournit un outil efficace et flexible pour décrire les relations complexes entre les données, en particulier dans le domaine de l'hydrologie [e.g. Campbell et Bates, 2001; Latraverse et al., 2002].

Le modèle proposé, la régression des quantiles par modèle additif (RQMA), présente une nouvelle optique de modélisation des extrêmes hydrologiques dans un cadre non stationnaire. Il s'agit d'un modèle statistique non paramétrique permettant de combiner et de percevoir les effets linéaires et non linéaires de plusieurs variables explicatives X sur la variable réponse Y. Ceci est réalisé via une combinaison linéaire de fonctions souvent non paramétriques dites fonctions de lissage f_i tel que :

$$Q_p(y \mid \mathbf{x}) = \mathbf{x}^T \mathbf{b}_p + \sum_i f_i(z_i) \qquad (13)$$

où z_i sont les nœuds du modèle.

Une grande variété de fonctions de lissage est disponible dans la littérature statistique [e.g. Mumford et Shah, 1989]. Une caractéristique commune aux différentes méthodes de lissage est le caractère local de l'estimation, autrement dit, pour produire une estimation en un point donné on n'utilise que les observations dans son voisinage. Cette propriété offre une grande flexibilité au modèle global. Une autre caractéristique de la modélisation additive est liée à la facilité d'interpréter les résultats. En pratique, ceci nous permet de percevoir le rôle de chaque variable séparément dans la prédiction de la variable réponse.

De nombreuses études ont été élaborées dans la littérature récente qui se sont servi du modèle RQMA pour étudier le retard de croissance chez les enfants en Inde [Fenske et al., 2013], identifier les facteurs de risque de la malnutrition infantile [Fenske et al., 2012], étudier la variation des prix des logements dans la ville de Munich ainsi que l'obésité infantile [Waldmann et al., 2013]. Outre l'aspect pratique, différents aspects théoriques du RQMA ont été également développés [Koenker, 2011; Yue et Rue, 2011]. Néanmoins, le potentiel de ce modèle n'a jamais été exploité dans des applications environnementales et hydrologiques.

Dans cette étude ce modèle est investigué dans le but de concevoir un modèle régional qui touche à la fois à l'aspect non linéaire et l'aspect non stationnaire du processus hydrologique.

4. Applications et résultats

Cette section inclut les principaux résultats d'application des approches proposées dans cette thèse.

4.1. Zones d'études et données

Dans le cadre des études de régionalisation, il est souvent préférable de traiter des jeux de données dont l'analyse locale préliminaire a été déjà réalisée dans des études antérieures. Ceci permet, non seulement de garantir la validité des hypothèses de base et la qualité des données traitées, mais également de se focaliser uniquement sur la partie de l'estimation dans des sites non jaugés. À ce titre, on dispose de trois bases de données provenant de trois régions en Amérique du Nord, à savoir la province du Québec (Canada) [Kouider et al., 2002], les états de l'Arkansas et du Texas (USA) [Tasker et al., 1996].

Dans chacune de ces trois bases de données, cinq variables physiographiques et deux/trois variables hydrologiques sont considérées. Les variables hydrologiques sont les quantiles du DMA normalisés par la superficie du bassin de chaque site afin d'éliminer l'effet d'échelle (des quantiles spécifiques). Les quantiles spécifiques relatifs à des périodes de retour de 10, 50 et 100 ans ont été utilisés dans ces travaux de thèse (Q_{ST}).

Pour des fins de comparaison, les approches proposées dans les chapitres 2, 3 et 4 [Ouali et al., 2015, 2016a, 2016b] ont été appliquées sur les mêmes bases de données, avec plus de détails pour celle du Québec. Les données de DMA relatives à cette région ont été acquises du Centre d'Expertise Hydrique du Québec (CEHQ) qui exploite un réseau d'environ 230 stations hydrométriques au Québec. En adoptant des critères tels que la taille minimum des séries

d'observation (15 ans) et le niveau de contrôle des stations (la proximité d'un régime naturel), 151 stations hydrométriques ont été retenues pour l'estimation des quantiles locaux pour une période d'observation entre 1900 et 2002. Les stations sélectionnées sont situées dans la partie Sud de la province de Québec, entre 45 et 55 °N (voir Figure 3). Dans Kouider et al., [2002], une AF locale a été réalisée dans chaque site jaugé incluant, entre autres, la vérification des hypothèses de base et le choix des distributions. Les tests effectués prouvent que les données utilisées sont généralement de bonne qualité. Les distributions de probabilités identifiées, qui ajustent au mieux les données de DMA, sont essentiellement les lois Gamma inverse et Log-Normale à deux paramètres. Ces dernières ont servi de base pour l'estimation des quantiles locaux dans chaque site de la région.

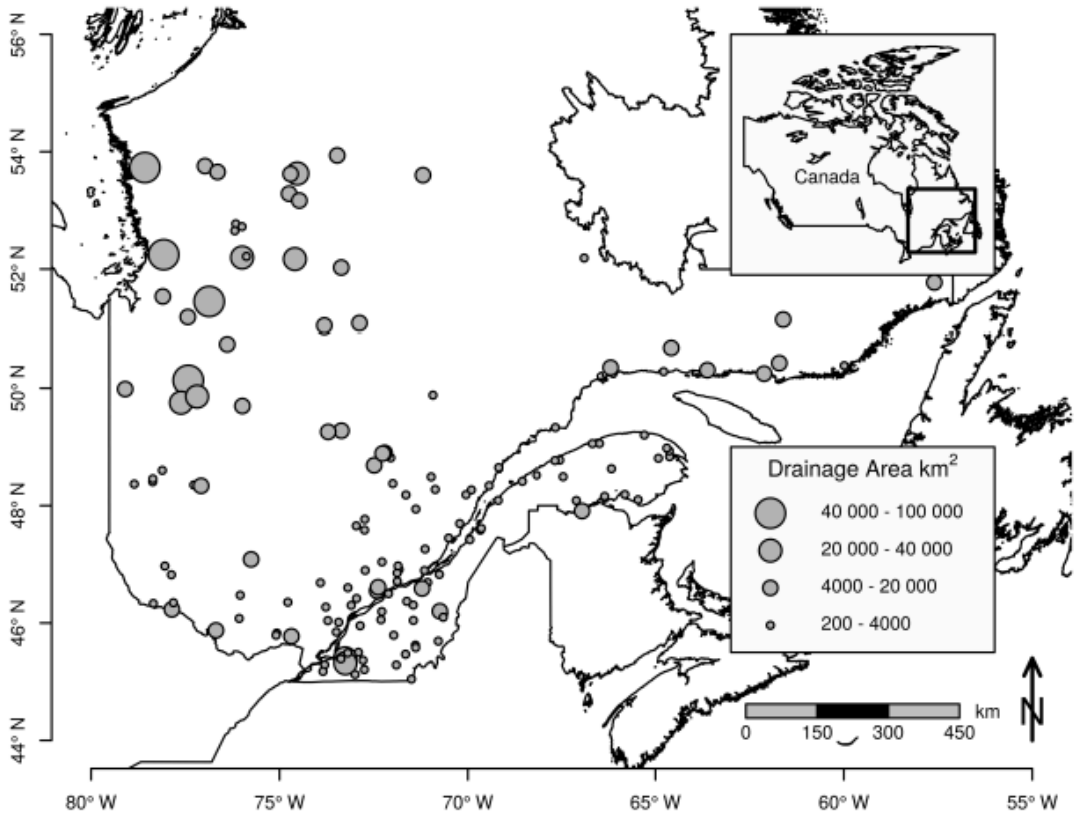


Figure 3. Localisation géographique des sites étudiés dans la partie sud de la province de Québec, Canada

Les variables météo-physiographiques ont été sélectionnées dans une étude précédente par Chokmani et Ouarda [2004]: la superficie du bassin versant (BV), la pente moyenne du bassin versant (MBS), la fraction de la superficie couverte par des lacs (FAL), les moyennes des précipitations totales annuelles (AMP), moyenne annuelle des degrés-jours supérieurs à 0°C (AMD). Les superficies des bassins versants drainées varient entre 200 km^2 et 100 000 km^2.
Plus de détails sur les deux autres régions sont présentés dans le chapitre 2 de ce rapport [Ouali et al., 2015].

Dans le dernier chapitre, correspondant à l'article Ouali et al. [2016c], les données utilisées proviennent de trois stations dont deux sont situées aux États-Unis (Arroyo Seco et Bear Creek) et la troisième est située au Québec, Canada (Dartmouth en Gaspésie). Les données utilisées pour les deux premières applications sont le DMA et l'indice d'oscillation australe (SOI pour Southern Oscillation Index). Les données utilisées dans la troisième application sont l'indice de l'oscillation Atlantique du Nord (NAO pour North Atlantic Oscillation) et la température moyenne annuelle (TMA). Plus d'informations sur ces cas d'études se trouvent dans le chapitre 5 de ce manuscrit de thèse.

4.2. Principaux résultats et discussions

Dans cette section, les principaux résultats associés à chaque étude effectuée dans le cadre de cette thèse sont présentés. Du fait du lien entre les deux études et pour éviter toute forme de redondance, les résultats des chapitres 2 et 3 sont fusionnés et présentés dans la sous-section suivante.

4.2.1. Résultats des approches non linéaires

La prise en charge de l'aspect non linéaire des processus hydrologiques dans un modèle régional se fait généralement par l'inclusion d'un modèle de régression non linéaire (comme par exemple

les RNA et les GAM) dans la partie de l'ER. Un de nos objectifs dans cette thèse est d'introduire la non-linéarité précisément dans l'étape de la DRH de l'AFR. A cet effet, une première méthode a été conçue dans Ouali et al. [2015] dans le but d'évaluer le potentiel de l'approche ACCNL dans l'étape de la DRH. L'ACCNL a ensuite permis de procéder à une étude comparative incluant une panoplie de modèles et de combinaisons visant à identifier la meilleure de ces combinaisons [Ouali et al., 2016a].

Initialement, une investigation des relations inter variables est réalisée par l'intermédiaire des nuages de points des quantiles de crue Q_{ST} et des variables météo-physiographiques. L'examen de ces nuages de points (Figure 4) montre différentes formes de relations entre les variables. On constate l'existence des relations non linéaires dont la plus remarquable est celle liant la variable superficie du bassin (BV) et le reste des variables.

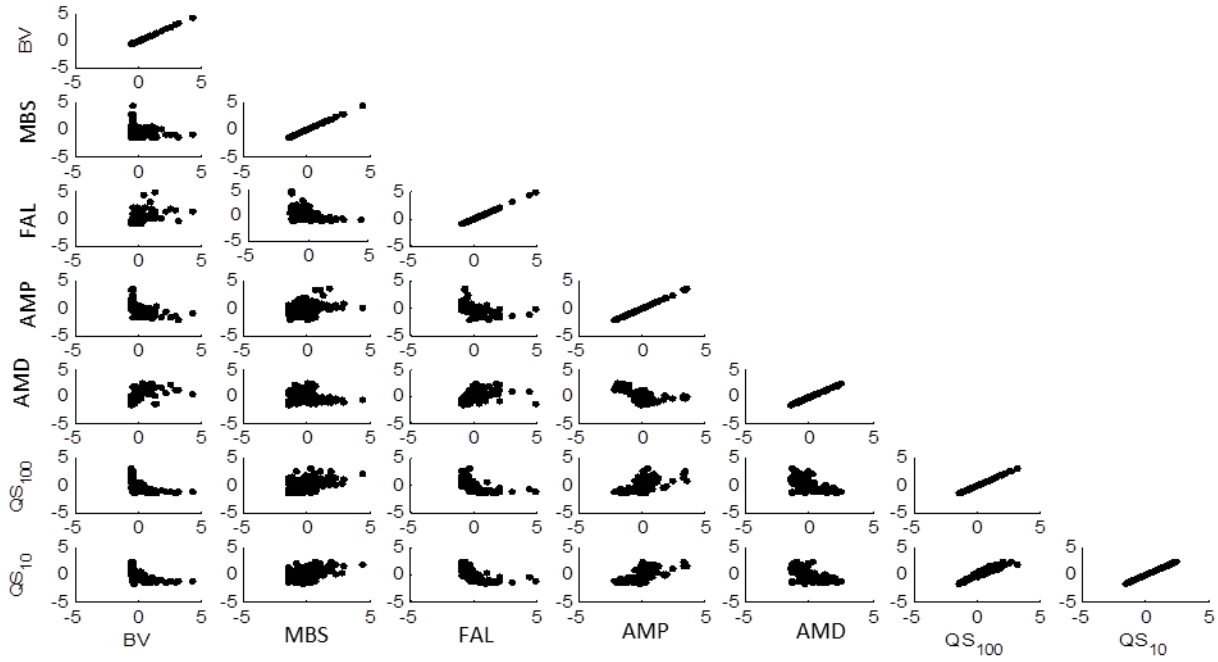


Figure 4. Diagramme de dispersion des caractéristiques physiographiques des bassins versants et des quantiles de crue (Québec).

Il importe aussi de noter que, malgré l'existence d'une corrélation linéaire négative relativement forte entre les quantiles et le pourcentage des lacs (FAL), d'une part, et positive entre les quantiles et les précipitations moyennes annuelles (AMP), d'autre part (Tableau 1), on s'aperçoit d'après les nuages de points associés que ces structures sont plutôt non linéaires.

Tableau 1. Corrélations entre les variables météo-physiographiques et hydrologiques (Québec)

	QS_{100}	QS_{10}
BV	-0.53	-0.55
MBS	0.44	0.45
FAL	-0.61	-0.65
AMP	0.58	0.65
AMD	-0.56	-0.57

Ces constatations justifient notre recours à des méthodes non linéaires pour reproduire les structures de corrélations complexes entre les variables. Le modèle adopté dans Ouali et al. [2015] a été considéré pour comparaison dans l'étude subséquente. En effet, dans le cadre d'une étude comparative menée dans Ouali et al. [2016a], différentes combinaisons linéaires, semi linéaires et purement non linéaires, présentées dans le Tableau 2, sont entreprises.

Notons que les choix effectués pour chacun des modèles proposés (choix des variables explicatives, des paramètres des modèles, …) font en sorte que les résultats sont comparables avec ceux des études antérieures.

Tableau 2. Modèles régionaux semi-linéaires et non linéaires adoptés

Étape / Modèle régional	**DRH**	**ER**
Modèles purement non linéaires		
ACCNL-RNA	ACCNL	RNA
ACCNL-RNE	ACCNL	RNE
ACCNL-GAM	ACCNL	GAM
Modèles semi-linéaires		
ACC-RNA	ACC	RNA
ACC-RNE	ACC	RNE
ACC-GAM	ACC	GAM
ACCNL-RL [Ouali et al., 2015]	ACCNL	RLM

L'évaluation des qualités prédictives des modèles régionaux est un point fondamental pour juger de l'adéquation des techniques utilisées. En effet, la meilleure approche régionale est associée à une erreur de prédiction minimale. Pour évaluer la performance de chaque approche, une procédure de validation croisée est adoptée. Cette procédure consiste à calibrer le modèle considéré en utilisant N-1 sites jaugés en enlevant le i^e site (supposé non jaugé) et de valider le modèle sur ce i^e site. Cette opération est répétée pour chaque site, donc N fois. L'erreur de prédiction est ensuite estimée en calculant des critères de performance tels que le RMSE ou le BIAS.

Les résultats obtenus de la procédure de validation croisée (Tableau 3) montrent que, en termes de NASH et RRMSE, les meilleures performances ont été obtenues par la combinaison non linéaire ACCNL-GAM. En effet, comparée à toutes les autres combinaisons, l'ACCNL-GAM fournit les

estimations les plus précises avec des valeurs de NASH les plus élevées (supérieures à 0.8) et des valeurs de RRMSE les plus faibles (28.35% pour Q_{S100}). En termes de RBIAS, les résultats montrent que, malgré que tous les modèles sous-estiment les quantiles de crues, l'ACC-GAM est le modèle le moins biaisé (-3.7% pour Q_{S100}). Cependant, en comparant ces valeurs avec celles de l'ACCNL-GAM on se retrouve avec une différence non significative (une différence de – 1.3% pour Q_{S100}).

Tableau 3. Résultats de la validation croisée des estimations des quantiles par les différents modèles adoptés.

Modèle	Variables Hydrologiques	NASH	RRMSE (%)	RBIAS (%)
CCA-LR	QS10	0.77	45.15	-6.28
	QS50	0.70	49.50	-5.81
	QS100	0.66	51.50	-5.81
CCA-RNA	QS10	0.70	45.15	-7.11
	QS50	0.66	48.45	-6.44
	QS100	0.62	50.19	-6.22
CCA-RNE	QS10	0.79	41.77	-6.49
	QS50	0.72	46.44	-6.16
	QS100	0.69	47.89	-5.94
CCA-GAM	QS10	0.80	34.3	**-3.3**
	QS50	0.74	37.9	**-3.6**
	QS100	0.70	40.3	**-3.7**
NLCCA-LR	QS10	0.79	33.9	-6.0
	QS50	0.74	39.0	-7.0
	QS100	0.71	41.4	-7.7
NLCCA-RNA	QS10	0.67	40.34	-8.59
	QS50	0.67	43.75	-9.54
	QS100	0.63	46.31	-10.07
NLCCA-RNE	QS10	0.76	35.80	-6.47
	QS50	0.71	40.44	-7.26
	QS100	0.70	41.36	-7.36
NLCCA-GAM	QS10	**0.87**	**23.47**	-4.41
	QS50	**0.84**	**26.76**	-4.72
	QS100	**0.82**	**28.35**	-5.03

Les meilleurs résultats sont présentés en caractère gras.

Comparé aux modèles basés sur les RNA, l'ACCNL-GAM se trouve plus avantageux non seulement en termes de critères mais également en termes d'interprétation des résultats, puisqu'il permet de séparer individuellement le rôle de chaque variable météo-physiographique à travers la visualisation des fonctions explicites.

D'autre part, les résultats mettent en évidence le potentiel important que révèle l'introduction d'une approche non linéaire dans la DRH en conduisant à des résultats meilleurs que les cas linéaires, selon les critères d'évaluation adoptés. Ceci est valide même en comparant nos résultats avec ceux des études antérieures. La figure 5 illustre à titre indicatif les régions homogènes d'un site cible (ID: 040830) en utilisant l'ACC et l'ACCNL. D'après cette figure, on remarque une réduction du nombre de sites inclus dans la région homogène du site cible en adoptant l'approche non linéaire.

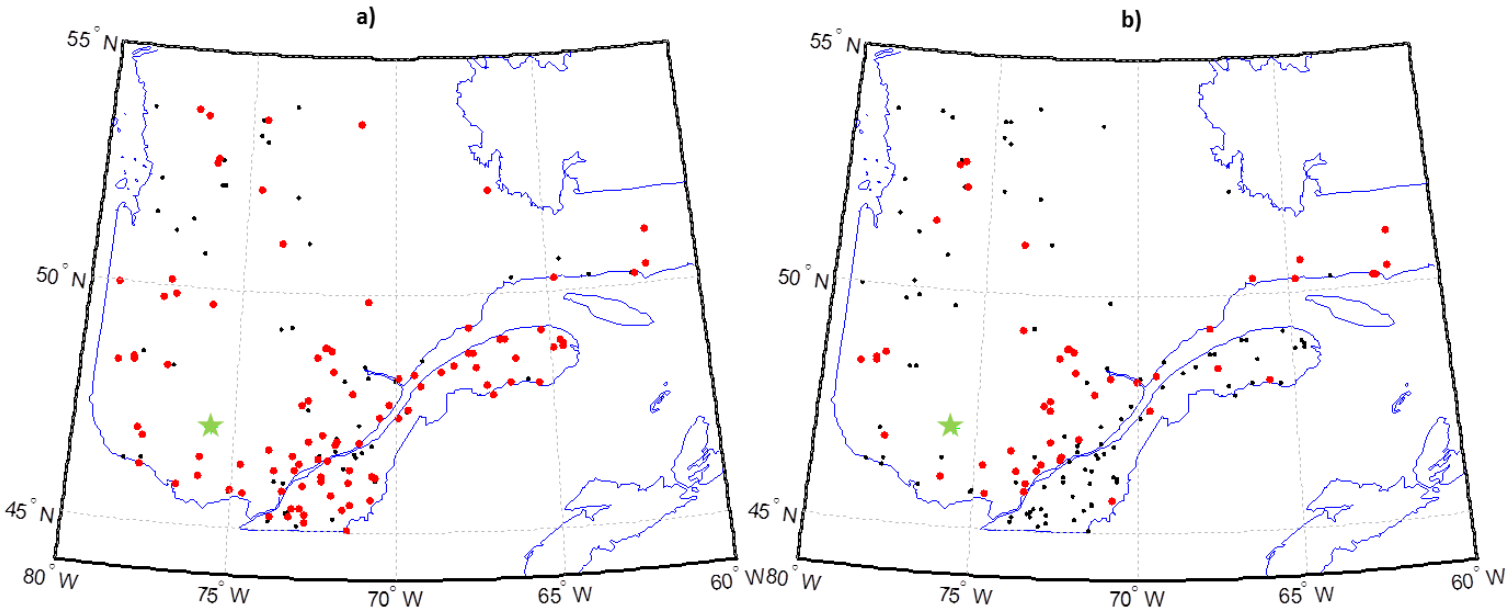


Figure 5. Résultats de la DRH en utilisant l'ACC (a) et l'ACCNL (b) pour la station Gatineau (ID: 040830), Québec. Le site cible est présenté par une étoile verte, les stations de tout le réseau hydrographique sont présentées en points noirs et les stations formant la région homogène du site cible sont présentées en points rouges.

En ce qui concerne l'importance de considérer la non-linéarité dans l'une ou l'autre des deux étapes de l'AFR (approches semi-linéaires), il a été constaté que les deux efforts aboutissent à des résultats comparables mais inférieurs à ceux des combinaisons purement non linéaires. En effet, l'amélioration de la performance globale du modèle nécessite l'intégration des techniques non linéaires dans les deux étapes.

En somme, les résultats obtenus exhibent le rôle de la considération d'une composante non linéaire aussi bien au niveau de la première étape (DRH) qu'au niveau de la deuxième étape (ER) du processus d'AFR. En fait, les techniques non linéaires permettent de refléter adéquatement les vraies relations qui existent entre les groupes de variables d'intérêts.

4.2.2. Résultats de l’approche régionale par RQ

Cette approche est appliquée sur la base de données du Québec. Ce jeu de données est composé de 151 stations hydrométriques ayant des séries allant de 15 jusqu’à 84 années. Notons également que, pour des raisons de simplification, uniquement l’étape de l’ER a été considérée. Les modèles régionaux considérés sont un modèle de RLM et un modèle de RQ direct non agrégé.

Pour des fins de comparaison, une procédure de validation croisée a été mise au point en se basant sur les critères d’évaluation classiques.

Les résultats des applications des deux modèles régionaux (RLM et RQ) sur l'ensemble des 151 sites sont présentés dans la Figure 6. Cette dernière illustre les nuages de points des quantiles régionaux en fonction des quantiles estimés localement à l’aide des lois de probabilité identifiées dans chaque site. Il est important de rappeler que ces derniers sont considérés comme des valeurs de références. Subséquemment, plus l'estimation régionale est précise plus elle se rapproche de l'estimation locale et non des vraies données observées.

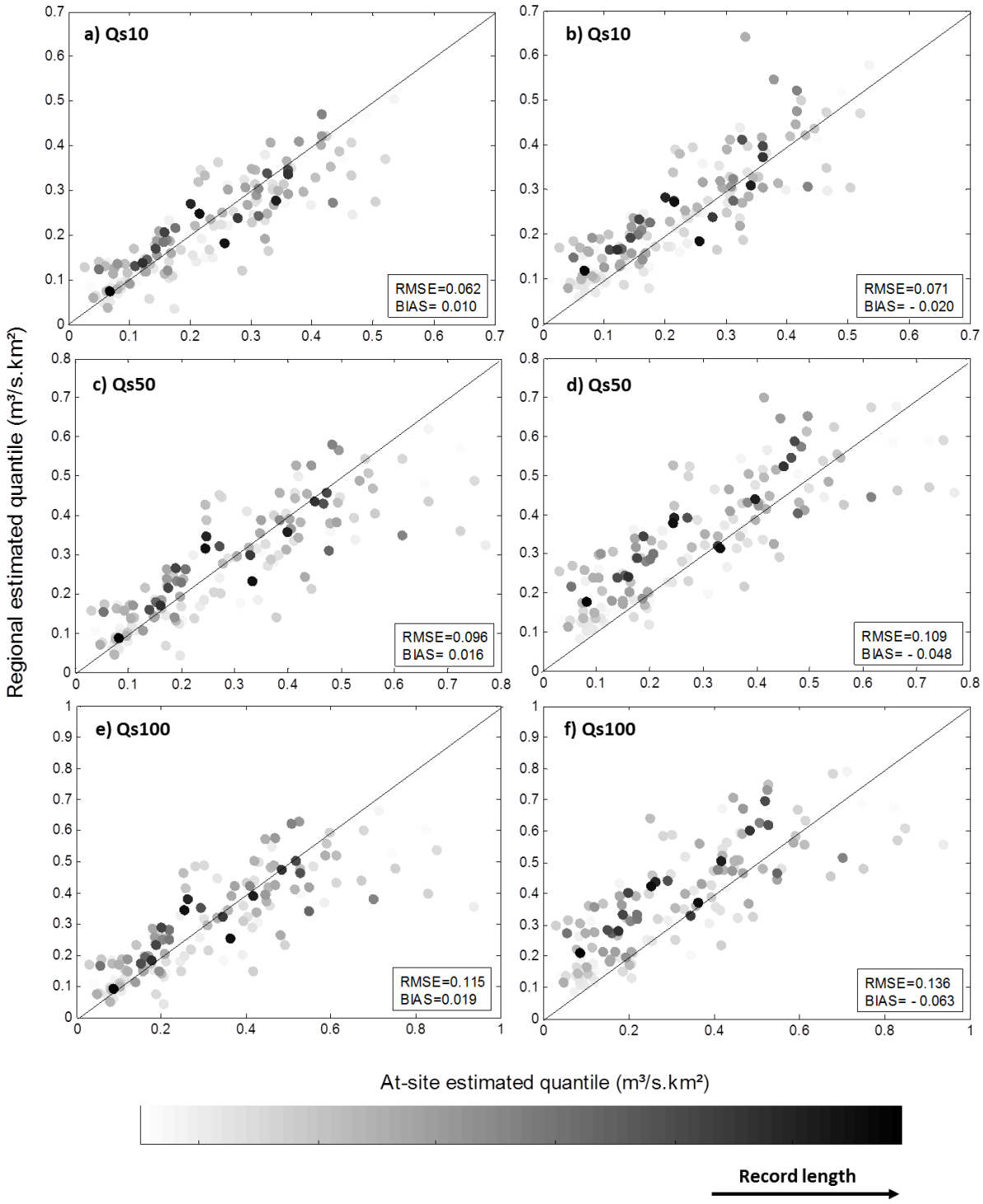


Figure 6. Diagrammes de dispersion des quantiles régionaux en fonction des quantiles estimés localement en utilisant le modèle RLM (première colonne) et le modèle RQ (deuxième colonne) pour les quantiles Q_{S10}, Q_{S50} et Q_{S100}. Les deux modèles sont calibrés et évalués en utilisant tous les sites. Les points foncés désignent les sites avec de longues séries de données.

Une comparaison des résultats obtenus révèle que la RLM reproduit plus adéquatement les quantiles de crue estimés localement que la RQ, qui tend généralement à les surestimer. En termes de BIAS et RMSE, on constate que les estimations produites par la RLM sont moins biaisées (BIAS plus faible) et plus précises (RMSE plus faible) que celles associées à la RQ. Ces constatations sont attendues et peuvent être expliquées par le concept même de ces critères qui sont, par construction, basés sur les quantiles estimés localement. Ainsi, la prise en compte de ces critères d'évaluation est en faveur de la RLM vu que tous les deux (le modèle et les critères) sont basés sur les valeurs estimées localement des quantiles de crues.

Concrètement, plus les séries de données sont longues plus l'estimation locale des quantiles de crue peut être considérée comme fiable. La longueur de chaque série a été également reportée sur la Figure 6 de manière à ce que les sites avec de courtes séries de données soient plus clairs que les sites avec de longues séries de données. Néanmoins, on constate que le RMSE et le BIAS sont insensibles aux longueurs d'enregistrements. Autrement dit, aussi bien les courtes que les longues séries sont traitées d'une façon similaire. À cet égard, il serait avantageux d'adopter un critère qui serait sensible à la qualité d'information introduite. Dans cette direction, le critère MPLF (défini à l'équation (1.12)) a été proposé pour évaluer les modèles considérés dans cette partie.

Partant du fait que la qualité de l'estimation locale pourrait être considérablement affectée par la longueur des séries des données, nous avons procédé à la calibration des deux modèles en faisant varier à chaque fois la longueur minimale des séries de données dans tous les sites. Comme le montrent les Figures 6a et 6b, qui représentent le RMSE de la RLM et de la RQ pour différentes données de calibration, l'approche RQ affiche une moins bonne performance comparativement à l'approche par RLM pour les différentes situations.

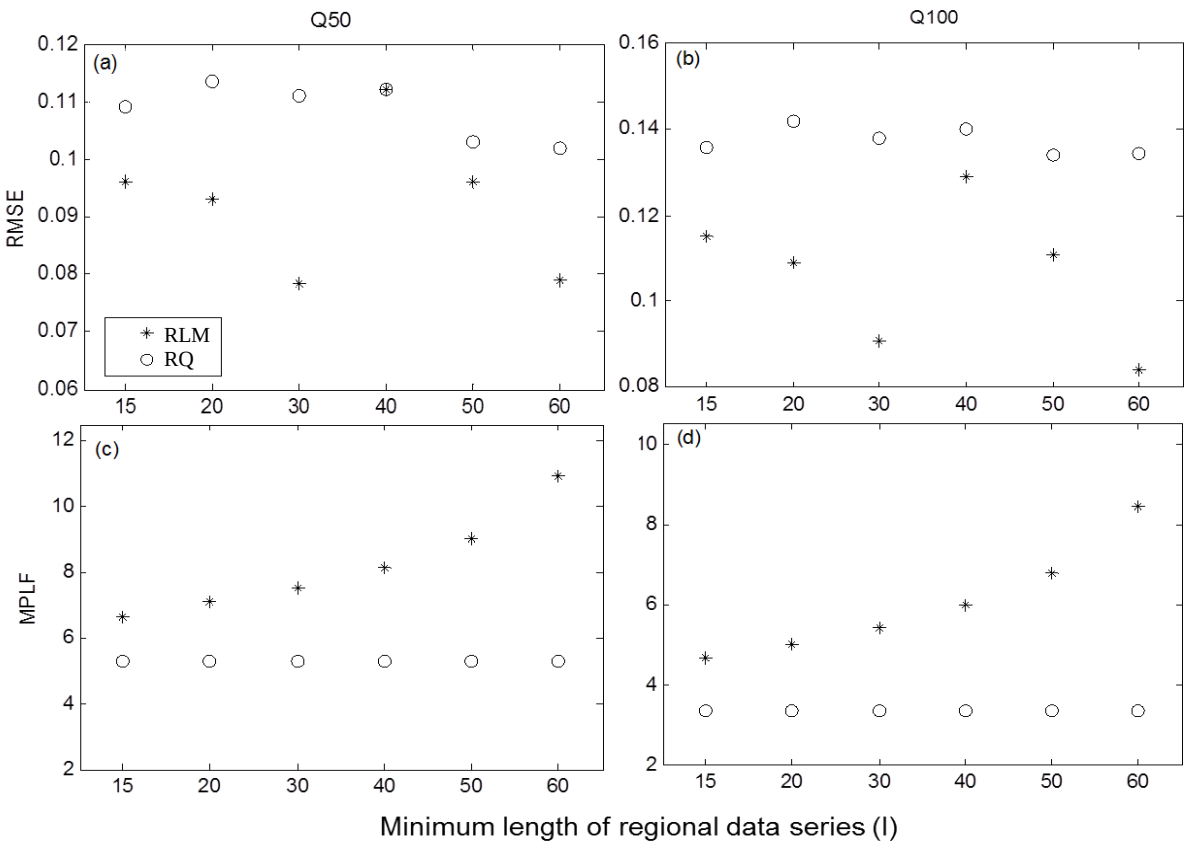


Figure 7. RMSE des estimations régionales de Q_{S50} (a) et Q_{S100} (b) ainsi que la MPLF des estimations régionales de Q_{S50} (c) et Q_{S100} (d) en fonction de la longueur des séries de données. Les deux modèles sont calibrés en utilisant des sites avec une longueur d'enregistrement dépassant *l* années, à l'exception de (c) et (d) où le modèle RQ a été calibré en utilisant toutes les données; la validation de la RQ et de la RLM se fait en utilisant tous les sites.

Les résultats des deux modèles en termes de MPLF sont présentés aux Figures 6c et 6d pour lesquelles différents cas sont considérés en fonction de la longueur minimale des séries de données utilisées pour la calibration du modèle RLM. On constate que la RQ montre une meilleure performance que le modèle classique (avec MPLF plus faible). Notons également que, pour des longueurs plus élevées (impliquant moins de sites considérés pour l'étape de la calibration), la performance de la RLM diminue. D'autre part, étant donné que la RQ et le critère MPLF ne

dépendent pas de l'estimation locale des quantiles, les valeurs du MPLF associées à l'approche RQ sont toujours constantes.

4.2.3. Résultats de l'AF locale non linéaire non stationnaire

Dans cette dernière partie de la thèse, on s'intéresse à explorer le potentiel de la RQ, particulièrement sous sa version non linéaire, pour l'estimation locale des extrêmes conditionnellement aux valeurs de co-variables. Ceci est réalisé en utilisant des données synthétiques simulées par la méthode de Monte Carlo, suivie par des études de cas réelles. Les résultats associés sont inclus dans le chapitre 5 de ce manuscrit de thèse [Ouali et al., 2016c].

Rappelons à ce stade que les approches classiques font face à quelques inconvénients particulièrement liés à la procédure d'estimation, à l'aspect non linéaire et à la relation indirecte entre les variables explicatives et la variable réponse. Afin de réduire l'impact de ces lacunes, une version non linéaire de la RQ est introduite dans ce travail. Il importe de mentionner qu'une approche linéaire par RQ (RQL) a été établie dans le cadre local par Sankarasubramanian et Lall [2003], dans lequel des échantillons aléatoires ont été générés à partir d'une loi Log-normale afin de comparer la performance de la RQL et d'une approche semi paramétrique.

En bref, l'approche par simulation adoptée dans cette étude consiste à générer des données à partir d'une distribution mère connue, et dont on connait les valeurs exactes des quantiles d'intérêt. Ensuite, pour évaluer la performance d'un modèle d'AF, on calcule les critères RMSE relative (RRMSE) et BIAS relative (RBIAS).

Différents modèles sont considérés dans cette étape soient les approches par RQ (RQL et RQMA) et les modèles non stationnaires classiques (GEV_{00}, GEV_{10}, GEV_{01}, GEV_{11}, GEV_{20}, et GEV_{21}). Cette étude par simulation vise à comparer les estimations fournies par les modèles considérés en calculant les quantiles de probabilité au non-dépassement 50, 90 et 99% de 1000 échantillons

aléatoires de taille 50 générés à partir d'une distribution mère. Dans le présent travail, cette dernière est un modèle non stationnaire dont uniquement le paramètre de position dépend d'une co-variable. Deux situations sont ainsi envisagées : i) une structure de dépendance linéaire entre le paramètre de position et une co-variable, $Y_t \sim GEV_{10}(\mu_t, \alpha, k)$ et ii) une structure de dépendance quadratique entre le paramètre de position et une co-variable, $Y_t \sim GEV_{20}(\mu_t, \alpha, k)$. Les paramètres de ces modèles ont été fixés comme suit : $\mu_t = 0.1\ t + 5; \alpha = 1; k = 0.1$ et $\mu_t = 0.1\ t^2 + 5; \alpha = 1; k = 0.1$ respectivement, où *t* désigne le temps. Les quantiles sont estimés conditionnellement à des valeurs particulières de la co-variable *t* soit la valeur médiane.

Les résultats de la comparaison entre les modèles GEV non stationnaires avec les modèles de RQ, pour les deux cas considérés, sont présentés au Tableau 4. On constate que globalement le RQMA performe bien pour les différents ordres de quantile p = 0.5, 0.9 et 0.99. Particulièrement, dans le cas où les échantillons sont générés à partir d'un modèle GEV_{10}, le RRMSE du modèle RQMA est le plus faible en comparaison avec celui de tous les autres modèles, pour les différentes probabilités de dépassement. Ceci n'est pas le cas en termes de RBIAS, où on note que le modèle RQMA est biaisé et que les modèles GEV_{10} et RQL sont les moins biaisés. Dans le cas où les échantillons sont générés à partir d'un modèle GEV_{20}, on constate que le modèle RQMA est le plus adéquat en termes de RRMSE et RBIAS. Ceci peut être dû à la structure non linéaire imposée entre les données.

Tableau 4. Résultats des simulations Monte-Carlo : RBIAS et RRMSE des quantiles estimés, conditionnellement à la co-variable, par le modèle GEV_{10} et GEV_{20}

Distribution mère	GEV_{10}		GEV_{20}	
	RRMSE (%)	RBIAS (%)	RRMSE (%)	RBIAS (%)
	p=0.50			
Gev_{00}	3.67	-2.42	3.50	-0.45
Gev_{01}	9.53	-0.45	3.58	-0.48
Gev_{10}	2.37	**-0.02**	8.59	-1.54
Gev_{11}	2.40	-0.07	8.59	-1.54
Gev_{20}	3.03	0.28	4.48	-0.53
Gev_{21}	3.32	0.40	4.49	-0.52
RQL	2.65	-0.35	4.07	-0.67
RQMA	**2.30**	-0.42	**3.13**	**-0.20**
	p=0.90			
Gev_{00}	10.14	-9.00	8.03	-4.62
Gev_{01}	24.73	-9.72	8.51	-3.82
Gev_{10}	4.48	0.46	10.93	-3.88
Gev_{11}	4.49	0.17	10.90	-3.85
Gev_{20}	4.72	0.82	9.36	-3.07
Gev_{21}	4.80	1.09	9.38	-3.91
RQL	5.70	**0.16**	9.85	-3.97
RQMA	**4.36**	1.32	**5.78**	**-0.36**
	p=0.99			
Gev0	12.97	-2.09	31.75	-18.16
Gev01	86.47	-18.97	37.40	-21.74
Gev10	14.40	**-0.76**	41.44	-24.79
Gev11	14.83	-1.83	36.46	-21.75
Gev20	14.87	-1.26	31.76	-20.72
Gev21	15.54	-1.40	35.68	-21.40
RQL	13.58	6.65	28.78	**-6.86**
RQMA	**12.60**	6.13	**23.35**	-14.31

Sur le plan pratique, trois applications ont été considérées dans le cadre de cette étude visant à analyser l'occurrence des évènements hydrologiques extrêmes conditionnellement à des indices (ou des variables) climatiques. En effet, ces derniers mesurent l'ampleur des phénomènes climatiques à grande échelle et peuvent expliquer une partie de la variabilité des débits.

Les deux premières zones d'études sont situées sur les côtes Ouest des États-Unis, soient : la station Arroyo Seco en Californie et la station de Bear Creek, Medford en Oregon. La troisième station est située sur la côte Est du Canada soit la Station Dartmouth en Gaspésie, Québec.

Vu l'emplacement géographique des sites en question, l'indice SOI a été choisi comme co-variable pour les deux premières applications. La Figure 7 présente les résultats de l'application des différents modèles (GEV_{00}, GEV_{01}, GEV_{20}, RQL et RQMA) sur le premier cas d'étude conditionnellement à différentes valeurs de SOI. D'après cette figure, on constate que pour la probabilité de non-dépassement 0.90 et pour des valeurs négatives de SOI, correspondantes aux valeurs extrêmes de DMA observés, le modèle RQMA conduit aux valeurs des quantiles les plus élevées. Notons également que, pour cette même fourchette qui correspond à une période d'El Nino, on constate un léger écart entre les quantiles conditionnels obtenus à partir des deux modèles GEV_{20} et RQL. Durant La Nina, c.-à-d. des valeurs positives de SOI, le modèle RQMA fournit les plus faibles estimations. Pour les plus grands quantiles correspondants aux probabilités de non-dépassement de 0.99, on constate que les modèles classiques donnent des valeurs qui sont nettement plus grandes que celles du modèle RQMA. Il importe également de noter que le modèle GEV_{20} n'arrive pas à bien exhiber l'effet du SOI sur la variation du débit. Par ailleurs, l'écart entre les quantiles conditionnels obtenus à partir de la RQL et de la RQMA est presque négligeable.

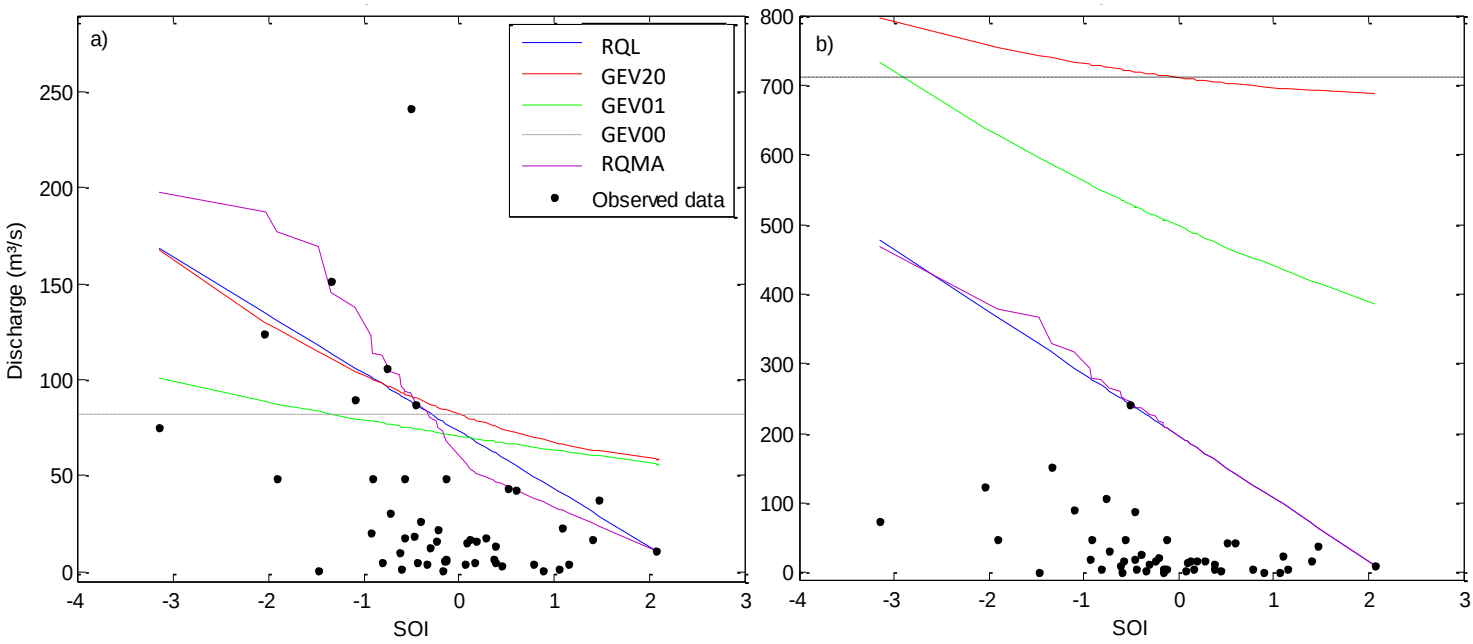


Figure 8. Estimations des quantiles associées aux probabilités de non-dépassement 0.90 (a) et 0.99 (b) par les modèles RQMA, RQL, GEV_{20}, GEV_{01} et GEV_{00}, conditionnellement aux valeurs du SOI, Arroyo Seco.

Pour évaluer la capacité des modèles utilisés à prédire les quantiles, la deuxième application a été entreprise avec une plus longue série de données, soit celle de la Station de Bear Creek, qui compte 98 années de mesures. L'idée consiste à diviser la période de mesures en une période de calibration (ou d'apprentissage) et une période de validation. Cette méthode, fréquemment utilisée en hydrologie [e.g. Oudin et al., 2005], est connue sous le nom "simple split-sample test". Environ 66% de l'ensemble des données observées est utilisé pour la calibration et le reste pour la validation. Les résultats de l'application des modèles considérés pour la probabilité au non-dépassement 0.90 sont présentés à la Figure 8. Les résultats durant la période de calibration montrent des comportements similaires des modèles GEV_{01} et du modèle stationnaire GEV_{00}. Par ailleurs, malgré qu'il prouve un comportement différent de celui du modèle stationnaire, on

remarque que le modèle GEV_{20} ne permet pas une nette distinction entre une période d'El Nino et celle de la Nina. D'autre part, durant la période de la validation, particulièrement durant La Nina, les modèles GEV_{01}, RQL, GEV_{20} et le modèle stationnaire GEV_{00} prouvent des comportements quasi-similaires.

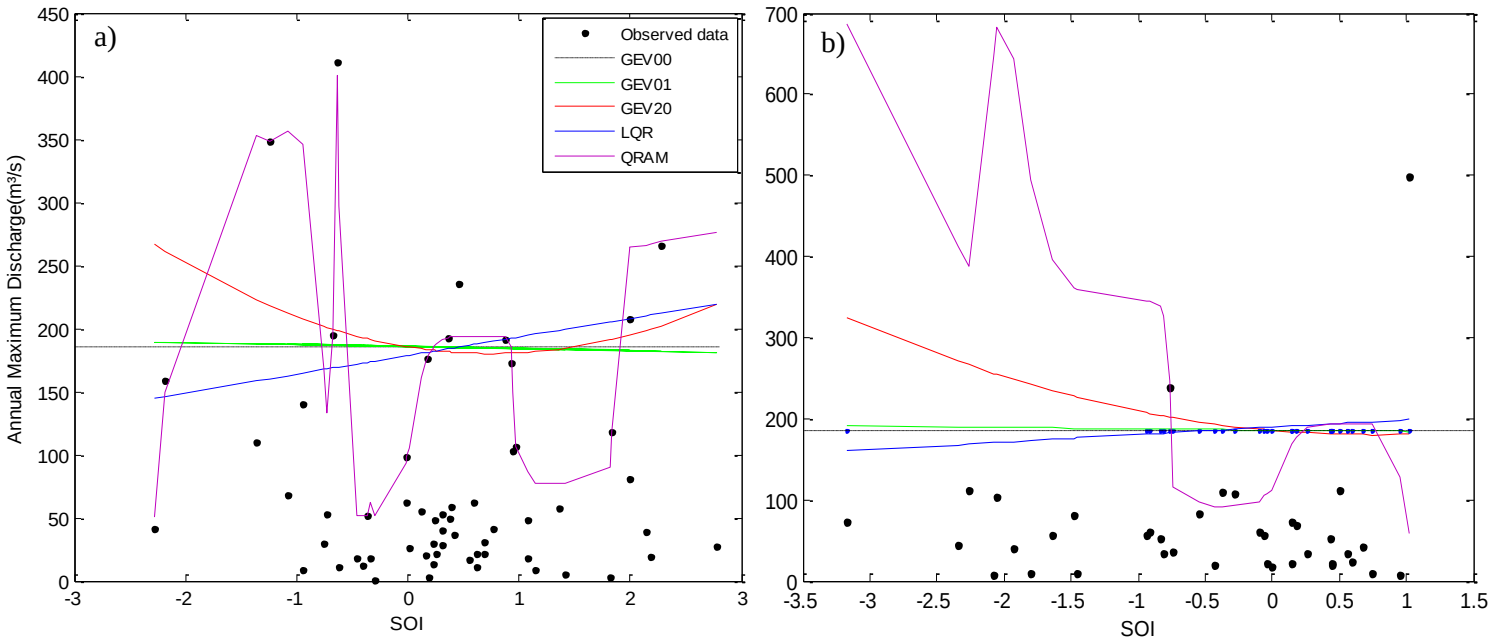
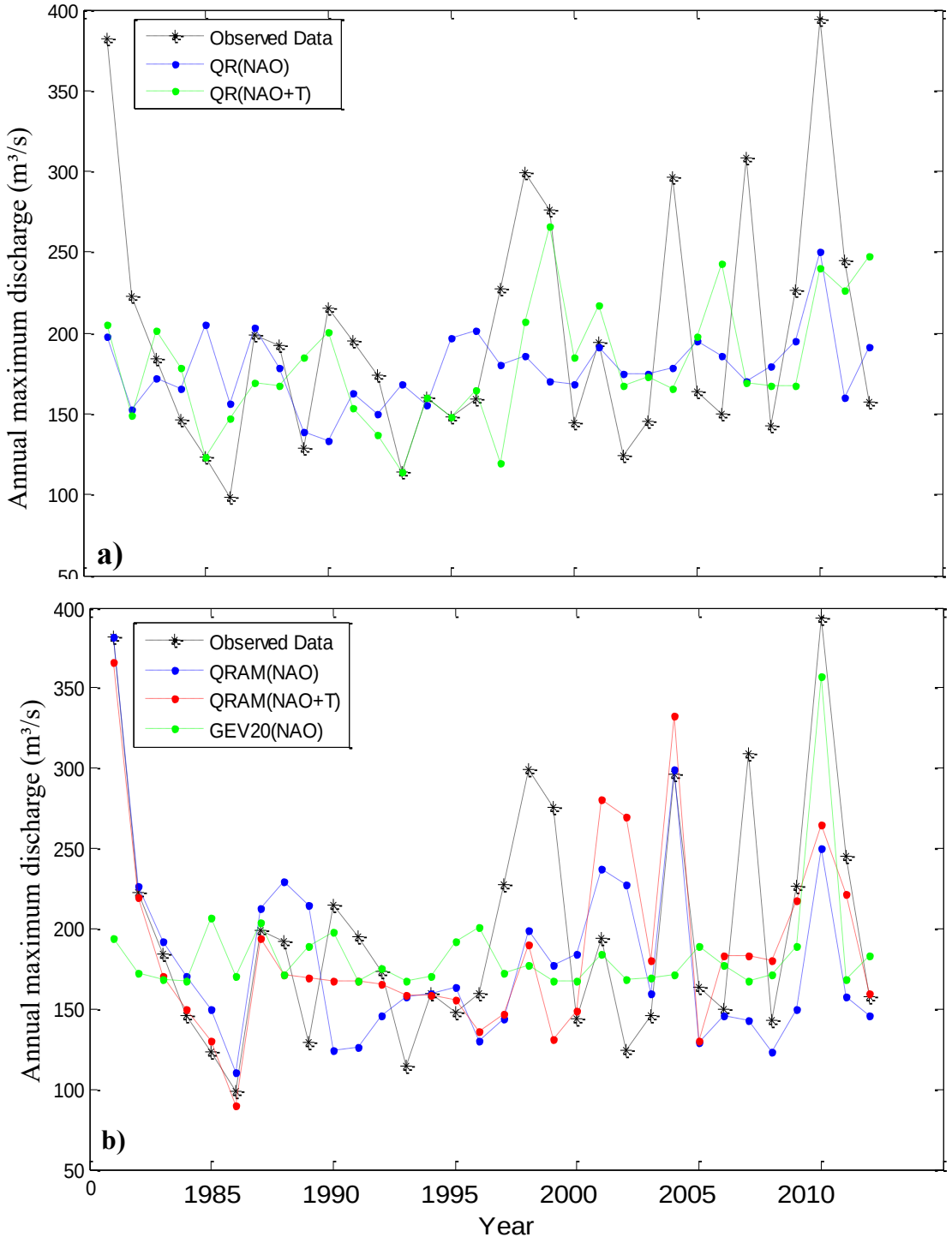


Figure 9. Estimations des quantiles associées aux probabilités de non-dépassement 0.90 par les modèles RQMA, RQL, GEV_{20}, GEV_{01} et GEV_{00}, conditionnellement aux valeurs du SOI, durant la période de calibration (a) et de validation (b), Bear Creek

La troisième application a été réalisée dans le but de mieux représenter et expliquer la variabilité de la variable réponse en intégrant plus d'information. Ceci est effectué en considérant, outre l'indice climatique NAO comme co-variable, la TMA comme variable explicative. L'hydrogramme de crues observé est illustré à la Figure 9 superposé aux formes médianes de l'hydrogramme estimées par les modèles RQL, GEV_{20} et RQMA avec et sans la variable explicative TMA. D'après cette figure, on constate que l'introduction d'une variable explicative supplémentaire améliore considérablement la reproduction de la forme générale de

l'hydrogramme, notamment en utilisant le modèle additif RQMA, ce qui n'est pas le cas dans les modèles traditionnels.



Figufre 10. Hydrogramme de crues de la Station Dartmouth, Gaspésie, superposé aux formes médianes (quantile 0.50) résultantes des modèles RQL (a), RQMA et GEV_{20} (b)

5. Conclusions et perspectives de la recherche

Les travaux réalisés au cours de la présente thèse ont permis d'aboutir aux conclusions générales suivantes mais également à soulever d'autres questions de recherche.

5.2. Conclusions générales

Cette étude a permis de développer et d'évaluer de nouvelles approches de modélisation de l'occurrence des évènements hydrologiques extrêmes pour l'estimation des quantiles dans des bassins versants jaugés et non jaugés. Dans le cas où on s'intéresse à estimer les quantiles dans des bassins versants non jaugés, la procédure utilisée, connue comme une analyse fréquentielle régionale (AFR), comporte deux principales étapes, la délimitation des régions homogènes (DRH) et l'estimation régionale (ER). Le principal objectif des méthodologies proposées dans cette thèse étant de remédier aux inconvénients des méthodes classiques d'AFR pour améliorer la qualité de l'estimation du risque lors de la construction des grands ouvrages hydrauliques, par exemple. Particulièrement, les approches proposées visent à traiter des aspects non courants qui touchent de près à la modélisation des crues en analyse fréquentielle (AF) locale et régionale à savoir la non-linéarité du processus hydrologique, l'exploitation des données disponibles, l'intégration de l'information météorologique et la non-stationnarité des crues. À la Figure 11, un sommaire des approches proposées correspondant aux diverses problématiques abordées, leurs relations avec les approches classiques ainsi que les techniques utilisées sont présentés.

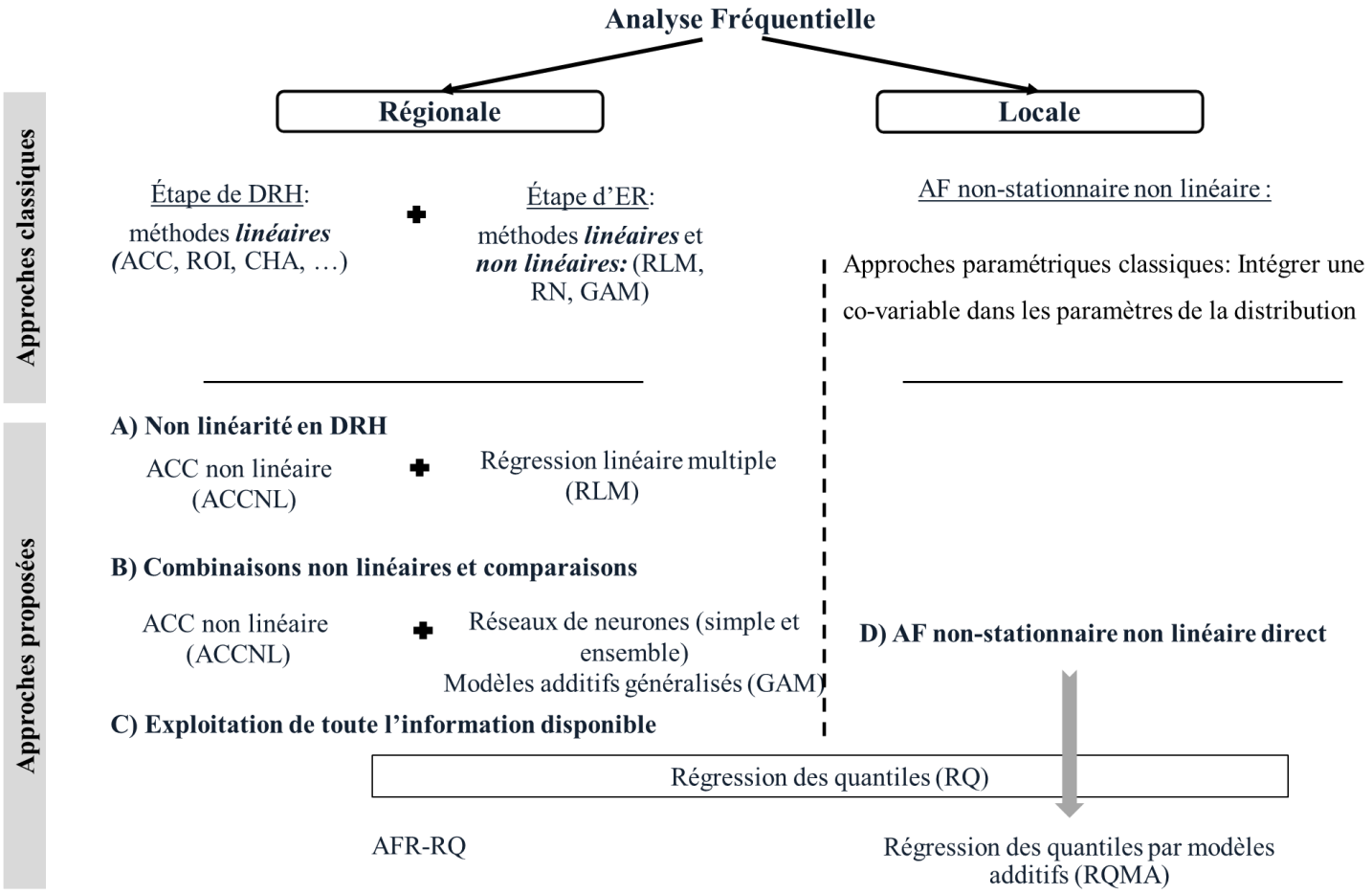


Figure 11. Approches classiques déjà existantes en AFR ainsi que les approches proposées dans cette recherche

Dans un premier temps, nous avons proposé une nouvelle méthode de DRH, l'ACC-NL, capable de reproduire l'aspect non linéaire du processus hydrologique. Différents cas d'études en Amérique du Nord ont permis de montrer que cette technique est prometteuse pour détecter les éventuelles relations non linéaires entre les variables hydrologiques d'une part et les variables météo-physiographiques d'autre part. Cette approche nous a permis également de concevoir d'autres modèles d'AFR en la combinant avec différentes méthodes d'ER.

Ainsi, des combinaisons linéaires, semi-linéaires et non linéaires relatives aux deux étapes de la procédure de régionalisation ont été établies. Une étude comparative a été réalisée dans le but d'identifier la meilleure combinaison (DRH-ER). Différents modèles ont été appliqués sur trois régions en Amérique du Nord. Les résultats obtenus montrent qu'il est avantageux de procéder avec des outils non linéaires dans les deux étapes de l'AFR.

Partant du constat que les modèles d'AFR classiques utilisent les quantiles estimés localement, et que ces derniers utilisent l'information disponible au sein d'une région d'une manière inadéquate, un modèle d'AFR basé sur la RQ a été proposé visant à bien exploiter toute l'information disponible. Dans cette même direction, un critère d'évaluation des estimations régionales a été proposé en utilisant les données observées plutôt que les quantiles estimés. Appliquée sur la base de données du Sud du Québec, l'approche s'est montrée prometteuse.

Malgré l'expansion des approches d'AF non stationnaires, il subsiste encore des déficiences plus ou moins sérieuses telles que, entre autres, la lourdeur de la procédure de modélisation, ainsi que l'absence d'un lien direct entre la réponse et les co-variables. Pour pallier à ces lacunes, un modèle d'AF locale non stationnaire et non linéaire, RQMA, a été proposé et appliqué sur des cas d'études réels. Les résultats montrent que le RQMA est plus approprié en particulier quand l'information météorologique y est incluse. Ainsi, cette direction semble être prometteuse non seulement en termes de performance mais surtout en termes de simplicité et d'interprétation.

5.3. Perspectives de recherche

Les principales perspectives de recherche qui peuvent être envisagées à l'issue de ces études de thèse sont :

1. Un modèle d'AFR non stationnaire par RQ : Dans la même direction que celle du modèle d'AF locale développé dans le dernier chapitre [Ouali et al., 2016c], un modèle régional

non stationnaire mérite une investigation. Partant des mêmes motivations qui expliquaient le développement des modèles AFR-QR et RQMA, particulièrement le fait de considérer les quantiles estimés localement comme des valeurs réelles, il serait sans doute intéressant de concevoir un modèle régional non stationnaire qui traite rigoureusement ces problématiques ;

2. Introduire l'étape de la DRH dans le modèle d'AFR-RQ : Le modèle régional basé sur la RQ, présenté dans le chapitre 3 [Ouali et al., 2016b], s'est limité uniquement à l'étape de l'ER. Afin d'exploiter l'approche a son plein potentiel, une direction future consiste à combiner la RQ comme outil d'ER avec une approche de DRH notamment l'ACC-NL. Une comparaison avec le modèle du point précédent est à envisager;
3. Un modèle d'AF locale bivariée non stationnaire en utilisant la RQ : Une extension de l'AF locale non stationnaire dans le cadre bivarié a été récemment introduite dans la littérature hydrométéorologique à l'aide des copules. Dans un premier temps, il serait intéressant de développer une approche qui combine les avantages de la RQ et des copules pour exploiter leurs potentiels à estimer des quantiles hydrologiques dans des conditions climatiques en évolution.

6. Références bibliographiques

Aissaoui-Fqayeh, I., S. El-Adlouni, T. B. M. J. Ouarda et A. St-Hilaire (2009). "Non-stationary lognormal model development anf comparison with non-stationary GEV model." Hydrological sciences journal **54**(6): 1141-1156.

Alobaidi, M. H., P. R. Marpu, T. B. M. J. Ouarda et F. Chebana (2015). "Regional frequency analysis at ungauged sites using a two-stage resampling generalized ensemble framework." Advances in Water Resources **84** 103-111.

Ashtiani, A., P. A. Mirzaei et F. Haghighat (2014). "Indoor thermal condition in urban heat island: Comparison of the artificial neural network and regression methods prediction." Energy and Buildings **76**: 597-604.

Austin, P. C. (2007). "A comparison of regression trees, logistic regression, generalized additive models, and multivariate adaptive regression splines for predicting AMI mortality." Statistics in medicine **26**(15): 2937-2957.

Aziz, K., A. Rahman, G. Fang et S. Shrestha (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.

Barnett, T. et R. Preisendorfer (1987). "Origins and levels of monthly and seasonal forecast skill for United States surface air temperatures determined by canonical correlation analysis." Monthly Weather Review **115**(9): 1825-1850.

Bayentin, L., S. El Adlouni, T. B. Ouarda, P. Gosselin, B. Doyon et F. Chebana (2010). "Spatial variability of climate effects on ischemic heart disease hospitalization rates for the period 1989-2006 in Quebec, Canada." International journal of health geographics **9**(1): 1.

Bekey, G. et K. Y. Goldberg (2012). Neural Networks in robotics, Springer Science & Business Media.

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

Benzer, R. et S. Benzer (2015). "Application of artificial neural network into the freshwater fish caught in Turkey." **2**(5): 341-346.

Buja, A., T. Hastie et R. Tibshirani (1989). "Linear smoothers and additive models." The Annals of Statistics: 453-510.

Burn, D. H. (1990). "Evaluation of regional flood frequency analysis with a region of influence approach." Water Resources Research **26**(10): 2257-2265.

Campbell, E. P. et B. C. Bates (2001). "Regionalization of rainfall-runoff model parameters using Markov chain Monte Carlo samples." WATER RESOURCES RESEARCH **37**(3): 731-739.

Campi, C., L. Parkkonen, R. Hari et A. Hyvärinen (2013). "Non-linear canonical correlation for joint analysis of MEG signals from two subjects." Frontiers in neuroscience **7**.

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

Castellarin, A., D. Burn et A. Brath (2001). "Assessing the effectiveness of hydrological similarity measures for flood frequency analysis." Journal of Hydrology **241**(3): 270-285.

Cavadias, G. (1990). "The canonical correlation approach to regional flood estimation." Regionalization in hydrology **191**: 171-178.

Chebana, F., C. Charron, T. B. M. J. Ouarda et B. Martel (2014). "Regional Frequency Analysis at Ungauged Sites with the Generalized Additive Model." Journal of Hydrometeorology **15**(6): 2418-2428.

Chen, P.-A., L.-C. Chang et F.-J. Chang (2013). "Reinforced recurrent neural networks for multi-step-ahead flood forecasts." Journal of Hydrology **497**: 71-79.

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

Chokmani, K., T. B. M. J. Ouarda, S. Hamilton, M. H. Ghedira et H. Gingras (2008). "Comparison of ice-affected streamflow estimates computed using artificial neural networks and multiple regression techniques." Journal of Hydrology **349**(3): 383-396.

Coad, P., B. Cathers, J. E. Ball et R. Kadluczka (2014). "Proactive management of estuarine algal blooms using an automated monitoring buoy coupled with an artificial neural network." Environmental Modelling & Software **61**: 393-409.

Dalrymple, T. (1960). "Flood frequency analysis." US Geological Survey Water Supply Paper 1543A: 11-51.

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

Dawson, C. et R. Wilby (2001). "Hydrological modelling using artificial neural networks." Progress in physical Geography **25**(1): 80-108.

El Adlouni, S., T. B. M. J. Ouarda, X. Zhang, R. Roy et B. Bobée (2007). "Generalized maximum likelihood estimators for the nonstationary generalized extreme value model." WATER RESOURCES RESEARCH **43**(3).

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

Fenske, N., J. Burns, T. Hothorn et E. A. Rehfuess (2013). "Understanding child stunting in India: a comprehensive analysis of socio-economic, nutritional and environmental determinants using additive quantile regression." PloS one **8**(11): e78692.

Fenske, N., T. Kneib et T. Hothorn (2012). "Identifying risk factors for severe childhood malnutrition by boosting additive quantile regression." Journal of the American Statistical Association.

Frie, K. G. et C. Janssen (2009). "Social inequality, lifestyles and health–a non-linear canonical correlation analysis based on the approach of Pierre Bourdieu." International journal of public health **54**(4): 213-221.

Friederichs, P. et A. Hense (2007). "Statistical downscaling of extreme precipitation events using censored quantile regression." Monthly weather review **135**(6): 2365-2378.

GREHYS (1996a). "Presentation and review of some methods for regional flood frequency analysis." Journal of Hydrology 186: 63-84.

GREHYS (1996b). "Inter-comparison of regional flood frequency procedures for Canadian rivers." Journal of Hydrology 186: 85-103.

Hamed, K. et A. R. Rao (1999). Flood frequency analysis, CRC press.

Hastie, T. et R. Tibshirani (1986). "Generalized additive models." Statistical science: 297-310.

Haykin, S. et R. Lippmann (1994). "Neural Networks, A Comprehensive Foundation." International Journal of Neural Systems **5**(4): 363-364.

He, Y., A. Bárdossy et J. Brommundt (2006). Non-stationary flood frequency analysis in southern Germany. The 7th International Conference on HydroScience and Engineering, Philadelphia.

Hsieh, W. W. (2000). "Nonlinear canonical correlation analysis by neural networks." Neural Networks **13**: 1095 -1105.

Hsieh, W. W. (2001). "Nonlinear Canonical Correlation Analysis of the Tropical Pacific Climate Variability Using a Neural Network Approach." Journal of climate **14**: 2528-2539.

Huo, Z., S. Feng, S. Kang, G. Huang, F. Wang et P. Guo (2012). "Integrated neural networks for monthly river flow estimation in arid inland basin of Northwest China." Journal of Hydrology **420**: 159-170.

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

Koenker, R. (2011). "Additive models for quantile regression: model selection and confidence bandaids." Brazilian Journal of Probability and Statistics **25**(3): 239-262.

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

Kouider, A., H. Gingras, T. Ouarda, Z. Ristic-Rudolf et B. Bobée (2002). "Analyse fréquentielle locale et régionale et cartographie des crues au Québec." Rep. R-627-el.

Latraverse, M., P. F. Rasmussen et B. Bobée (2002). "Regional estimation of flood quantiles: Parametric versus nonparametric regression models." WATER RESOURCES RESEARCH **38**(6).

López-Moreno, J. I. et D. Nogués-Bravo (2005). "A generalized additive model for the spatial distribution of snowpack in the Spanish Pyrenees." Hydrological Processes **19**(16): 3167-3176.

López, J. et F. Francés (2013). "Non-stationary flood frequency analysis in continental Spanish rivers, using climate and reservoir indices as external covariates." Hydrology and Earth System Sciences Discussions **17**(8): 3103-3142.

Marco, J. B., R. Harboe et J. D. Salas (2012). Stochastic hydrology and its use in water resources systems simulation and optimization, Springer Science & Business Media.

Martins, E. S. et J. R. Stedinger (2000). "Generalized maximum-likelihood generalized extreme-value quantile estimators for hydrologic data." Water Resources Research **36**(3): 737-744.

McCullagh, P. et J. A. Nelder (1989). Generalized linear models, CRC press.

Michael, A. G. et C. P. Raymond (2003). "Using traffic conviction correlates to identify high accident-risk drivers." Accident Analysis and Prevention **35**(6): 903-912.

Mumford, D. et J. Shah (1989). "Optimal approximations by piecewise smooth functions and associated variational problems." Communications on pure and applied mathematics **42**(5): 577-685.

Nohair, M., A. St-Hilaire et T. B. M. J. Ouarda (2008). "The Bayesian-Regularized neural network approach to model daily water temperature in a small stream." Revue des sciences de l'eau **21**(3).

Ouali, D., F. Chebana et T. B. M. J. Ouarda (2015). "Non-linear canonical correlation analysis in regional frequency analysis." Stochastic Environmental Research and Risk Assessment: 1-14.

Ouali, D., F. Chebana et T. B. M. J. Ouarda (2016a). "Fully nonlinear regional hydrological frequency analysis." Submitted.

Ouali, D., F. Chebana et T. B. M. J. Ouarda (2016b). "Quantile regression in regional frequency analysis: a better exploitation of the available information." Journal of Hydrometeorology(2016).

Ouali, D., F. Chebana et T. B. M. J. Ouarda (2016c). "Frequency analysis of hydro-meteorological extremes in a changing climate using additive quantile regression." To be submitted.

Ouarda, T. B., C. Charron et A. St-Hilaire (2008a). "Statistical models and the estimation of low flows." Canadian Water Resources Journal **33**(2): 195-206.

Ouarda, T. B. M. J. (2013). "Hydrological frequency analysis, regional." Encyclopedia of Environmetrics.

Ouarda, T. B. M. J., C. Girard, G. S. Cavadias et B. Bobée (2001). "Regional flood frequency estimation with canonical correlation analysis." Journal of Hydrology **254**(1): 157-173.

Ouarda, T. B. M. J. et C. Shu (2009). "Regional low-flow frequency analysis using single and ensemble artificial neural networks." Water Resources Research **45**(11).

Ouarda, T. B. M. J., A. St-Hilaire et B. Bobée (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., F. Hervieu, C. Michel, C. Perrin, V. Andréassian, F. Anctil et C. Loumagne (2005). "Which potential evapotranspiration input for a lumped rainfall–runoff model?: Part 2—Towards a simple and efficient potential evapotranspiration model for rainfall–runoff modelling." Journal of Hydrology **303**(1): 290-306.

Pandey, G. et V.-T.-V. Nguyen (1999). "A comparative study of regression based methods in regional flood frequency analysis." Journal of Hydrology **225**(1): 92-101.

Riad, S. et J. Mania (2004). "Rainfall-Runoff Model Using an Artificial Neural Network Approach." Mathematical and Computer Modelling **40**: 839-846.

Ribeiro-Corréa, J., G. Cavadias, B. Clement et J. Rousselle (1995). "Identification of hydrological neighborhoods using canonical correlation analysis." Journal of Hydrology **173**(1): 71-89.

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

Shu, C. et D. H. Burn (2004). "Artificial neural network ensembles and their application in pooled flood frequency analysis." Water Resources Research **40**(9).

Shu, C. et 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**(07).

Tasker, G. D., S. A. Hodge et C. S. Barks (1996). "Region of influence regression for estimating the 50-year flood at ungaged sites." JAWRA Journal of the American Water Resources Association **32**(1): 163-170.

Tasker, G. D. et M. E. Moss (1979). "Analysis of Arizona flood data network for regional information." Water Resources Research **15**(6): 1791-1796.

Tishler, A. et S. Lipovetsky (1996). "Canonical correlation analyses for three data sets: a unified framework with application to management." Computers & operations research **23**(7): 667-679.

Tsakiris, G., I. Nalbantis et G. Cavadias (2011). "Regionalization of low flows based on canonical correlation analysis." Advances in Water Resources **34**(7): 865-872.

Waldmann, E., T. Kneib, Y. R. Yue, S. Lang et C. Flexeder (2013). "Bayesian semiparametric additive quantile regression." Statistical Modelling **13**(3): 223-252.

Wang, D., L. Shi, D. S. Yeung et E. Tsang (2005). Nonlinear canonical correlation analysis of fMRI signals using HDR models. Engineering in Medicine and Biology Society, 2005. IEEE-EMBS 2005. 27th Annual International Conference of the, IEEE.

Wang, W.-c., K.-w. Chau, L. Qiu et Y.-b. Chen (2015). "Improving forecasting accuracy of medium and long-term runoff using artificial neural network based on EEMD decomposition." Environmental research **139**: 46-54.

Wood, S. (2006). Generalized additive models: an introduction with R, CRC press.

Wu, A. et W. W. Hsieh (2003). "Nonlinear interdecadal changes of the El Nino-Southern Oscillation " Climate Dynamics **21**: 719–730.

Yue, Y. R. et H. Rue (2011). "Bayesian inference for additive mixed quantile regression models." Computational Statistics & Data Analysis **55**(1): 84-96.

Zhihua, J. et Y. Zhen (2010). "On using non-linear canonical correlation analysis for voice conversion based on Gaussian mixture model." Journal of Electronics **27**(1): 1-7.

CHAPITRE 2

NON-LINEAR CANONICAL CORRELATION ANALYSIS IN REGIONAL FREQUENCY ANALYSIS

Non-linear canonical correlation analysis in regional frequency analysis

D. Ouali[1], F. Chebana[1], T. B.M.J. Ouarda[2, 1]

[1]*Institut National de la Recherche Scientifique, Centre Eau Terre et Environnement,*

490, rue de la Couronne, Québec (Québec), G1K 9A9, Canada.

[2]*Institute Centre for Water Advanced Technology and Environmental Research*

P.O. Box 54224, Abu Dhabi, UAE

[*]**Corresponding author:** Tel: +1 (418) 654 2530#4477

Email: dhouha.ouali@ete.inrs.ca

April 2015

Abstract

Hydrological processes are complex non-linear phenomena. Canonical Correlation analysis (CCA) is frequently used in regional frequency analysis (RFA) to delineate hydrological neighborhoods. Although non-linear CCA (NL-CCA) is widely used in several fields, it has not been used in hydrology, particularly in RFA. This paper presents an overview of techniques used to reproduce non-linear relationships between two sets of variables. The approaches considered in this work are based on NL-CCA using neural networks (CCA-NN), coupled to a log-linear regression model for flood quantile estimation. In order to demonstrate the usefulness of these approaches in RFA, a comparative study between the latter and linear CCA is performed using three different databases from North America. Results show that CCA-NN is more robust and can better reproduce the non-linear relationship structures between physiographical and hydrological variables. This reflects the high flexibility of this approach. Results indicate that for all three databases, it is more advantageous to proceed with the non-linear CCA approach.

Keywords: Non-linear canonical correlation analysis, neural network, regional frequency analysis, homogeneous region, hydrological neighborhood, ungauged basins.

1. Introduction and literature review

One of the main objectives of regional frequency analysis (RFA) is the estimation of extreme event quantiles (e.g. floods and droughts) at sites where little or no hydrological data is available. In general, RFA procedures have two main steps, namely the delineation of homogeneous regions (DHR) and regional estimation (RE) (e.g. Chebana and Ouarda 2007; Chebana and Ouarda 2008; Ouarda et al. 2008). For each of these two steps, a large number of methodologies have been proposed (Ouarda et al. 2008). Canonical correlation analysis (CCA) is one of the most commonly used methods for DHR where it consists in identifying linear combinations of variables within the same group, for which the canonical correlation is maximal. Ouarda et al. (2008) demonstrated the advantages of CCA by comparing its performance to other techniques such as the hierarchical cluster analysis approach**.** However, note that in Shu and Ouarda (2007), CCA was used not for the DHR step, but to form a canonical physiographic space over which an artificial neuronal network (ANN) is then employed to estimate flood quantile.

CCA is an important statistical tool for multivariate data analysis. However, it presents a drawback in the interpretation of results, which seems to be often difficult. In addition, this approach is based on a linear foundation and, hence, is not able to adequately describe non-linear relationships between variables. Therefore, CCA may not be suitable for representing hydrological processes in the DHR step. Two groups of variables are usually considered in RFA: i) hydrological variables and ii) meteorological and/or physiographical characteristics of the watersheds (Ouarda 2013). Hydrological processes are relatively complex because of the variability in the response of watersheds which does not generally result from a linear relationship between the hydrological and the physiographical characteristics (e.g. Chen et al.

2008; Xu et al. 2010; Chebana et al. 2014). Hydrological processes and their inherent non-linearities could not be adequately represented by linear relationships. One aspect of the non-linearity is represented by the rainfall-runoff relationship. Indeed, the variations of meteorological variables and flows are linked by a non-linear relationship (Riad and Mania 2004). This non-linear behavior depends strongly on the physiographic characteristics of the watersheds. For instance, surface runoff is strongly influenced by the soil storage capacity and soil infiltration.

A number of statistical tools have been proposed in the literature to deal with the additional complexity associated to non-linearity in a variety of fields (e.g. Bolton et al. 2003; Yin 2007). Among the proposed techniques, we can mention non-linear principal component analysis (NL-PCA) (Rumelhart et al. 1985; Kramer 1991) and non-linear CCA (NL-CCA) (Dauxois and Nkiet 1998; Hsieh 2000). NL-PCA has been applied in various fields such as chemistry (Kramer 1991), image processing (Botelho et al. 2005) and atmospheric sciences (e.g. Sengupta and Boyle 1995; Monahan 2000). Sengupta and Boyle (1995) applied NL-PCA to average monthly rainfall data in the United States. Compared to conventional PCA, results showed that the non-linear approach is a more effective data reduction tool. It was also demonstrated that NL-PCA represented better the variation of variables than ordinary PCA. However, this method presents some technical drawbacks (Malthouse 1998).

Although the above constraints of the NL-PCA also persist for NL-CCA (Hsieh 2000), the latter seems to provide better results than the CCA. NL-CCA was used in several fields, such as analysis of voice conversion (e.g. Zhihua and Zhen 2010), biomedicine (e.g. Campi et al. 2013), medicine (e.g. Wang et al. 2005) and sociology (e.g. Frie and Janssen 2009). A number of techniques related to NL-CCA have been proposed in the literature. For instance, Dauxois and

Nkiet (1998) introduced measures of association between two random variables based on NL-CCA. Among the most studied non-linear methods associated to CCA, we can mention the neural network approach (NN) (Hsieh 2000), genetic algorithms (GA) (Kruger et al. 2004) and Kernel based methods (Akaho 2001; Hardoon and Shawe-Taylor 2009). Recently, Nagai (2013) proposed an optimization approach based on cross validation to optimize the NL-CCA parameters. In terms of applications, the non-linear method based on NN was adopted in a number of studies in meteorology and climatology. For example, Wu and Hsieh (2002) studied the El Nino Southern oscillation using NL-CCA based on the NN approach (CCA-NN). They showed the ability of CCA-NN to detect non-linearity between surface wind stress and sea surface temperature. Hsieh (2001) also applied CCA-NN to study the relationship between sea level pressure in the tropical Pacific and sea surface temperature. Results revealed the ability of this model to characterize non-linearity between variables, which was not the case with the conventional CCA.

Other studies in the past were interested by treating non-linear aspects of categorical variables (qualitative). Gifi (1990) presented two different techniques and algorithms, mainly OVERALS and CANALS to deal with such qualitative variables. However, the treated variables in RFA are quantitative and continuous. Therefore, the latter methods are not applicable in the context of the present study. In Table 1, all non-linear approaches discussed previously are summarised including their advantages and drawbacks. Note that methods designed for quantitative variables are more flexible than those for categorical ones.

Despite strong evidence concerning the non-linearity of hydrological processes, NL-CCA approaches have not yet been considered in hydrology. In RFA, non-linear approaches can account for possible non-linearities in order to determine the most representative homogeneous

regions and lead to a better regional estimation. The purpose of the present paper is to deal with the issue of non-linearity in RFA by introducing NL-CCA in the DHR step in order to improve its performance and representativeness.

The present paper is organized as follows: In the following section, the potential of NL-CCA in the DHR step is developed. In order to verify and validate the usefulness of the NL-CCA approach for the modelling of hydrological processes, a comparative study is carried out in section 3 using three different datasets from North America (Quebec, Arkansas and Texas). These approaches are used in the delineation of hydrological neighborhoods where the obtained results are presented and discussed in section 4. The conclusions of this work are reported in Section 5.

2. Background and methodology

In this section, we present a brief description of the use of CCA in RFA, as well as a description of the NL-CCA method and its application to RFA.

2.1. Canonical Correlation Analysis in RFA

CCA is a multivariate analysis method used to identify the correlations that may exist between two groups of variables. It has been applied in a number of fields, such as seasonal climate forecasting (e.g. Barnett and Preinsendorfer 1987), management science (e.g. Tishlert and Lipovetsky 1996), forecasting of accident risk modeling (e.g. Michael and Raymond 2003), river thermal regime modeling (e.g. Guillemette et al. 2009), water quality estimation (e.g. Khalil et al. 2011) and especially flood frequency estimation (e.g. Ouarda et al. 2001).

As mentioned above, in RFA, variables of interest are mainly hydrological and physiographical variables. We denote Y the vector describing hydrological variables, and X the vector containing meteorological and/or physiographical variables. Considering linear combinations of variables $X_1, X_2, ..., X_q$ and $Y_1, Y_2, ..., Y_r$, we obtain a new canonical space composed by canonical vectors U_i and V_i such as:

$$U_i = a_{i1}X_1 + a_{i2}X_2 + ... + a_{iq}X_q \quad (1)$$

$$V_i = b_{i1}Y_1 + b_{i2}Y_2 + ...b_{ir}Y_r \quad (2)$$

where $i = 1, ..., p$ with $p = \min(r, q)$. The canonical space is built under constraints of unit variance and maximum correlation between pairs of canonical variables. Let Λ be a *p*-by-*p* diagonal matrix composed of canonical correlation coefficients given by:

$$\lambda_i = corr(U_i, V_i) \quad ; \quad i = 1, ..., p \quad (3)$$

Once the first pair of canonical variables $(U_1, V_1)_p$ is obtained, other canonical pairs are obtained subject to the constraint $corr(U_i, V_j) = 0$ for $i \neq j$. Note that all distinct hydrological canonical variables (as well as distinct physiographical variables) are also uncorrelated (Ouarda et al. 2001).

In order to improve quantile estimations in RFA, CCA is commonly used for the determination of neighborhoods of target sites. For an ungauged site, the canonical meteorological-physiological information U_0 is usually known but the hydrological information V_0 is not available. The

hydrological mean position of the target site *S* is given by ΛU_0. Hence, a 100(1-α) % confidence level neighborhood is identified by the Mahalanobis distance. It is considered between the mean position of target site ΛU_0 and positions of other sites *V*, such that:

$$\left(V-\Lambda U_0\right)'\left(I_p-\Lambda^2\right)^{-1}\left(V-\Lambda U_0\right)\le \chi^2_{\alpha,p} \qquad (4)$$

where $P(\chi^2_p \le \chi^2_{\alpha,p})=1-\alpha$ and χ^2_p has a chi-squared distribution with *p* degrees of freedom. Expression (4) is used to define an ellipsoid representing the neighborhood region for the ungauged site associated to ΛU_0 (Ouarda et al. 2001).

The equation of the ellipsoid has the following form:

$$\frac{\left(V_1-\Lambda_1 U_{01}\right)^2}{a^2}+\frac{\left(V_2-\Lambda_2 U_{02}\right)^2}{b^2}=1 \qquad (5)$$

where V_1 and V_2 denote the hydrological canonical variables, Λ_1 and Λ_2 are the canonical correlation coefficients, $(\Lambda_1 U_{01}, \Lambda_2 U_{02})$ are the coordinates of the center of the ellipsoid and *a* and *b* denote respectively the semi-major axis (or focal) and the semi-minor axis (Ballard 1981). Expression (5) is the equation of an ellipsoid in an orthonormal base (two orthogonal unit vectors), where axes are parallel to the coordinate system axes.

2.2. Nonlinear CCA using a Neural Network approach (CCA-NN)

An artificial neural network ANN is a fairly simple mathematical model compared to the natural biological evolution, with a running-inspired design of biological neurons (Bishop 1995). It consists essentially in several neurons generally organized in layers. The output of each neuron

results from the weighted sum of inputs, and transformed by a transfer function. Different transfer functions can be used (Duch and Jankowski 1999). ANNs have been widely used in a number of fields, such as in geology where Li et al. (2014) utilized the back-propagation (BP) neural network approach to forecast the geological hazard linked to bank destruction and landslides, and in hydrology where Zaier et al. (2010) used ANNs to model lake ice thickness, and Chen et al. (2014) used ANNs to model the rainfall-runoff relationship. As previously indicated, ANNs were integrated in RFA for instance by Ouarda and Shu (2009) and by Aziz et al. (2014) for the estimation of flood quantiles at ungauged sites.

In the meteorological field, Hsieh (2000) developed a NL-CCA version based on ANN (CCA-NN). The CCA-NN approach consists on establishing non-linear combinations between groups of original variables (*X* and *Y*) and the new canonical variables (*U* and *V)* via a transfer function. Consider the following hidden layer:

$$h_k^{(x)} = f\left(\left(W^{(x)}x + b^{(x)}\right)_k\right) \quad ; \quad k \text{ and } n = 1...l \tag{6}$$

$$h_n^{(y)} = f\left(\left(W^{(y)}y + b^{(y)}\right)_n\right) \tag{7}$$

where $W^{(x)}$ and $W^{(y)}$ are weight matrices, $b^{(x)}$ and $b^{(y)}$ are vectors of biased parameters, *k* and *n* denote respectively the indexes of the vector's elements $h^{(x)}$ and $h^{(y)}$ and l denotes the number of hidden neurons. The transfer function f, the same for *x* and *y*, is generally set to the hyperbolic tangent function (Hsieh 2000):

$$f(x) = \frac{e^x - e^{-x}}{e^x + e^{-x}} \tag{8}$$

Multivariate canonical neurons U and V are determined from a linear combination of respective neurons $h^{(x)}$ and $h^{(y)}$ (but from a non-linear combination with respect to *x* and *y*):

$$U = w^{(x)}h^{(x)} + \bar{b}^{(x)} \tag{9}$$

$$V = w^{(y)}h^{(y)} + \bar{b}^{(y)} \tag{10}$$

Without loss of generality, *U* and V are assumed to have zero mean. Thus, we have

$$\bar{b}^{(x)} = -\left\langle w^{(x)}h^{(x)} \right\rangle \text{ and } \bar{b}^{(y)} = -\left\langle w^{(y)}h^{(y)} \right\rangle \tag{11}$$

where $\langle z \rangle$ is the empirical mean of variable *z*.

A limitation of the CCA-NN is that, once applied to the original data, it provides only one pair of canonical variables, i.e. one for the physiographical variables and one for the hydrological variables. This may lead to ignoring a part of the information since it is not guaranteed that the first pair of canonical variables covers a significant part of the explained variance. To overcome this problem, the notion of modes was considered (Hsieh 2000). It consists in applying CCA-NN on the original datasets. The obtained result, denoted x', is related to the first mode. For the second mode, the CCA-NN is applied to the initial data, i.e. the set *x*, excluding the first mode. In other words, we determine the unexplained information in the previous mode by reapplying the procedure on the new variables:

$$I_2 = x - x' \tag{12}$$

Based on equation (12) we get:

$$J_2 = y - y' \tag{13}$$

where *y'* is the result of the first iteration, *y* is the matrix of original data.

The same procedure applies for higher order modes by considering each time the residual of the previous mode as input. The number of iterations, *m*, should be at least equal to the lowest number of variables, *p* in our case. The final result consists in summing up the results of all considered iterations:

$$x_{estimated} = x' + x'' + ... + x^m \tag{14}$$

where x^m is the result of the m^{ith} iteration, $m \geq p$. Therefore, the use of several modes may increase the percentage of the information contained in the resulting canonical variables.

2.3. Adaptation of CCA-NN to regional frequency analysis

For more clarity and to avoid confusion, it is important to note that in the approach proposed by Shu and Ouarda (2007), a CCA-based ANN model is used for flood quantile estimation without considering the DHR step and in which the employed CCA is the linear one. The aim of the linear CCA in Shu and Ouarda (2007) is to filter the signal from the original data and apply the ANN model on the canonical variables. However, in the present work the non-linear version of CCA using ANN (CCA-NN) is introduced in order to identify homogeneous regions, while a log linear regression model is used in the RE step.

Several versions of CCA-NN may be considered depending on the selected cost functions (canonical correlation, mean square error MSE, mean absolute error MAE). Indeed, Cannon

(2008) introduced a robust version of CCA-NN based on the biweight midcorrelation coefficient as a new measure of correlation instead of the Pearson correlation. After choosing the cost functions, canonical variables can be obtained and hence one can determine the hydrological neighborhood for an ungauged site. In the non-linear case, the variables V_1 and V_2 denote the hydrological canonical variables of the first and second mode, respectively, and Λ_1 and Λ_2 are the canonical correlation coefficients of the two modes. Identifying the physiographical coordinates of an ungauged site, U_{01} and U_{02}, is performed using relation (9).

Similarly to the neighborhood of the linear case, the non-linear one can be obtained using the same constraint. However, the equation of the ellipsoid is different from the linear case (5), since the axes are not parallel to those of the coordinate system.

Let Y denote an array of hydrological data and V the corresponding canonical variable, thus we can write:

$$Y = h(V) \tag{15}$$

Therefore by substituting (15) in (13) we obtain:

$$h_2(V_2) = h_1(V_1) - y' \tag{16}$$

Note that h, h_1 and h_2 are known non-linear functions.

Hence, the angle $\theta = (V_1, V_2)$ is different from $\pi/2$. Since the axes of the ellipsoid are always perpendicular, the ellipsoid is then rotated through an angle ϕ relative to the coordinate system

(V_1, Z). As illustrated in Figure 1, (V_1, Z) is an orthonormal basis with Z= sin(θ)V_2. The equation of the ellipsoid in the non-linear canonical space is given by:

$$\frac{\left(P_1 - \Lambda_1 U_{01}\right)^2}{a_1^2} + \frac{\left(P_2 - \Lambda_2 U_{02}\right)^2}{b_1^2} = 1 \qquad (17)$$

where:

$$P_1 = V_1 \cos\phi - Z \sin\phi \quad \text{and} \quad P_2 = V_1 \sin\phi + Z \cos\phi \qquad (18)$$

Note that the angle ϕ is the same for all sites and with different values of α. It depends only on θ: $\phi = f(\theta)$.

Equation (5) related to the linear CCA is a special case of (17) with a zero angle of rotation ϕ and $\theta = \pi/2$.

Similarly to CCA, the objective of NL-CCA consists in reducing the dimensions of hydrological and physiographical/meteorological spaces by taking into account the relationships between the considered variables. However, the construction of CCA reflects only linear relationships. The use of NL-CCA is necessary especially in the presence of non-linear structures. Note that the non-linearity in the hydrological processes is related to the non-linearity treated in NL-CCA.

To get a clear view of the correlation structure, it is essential to locate the source of interactions between variables. Note that the non-linearity in NL-CCA exists between the canonical and original variables of the same set, e.g. between U and physiographic variables. However, the non-linearity that occurs through the hydrological process is between hydrological variables Y

and physiographical ones *X*. We show that these two types of nonlinearities are connected. Indeed, in the NL-CCA context, the canonical variables can be written as:

$$U_i = f_1(X_i) \quad \text{and} \quad V_i = f_2(Y_i) \tag{19}$$

where f_1 and f_2 are non-linear functions (or linear in the case of CCA) and $i = 1, ..., p$. The simplest situation is the linear case, where more complex relations load to the same correlation:

$$U_i \approx \lambda_i V_i \tag{20}$$

The symbol $\approx$ indicates that both sides are approximately equal. Using relation (19), we obtain:

$$U_i \approx \lambda_i f_2(Y_i) \approx h(Y_i) \tag{21}$$

Substituting equation (19) into (21), we get:

$$h(Y_i) \approx f_1(X_i) \tag{22}$$

which leads to

$$Y_i \approx k(X_i) \tag{23}$$

where *k (.)* is a general function (if *h* is invertible *k* would be equal to $h^{-1} o f_1$).

Thus non-linear relations described by (19) are equivalent to non-linear relationships between the two groups of original variables (23). On the other hand, the presence of non-linearity in hydrological processes, between *X* and *Y*, leads to a non-linearity between canonical variables. Therefore, it is necessary to use the nonlinear approach in the context of RFA.

2.4. Regional estimation

Among the various RE methods, the most popular ones are the index-flood and regression models (Ouarda 2013). In this paper, we focus on the multivariate log-linear regression model, since it is more appropriate to use with CCA and with the available datasets. The relationship between flood quantiles (Y) and the physiographical/meteorological characteristics (X) is generally described by a power product model. With a log-transformation, the following log-linear model is obtained:

$$\log(Y) = \beta \log(X) + \varepsilon \qquad (24)$$

where β is a vector of parameters and $\boldsymbol{\varepsilon}$ represents the error (see Pandey and Nguyen (1999) for instance).

2.5. Evaluation criteria

To assess the performance of the proposed techniques, different criteria are used. Each model is evaluated using the following five indices: the Nash criterion (NASH) which provides a general evaluation of the quality estimation, the root mean squared error (RMSE) providing information about the accuracy of the estimator in an absolute scale, the relative RMSE (RMSEr) which is related to the relative scale, the mean bias (BIAS) and the relative mean bias (BIASr) provide a measure of the magnitude of overestimation or underestimation of a model. These indices are estimated based on a jackknife resampling procedure (e.g. Ouarda et al. 2001). It consists in removing temporarily each site and considering it as an ungauged one. The regional estimate is thus compared to the local estimate and the ability of each method is then evaluated.

The correlation coefficient and the proportion of explained variance are also used as evaluation criteria in the present work. The explained variance is deduced from the correlations between canonical components and initial variables, (Van Den Wollenberg 1977):

$$\sigma_E{}^2\left(U_i\right) = \frac{1}{q}\sum_{j=1}^{q}[corr\,(U_i, X_j)]^2 \qquad (25)$$

In a similar way, expression (25) is also valid for hydrological variables $Y_{j,\ j=1,..,r}$ and canonical variables $V_{i,\ i=1,..,2}$.

3. Case study

3.1. Data

The data used in this study covers three regions in North America, namely the province of Quebec (Canada), and the states of Arkansas and Texas (USA). The data from Arkansas and Texas are available in Tasker et al. (1996).

The first region includes 151 hydrometric stations and is located in the southern part of the province of Quebec, between 45° and 55° N. The considered physiographical and meteorological variables are those used previously by Chokmani and Ouarda (2004) : the mean basin slope (PMBV), the basin area (BV), the proportion of the basin area covered with lakes (PLAC), the annual mean total precipitation (PTMA) and the annual mean degree-days (DJBZ). Hydrological variables are at-site flood quantiles standardized by basin area to eliminate the scale effect (specific quantiles), denoted Q_{ST} for a return period T. For each site, the most appropriate statistical distribution has been identified in order to estimate the quantiles corresponding to different return periods. Two specific quantiles are selected for this study, namely the 10-year and the 100-year quantiles.

The second case-study concerns data from the state of Arkansas in the southern United States. Data stems from a hydrometric network composed of 204 gauging stations with drainage areas ranging from 0.13 km^2 to 6890 km^2. The same data was used by Tasker et al. (1996), namely the area (A), the slope of the main channel (S), the mean annual precipitation (P), the mean elevation of the watershed (EL), the length of the main stream (L), and estimated flood quantiles, Q_{ST}, corresponding to return periods of T = 2, 5, 10, 25 and 50.

The last region covers a hydrometric network of 69 stations in the state of Texas. Basin areas range between 86 Km^2 and 101,000 Km^2. The variables used are those indicated in Tasker et al. (1996) i.e. five physiographic variables (A, S, P, EL and L) and five flood quantiles which are the same as those considered in the Arkansas case study.

3.2. Model Design

In order to determine the homogeneous region, both CCA and CCA-NN analysis were carried out in the DHR step using r = 2 hydrological variables and q = 5 physiographical variables for all case studies (Quebec, Arkansas and Texas).

To build a model able to provide flood quantile estimation using the neighborhood approach, the CCA and CCA-NN approaches are coupled to a log-linear regression (24) in the RE step (denoted CCA & LR and CCA-NN & LR respectively). For comparison purposes, two regression models are considered in the non-linear case, according to the explanatory input variables, either directly using the initial data (X) or using the physiographical canonical variables (U_1, U_2). The latter is denoted CCA-NN & CLR and has the advantage of considering only the useful information with a smaller number of variables.

To compare the obtained results with different approaches presented in Chebana and Ouarda (2008), we discuss essentially results related to Quebec. Results associated to the other two regions will be presented briefly. Actually, several versions of CCA-NN with different cost functions were treated (Correlation coefficient/Mean absolute error COR/MAE, biweight midcorrelation coefficient/Mean absolute error BICOR/MAE and biweight midcorrelation coefficient /Mean square error BICOR/MSE). In the section below, only the results associated to BICOR/MSE are presented and discussed since this version provides the lowest evaluation criteria values. This finding is in concordance with the conclusion presented by Cannon (2008). In addition, it should be noted that the choice of the transfer function is an important step in ANN modeling, as it can significantly affect the results. In the hydrological literature, the sigmoid and the hyperbolic tangent functions are most commonly used as nonlinear transfer functions (Dawson and Wilby 2001; Yonaba et al. 2010). In this regard, several transfer functions belonging to the sigmoid function class were tested (the arctangent, the hyperbolic tangent and the sigmoid), and the hyperbolic tangent function yielded the best results. Hence, this transfer function (8) is employed for all case studies in the neurons of the hidden layers. The outputs of this model are canonical variables when the model is designed to forward mapping, and original variables in the case of inverse mapping. In the current application, three NNs were considered where the first ensures the forward mapping, while the second and the third are relative to the inverse mapping.

After extracting the first CCA-NN mode, the extraction of second mode is carried out by taking the residual as input, i.e., the original data minus the first CCA-NN mode, as in (12). Hence, we obtain the canonical variables in the non-linear space. Based on the Mahalanobis distance (4), the hydrological neighborhood of each ungauged site is determined.

4. Results

In this section, we present the results of the regional flood estimation procedure where the CCA-NN approach is considered for the DHR step. First, preliminary results are presented in order to study the relationships between variables. Figure 2 presents scatter plots of flood quantiles and physiographical/meteorological variables for Quebec. The examination of the scatter plots shows different forms of relationships between variables. We note, for instance, the existence of non-linear relations. The most notable ones are those between the variable basin area (BV) and the rest of the variables. Table2 presents the correlation coefficients between the hydrological and the physiographical variables. Despite the existence of a relatively strong positive correlation between flood quantiles and PLAC on one hand, and negative linear correlation between quantiles and PTMA on the other hand, we can observe from Figure 2 that these structures are rather non-linear. Further correlation measures are also evaluated between these variables. Figure 3 shows the correlation coefficients obtained by other correlation measures with respect to the Pearson correlation. This empirical comparison shows differences between measures, expressed by values higher or lower than those based on Pearson correlation. These behaviors indicate the existence of other dependence structures that are more complex than linearity.

By carrying out a linear CCA, the canonical correlation coefficients (3) are $\lambda_1 = 0.81$ and $\lambda_2 = 0.27$. In Chebana and Ouarda (2008), representations of data in the canonical spaces (not presented here to avoid repetition) show that the relationship between the first two canonical variables (U_1, V_1) can be considered to be linear, unlike variables (U_2, V_2) where linearity is relatively low.

In the following, results related to the CCA-NN are presented and discussed. Figure 4 presents the scatterplot of the study sites in the non-linear canonical spaces: physiographical (U_1, U_2) and hydrological (V_1, V_2). It is also convenient to present data in the spaces (U_1, V_1) and (U_2,V_2) to get prior information about the estimation error (Chebana and Ouarda 2008). This is illustrated in Figure 5 for the non-linear case. A nearly linear relationship is observed between the two canonical variables (U_1, V_1). This is not the case for the couple (U_2, V_2). However, the CCA-NN scatterplot seems to be more linear than the scatterplot of the data set in the linear space (U_2, V_2) presented in Chebana and Ouarda (2008). This may be explained by the fact that the canonical correlation coefficients obtained from CCA-NN ($\lambda_1 = 0.90$ and $\lambda_2 = 0.36$ using (3) and (20)) are higher than their counter parts deduced from CCA.

The explained variance (25), for the two first components, is respectively 51.16% and 97.36% (versus 56.92% and 99% in the linear CCA). Therefore, the canonical variables deduced from the linear CCA explain slightly better the variance of variables than those corresponding to CCA-NN. This may be due to the linearity induced by the correlation coefficient in the expression of the explained variance. However, this does not affect the results significantly since the selection of canonical variables is based essentially on the canonical correlation coefficients.

In the following we study the difference between the linear and non-linear approaches in identifying the hydrological neighborhood. The neighborhoods of selected stations are presented for both CCA and CCA-NN approaches in Figure 6. We observe a remarkable difference between the two approaches. Indeed, using the CCA, the neighborhood of each site is an ellipsoid with a zero angle of rotation. The non-linear method identified a rotated ellipsoid with a rotation angle ϕ~21°. Unlike CCA, the orientation of the CCA-NN ellipsoid tends to follow the shape of

the data dispersion. For instance, the non-linear neighborhood of station 030340 (n=45) identified 31 neighboring stations while the linear one identified a classical neighborhood with 39 stations, for the same value of α, $\alpha_{CCA-NN} = 0.2$. This means that the CCA-NN requires a smaller number of stations to reach the same RMSE as CCA. The optimal value of α corresponds to the minimum RMSEr. Figure 7 presents the variation of RMSEr for different values of α using CCA-NN. It can be seen that the optimal value $\alpha_{CCA\text{-}NN}$ is 0.2. Note that for high values of α, the performance criterion tends to infinity. To assess the magnitude of obtained results and their impact on RFA, we proceed to the RE step. Table 3 illustrates the jackknife results for all considered approaches through the criteria cited above. It can be seen that the NASH of the linear and non-linear models are substantially equal and sufficiently high to present acceptable results. For instance, for a return period of 100 years, the NASH of CCA is equal to 0.70 while it is equal to 0.71 for the non-linear case. Results indicate also that the RMSE of CCA-NN & LR and CCA & LR are almost equal whereas the RMSEr of the estimates computed by the CCA-NN & LR model are considerably lower than the linear model. By comparing the results with those obtained with the iterative procedure in Chebana and Ouarda (2008) and Wazneh et al. (2013) for the same data set, it can be seen that the proposed model, CCA-NN & LR, leads to best results among all models in terms of RMSEr. Indeed, while the linear approaches resulted in an RMSEr value of about 38% for the quantile QS10 and 44% for the quantile QS100, the CCA-NN & LR RMSEr values are around 34% for the quantile QS10 and 41% for the quantile QS100. It is also observed that the CCA-NN & LR results in both spaces, canonical and original, are very similar and are significantly better than the other models, i.e. the linear approach and the iterative procedure.

For all considered models, the BIAS is very close to zero with a slight improvement with the CCA-NN & LR approach. According to the BIASr criterion, the CCA-NN & CLR leads to the

best results. However, in comparison with results reported in Wazneh et al. (2013), the BIASr of the proposed models is higher (for values of QS_{100} and QS_{10} BIASr values are about -6 % and -7 % respectively using the CCA-NN & LR, versus around -2 % and -3% with the iterative procedure). This may be explained by the choice of the ANN parameters in the CCA-NN method. In fact, different parameters must be fixed from the beginning to guarantee optimum solution, such as penalty parameters which are chosen in such a way to avoid over-fitting. Optimization of these parameters is performed based on the RMSEr criterion. Consequently, the model loses in terms of BIASr but this latter remains in the same order of magnitude as the linear approaches.

Figure 8 presents the estimation error for flood quantiles QS_{100}, and QS_{10} using both the CCA & LR and the CCA-NN & LR models. One can observe that, overall, the CCA-NN & LR leads to smaller estimation errors than the linear model, CCA & LR. Particularly, the improvement for some sites is significant. For instance, for site 66 which has a particular location in both linear and non-linear canonical spaces, the estimation error goes from -4.13 using CCA & LR to -2.3 using CCA-NN & LR.

In the following, selected results related to Arkansas and Texas are presented. Without loss of generality, we will focus on specific quantiles corresponding to return periods of 10 and 50 years. Table 4 presents canonical correlation coefficients as well as percentages of explained variance for these two regions resulting from linear and non-linear CCA. Results indicate that, similarly to the region of Quebec, the canonical correlation coefficients are more important using a CCA-NN than using a CCA. This means that the non-linear components capture more information than the linear ones. However, as it was the case for Quebec case study, the explained variance of CCA is slightly higher than that of CCA-NN.

Table 5 summarises the results of the jackknife procedure using linear and non-linear analysis for these two regions. These results confirm the superiority of the non-linear approach. Indeed, when proceeding with CCA-NN & CLR applied to data of Arkansas, this model improves the RMSEr of QS_{10} by about 2% over the linear model CCA-LR and about 10% for QS_{50}. Similarly, results for the Texas region indicate that non-linear models perform better than CCA. The improvement of the RMSEr is even more important for Texas than for the Arkansas case study, with a significant improvement of BIASr.

5. Conclusions

This study has focused on the use of CCA-NN & LR methods in the context of RFA. The CCA approach has been successfully used for the delineation of homogeneous regions in RFA. However, this approach is not capable of representing the possible non-linear relationships between the variables of interest. To overcome the CCA limitations, several non-linear methods have been developed and used in other fields. CCA-NN and CCA-K are among the most prominent and most commonly used non-linear CCA methods.

In the current work, CCA-NN is presented and adapted to the RFA context. The method is also applied to three different regions to study its robustness in dealing with the nonlinearity of hydrological processes. In order to assess the performance of this method, its results are compared to those of linear CCA. Results show that CCA-NN can be adopted to represent the non-linear behavior of hydrological process and provide a more accurate and flexible delineation of homogeneous neighborhoods leading to a better regional estimation. However, this method has a number of drawbacks similarly to other ANN-based approaches, such as the identification of

optimum parameters and the selection of the transfer function. This latter requires the non-linear relationship to be empirical, i.e., dependent on the data, whereas in the current work and previous works, the hyperbolic tangent function was considered.

Acknowledgments

The authors thank M.A. Ben Alaya, for his valuable help and input into the present work. The authors wish also to express their appreciation to the reviewers, the Associate Editor and the Editor in Chief for their invaluable comments and suggestions. Financial support for the present study was provided by the National Sciences and Engineering Research Council of Canada (NSERC). To get access to the data used in this study, reader may refer to the report of A. Kouider (http://espace.inrs.ca/365/1/T000342.pdf).

6. References

Akaho, S. (2001). A kernel method for canonical correlation analysis. In Proceedings of the International Meeting of Psychometric Society (IMPS). University Convention Center-Osaka, Japan.

Aziz, K., A. Rahman, G. Fang and S. Shrestha (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.

Ballard, D. H. (1981). "Generalizing the Hough transform to detect arbitrary shapes." Pattern recognition **13**(2): 111-122.

Barnett, T. P. and R. Preinsendorfer (1987). "Origins and levels of monthly and seasonal forecast skill for United States surface air temperatures determined by canonical correlation analysis." Monthly Weather Review **115**: 1825-1850.

Bishop, C. M. (1995). Neural Networks for Pattern Recognition. Clarendon,New York, Oxford University Press.

Bolton, R. J., D. J. Hand and A. R.Webb (2003). "Projection techniques for nonlinear principal component analysis." Statistics and Computing **13**(3): 267-276.

Botelho, S. S. d. C., R. A. d. Bem, Í. L. d. Almeida and M. M. Mata (2005). C-nlpca: Extracting non-linear principal components of image datasets. Anais XII Simposio Brasileiro de Sensoriamento Remoto, Goiania, Brasil**:** 3495-3502.

Campi, C., L. Parkkonen, R. Hari and A. Hyvärinen (2013). "Non-linear canonical correlation for joint analysis of MEG signals from two subjects." Frontiers in neuroscience **7**.

Cannon, A. J. (2008). Multivariate statistical models for seasonal climate prediction and climate downscaling. Atmospheric Science, University Of British Columbia. **Doctor of philosophye:** 141.

Chebana, F., C. Charron, T. B. M. J. Ouarda and B. Martel (2014). "Regional frequency analysis at ungauged sites with the generalized additive model " In press. J. of Hydrometeorology.

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

Chebana, F. and T. B. M. J. Ouarda (2008). "Depth and homogeneity in regional flood frequency analysis." Water Resources Research **44**(11).

Chen, C.-S., C.-H. Liu and H.-C. Su (2008). "A nonlinear time series analysis using two-stage genetic algorithms for streamflow forecasting." Hydrological Processes **22**: 3697–3711.

Chen, L., V. P. Singh, S. Guo, J. Zhou and L. Ye (2014). "Copula entropy coupled with artificial neural network for rainfall–runoff simulation." Stochastic Environmental Research and Risk Assessment **28**(7): 1755-1767.

Chokmani, K. and T. B. M. J. Ouarda (2004). "Physiographical space-based kriging for regional flood frequency estimation at ungauged sites." Water Resources Research **40**(12).

Dauxois, J. and G. M. Nkiet (1998). "Nonlinear canonical analysis and independence tests." The Annals of Statistics **26**(4): 1254-1278.

Dawson, C. and R. Wilby (2001). "Hydrological modelling using artificial neural networks." Progress in physical Geography **25**(1): 80-108.

Duch, W. and N. Jankowski (1999). "Survey of neural transfer functions." Neural Computing Surveys **2**(1): 163-212.

Frie, K. G. and C. Janssen (2009). "Social inequality, lifestyles and health–a non-linear canonical correlation analysis based on the approach of Pierre Bourdieu." International journal of public health **54**(4): 213-221.

Gifi, A. (1990). Nonlinear multivariate analysis, Wiley (Chichester and New York)**:** 579 p.

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

Hardoon, D. R. and J. Shawe-Taylor (2009). "Convergence analysis of kernel Canonical Correlation Analysis: theory and practice." Mach Learn **74**: 23–38.

Hsieh, W. W. (2000). "Nonlinear canonical correlation analysis by neural networks." Neural Networks **13**: 1095 -1105.

Hsieh, W. W. (2001). "Nonlinear Canonical Correlation Analysis of the Tropical Pacific Climate Variability Using a Neural Network Approach." Journal of climate **14**: 2528-2539.

Khalil, B., T. Ouarda and A. St-Hilaire (2011). "Estimation of water quality characteristics at ungauged sites using artificial neural networks and canonical correlation analysis." Journal of Hydrology **405**(3): 277-287.

Kramer, M. A. (1991). "Nonlinear principal component analysis using autoassociative neural networks." American Institute of Chemical Engineers Journal **37**(2): 233-243.

Kruger, U., S. K. Sharma and G. W. Irwin (2004). Improved nonlinear canonical correlation analysis using genetic strategies. UKACC Control. University of Bath, UK.

Li, C., H. Tang, Y. Ge, X. Hu and L. Wang (2014). "Application of back-propagation neural network on bank destruction forecasting for accumulative landslides in the three Gorges Reservoir Region, China." Stochastic Environmental Research and Risk Assessment **28**(6): 1465-1477.

Malthouse, E. C. (1998). "Limitations of Nonlinear PCA as Performed with Generic Neural Networks." IEEE Transactions on Neural Networks **9**(1): 165-173.

Michael, A. G. and C. P. Raymond (2003). "Using traffic conviction correlates to identify high accident-risk drivers." Accident Analysis and Prevention **35**(6): 903-912.

Monahan, A. H. (2000). "Nonlinear Principal Component Analysis by Neural Networks: Theory and Application to the Lorenz System." Journal of climate **13**: 821-835.

Nagai, I. (2013). "Optimization using Cross-Validation for Penalized Nonlinear Canonical Correlation Analysis." Graduate School of Science and Technology, Kwansei Gakuin University 2-1 Gakuen, Sanda,Japan: 669-1337.

Ouarda, T., K. Bâ, C. Diaz-Delgado, A. Cârsteanu, K. Chokmani, H. Gingras, E. Quentin, E. Trujillo and B. Bobée (2008). "Intercomparison of regional flood frequency estimation methods at ungauged sites for a Mexican case study." Journal of Hydrology **348**(1): 40-58.

Ouarda, T. B. M. J. (2013). "Hydrological Frequency Analysis, Regional." Encyclopedia of Environmetrics: DOI:10.1002/9780470057339.vnn9780470057043.

Ouarda, T. B. M. J., C. Girard, G. S. Cavadias and B. Bobée (2001). "Regional flood frequency estimation with canonical correlation analysis." Journal of Hydrology **254**(1-4): 157-173.

Ouarda, T. B. M. J. and C. Shu (2009). "Regional low-flow frequency analysis using single and ensemble artificial neural networks." Water Resources Research **45**(11).

Ouarda, T. B. M. J., A. St-Hilaire and B. Bobée (2008). "Synthèse des développements récents en analyse régionale des extrêmes hydrologiques." Revue des sciences de l'eau **21**(2): 219-232.

Pandey, G. and V.-T.-V. Nguyen (1999). "A comparative study of regression based methods in regional flood frequency analysis." Journal of Hydrology **225**(1): 92-101.

Riad, S. and J. Mania (2004). "Rainfall-Runoff Model Using an Artificial Neural Network Approach." Mathematical and Computer Modelling **40**: 839-846.

Rumelhart, D. E., G. E. Hinton and R. J. Williams (1985). "Learning internal representations by error propagation." Rumelhart, J. L. McClelland & P. R. Group, **1**: 318-362.

Sengupta, S. and J. Boyle (1995). Non-linear principal component analysis of climate data. PCMDI.

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**(07).

Tasker, G. D. H., S.A.N. and C. S. Barks (1996). "Region of influence regression for estimating the 50-year flood at ungauged sites " Water Resources Research **1**(32): 163–170.

Tishlert, A. and S. Lipovetsky (1996). "Canonical correlation analyses for three data sets: a unified framework with application to management." Computers & Operations Research **23**(7): 667–679.

Van Den Wollenberg, A. L. (1977). "Redundancy analysis an alternative for canonical correlation analysis." Psychometrika **42**(2): 207-219.

Wang, D., L. Shi, D. S. Yeung and E. Tsang (2005). Nonlinear canonical correlation analysis of fMRI signals using HDR models. Engineering in Medicine and Biology Society, 2005. IEEE-EMBS 2005. 27th Annual International Conference of the, IEEE.

Wazneh, H., F. Chebana and T. Ouarda (2013). "Optimal depth-based regional frequency analysis." Hydrology and Earth System Sciences **17**(6): 2281-2296.

Wu, A. and W. W. Hsieh (2002). "Nonlinear canonical correlation analysis of the tropical Pacific wind stress and sea surface temperature." Climate Dynamics **19**: 713–722.

Xu, J., W. Li, M. Ji, F. Lu and S. Dong (2010). "A comprehensive approach to characterization of the nonlinearity of runoff in the headwaters of the Tarim River, Western China." Hydrological Processes **24**: 136–146.

Yin, H. (2007). "Nonlinear Dimensionality Reduction and Data Visualization: A Review." International Journal of Automation and Computing **4**(3): 294-303.

Yonaba, H., F. Anctil and V. Fortin (2010). "Comparing sigmoid transfer functions for neural network multistep ahead streamflow forecasting." Journal of Hydrologic Engineering **15**(4): 275-283.

Zaier, I., C. Shu, T. Ouarda, O. Seidou and F. Chebana (2010). "Estimation of ice thickness on lakes using artificial neural network ensembles." Journal of Hydrology **383**(3): 330-340.

Zhihua, J. and Y. Zhen (2010). "On using non-linear canonical correlation analysis for voice conversion based on Gaussian mixture model." Journal of Electronics **27**(1): 1-7.

List of tables

Table 1. Summary of common methods of NL-CCA 89
Table2. Correlation between hydrological and physiographical variables-Quebec.....89
Table 3. Jackknife validation results-Quebec 90
Table 4. Correlation coefficients and percentage of explained variance for CCA & LR and CCA-NN & LR relative to Arkansas and Texas.....90
Table 5.Jackknife validation results.....91

Table 1. Summary of common methods of NL-CCA

Variables	Method	Advantages	Drawbacks
Categorical	CANALS		-Requires only two sets of variables -Allows only a small number of possible values
	OVERALS	- Ability to treat **k** groups of variables	-Allows only a small number of possible values
Quantitative	CCA-NN		-Significant computation time -Black box -A fairly complex mathematical structure
	CCA-K	- Flexibility - Low computation time -No local optima	- Difficult to interpret
	CCA-GA	-A parsimonious technique - Easier to interpret	

Table2. Correlation between hydrological and physiographical variables-Quebec

	QS100	**QS10**
BV	-0.43	-0.46
PMBV	0.45	0.47
PLAC	-0.63	-0.67
PTMA	0.61	0.68
DJBZ	-0.59	-0.60

Table 3. Jackknife validation results-Quebec

	Variables	CCA-NN& CLR	CCA-NN& LR	CCA & LR
NASH	QS_{100}	0.672	**0.710**	0.700
	QS_{10}	0.728	**0.793**	0.790
RMSE	QS_{100}	0.114	**0.107**	0.109
($m^3/s.km^2$)	QS_{10}	0.066	**0.058**	**0.057**
RMSEr	QS_{100}	42.250	**41.400**	51.030
(%)	QS_{10}	35.696	**33.903**	44.870
BIAS	QS_{100}	0.014	**0.010**	0.017
($m^3/s.km^2$)	QS_{10}	0.006	**0.002**	0.005
BIASr	QS_{100}	-7.953	**-7.747**	-8.390
(%)	QS_{10}	-6.114	**-6.026**	-7.880

Best results are shown in bold character.

Table 4. Correlation coefficients and percentage of explained variance for CCA & LR and CCA-NN & LR relative to Arkansas and Texas

		Arkansas		Texas	
		CCA-NN & LR	CCA & LR	CCA-NN & LR	CCA & LR
Correlations	(U_1,V_1)	0.96	0.93	0.90	0.90
	(U_2,V_2)	0.45	0.37	0.61	0.50
Explained variance (%)	U_1	39.96	46.09	40.93	42.19
	V_1	79.97	65.11	61.29	62.99

Table 5.Jackknife validation results

		Region					
		Arkansas (USA)			Texas (USA)		
	Variables	**CCA-NN& CLR**	**CCA-NN & LR**	**CCA & LR**	**CCA-NN& CLR**	**CCA-NN & LR**	**CCA & LR**
NASH	**QS_{50}**	0.733	**0.748**	0.735	**0.552**	0.389	0.136
	QS_{10}	0.732	**0.761**	0.755	**0.577**	0.499	0.351
RMSE	**QS_{50}**	2.923	**2.839**	2.913	**0.255**	0.298	0.355
($m^3/s.km^2$)	**QS_{10}**	1.685	**1.592**	1.610	**0.119**	0.129	0.147
RMSEr	**QS_{50}**	**55.104**	59.308	61.360	**39.309**	50.757	54.887
(%)	**QS_{10}**	**46.786**	47.083	47.705	**35.599**	42.114	44.759
BIAS	**QS_{50}**	0.790	**0.610**	0.627	0.017	**0.005**	0.008
($m^3/s.km^2$)	**QS_{10}**	0.464	**0.336**	0.337	**0.000**	0.002	0.008
BIASr	**QS_{50}**	**1.557**	-3.759	-5.762	-5.168	-11.225	-4.119
(%)	**QS_{10}**	3.390	**-1.371**	-3.046	**-5.841**	-6.261	-7.567

Best results are shown in bold character.

Abbreviations

DHR: Delineation of homogeneous regions

RE: Regional estimation

CCA & LR: CCA associated to a log-linear regression

CCA-NN & LR : Non-linear CCA based on Neural Network in DHR step associated to a log-linear regression in the RE step

CCA-NN & CLR: Non-linear CCA based on Neural Network in DHR step associated to a log-linear regression in the canonical space in the RE step

List of Figures

Figure 1. Geometrical definition of the homogeneous region in the non-linear canonical space..93
Figure 2. Scatter plot of physiographical and hydrological variables- Quebec 93
Figure 3. Empirical comparison between the Pearson correlation and other measures of correlation (the Kendall tau, the Spearman Rho and the biweight midcorrelation) - Quebec 94
Figure 4. Data set in the non-linear canonical spaces: (a) physiographical and (b) hydrological - Quebec.....94
Figure 5. Data set in the non-linear canonical spaces: (a) (U_1, V_1) and (b) (U_2,V_2)- Quebec 95
Figure 6. DHR results shown for stations 030340, 030420 and 02717 using: a) CCA and b) CCA-NN approaches, n=45, 49 and 150 respectively- Quebec. 96
Figure 7. RMSEr variation as a function of the α parameter for hydrological variables Q_{S10} and Q_{S100}-Quebec 97
Figure 8. Estimation error resulting from the CCA & LR and CCA-NN& LR models- Quebec..98

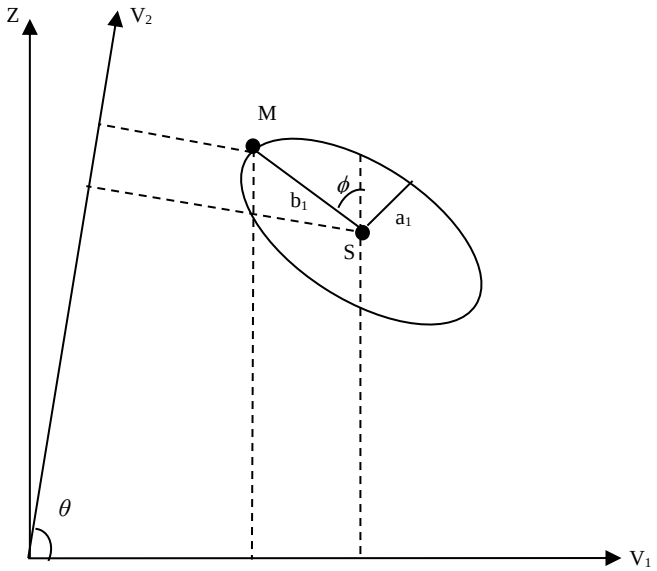


Figure 1. Geometrical definition of the homogeneous region in the non-linear canonical space

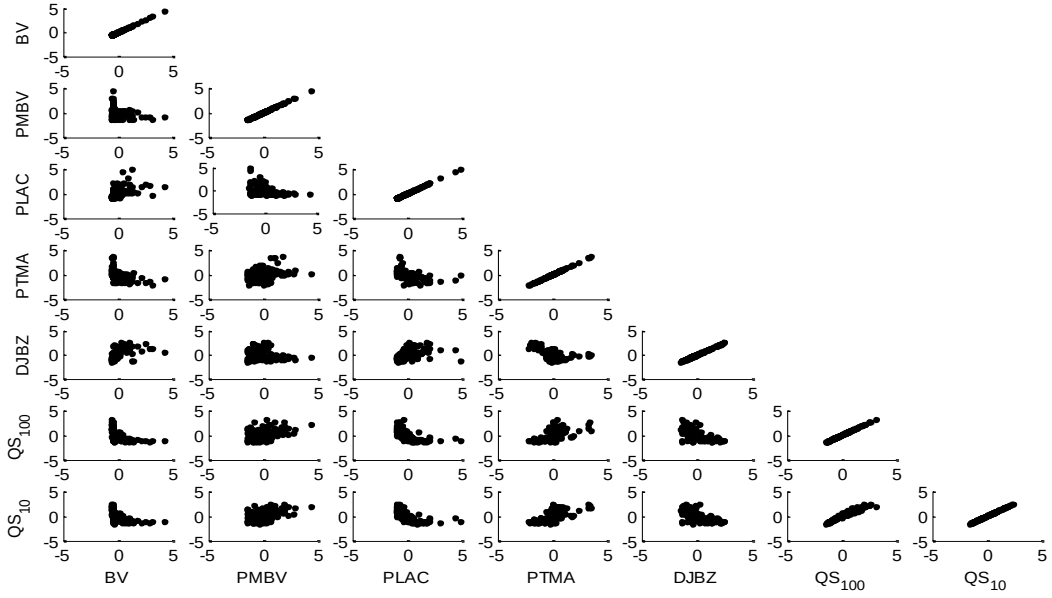


Figure 2. Scatter plot of physiographical and hydrological variables- Quebec

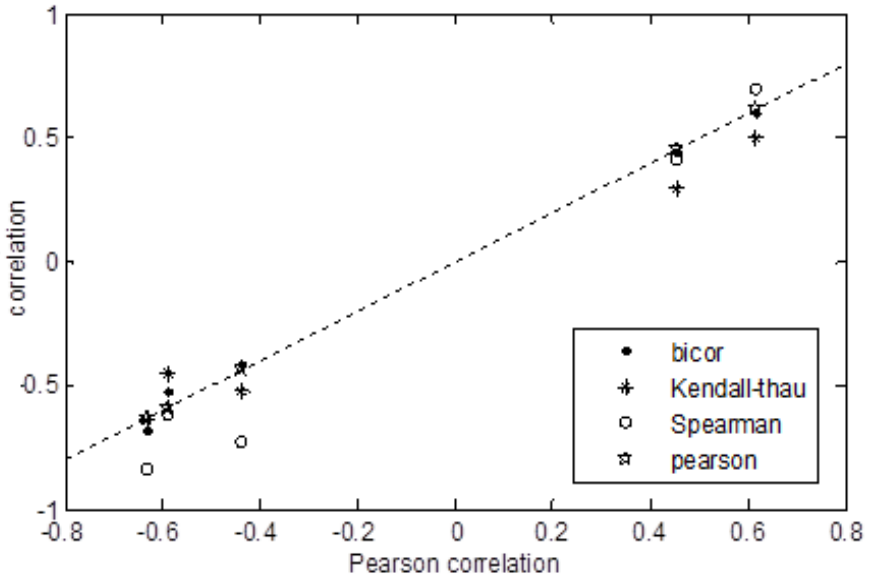


Figure 3. Empirical comparison between the Pearson correlation and other measures of correlation (the Kendall tau, the Spearman Rho and the biweight midcorrelation) - Quebec

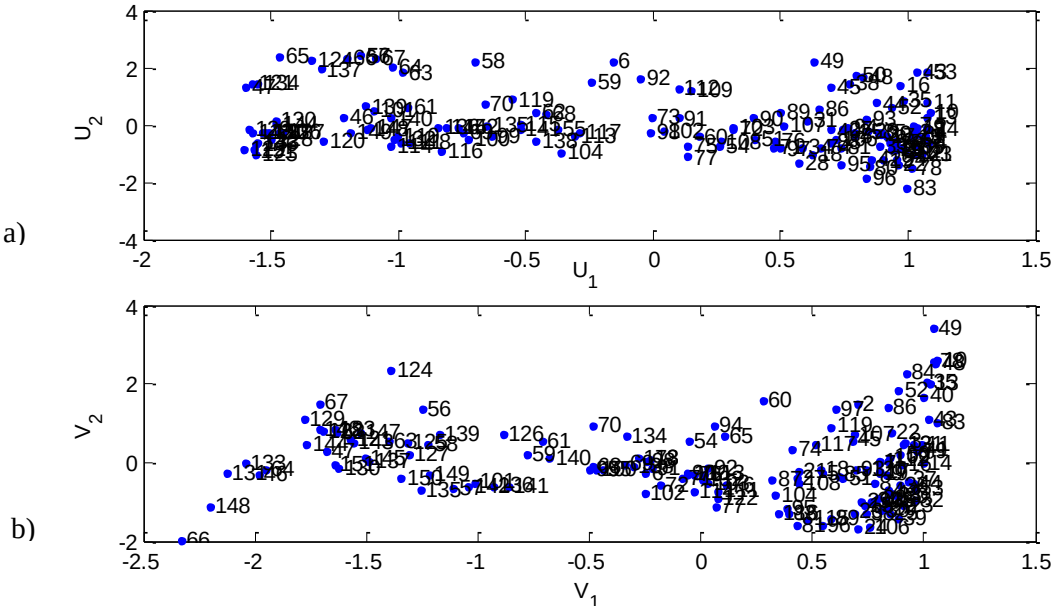


Figure 4. Data set in the non-linear canonical spaces: (a) physiographical and (b) hydrological - Quebec

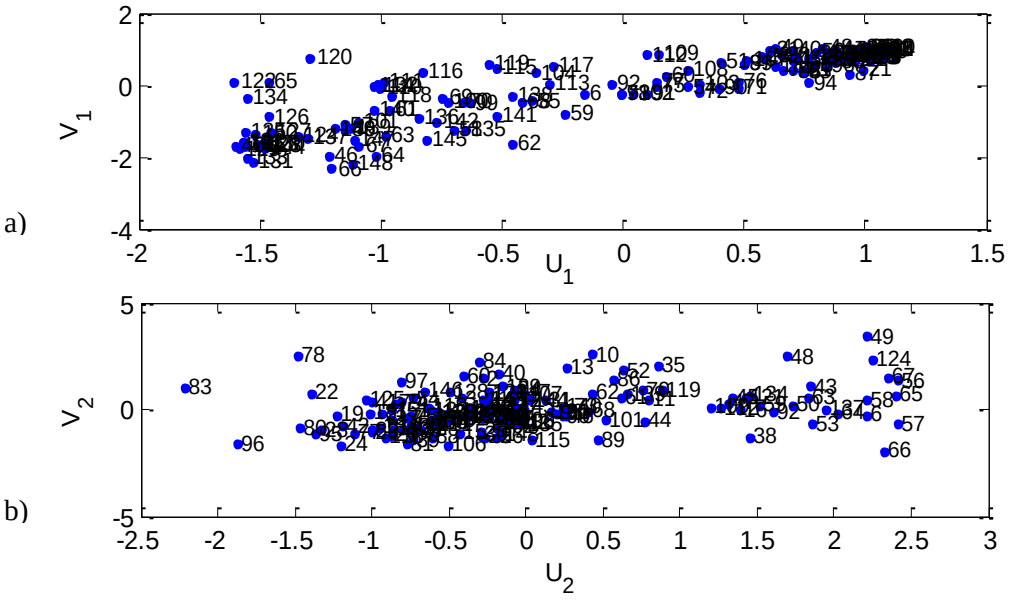
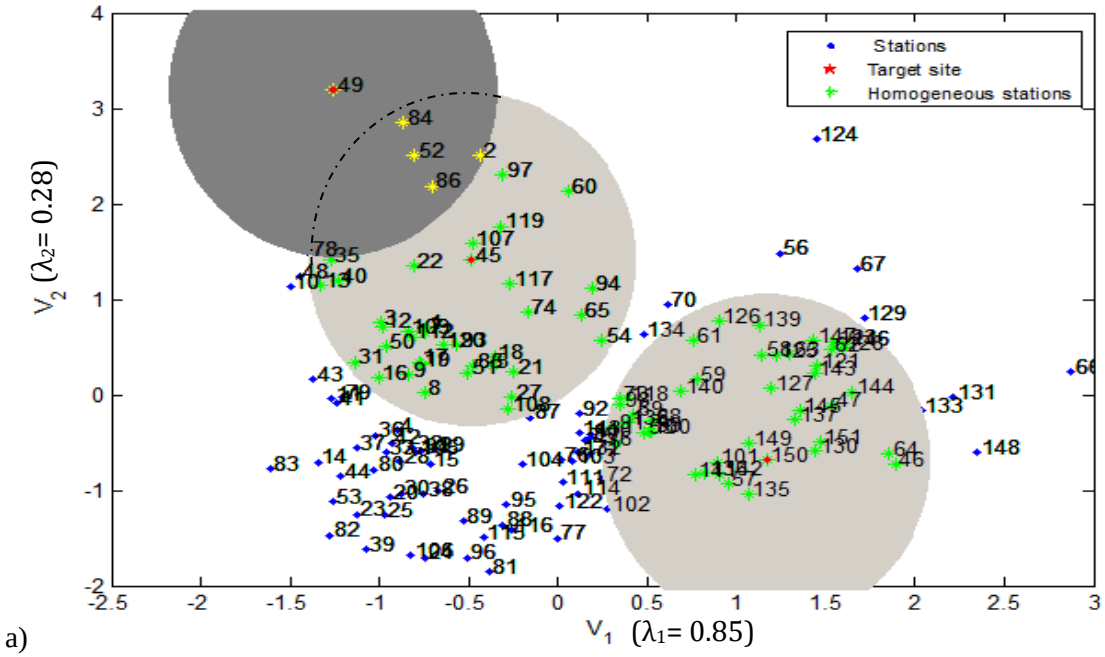
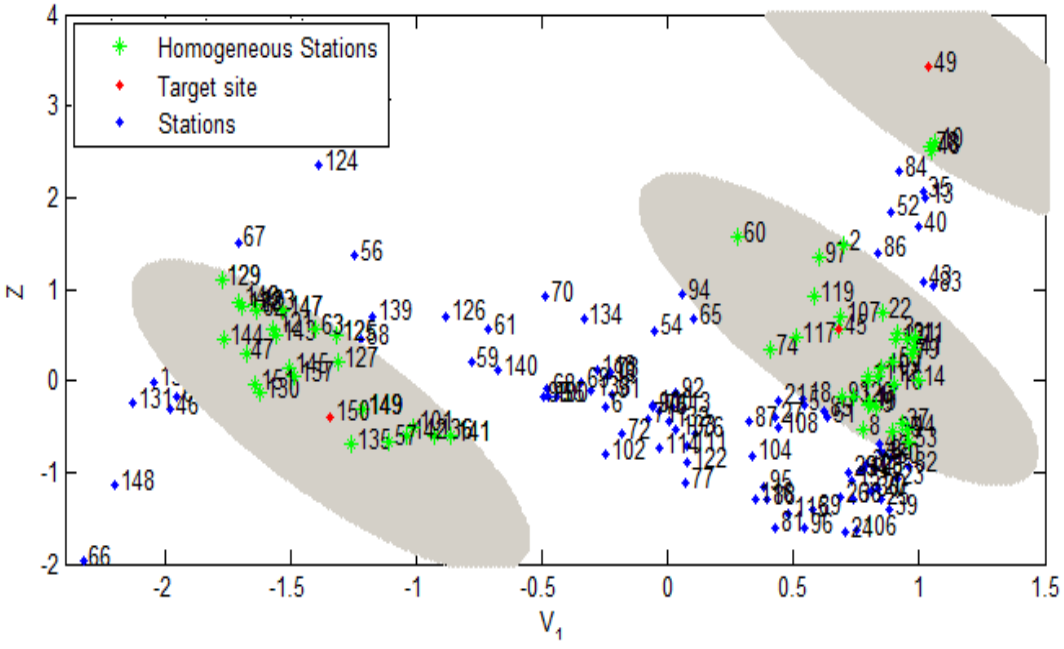


Figure 5. Data set in the non-linear canonical spaces: (a) (U_1, V_1) and (b) (U_2,V_2)- Quebec



a)



b)

Figure 6. DHR results shown for stations 030340, 030420 and 02717 using: a) CCA and b) CCA-NN approaches, n=45, 49 and 150 respectively- Quebec.

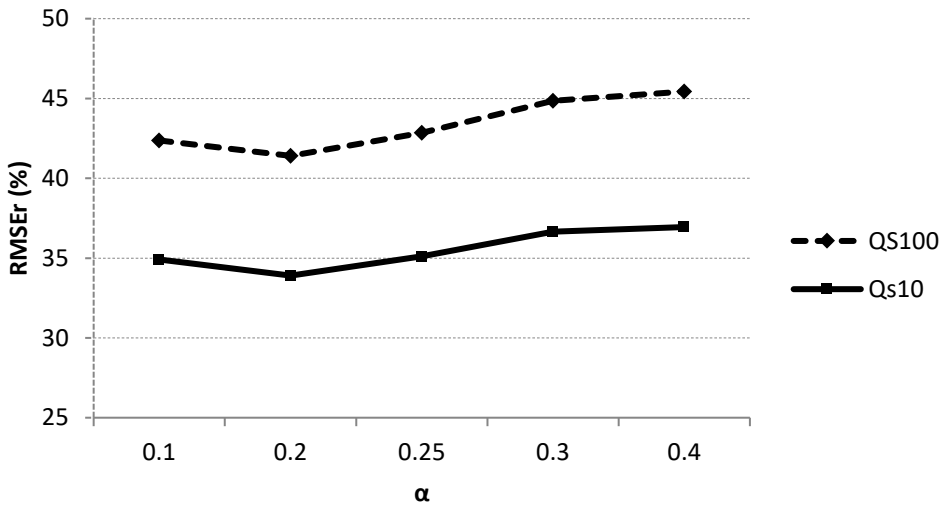


Figure 7. RMSEr variation as a function of the α parameter for hydrological variables Q_{S10} and Q_{S100}- Quebec

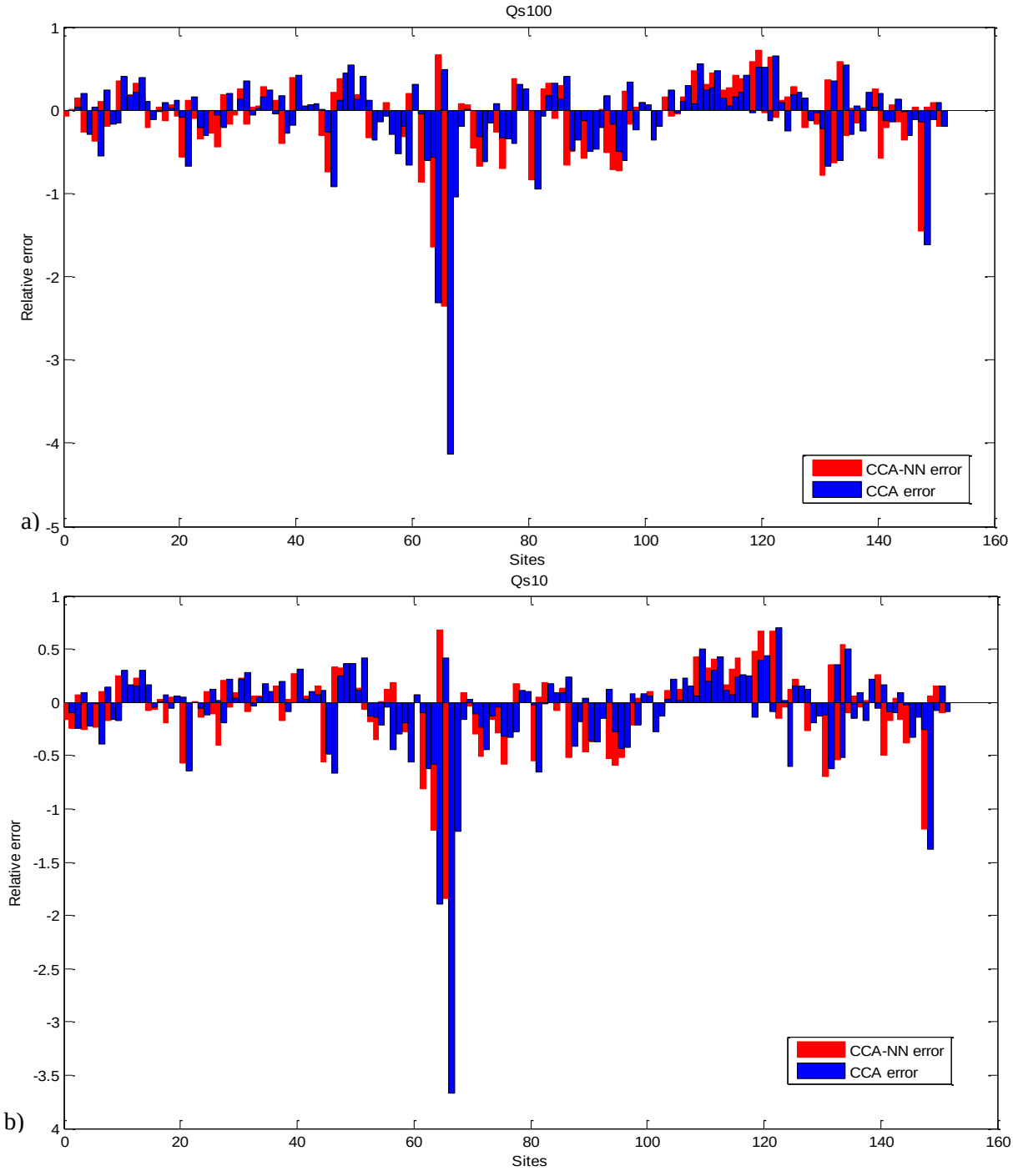


Figure 8. Estimation error resulting from the CCA & LR and CCA-NN& LR models- Quebec

CHAPITRE 3

FULLY NONLINEAR REGIONAL HYDROLOGICAL FREQUENCY ANALYSIS

Fully nonlinear regional hydrological frequency analysis

D. Ouali[1,*], F. Chebana[1], T. B.M.J. Ouarda[2, 1]

[1]*Institut National de la Recherche Scientifique, Centre Eau Terre et Environnement,*

490, rue de la Couronne, Québec (Québec), G1K 9A9, Canada.

[2]*Institute Centre for Water Advanced Technology and Environmental Research,*

Masdar Institute of Science and Technology

P.O. Box 54224, Abu Dhabi, UAE

[*]**Corresponding author:** Tel: +1 (418) 654 2530#4477

Email: dhouha.ouali@ete.inrs.ca

10 January 2016

Abstract

The high complexity of hydrological systems has long been recognized. Despite the increasing number of statistical techniques that aim to estimate hydrological quantiles at ungagged sites, few approaches were designed to account for the possible nonlinear connections between hydrological variables and catchments characteristics. Recently, a number of nonlinear tools have received attention in regional frequency analysis (RFA) applications especially for estimation purposes. In this paper, the aim is to study nonlinearity related aspects in the RFA of hydrological variables. To this end, a variety of combinations of linear and nonlinear approaches are considered in the main RFA steps (delineation and estimation). Artificial neural networks (ANN) and generalized additive models (GAM) are combined to a non-linear ANN-based canonical correlation analysis (NLCCA) procedure to ensure an appropriate nonlinear modelling of the complex processes involved. A comparison is carried out between classical linear combinations (CCA combined with linear regression model, LR), semi-linear combinations (e.g. NLCCA with LR) and fully nonlinear combinations (e.g. NLCCA with GAM). The considered combinations are applied to three different datasets located in North America. Results indicate that fully nonlinear combinations (in both RFA steps) are the most appropriate since they provide best performances and a more realistic description of the physical processes involved, even though they are relatively more complex than linear ones. On the other hand, semi-linear combinations which consider non-linearity either in the delineation or estimation steps showed little improvement over linear models. The linear approaches provided the lowest performances.

Keywords: Non-linear canonical correlation analysis, generalized additive models, artificial neural network, regional frequency analysis, quantile estimation.

1. Introduction and literature revue

Appropriate estimation of the occurrence frequency of hydrological extreme events, such as droughts and floods is of extreme importance for the adequate design and operation of water resources systems and to ensure public safety. To this end, frequency analysis of hydrological variables is a widely used approach when hydrological information is available at a given target site. Nevertheless, it is often required to estimate extreme events at ungauged sites where no hydrological observations are available. Regional frequency analysis (RFA) is a commonly used approach that aims to estimate hydrological quantiles at ungagged sites. It consists in two main steps, namely the identification of homogeneous regions and the transfer of hydrological information within the same homogeneous region (e.g. Hosking and Wallis 1997).

A large number of techniques were proposed in the literature for each step assuming generally linear relationships between flood quantiles and catchment characteristics (Pandey and Nguyen 1999; Ouarda et al. 2000). However, as hydrological systems involve complex processes, it is irrational to assume a linear coupling between hydrological and physio-meteorological variables. Indeed, the linkage between these variables is generally characterised by a strong nonlinearity (e.g. Sivakumar and Singh 2012). Therefore, a number of techniques have been proposed in the literature to account for possible nonlinearities in the relationships between variables. Recently, artificial neural networks (ANN) and generalized additive models (GAM) have known increasing popularity in a number of fields including hydrology. These two nonlinear models have also attracted significant attention in hydrological modeling as alternatives to classical regressive models (Shu and Burn 2004; Chebana et al. 2014).

An ANN is a nonparametric computing and modeling approach inspired by the biological functioning of the human brain (e.g. Rumelhart et al. 1986). Due to its capacity to detect complex

nonlinear relationships, ANN has been widely adopted for simulating and forecasting hydrological processes. Different ANN configurations were used for solving numerous hydrological problems such as rainfall-runoff modelling, groundwater flow analysis, river ice modelling and streamflow forecasting (Dawson and Wilby 2001; Zhang and Govindaraju 2003; Seidou et al. 2006; Nohair et al. 2008; Gao et al. 2010; Huo et al. 2012; Aziz et al. 2014).

Despite the extensive use of ANNs in the hydrological framework, their adoptions in RFA have been more modest. For instance, in Shu and Burn (2004) six various approaches have been applied using ANN ensembles, and compared to the single ANN model to estimate the index flood and the 10-year flood quantile. The application of the above models to some selected catchments indicated their ability to take into account nonlinear structures. In another study, Dawson et al. (2006) exploited the ANN ability to estimate the T-year flood events at ungagged sites. Shu and Ouarda (2007) introduced a one-step estimation model based on physiographical canonical variables produced by canonical correlation analysis (CCA), as inputs to ANN models (single and ensemble). Results showed that this technique provided superior estimations to those obtained in previous studies such as Chokmani and Ouarda (2004) and Ouarda et al. (2001). In Shu and Ouarda (2008), the adaptive neurofuzzy inference system model was applied to 151 catchments in the province of Quebec, Canada, and compared to the single ANN model and the power-form nonlinear regression model. Results of this study suggested that the proposed model outperforms the nonlinear regression model and has a comparable performance to the ANN based approach. The ANN approach has also been considered in Aziz et al. (2014) on an extensive dataset of 452 gauged catchments in Australia. The authors found that the ANN-based model presents the best performance among all employed models. Several other relevant studies used ANN models to obtain flood (or low-flow) estimations at ungagged sites, such as Hall and Minns (1998); Ouarda and Shu (2009); Besaw et al. (2010); Alobaidi et al. (2015); and Kumar et al.

(2015). A major drawback of ANN modelling, as a machine learning method, is the requirement of a large dataset to obtain the expected performances (Dawson et al. 2006). Furthermore, ANN calibration is a somewhat complex task which requires some subjective choices since no explicit regression equations can be given.

As opposed to the ANN model, the Generalized Additive Model (GAM) is an effective nonlinear tool defined using an explicit formulation (Hastie and Tibshirani 1990). Due to its considerable flexibility, it has been successfully applied in different fields such as medicine (e.g. Austin 2007), environment (e.g. Guisan et al. 2002), finance (e.g. Taylan et al. 2007) and hydrology (e.g. López-Moreno and Nogués-Bravo 2005). For regional estimation purposes, GAM was introduced in the RFA context by Chebana et al. (2014) who showed that the GAM-based approaches outperformed the classical ones and provided an explicit description of nonlinearities.

However, most of the current RFA literature, including the above mentioned studies, pays particular attention to the integration of nonlinearity in the estimation step. There have been very few studies dealing with the integration of nonlinear approaches in the delineation step. For instance, Lin and Chen (2006) applied the Self-organizing map to identify hydrological regions. It was shown that the Self-organizing map approach is an effective and robust tool providing accurate hydrological neighbourhoods. Recently, a nonlinear Canonical Correlation Analysis (NLCCA) approach was introduced by Ouali et al. (2015). The authors combined CCA and ANN approaches to identify hydrological neighborhoods, and then combined the proposed approach to a classical log-linear regression model for the estimation step. The obtained results showed the importance of accounting for nonlinear connections in the delineation step which also improved estimation performances.

Despite previous research efforts, it is important to note that the nonlinear models have not yet been considered simultaneously in both RFA steps. The main goal of the present paper is to

consider a variety of combinations of linear and nonlinear methods in both RFA steps in order to identify which step is more affected by nonlinearity. Therefore, new nonlinear combinations are proposed, assessed and compared.

The remainder of the present paper is organized as follows. The theoretical background of the techniques used in this study is given in section 2. In section 3, the description of the three case studies as well as the details of the implementation approaches of ANNs and GAMs in RFA are provided. In section 4, the results of the application of the proposed approaches are presented and discussed. Finally, section 5 summarizes the main conclusions.

2. Theoretical background

The present paper deals with the nonlinear aspects of complex hydrological systems. Unlike previous RFA studies, which treated the nonlinearity in only one RFA step, either the delineation or the estimation, all employed estimation tools herein (for both steps) are nonlinear techniques. In this section, we briefly present the theoretical background of the adopted approaches in each step.

2.1. Regional hydrological quantile estimation

In this subsection, an overview of the estimation approaches adopted in the current work is presented, namely the ANN (single and ensemble) and the GAM models.

2.1.1. Single and Ensemble ANN

To date, a number of ANN models have been developed and introduced allowing solving large complex problems especially in environmental concerns (Eissa et al. 2013; Anmala et al. 2014; Ashtiani et al. 2014; Coad et al. 2014; Benzer and Benzer 2015; and Wang et al. 2015). The differences between various ANN classes may reside, for instance, in the model topology, the training algorithm and the transfer function used. Among the various ANN types that are

available, the multilayer perceptron (MLP), also known as the multilayer feed-forward network is, so far, the most commonly used model for hydrological applications (Chokmani et al. 2008; e.g. Pramanik and Panda 2009; Zaier et al. 2010; Wu and Chau 2011; Kia et al. 2012; Chen et al. 2013).

A typical architecture of a MLP network is characterized by an input layer, one or more hidden layers and an output layer. Each layer contains computational units directly interconnected in a feed-forward way. Connections between neurons of two succeeding layers are performed using transfer functions designed through estimating appropriate parameters. Indeed, during the training process, the ANN parameters are estimated using an optimisation procedure. A number of training algorithms for MLP network are proposed in the literature among which the basic back propagation algorithm is the most popular (Shu and Burn 2004). More technical details about this algorithm are provided in Haykin and Lippmann (1994) and Werbos (1994).

A generalization of the single ANN abilities may show a significant improvement in its robustness and reliabilities by combining several ANNs into an Ensemble of ANNs (EANN). The EANN approach has received considerable attention in the hydrological literature (e.g. Cannon and Whitfield 2002; Araghinejad et al. 2011; Demirel et al. 2015). Although combining identical single ANNs may appear redundant, this generalized approach offers a better performance than the single ANN (Shu and Burn 2004). The principal idea is to train each network differently through, for example, considering different training sets, and then to combine all ANN estimations to provide a single output. To this end, boosting and bagging approaches are two popular training methods. Several ways to combine all network outputs were proposed in the literature, such as averaging and stacking. For more details about these techniques, the reader is referred to Schwenk and Bengio (2000), Breiman (1996), Bishop (1995) and Wolpert (1992).

2.1.2. Generalized Additive Model (GAM)

Before presenting the Generalized Additive Model (GAM), it is of interest to introduce the Generalized Linear Model (GLM). The latter is a flexible extension of the ordinary linear regression model allowing for the response distribution to be non-Gaussian and relating a response variable Y to explanatory variables X via a link function g (McCullagh and Nelder 1989). GAMs, initially introduced by Hastie and Tibshirani (1986), are an extension of GLMs linking, via a link function g, a non-Gaussian response to a sum of (nonlinear) smooth functions of explanatory variables.

The basic model formulation is explicitly given by (Wood 2006):

$$g\{Y\} = \alpha + \sum_{i=1}^{m} f_i(X_i) + \varepsilon \qquad (1)$$

where g is a monotonic link function and f_i is a smooth function of explanatory variable X_i.

This model allows accounting for nonlinear connections between response and explanatory variables through the smooth functions. Accordingly, the first step in GAM estimation is to estimate the smooth functions such that:

$$f_i(x) = \sum_{j=1}^{q} \beta_{ij} b_{ij}(x) \qquad (2)$$

where b_{ij} are basis functions and β_{ij} are parameters to be estimated. Typically, smooth functions can take both parametric and nonparametric forms. Note that Spline functions are the most commonly used basis to characterize smooth functions (Wahba 1990). Overall, the ability to consider non-parametric fitting with relaxed linear as well as Gaussian assumptions provides the potential for GAM to better describe regression relationships.

2.2. Delineation of homogeneous regions

CCA is one of the most recommended approaches adopted in RFA for identifying hydrological neighborhoods (Ouarda et al. 2001). To represent the relationship between two groups of variables, this technique consists on constructing new canonical variables resulting from *linear* combinations of physiographical and hydrological variables (*X* and *Y* respectively).

Recent research efforts have shown increased interest in the nonlinear dynamics of hydrological processes. In this regard, the nonlinear CCA (NLCCA) based on ANN approach (Ouali et al. 2015) is considered in the current study. This method consists in establishing non-linear combinations between original variables (*X* and *Y*) and the new canonical variables (*U* and *V)* via a transfer function. Consider the following hidden layer:

$$h_k^{(x)} = f\left(\left(W^{(x)}x + b^{(x)}\right)_k\right) \quad ; \quad n = 1,\ldots l \tag{3}$$

$$h_n^{(y)} = f\left(\left(W^{(y)}y + b^{(y)}\right)_n\right) \quad ; \quad k = 1,\ldots l \tag{4}$$

where $W^{(x)}$ and $W^{(y)}$ are weight matrices, $b^{(x)}$ and $b^{(y)}$ are vectors of biased parameters, *k* and *n* denote respectively the indices of the vector's elements $h^{(x)}$ and $h^{(y)}$ and l denotes the number of hidden neurons. Therefore, canonical variables U and V are determined from a linear combination of $h^{(x)}$ and $h^{(y)}$ as:

$$U = w^{(x)}h^{(x)} + \bar{b}^{(x)} \tag{5}$$

$$V = w^{(y)}h^{(y)} + \bar{b}^{(y)} \tag{6}$$

A more detailed description of the properties of NLCCA can be found in Hsieh (2000) whereas the adaptation and application to the RFA context can be found in Ouali et al. (2015).

3. Application and implementations

In this section we present the datasets used in this work as well as the study design.

3.1. Datasets

In this work, the proposed models and methods are applied to real-world case studies and each model performance is then compared to the performance of a number of classical approaches. For comparison purposes, case studies already used in previous studies are also adopted in the present study.

The first considered data base is inherent from the hydrometric station network of the southern part of the province of Quebec, Canada. A total of 151 stations located between the 45^o N and the 55^o N were selected (Chokmani and Ouarda 2004). Three types of variables are identified namely physiographical, meteorological and hydrological. The physiographical variables, as identified in Chokmani and Ouarda (2004), are the basin area (BV), mean basin slope (MBS) and the fraction of the basin area covered with lakes (FAL). The meteorological variables are the annual mean total precipitation (AMP) and the annual mean degree days over 0^o C (AMD). The hydrological variables correspond to the specific at-site flood quantiles Q_{ST} corresponding to a given return period *T*. A summary of all data statistics is provided in Table 1.

Two other case studies are also considered in this work, namely the hydrometric networks of the states of Arkansas and Texas in the United States with 204 and 69 catchments respectively. The employed basin characteristics are the same as in Ouali et al. (2015), explicitly, the basin area (BV), the slope of the main channel (S), the annual mean total precipitation (AMP), the mean

basin elevation (EL) and the length of the main channel (L). The hydrological variables are the specific at-site flood quantiles , Q_{ST}, corresponding to 10, and 50 years return periods.

3.2. Model designs for RFA

One important issue to address in this study is the use of nonlinear techniques in both delineation and estimation steps. Consequently, several combinations will be treated. Because of space limitations, model implementations and results associated to the Quebec case study are reported in details whereas those of Texas and Arkansas are briefly presented. In Table 2 a summary of all adopted regional models as well as the list of selected explanatory variables for the Quebec case study are presented. It is worthwhile to note that NLCCA implementation is carried out as in Ouali et al. (2015).

a. ANN and EANN implementation

In the present study, the MLP was selected to design both the ANN and EANN models. The model inputs are the standardized catchment characteristics (BV, MBS, FAL, AMP and AMD) that may affect the watershed hydrological behaviour. Model outputs are the log-transformed at-site estimated specific quantiles. According to the literature, the adopted transfer functions for both the hidden and the output layers are respectively the tan-sigmoid and the linear function. As mentioned in previous studies (Shu and Ouarda 2007; Ouarda and Shu 2009), seeking the optimal number of neurons in the hidden layer is a crucial step when designing an ANN model. Indeed, this number should neither be too high, to avoid overfitting, nor too low to avoid underfitting. In Shu and Burn (2004), five neurones in the hidden layer lead to accurate results.

After testing several ANN configurations including for instance varying the number of hidden neurons from 1 to 15, models using 4 neurones were selected since they allowed optimising the mean squared error (MSE) criterion. The Levenberg-Marquardt (LM) training algorithm (Hagan and Menhaj, 1994) was employed for training both ANN and EANN. Although it requires more

memory than other algorithms, it is much faster and more efficient than the basic back-propagation algorithm. It has also the ability to resolve several complex problems through proposing optimal solutions (Shu and Burn 2004). Depending on the initial value of the learning parameter μ, which appears in the LM algorithm weights, the LM algorithm behaves as a gradient descent method for large values of μ and as the Gauss-Newton method when μ is close to zero (Ouarda and Shu 2009). Similar to Shu and Ouarda (2007), an initial value of μ is given in the current work as 0.005.

For the EANN, a bagging with averaging approach is adopted. To achieve sufficient generalization ability, the ensemble size should be well selected. Indeed, if the size is too large, the training time increases whereas if the size is too small, no significant improvement in the generalization ability can be obtained (Shu and Ouarda 2007). In previous studies by Agrafiotis et al. (2002) and Shu and Burn (2004), an ensemble size of 10 was found to lead to satisfactory results. In Shu and Ouarda (2007), where the same case study of the province of Quebec was treated, an ensemble of 14 ANNs yielded the best results. In the current work, using the bagging-averaging configuration, different network sizes were trained (including 14 and 10 ANNs) with randomly sampled training data. The ensemble output is obtained after averaging all ANN outputs. According to the criteria presented hereafter, results indicated that using a network size of 10 achieved the best performance.

b. GAM implementation

In this application, GAM was implemented based on the **mgcv** package in the R language and environment (Wood 2006). Due to their theoretical motivations, the thin plate regression splines, which are a generalisation of cubic splines, are considered as basis b_{ij} in the smoothing functions f_i in (2). Note that this class of basis is characterized by its high computational speed and

includes a reduced number of parameters compared to other smoothing functions (Wood 2003). The considered link function g in (1) is the identity function since the log-transformed quantiles are approximately normally distributed (as in Chebana et al. (2014)).

A critical task when dealing with GAMs consists in selecting the appropriate smoothing level for each explanatory variable. This is achieved using the concept of effective degrees of freedom (edf) (Guisan et al. 2002). The total edf number used for all explanatory variables must be lower than the total number of observations (in the RFA context, it corresponds to the number of sites belonging to a given homogeneous region). In the current work, edf values are estimated using a stepwise procedure.

A stepwise selection procedure was also carried out to ensure an objective selection of the explanatory variables. As indicated in Chebana et al. (2014), the correlation-based selection method is a linear tool which seems to be more adequate with the CCA concept. Accordingly, in this study significant variables were selected using an automatic stepwise procedure within GAM. Prediction error criteria such as the Generalized Cross Validation score (GCV) and the Akaike Information Criterion (AIC) are adopted to select appropriate variables. As a result, identified variables were found to be the same as in Chebana et al. (2014), namely BV, FAL, AMD, LAT and LONG with edf respectively 1, 4, 4, 1 and 2.

Once a RFA model is established, a cross validation procedure (also called jackknife or leave-one-out procedure) is used to assess model performance. To this end, the following evaluation indices are used:

Nash:
$$Nash = 1 - \frac{\sum_{i=1}^{N} \left(y_i - \hat{y}_i \right)^2}{\sum_{i=1}^{N} \left(y_i - \bar{y} \right)^2} \qquad (7)$$

Relative root mean square error: $$RRMSE = \sqrt{\frac{1}{N}\sum_{i=1}^{N}\left(\frac{y_i - \hat{y}_i}{y_i}\right)^2} \quad (8)$$

Relative bias: $$RBIAS = \frac{1}{N}\sum_{i=1}^{N}\left(\frac{y_i - \hat{y}_i}{y_i}\right) \quad (9)$$

4. Results and discussion

Both ANN and GAM models were combined to the NLCCA in the delineation step and applied to the three considered datasets.

4.1. Results of the Quebec case study

The obtained results for the province of Quebec, using the cross-validation procedure for all considered combinations, are presented in Figure 1. Accordingly, the best overall performances are those obtained from the full nonlinear combination NLCCA-GAM first when the explanatory variables are those identified in Chokmani and Ouarda (2004), followed by the case when variables are selected with a stepwise technique (mainly in terms of RRMSE). In the following, we denote by NLCCA-GAM the model using BV, MBS, AMP, FAL and AMD variables. According to the high NASH values (more than 0.8) and the lowest RRMSE values (28.35% for Q_{S100}), the NLCCA-GAM combination provides the most accurate estimates compared to all other approaches. Based on the RBIAS, results show that, even though all models underestimate flood quantiles, the CCA-GAM is the least biased model (-3.7 % for Q_{S100}). However, compared to the NLCCA-GAM approach, the difference is not significant (a difference of – 1.3 % for Q_{S100}).

Results indicate also that the NLCCA-GAM combination yields more accurate estimates when compared to the same approach using variables identified by stepwise, despite the fact that the

difference is not too large. This may be explained by the fact that criteria used to select the variables (GCV and AIC) are not the same criteria used to evaluate model performances (NASH, RRMSE, RBIAS). In addition, in the case of the stepwise based combination, the used NLCCA solution is the same as in the NLCCA-GAM. Hence, through a more advanced NLCCA parameterization, better results could be achieved by using the stepwise approach.

Moreover, the obtained results reveal that, when adopting the same delineation method, GAM outperforms ANN-based approaches (ANN and EANN) in terms of all evaluation criteria. This may be attributable not only to the flexibility of GAM and its ability to adequately account for the nonlinearities, but also to the data size. Indeed, since the considered dataset is relatively not too large (151 catchments), the ANN-based models might not be properly trained. On the other hand, an expected finding is that, overall, the NLCCA-EANN approach outperforms ANN-based approaches (CCA-ANN, CCA-EANN and NLCCA-ANN). This is due to the combination of the advantages of the nonlinear delineation method and the generalization ability of the nonlinear estimation method.

Note that the use of the NLCCA approach for the identification of homogeneous regions leads to significant improvements in regional flood estimation when compared to the use of the linear CCA approach. More precisely, based on the RRMSE criterion (Figure 1-b), a relative improvement of 20% for the Q_{S100} estimates is obtained when considering the NLCCA-LR approach compared to the basic CCA-LR model. Regarding the estimation step, a relative improvement of 22% is recorded when considering CCA-GAM, and only 7% when considering the CCA-EANN approach. However, when we account for nonlinearity in both RFA steps, especially NLCCA combined with GAM, the gain reaches 45% (compared to the full linear

model CCA-LR). This illustrates clearly the importance of using nonlinear tools in both RFA steps.

The comparison can also be extended to other regional models in the literature, such as the depth-based approach (Wazneh et al. 2013), the EANN in the CCA space (Shu and Ouarda 2007; Khalil et al. 2011) and the projection pursuit regression approach (Durocher et al. 2015). In the two latter approaches the delineation step is not considered (one-step RFA models) and, in addition, the estimation models are nonlinear. Table 3 reports the results of the above-listed studies. It indicates that, in terms of RRMSE and NASH, the NLCCA-GAM combination outperforms considerably all approaches. However, the RBIAS values indicate that the depth-based approach performs slightly better. Further investigation of Table 3 reveals that, in terms of RRMSE, both CCA-GAM and projection pursuit regression perform similarly. It confirms the uselessness of a linear delineation tool (CCA) since both models are of the same nature.

To further explain the above results, the relative errors over sites associated to the best model in each category of combinations, CCA-GAM, NLCCA-LR, NLCCA-EANN and NLCCA-GAM, are presented in Figure 2. One can notice that the lowest errors are associated to the full nonlinear combination NLCCA-GAM. Note also that, for some sites, the NLCCA-LR and NLCCA-EANN approaches show comparable performances. Furthermore, for a small number of sites the CCA-GAM approach seems to perform poorly. Indeed, a number of problematic stations corresponding to atypically large relative errors for most of the considered approaches have been identified (stations with identification numbers: 030401, 041901, 041903, 042607, 050701, 076601, 081002 and 092711). Some of these stations (030401, 041903 and 042607) were also identified in previous studies treating the same case study (Chokmani and Ouarda 2004; Durocher et al. 2015). These sites were found to have under-evaluated areas (Chokmani and Ouarda 2004).

Using the NLCCA-GAM approach, the estimations corresponding to these particular sites are significantly improved as shown in Figure 2.

On the other hand, the exploration of the variability of errors as a function of the at-site Q_{S100} is shown in Figure 3 (because of space limitations and the similarity between results, those corresponding to Q_{S10} and Q_{S50} are not presented). One can see that the lowest specific quantile values are poorly estimated by all approaches except when using the full nonlinear combination, NLCCA-GAM, which provides accurate estimates.

At-site versus regional quantile estimates are presented in Figure 4 for Q_{S100}. To this end, five combinations are considered (CCA-LR, NLCCA-LR, NLCCA-EANN, CCA-GAM, and NLCCA-GAM) where LR-based ones (CCA-LR and NLCCA-LR) are considered as benchmarks and the NLCCA-EANN is selected as the best ANN-based model. According to Figure 4 the full nonlinear models show better overall performances (NLCCA-GAM followed by NLCCA-EANN). Indeed, associated at-site and regional estimations are very close since the points are less dispersed around the diagonal line. Moreover, higher specific quantile values are somewhat underestimated leading to the above obtained negative RBIAS values. These large quantile values were found to be associated to small basins (less than 800 km²), which seems to be systematically explained by their sharp hydrological responses. On the other hand, one can see, again, that the lowest specific quantile values are often overestimated except when using the full nonlinear combination, NLCCA-GAM, by which they are well estimated. Note that these sites are the same ones identified as problematic in Figure 2. These sites, whose geographical locations are indicated in Figure 5, were found to have large basin areas (such as sites 030401, 076601, 081002 and 092711) or to be located in the limit of the province with medium size catchments (041901, 041903, 042607 and 050701).

As opposed to previous studies, where problematic sites were often removed to improve the model and the overall estimation results, in this work these stations are preserved. Figure 6 shows specifically relative errors for these sites. It indicates that the NLCCA-GAM model yields the best estimations for these particular sites and significant improvements are obtained which explain the overall high performance. In particular, the NLCCA-GAM combination leads to a remarkable accurate estimate at site 042607 which is the most notable station in previous studies and models. This site has the lowest at-site quantile values for all return periods (64 m^3/s for Q_{S100}). Hence, the high flexibility offered by GAM leads to a better modelling of the complex hydrological phenomena and to a much improved estimation. This finding points out a significant advantage of nonlinear models, in particular the NLCCA-GAM approach. Indeed, it shows that there is no need to develop specific models for different classes of basins according to their size, slope, or streamflow magnitude.

4.2. Results of the Arkansas and Texas case studies

Results of the Arkansas and Texas case studies are presented in Tables 4 and 5. It can be seen that, again, the NLCCA-GAM approach provides the most accurate estimates especially in terms of RRMSE. In fact, for Texas, the NLCCA-GAM performs well in terms of all evaluation criteria. However, the relative improvement was not as large as in the case of the province of Quebec. Indeed, the comparison of NLCCA-GAM and CCA-LR shows that a relative improvement of only 31% has been achieved in Texas for Q_{S10} against 48% in the case of Quebec. Results associated to Arkansas reveal that the NLCCA-GAM approach is recommended when considering the RRMSE which is the most important criterion (Hosking and Wallis 1997). Compared to the fully linear combination, the relative improvement reaches 35% for Q_{S10}. This large difference between the NLCCA-GAM results in the three considered case studies can be

explained by the fact that the nonlinearity is not as pronounced in the Arkansas and Texas case studies as it is the case of Quebec.

Figure 7 illustrates the smooth functions of the response variables as a function of the explanatory variables for the three considered case studies. One can effectively notice the difference in the degree of nonlinearity between the three regions. Indeed, the most complex relations between explanatory and response variables appear in the case of the province of Quebec which explains the high gain recorded when using the fully nonlinear combination NLCCA-GAM (48%). Note also, from these figures, that the Texas region seems to represent the most linear case study (linear smooth function curves and low edf values) which justifies the smallest relative improvement (31%).

5. Conclusions

The main objective of this study is to present regional approaches that model nonlinear hydrological processes in both RFA steps. To this end, a number of combinations of delineation methods (CCA and NLCCA) and regional estimation models (LR, GAM, ANN and EANN) are considered. These combinations were also applied to three different case studies in North America.

The results show that it is important to consider nonlinear techniques in both RFA steps, in particular NLCCA for the delineation step and GAM for the estimation step. The NLCCA-EANN approach was found to be the second best model for Quebec case study. This is due to the fact that a satisfactory performance of ANN-based model requires large datasets to be trained which is not the case for the three studied regions.

Regarding the importance of considering nonlinearity in the delineation or in the estimation step, it was found that both efforts lead to comparable results. Indeed, improvement in the overall

model performance requires the integration of nonlinear models in both steps. In summary, despite the relative complexity of the NLCCA-GAM approach, it is worthwhile to consider such model to adequately account for the nonlinearities of complex hydrological phenomena.

In this study, the focus was on assessing the performances of the ANN and GAM models, combined to the NLCCA approach. In further efforts, it may be of interest to proceed with other combinations such as the projection pursuit regression, as a generalization of these two techniques, coupled to a delineation approach.

Acknowledgments

The authors thank Claude Onikpo for his valuable help and input. Financial support for the present study was graciously provided by the Natural Sciences and Engineering Research Council of Canada (NSERC). To get access to the data used in this study, reader may refer to the report of A. Kouider (http://espace.inrs.ca/365/1/T000342.pdf).

References

Agrafiotis, D. K., W. Cedeno and V. S. Lobanov (2002). "On the use of neural network ensembles in QSAR and QSPR." Journal of chemical information and computer sciences **42**(4): 903-911.

Alobaidi, M. H., P. R. Marpu, T. B. M. J. Ouarda and F. Chebana (2015). "Regional frequency analysis at ungauged sites using a two-stage resampling generalized ensemble framework." Advances in Water Resources **84** 103-111.

Anmala, J., O. W. Meier, A. J. Meier and S. Grubbs (2014). "GIS and Artificial Neural Network–Based Water Quality Model for a Stream Network in the Upper Green River Basin, Kentucky, USA." Journal of Environmental Engineering **141**(5): 04014082.

Araghinejad, S., M. Azmi and M. Kholghi (2011). "Application of artificial neural network ensembles in probabilistic hydrological forecasting." Journal of Hydrology **407**(1): 94-104.

Ashtiani, A., P. A. Mirzaei and F. Haghighat (2014). "Indoor thermal condition in urban heat island: Comparison of the artificial neural network and regression methods prediction." Energy and Buildings **76**: 597-604.

Austin, P. C. (2007). "A comparison of regression trees, logistic regression, generalized additive models, and multivariate adaptive regression splines for predicting AMI mortality." Statistics in medicine **26**(15): 2937-2957.

Aziz, K., A. Rahman, G. Fang and S. Shrestha (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.

Benzer, R. and S. Benzer (2015). "Application of artificial neural network into the freshwater fish caught in Turkey." **2**(5): 341-346.

Besaw, L. E., D. M. Rizzo, P. R. Bierman and W. R. Hackett (2010). "Advances in ungauged streamflow prediction using artificial neural networks." Journal of Hydrology **386**(1): 27-37.

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

Breiman, L. (1996). "Bagging predictors." Machine learning **24**(2): 123-140.

Cannon, A. J. and P. H. Whitfield (2002). "Downscaling recent streamflow conditions in British Columbia, Canada using ensemble neural network models." Journal of Hydrology **259**(1): 136-151.

Chebana, F., C. Charron, T. B. M. J. Ouarda and B. Martel (2014). "Regional Frequency Analysis at Ungauged Sites with the Generalized Additive Model." Journal of Hydrometeorology **15**(6): 2418-2428.

Chen, P.-A., L.-C. Chang and F.-J. Chang (2013). "Reinforced recurrent neural networks for multi-step-ahead flood forecasts." Journal of Hydrology **497**: 71-79.

Chokmani, K. and T. B. M. J. Ouarda (2004). "Physiographical space-based kriging for regional flood frequency estimation at ungauged sites." Water Resources Research **40**(12).

Chokmani, K., T. B. M. J. Ouarda, S. Hamilton, M. H. Ghedira and H. Gingras (2008). "Comparison of ice-affected streamflow estimates computed using artificial neural networks and multiple regression techniques." Journal of Hydrology **349**(3): 383-396.

Coad, P., B. Cathers, J. E. Ball and R. Kadluczka (2014). "Proactive management of estuarine algal blooms using an automated monitoring buoy coupled with an artificial neural network." Environmental Modelling & Software **61**: 393-409.

Dawson, C. and R. Wilby (2001). "Hydrological modelling using artificial neural networks." Progress in physical Geography **25**(1): 80-108.

Dawson, C. W., R. J. Abrahart, A. Y. Shamseldin and R. L. Wilby (2006). "Flood estimation at ungauged sites using artificial neural networks." Journal of Hydrology **319**(1): 391-409.

Demirel, M. C., M. Booij and A. Hoekstra (2015). "The skill of seasonal ensemble low-flow forecasts in the Moselle River for three different hydrological models." Hydrological Earth System Science **19**: 275–291.

Durocher, M., F. Chebana and T. B. M. J. Ouarda (2015). "A Nonlinear Approach to Regional Flood Frequency Analysis Using Projection Pursuit Regression." Journal of Hydrometeorology **16**(4): 1561-1574.

Eissa, Y., P. R. Marpu, I. Gherboudj, H. Ghedira, T. B. M. J. Ouarda and M. Chiesa (2013). "Artificial neural network based model for retrieval of the direct normal, diffuse horizontal and global horizontal irradiances using SEVIRI images." Solar Energy **89**: 1-16.

Gao, C., M. Gemmer, X. Zeng, B. Liu, B. Su and Y. Wen (2010). "Projected streamflow in the Huaihe River Basin (2010–2100) using artificial neural network." Stochastic Environmental Research and Risk Assessment **24**(5): 685-697.

Guisan, A., T. C. Edwards and T. Hastie (2002). "Generalized linear and generalized additive models in studies of species distributions: setting the scene." Ecological modelling **157**(2): 89-100.

Hall, M. and A. Minns (1998). Regional flood frequency analysis using artificial neural networks. Hydroinformatics Conference. V. B. C. L. Larsen. Copenhagen, Denmark, A.A.Balkema. **2:** 759–763.

Hastie, T. and R. Tibshirani (1986). "Generalized additive models." Statistical science: 297-310.

Hastie, T. J. and R. J. Tibshirani (1990). Generalized additive models, CRC Press.

Haykin, S. and R. Lippmann (1994). "Neural Networks, A Comprehensive Foundation." International Journal of Neural Systems **5**(4): 363-364.

Hosking, J. and J. Wallis (1997). Regional Frequency Analysis. An Approach Based on L–moments. Cambridge, United Kingdom., Cambridge University Press. .

Hsieh, W. W. (2000). "Nonlinear canonical correlation analysis by neural networks." Neural Networks **13**: 1095 -1105.

Huo, Z., S. Feng, S. Kang, G. Huang, F. Wang and P. Guo (2012). "Integrated neural networks for monthly river flow estimation in arid inland basin of Northwest China." Journal of Hydrology **420**: 159-170.

Khalil, B., T. B. M. J. Ouarda and A. St-Hilaire (2011). "Estimation of water quality characteristics at ungauged sites using artificial neural networks and canonical correlation analysis." Journal of Hydrology **405**(3): 277-287.

Kia, M. B., S. Pirasteh, B. Pradhan, A. R. Mahmud, W. N. A. Sulaiman and A. Moradi (2012). "An artificial neural network model for flood simulation using GIS: Johor River Basin, Malaysia." Environmental Earth Sciences **67**(1): 251-264.

Kumar, R., N. K. Goel, C. Chatterjee and P. C. Nayak (2015). "Regional Flood Frequency Analysis using Soft Computing Techniques." Water Resources Management **29**(6): 1965-1978.

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

López-Moreno, J. I. and D. Nogués-Bravo (2005). "A generalized additive model for the spatial distribution of snowpack in the Spanish Pyrenees." Hydrological Processes **19**(16): 3167-3176.

McCullagh, P. and J. A. Nelder (1989). Generalized linear models, CRC press.

Nohair, M., A. St-Hilaire and T. B. M. J. Ouarda (2008). "The Bayesian-Regularized neural network approach to model daily water temperature in a small stream." Revue des sciences de l'eau **21**(3).

Ouali, D., F. Chebana and T. B. M. J. Ouarda (2015). "Non-linear canonical correlation analysis in regional frequency analysis." Stochastic Environmental Research and Risk Assessment: 1-14.

Ouarda, T. B. M. J., C. Girard, G. S. Cavadias and B. Bobée (2001). "Regional flood frequency estimation with canonical correlation analysis." Journal of Hydrology **254**(1): 157-173.

Ouarda, T. B. M. J., M. Haché, P. Bruneau and B. Bobée (2000). "Regional flood peak and volume estimation in northern Canadian basin." Journal of Cold Regions Engineering **14**(4): 176-191.

Ouarda, T. B. M. J. and C. Shu (2009). "Regional low-flow frequency analysis using single and ensemble artificial neural networks." Water Resources Research **45**(11).

Pandey, G. and V.-T.-V. Nguyen (1999). "A comparative study of regression based methods in regional flood frequency analysis." Journal of Hydrology **225**(1): 92-101.

Pramanik, N. and R. K. Panda (2009). "Application of neural network and adaptive neuro-fuzzy inference systems for river flow prediction." Hydrological Sciences Journal **54**(2): 247-260.

Rumelhart, D., G. Hinton and R. Williams (1986). Learning Internal Representations by Error Propagation, Parallel Distributed Processing, Explorations in the Microstructure of Cognition, ed. DE Rumelhart and J. McClelland. Vol. 1. 1986, Cambridge, MA: MIT Press.

Schwenk, H. and Y. Bengio (2000). "Boosting neural networks." Neural Computation **12**(8): 1869-1887.

Seidou, O., T. B. M. J. Ouarda, L. Bilodeau, M. Hessami, A. St-Hilaire and P. Bruneau (2006). "Modeling ice growth on Canadian lakes using artificial neural networks." Water Resources Research **42**(11).

Shu, C. and D. H. Burn (2004). "Artificial neural network ensembles and their application in pooled flood frequency analysis." Water Resources Research **40**(9).

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**(07).

Shu, C. and T. B. M. J. Ouarda (2008). "Regional flood frequency analysis at ungauged sites using the adaptive neuro-fuzzy inference system." Journal of Hydrology **349**: 31– 43.

Sivakumar, B. and V. Singh (2012). "Hydrologic system complexity and nonlinear dynamic concepts for a catchment classification framework." Hydrology and Earth System Sciences **16**(11): 4119-4131.

Taylan, P., G.-W. Weber and A. Beck (2007). "New approaches to regression by generalized additive models and continuous optimization for modern applications in finance, science and technology." Optimization **56**(5-6): 675-698.

Wahba, G. (1990). Spline models for observational data. Philadelphia, SIAM.

Wang, W.-c., K.-w. Chau, L. Qiu and Y.-b. Chen (2015). "Improving forecasting accuracy of medium and long-term runoff using artificial neural network based on EEMD decomposition." Environmental research **139**: 46-54.

Wazneh, H., F. Chebana and T. B. M. J. Ouarda (2013). "Optimal depth-based regional frequency analysis." Hydrology and Earth System Sciences **17**(6): 2281-2296.

Werbos, P. J. (1994). The roots of backpropagation: from ordered derivatives to neural networks and political forecasting, John Wiley & Sons.

Wolpert, D. H. (1992). "Stacked generalization." Neural networks **5**(2): 241-259.

Wood, S. (2006). Generalized additive models: an introduction with R, CRC press.

Wood, S. N. (2003). "Thin plate regression splines." Journal of the Royal Statistical Society: Series B (Statistical Methodology) **65**(1): 95-114.

Wu, C. and K. Chau (2011). "Rainfall–runoff modeling using artificial neural network coupled with singular spectrum analysis." Journal of Hydrology **399**(3): 394-409.

Zaier, I., C. Shu, T. B. M. J. Ouarda, O. Seidou and F. Chebana (2010). "Estimation of ice thickness on lakes using artificial neural network ensembles." Journal of Hydrology **383**(3): 330-340.

Zhang, B. and R. S. Govindaraju (2003). "Geomorphology-based artificial neural networks (GANNs) for estimation of direct runoff over watersheds." Journal of Hydrology **273**(1): 18-34.

List of tables

Table 1. Descriptive statistics of hydrological and physio-meteorological variables-Quebec 127

Table 2. Summary of all considered regional models. 127

Table 3. Comparison of NLCCA-GAM with a number of RFA approaches from previous studies applied to the same dataset, Quebec. 128

Table 4. Jackknife Validation Results- Arkansas. 128

Table 5. Jackknife Validation Results- Texas. 129

Table 1. Descriptive statistics of hydrological and physio-meteorological variables-Quebec

Variable	Min	Mean	Max	STD
Mean Basin Slope (MBS) (%)	0.96	2.43	6.81	0.99
Fraction of the basin area covered with lakes (FAL) (%)	0.00	7.72	47.00	7.99
Annual mean total precipitation (AMP) (mm)	646	988	1534	154
Annual mean degree days over 0° (AMD) (°C)	8589	16346	29631	5382
Basin area (BV) (km²)	208	6255	96600	11716
Latitude (LAT) (°N)	45	48	54	2
Longitude (LONG) (°W)	58	72	79	4
Flood quantile of 10 year return period (m³/s)	53	698	5649	828
Flood quantile of 50 year return period (m³/s)	61	851	6642	985
Flood quantile of 100 year return period (m³/s)	64	913	7013	1048

Table 2. Summary of all considered regional models.

	Delineation step (D)	**Estimation step (E)**	**Regional model notation**	**Reference**	**Physiographical variables**
Linear D & E	CCA	LR	CCA-LR	Ouarda et al. (2001)	BV, MBS, FAL, AMP, AMD
Linear D & nonlinear E	CCA	ANN	CCA-ANN	Current work	BV, MBS, FAL, AMP, AMD
	CCA	EANN	CCA-EANN		
	CCA	GAM	CCA-GAM	Chebana et al. (2014)	
Nonlinear D & linear E	NLCCA	LR	NLCCA-LR	Ouali et al. (2015)	BV, MBS, FAL, AMP, AMD
Nonlinear D & E	NLCCA	ANN	NLCCA-ANN	Current work	BV, MBS, FAL, AMP, AMD
	NLCCA	EANN	NLCCA-EANN		
	NLCCA	GAM	NLCCA-GAM		
	NLCCA	GAM	NLCCA-GAM/STPW		BV, FAL, AMD, LAT, LONG

Table 3. Comparison of NLCCA-GAM with a number of RFA approaches from previous studies applied to the same dataset, Quebec.

Regional model	Hydrological variables	NASH	RRMSE (%)	RBIAS (%)
ANN- Linear CCA (Shu and Ouarda 2007)	QS10	0.84	37	-5
	QS100	0.78	45	-6
Optimal depth-based approach (Wazneh et al. 2013)	QS10	-	38	**-3**
	QS100	-	44	**-2**
Projection pursuit regression_STPW (Durocher et al. 2015)	QS10	0.82	34	-4
	QS100	0.79	40	-6
NLCCA-GAM	QS10	**0.87**	**23**	-4
	QS100	**0.82**	**28**	-5

Best results are in bold character.

Table 4. Jackknife Validation Results- Arkansas.

Regional model	Hydrological variables	NASH	RRMSE (%)	RBIAS (%)
CCA-LR	QS10	**0.75**	47.70	**-3.04**
	QS50	**0.73**	61.36	-5.76
CCA-ANN	QS10	0.71	63.58	-19.35
	QS50	0.71	66.27	-14.80
CCA-EANN	QS10	0.74	61.63	-16.83
	QS50	**0.73**	68.97	-19.12
CCA-GAM	QS10	0.74	40.54	-10.39
	QS50	0.72	52.37	-13.50
NLCCA-LR	QS10	0.72	37.23	6.27
	QS50	0.71	44.78	5.54
NLCCA-GAM	QS10	0.73	**31.10**	8.70
	QS50	0.72	**34.50**	8.40
NLCCA-ANN	QS10	0.65	49.71	8.16
	QS50	0.67	51.30	**2.71**
NLCCA-EANN	QS10	0.69	41.35	4.14
	QS50	0.70	45.93	3.66

Best results are in bold character.

Table 5. Jackknife Validation Results- Texas.

Regional model	**Hydrological variables**	NASH	RRMSE(%)	RBIAS(%)
CCA-LR	QS10	0.35	44.75	-7.56
	QS50	0.13	54.88	-4.11
CCA-ANN	QS10	0.49	52.46	-10.85
	QS50	0.46	58.82	-14.91
CCA-EANN	QS10	0.53	44.92	-14.90
	QS50	0.41	56.52	-18.75
CCA-GAM	QS10	0.55	40.24	-3.49
	QS50	0.49	44.72	-6.72
NLCCA-LR	QS10	0.53	42.85	-5.64
	QS50	0.44	51.11	-7.09
NLCCA-GAM	QS10	**0.68**	**30.7**	**-2.9**
	QS50	**0.61**	**38.4**	**-5.2**
NLCCA-ANN	QS10	0.56	43.26	-9.11
	QS50	0.53	45.70	-7.33
NLCCA-EANN	QS10	0.57	41.90	-12.65
	QS50	0.46	52.73	-16.40

Best results are in bold character.

List of Figures

Figure 1. Jackknife validation Results- Quebec. 132

Figure 2. Relative errors associated to Q_{S100} calculated at each site using CCA-GAM, NLCCA-LR, NLCCA-EANN and NLCCA-GAM. 133

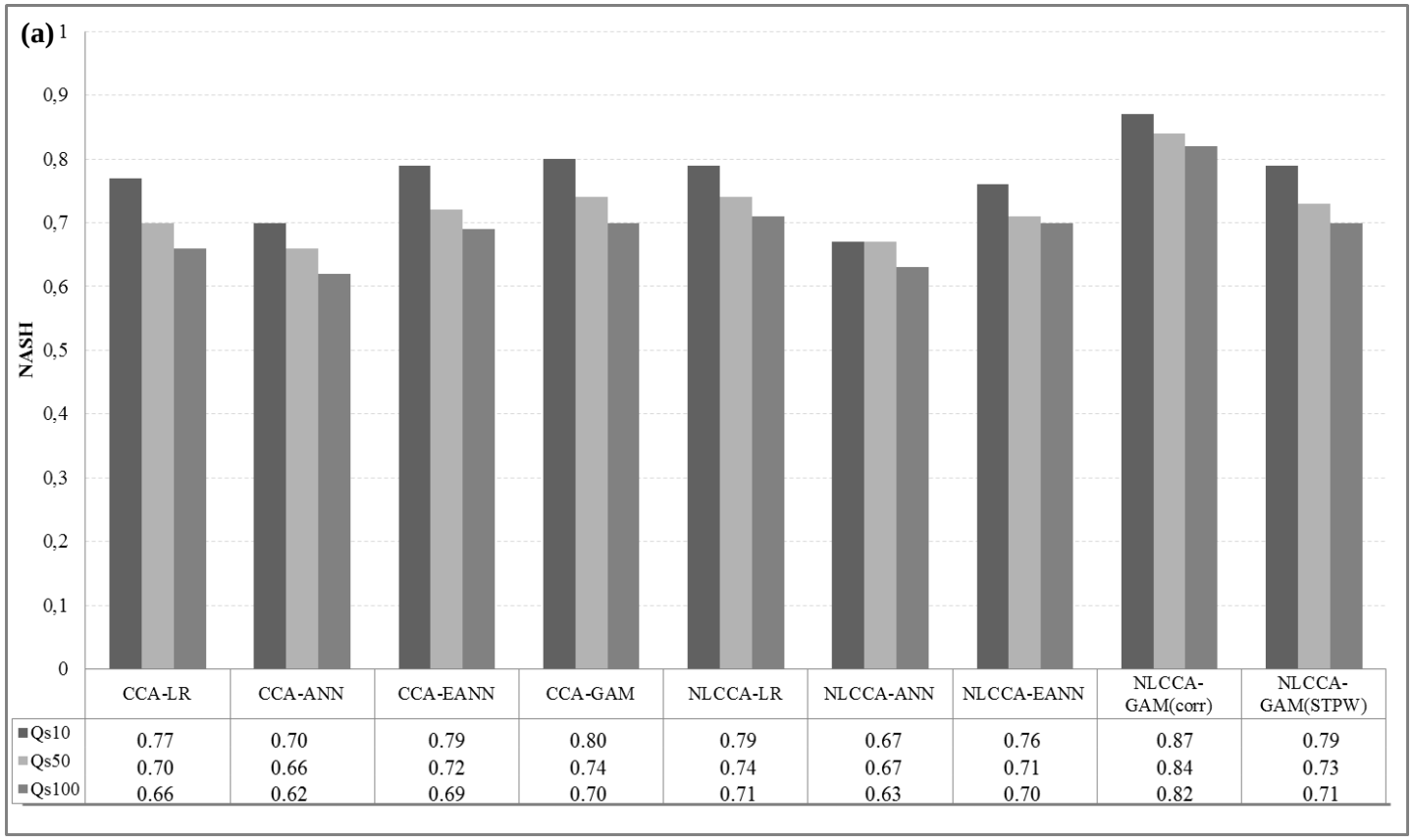
Figure 3. Relative errors using CCA-LR, CCA-GAM, NLCCA-LR and NLCCA-GAM as a function of Q_{S100} for Quebec. 133

Figure 4. Jackknife estimation using the CCA-LR, CCA-GAM, NLCCA-LR, NLCCA-EANN, and the NLCCA-GAM approaches for Q_{S100}. Red asterisks are associated to estimations at particular sites. 134

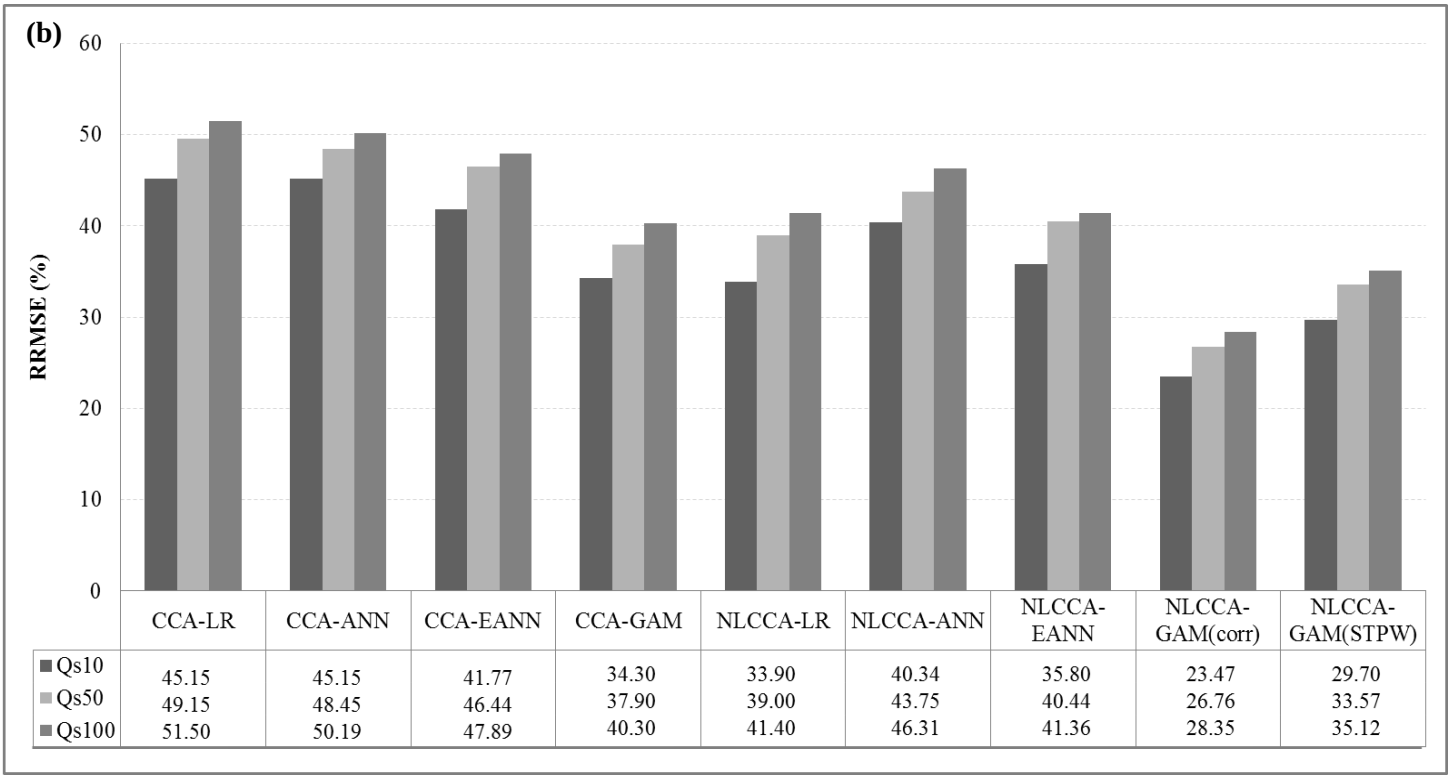
Figure 5. Geographical location of the identified particular stations in southern Quebec, Canada 135

Figure 6. Relative errors for identified problematic sites using several approaches, Q_{S100}. 135

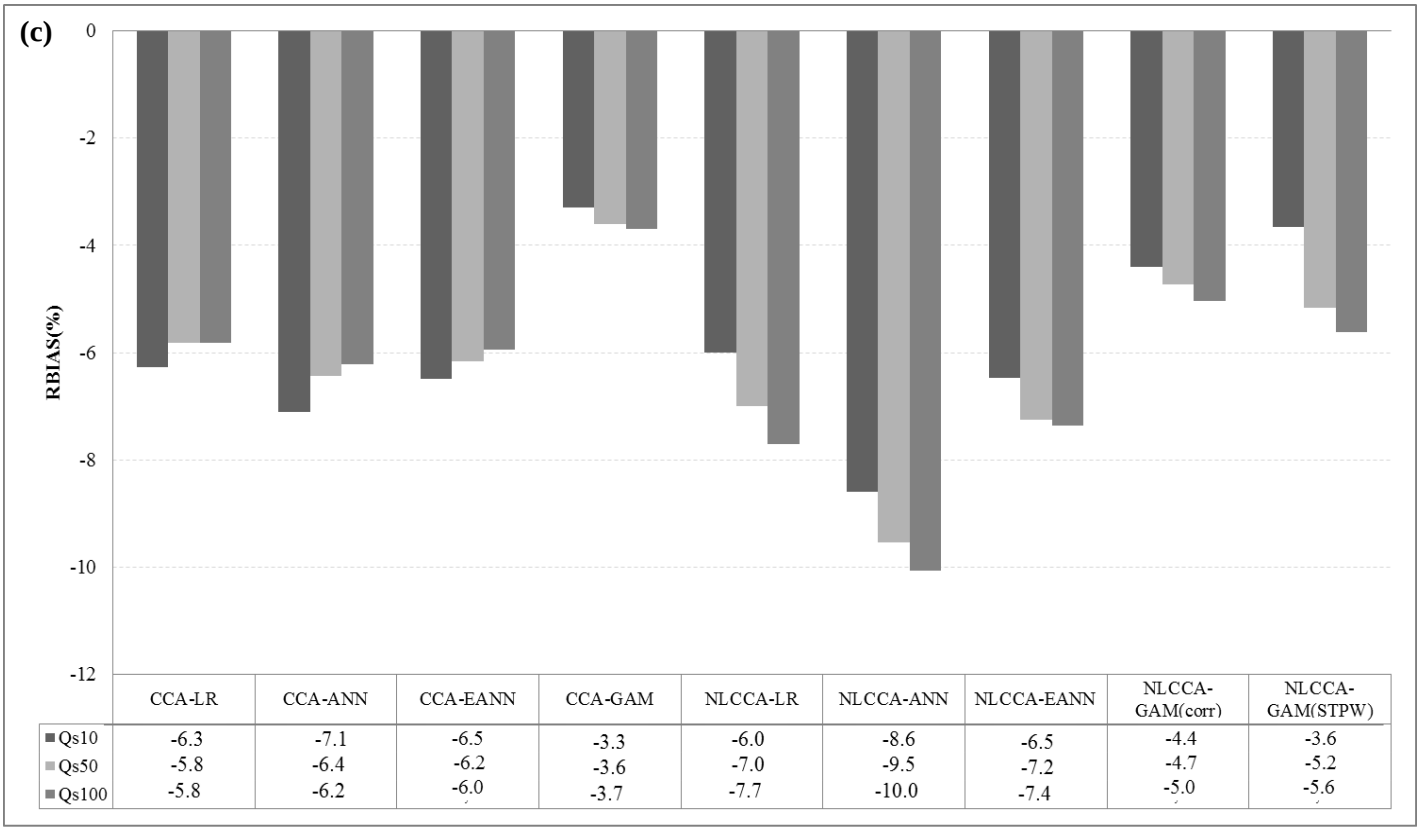
Figure 7. Smooth functions of Q_{S10} as a function of the explanatory variables included in the regional model NLCCA-GAM for Quebec, Arkansas and Texas. The dotted lines represent the 95% confidence intervals. The vertical axes are labelled s(var, edf), where var denotes the explanatory variable and edf denotes the estimated degree of freedom of the smooth. 136



	CCA-LR	CCA-ANN	CCA-EANN	CCA-GAM	NLCCA-LR	NLCCA-ANN	NLCCA-EANN	NLCCA-GAM(corr)	NLCCA-GAM(STPW)
Qs10	0.77	0.70	0.79	0.80	0.79	0.67	0.76	0.87	0.79
Qs50	0.70	0.66	0.72	0.74	0.74	0.67	0.71	0.84	0.73
Qs100	0.66	0.62	0.69	0.70	0.71	0.63	0.70	0.82	0.71



	CCA-LR	CCA-ANN	CCA-EANN	CCA-GAM	NLCCA-LR	NLCCA-ANN	NLCCA-EANN	NLCCA-GAM(corr)	NLCCA-GAM(STPW)
Qs10	45.15	45.15	41.77	34.30	33.90	40.34	35.80	23.47	29.70
Qs50	49.15	48.45	46.44	37.90	39.00	43.75	40.44	26.76	33.57
Qs100	51.50	50.19	47.89	40.30	41.40	46.31	41.36	28.35	35.12



	CCA-LR	CCA-ANN	CCA-EANN	CCA-GAM	NLCCA-LR	NLCCA-ANN	NLCCA-EANN	NLCCA-GAM(corr)	NLCCA-GAM(STPW)
Qs10	-6.3	-7.1	-6.5	-3.3	-6.0	-8.6	-6.5	-4.4	-3.6
Qs50	-5.8	-6.4	-6.2	-3.6	-7.0	-9.5	-7.2	-4.7	-5.2
Qs100	-5.8	-6.2	-6.0	-3.7	-7.7	-10.0	-7.4	-5.0	-5.6

Figure 1. Jackknife validation Results- Quebec.

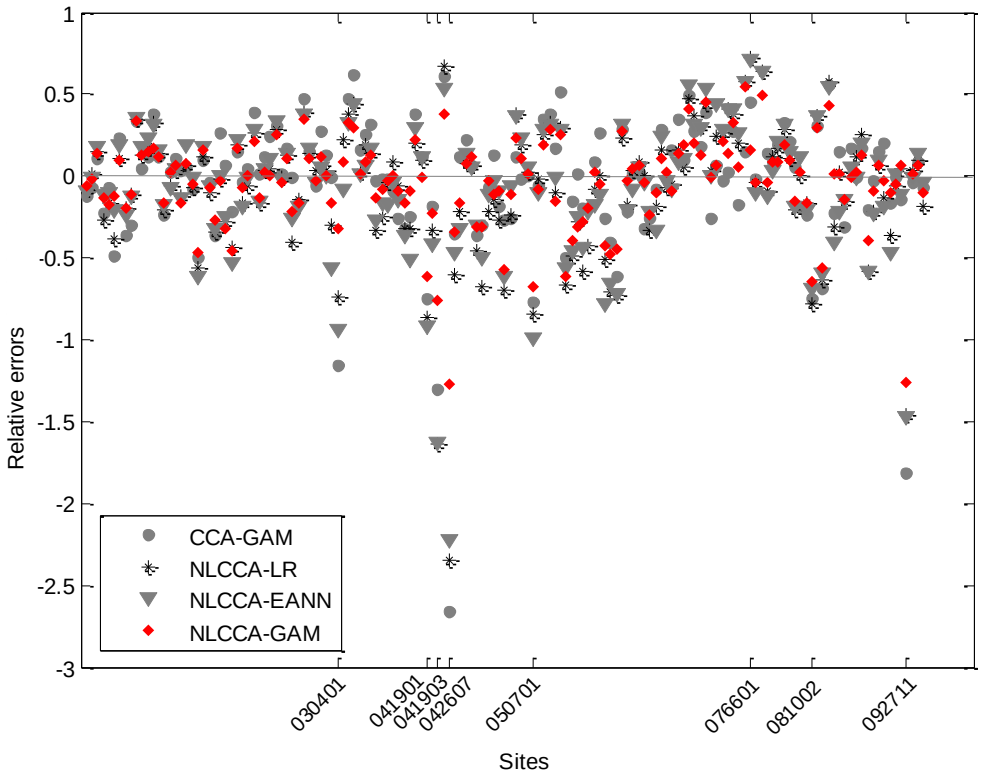


Figure 2. Relative errors associated to Q_{S100} calculated at each site using CCA-GAM, NLCCA-LR, NLCCA-EANN and NLCCA-GAM.

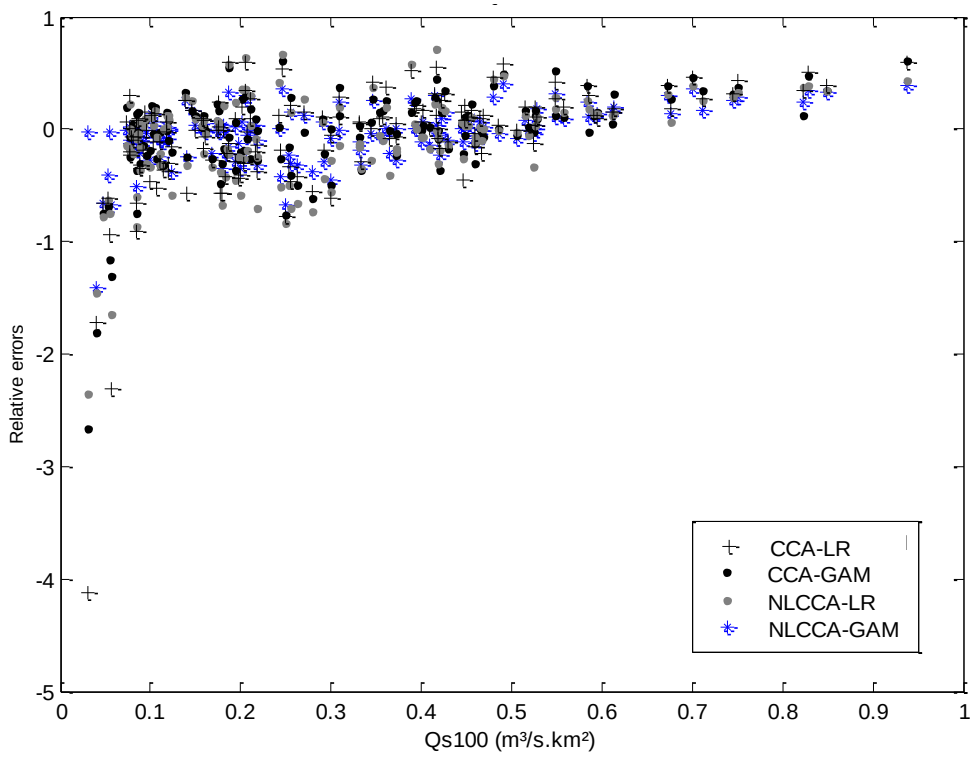


Figure 3. Relative errors using CCA-LR, CCA-GAM, NLCCA-LR and NLCCA-GAM as a function of Q_{S100} for Quebec.

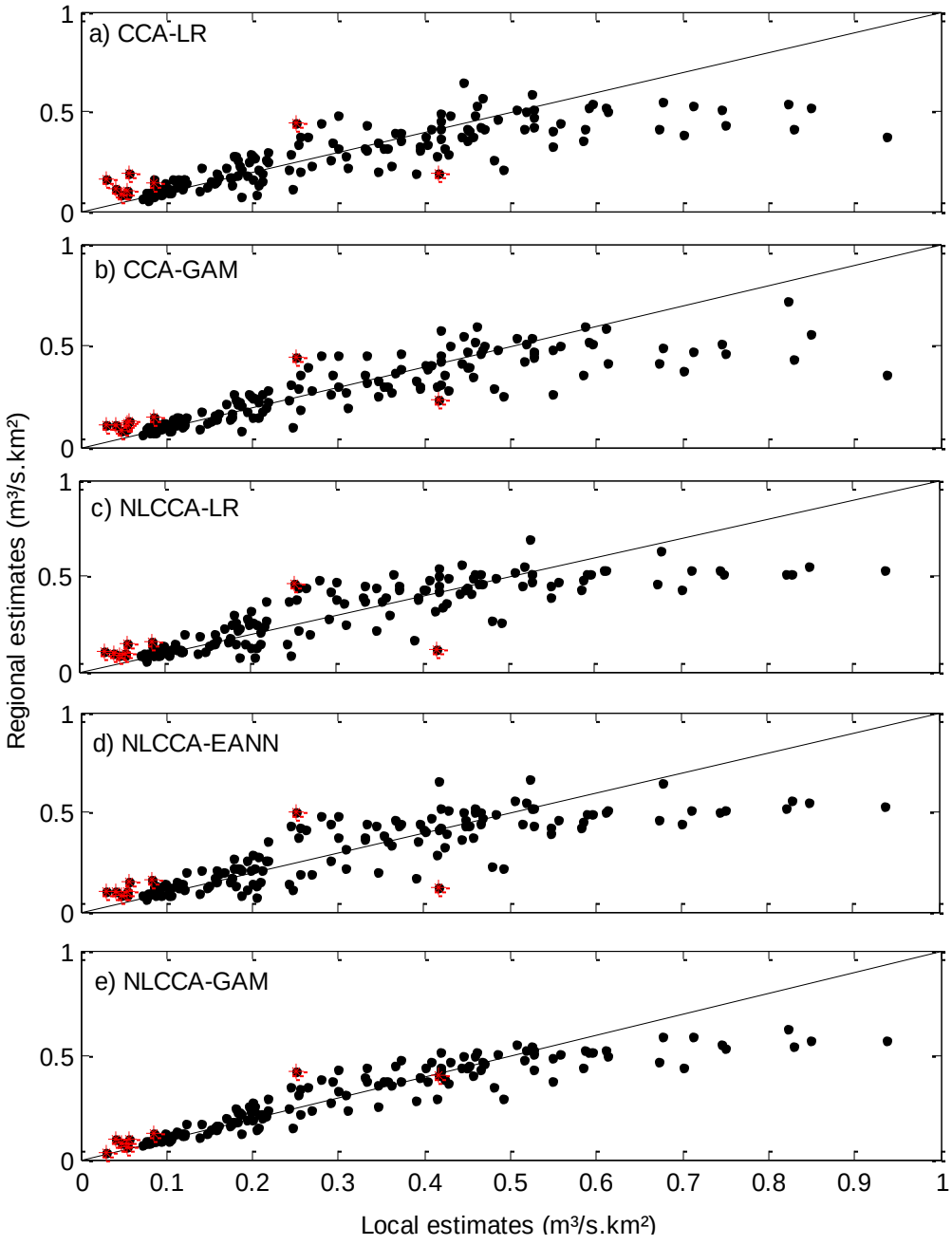


Figure 4. Jackknife estimation using the CCA-LR, CCA-GAM, NLCCA-LR, NLCCA-EANN, and the NLCCA-GAM approaches for Q_{S100}. Red asterisks are associated to estimations at particular sites.

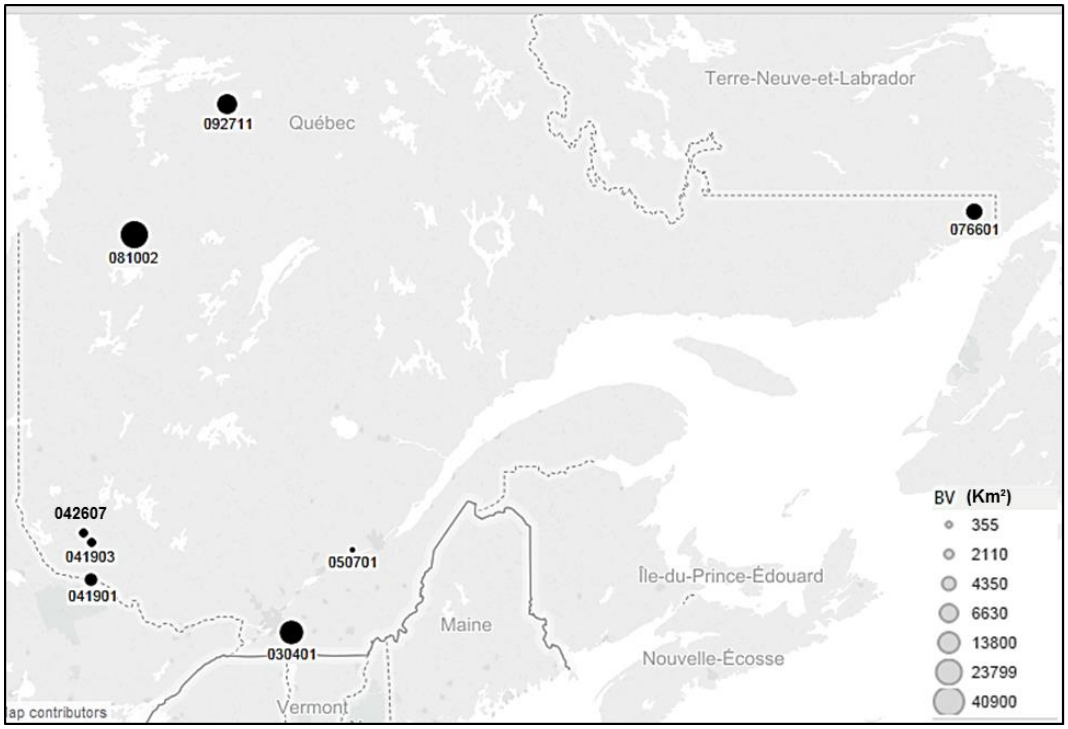


Figure 5. Geographical location of the identified particular stations in southern Quebec, Canada

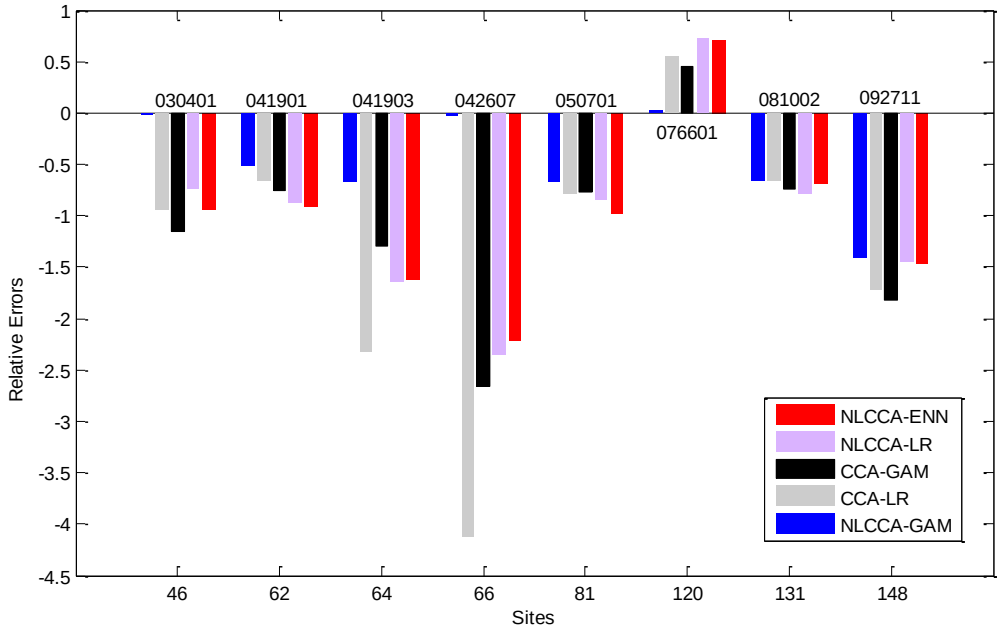


Figure 6. Relative errors for identified problematic sites using several approaches, Q_{S100}.

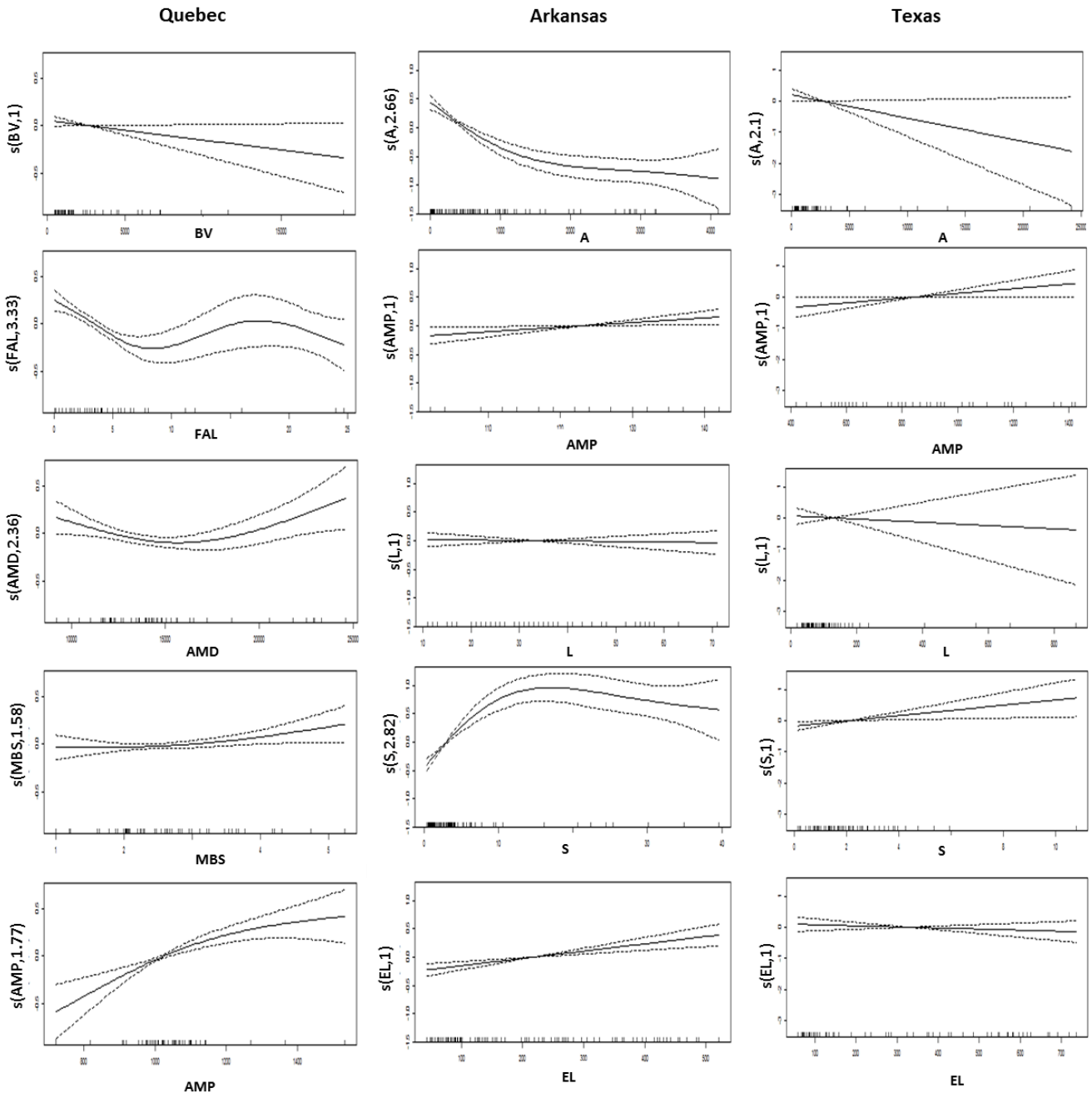


Figure 7. Smooth functions of Q_{S10} as a function of the explanatory variables included in the regional model NLCCA-GAM for Quebec, Arkansas and Texas. The dotted lines represent the 95% confidence intervals. The vertical axes are labelled s(var, edf), where var denotes the explanatory variable and edf denotes the estimated degree of freedom of the smooth.

CHAPITRE 4

QUANTILE REGRESSION IN REGIONAL FREQUENCY ANALYSIS: A BETTER EXPLOITATION OF THE AVAILABLE INFORMATION

Quantile regression in regional frequency analysis: a better exploitation of the available information

D. Ouali[*,1], F. Chebana[1], T.B.M.J. Ouarda[2,1]

[1]*Institut National de la Recherche Scientifique, Centre Eau Terre et Environnement,*

490, rue de la Couronne, Québec (Québec), G1K 9A9, Canada.

[2]*Institute Centre for Water Advanced Technology and Environmental Research (IWATER)*

Masdar Institute of science and technology

P.O. Box 54224, Abu Dhabi, UAE

[*]**Corresponding author:** Tel: +1 (418) 654 2530#4477

Email: dhouha.ouali@ete.inrs.ca

2016-02-05

Abstract

Classical regression models are widely used in hydrological regional frequency analysis (RFA) in order to provide quantile estimates at ungauged sites given physio-meteorological information. Since classical regression-based methods only provide the conditional mean of the response variable, estimated at-site quantiles at gauged sites are commonly used to calibrate the regression models in RFA. Generally, only at-site quantiles estimated with long data records are retained for the calibration and the evaluation steps, whereas hydrological information from stations with few data is ignored. In addition, even if the at-site quantiles are estimated with long data series, they are always subject to model selection and parameter estimation. Hence, their use for the calibration of the RFA models may induce significant uncertainties in the modeled relationships. The aim of this paper is to propose a quantile regression (QR) model that gives directly the conditional quantile for RFA, and avoids using at-site estimated quantiles in the calibration step. The proposed model presents another advantage where all the available hydrological information can be used in the calibration step including stations with very short data records. An evaluation criterion using observed data is also proposed in a cross-validation procedure. The proposed QR model is applied on a data set representing 151 hydrometric stations from the province of Quebec, and compared with a classical regression model. According to the proposed evaluation criterion, the QR is shown to be a viable model for regional estimations. Indeed, the proposed model proved to be robust and flexible, allowing considering all the region's sites, even those with extremely short flood records.

Keywords: quantile regression; classical linear regression; Koenker function; regional estimation, model evaluation, ungauged site.

1. Introduction and literature review

Frequency analysis (FA) is an operational tool commonly used in hydrological analysis. It is a crucial step in the analysis of hydrological risk enabling optimal water resource management and design of hydraulic structures. The procedure consists generally in identifying the probability distribution that fits best the observed data and hence provides adequate estimates of quantiles associated to specified return periods. In practice, this approach is privileged when enough hydrological information is available at the site of interest. However, the use of this technique becomes inefficient at-sites where little or no data are available. Regional FA (RFA) is rather used in such a case to estimate quantiles at ungauged sites (e.g. Cunnane 1988; Burn 1990; Castellarin et al. 2001; Hosking and Wallis 2005). In a RFA, to mitigate the lack of data, regional flood quantiles estimation is achieved via transferring information from gauged sites to the ungauged site (e.g. Burn 1990; Guse et al. 2010; Ouarda 2013; Chebana et al. 2014).

Ouarda et al. (2008) provided an overview of the various available RFA methods. Among regional flood quantile estimation methods, regression and index-flood models are equivalent and are superior to other models. The index-flood method makes the basic assumptions that data at different sites within a homogeneous region are independent and follow the same statistical distribution apart from a scale parameter that characterizes each site (e.g. Brath et al. 2001; Sveinsson et al. 2001; Javelle et al. 2002; Chebana and Ouarda 2009). Conversely, the regression model is a simple approach that allows the use of different distributions for different sites in the region (e.g. Pandey and Nguyen 1999; Shu and Burn 2004; Ouarda et al. 2006). Regression models use a transfer function to find a direct relationship between at-site quantiles (outputs) and physio-meteorological variables (predictors or inputs). They are commonly used in RFA because

of their ease of implementation, their rapidity and their good performance. In this regard, numerous models were proposed in RFA using different transfer functions including the linear regression model (e.g. Holder 1985; Phien et al. 1990; Pandey and Nguyen 1999; Prinzio et al. 2011), the generalized linear model (e.g. Nelder and Baker 1972), the generalized additive model (e.g. Chebana et al. 2014), the artificial neural networks (e.g. Abrahart and See 2007; Shu and Ouarda 2007).

The major drawback of regression-based methods is that they generally provide only the mean or the central part of the at-site quantiles. As a result, most regression methods are applied in RFA to provide the *conditional mean of the quantile* at ungauged sites given the physiographical variables (e.g. Pandey and Nguyen 1999; Ouarda 2013; Wazneh et al. 2013; Ouali et al. 2015). Hence, estimated quantiles at gauged sites are commonly used to calibrate the transfer function of the regression model in RFA and are not directly derived from the hydrological observations. For each quantile p, a regression model has to be performed including variable selection and parameter estimation. Generally, only quantiles estimated with long data series are retained for the calibration and the evaluation of the RFA model, while regional information from sites with few data is ignored. In addition, even if the at-site quantiles are estimated with long data series they are always inaccurate since different sources of uncertainties may occur (Arnell 1989; Girard et al. 2004; Hamed and Rao 2010). Hence, the use of at-site estimated quantiles for the calibration of the RFA model may induce significant biases in the modeled relationships. This makes the evaluation of the model performance more difficult, especially when statistical evaluation criteria such as the mean errors (ME) and the root mean square errors (RMSE) are computed using the at-site estimated quantiles. The availability of regression techniques that provide directly the conditional quantile (instead of the conditional mean) would be useful in

RFA in order to use the raw data for the model calibration (rather than using the estimated quantiles at gauged sites) as well as the appropriate evaluation criteria.

For this purpose, a quantile regression (QR) model is presented for RFA in the present paper. Unlike classical regression approaches, there is no need to perform at-site studies as a step to provide quantiles for the calibration of the regression model. The model proposed in this work presents another advantage: all regional information can be used to calibrate the QR model, including sites with short data records. Actually, even a single local observation can be considered in the QR approach. QR was developed by Koenker and Bassett (1978) in order to model the functional relationship between the predictors and conditional quantiles of the response distribution. More than a decade after, this technique started to receive considerable attention and a variety of applications were carried out in several fields (Koenker and Hallock 2001; Coad and Rao 2008). Examples of such applications fields include meteorology (e.g. Ben Alaya et al. 2015), economy (e.g. Melly 2005), medicine (e.g. Gebregziabher et al. 2011), ecology (e.g. Planque and Buffaz 2008) and education (e.g. Hartog et al. 2001). In term of statistical development, a number of variations of the QR model were proposed in the literature, such as the Single-index QR (Wu et al. 2010) and the penalized Single-index QR (Alkenani and Yu 2013). Several other studies including for instance Cheng et al. (2011), Koenker (2011) and Hu et al. (2012) presented advances in QR modeling. Despite all this diversity and progress in the QR literature, little attention was devoted to this approach in the water resources literature. For instance, Sankarasubramanian and Lall (2003) used QR model with both synthetic and real data to estimate flood quantiles under climate change circumstances. Cannon (2011) developed a neural network QR model in statistical downscaling of precipitations in order to identify the conditional distribution of a given day. In Villarini et al. (2011) authors performed a FA of the

annual maximum daily precipitation records where the QR approach has been used to investigate the stationarity assumption. Few other relevant studies have also integrated the QR tool when dealing with precipitation analysis, such as Tareghian and Rasmussen (2013) and Choi et al. (2014). In the present paper, the aim is to investigate the applicability, potential and benefits of the QR technique in the RFA context. The performance of the proposed approach is evaluated through a rigorous comparison with the classical regression model.

To avoid confusion, note that in some studies, for instance Palmen et al. (2011); Haddad and Rahman (2012), the terminology of QR refers rather to the classical regression model which require estimated at-site quantiles as inputs to provide quantiles estimates at ungauged sites.

The hydrological literature abounds with studies dealing with the development of new RFA models. However, much less attention has been dedicated to the development of new evaluation criteria. Traditional evaluation criteria such as the RMSE and the ME are commonly defined in a cross-validation procedure. Such statistical criteria require the availability of at-site quantiles. Therefore, since they are estimated they are not suitable in the cross-validation framework. For this purpose, an evaluation criterion directly based on raw data (rather than estimated at-site quantiles) is proposed using the Koenker loss function.

The remainder of this paper is organized as follows. Section 2 presents the theoretical background of linear regression as well as the QR models. In section 3, the adopted methodology for the adaptation of the QR model to the RFA framework is presented. In order to assess the potential of the proposed method, an evaluation criterion is developed. The QR model is then applied to a case study of the southern part of the province of Quebec, Canada. The considered data set is described in section 4. Obtained results of flood quantiles estimation corresponding to

the 10, 50 and 100 years return periods using both LR and QR models are given in Section 5. Section 6 includes a discussion of the implication of the findings to future research into the same field. The last section of the paper is dedicated to concluding remarks.

2. Theoretical Background

In this section, the adopted procedures in the current paper are described. A brief description of the statistical background of regression models is provided herein.

2.1. Linear regression model

Let $\mathbf{x} = (x_1, x_2, \ldots, x_m)$ denote a set of an observed explanatory variables and y be a given predictand or response variable. The linear regression model has the form:

$$y = \mathbf{x}^T \mathbf{b} + \varepsilon \tag{1}$$

where $\mathbf{b}$ is a vector of parameters and ε is the model error. The intercept of the vector of parameters can be included by adding 1 in the first element of $\mathbf{x}$. The vector $\mathbf{b}$ is generally estimated using the Ordinary Least Square (OLS) method (e.g. Hao and Naiman 2007). Under the assumption that errors are independent, unbiased and homoscedastic, the OLS method is the optimal predictor which yields the maximum likelihood estimates (Johnston and DiNardo 1972; Bro et al. 2002). Given a set of observations $(\mathbf{x}_i, y_i)$ for $i = 1, \ldots, n$, estimation using the OLS method consists in minimising with respect to $\mathbf{b}$ the sum of square errors (SSE):

$$SSE = \frac{1}{n} \sum_{i=1}^{n} \left(y_i - (\mathbf{x}_i^T \mathbf{b}) \right)^2 \tag{2}$$

The solution to this problem converges to the conditional mean of the response:

$$E(y_i \mid \mathbf{x}_i) = \mathbf{x}_i^T \mathbf{b} \tag{3}$$

Analogous to the conditional mean function of linear regression developed above, one may estimate the vector of parameters $\mathbf{b}$ by minimizing the sum of absolute errors (SAE):

$$SAE = \frac{1}{n} \sum_{i=1}^{n} \left| y_i - \mathbf{x}_i^T \mathbf{b} \right| \tag{4}$$

The solution of this optimization problem is given by the conditional median. In such a case, we generally speak about the median regression, or the least-absolute-deviations (LAD) regression (Ying et al. 1995). This kind of regression is more robust to outliers and non-normal errors than the OLS regression. Hence, a question which arises at this level is « Since the median corresponds to the 0.5 quantile, why not use other quantiles *p* in (0, 1)? ».

2.2. Quantile regression (QR)

Let us consider the following question: If the sample mean is the solution to the problem of minimizing a sum of squared errors (2), and the sample median is the solution to the problem of minimizing a sum of absolute residuals (4) , which optimization problem can have, as a solution, a given sample quantile of order *p*? By looking for the answer to this question, Koenker and Bassett (1978) introduced a new regression technique called quantile regression (QR) which provides the conditional quantile of the response variable given a set of predictors. Several applications of QR were carried out in the environmental sciences such as climatic change detection (Chamaillé-Jammes et al. 2007), air pollution prediction (Sousa et al. 2009) and

statistical precipitation downscaling (Friederichs and Hense 2007; Cannon 2011). In this paper, we adopt the QR model in the RFA context.

The regression equation of the conditional quantile is written as follows:

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

where $\mathbf{b}_p$ is a vector of parameters related to the $p^{th}(0 < p < 1)$ quantile $Q_p(y \mid \mathbf{x})$ of the conditional distribution of y given $\mathbf{x}$. Regression parameters are estimated through the minimization of the absolute deviation between observations and regression estimates weighted by the quantile *p*, denoted by the Koenker function (KF):

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

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

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

It is also known as the piecewise quantile loss function or the pinball loss function. The minimization problem in (6) can be conveniently solved by conventional linear programming methods.

Regarding the model assumptions, the basic QR model assumes residuals to be independent and identically distributed (iid) (Lee et al. 2014) and follow an Asymmetric Laplace Distribution (ALD) with a probability density function given by (Yu and Moyeed 2001) :

$$f(z) = p(1-p)\exp\{-\rho_p(z)\} \quad (8)$$

Note that, unlike the normal distribution, the ALD is more appropriate for high peaks and thick tail data. The reader is referred to Kozubowski and Podgórski (1999) for more details.

The QR model presents as well a number of attractive statistical properties that are absent in the LR model such as the invariance with respect to any monotonic transformation and the robustness against outliers. More theoretical aspects are detailed in Koenker (2005). Another important advantage of the QR model is the simultaneous estimation of quantiles with different orders p (e.g. Tokdar and Kadane 2011; Reich and Smith 2013). The simultaneous QR estimation can be seen as a solution to the crossing QR problem by imposing simultaneous non-crossing constraints (Liu and Wu 2011). Unlike the proposed QR-based approach, the classical regression modelling in RFA focuses only on few values of *p* and also requires conducting a new analysis for each new value of *p* (including each time variable selection, transformations, assumptions checking). Table 1summarizes the main differences between the classical LR model and the QR model.

3. Application to RFA

In hydrological RFA, the aim is to estimate flood quantiles at ungauged sites. However, classical regression models provide estimates of the conditional mean of the response variable. Consequently, these models are calibrated using estimated quantiles derived from a local FA which may generate unreliable results. In this section, the adaptation of the above statistical tools within the RFA context is briefly presented.

3.1. Regional models

Given *N* hydrologic stations, for a given station *i*, let Y_i be the hydrological vector of the series of maximum annual streamflows of length n_i, and y_{ij} be the j^{th} observation (maximum annual streamflows at year j for station i) where $j=1,\ldots, n_i$. For each station, let us define the vector of physio-meteorological variables $\mathbf{x}_i = (x_{i1}, x_{i2}, \ldots, x_{im})$ of dimension *m*.

Let q_{ip}^L be the quantile of order *p* estimated using local FA at a site *i* when the series of observations is of adequate length. Typically, a classical LR is performed using the log-linear regression function:

$$E\left(\log(\hat{q}_{ip}^L) \mid \mathbf{x}_i\right) = \mathbf{x}_i^T \mathbf{b}_p \tag{9}$$

where the vector of parameter $\mathbf{b}_p$ is estimated using the OLS estimator by considering the response variable $\log(q_i^L)$ rather than using observed data Y_i at a given site *i*. Once the vector of parameter $\mathbf{b}_p$ is estimated, the regional quantile at ungauged site *i'* can be estimated given the physio-meteorological variables $\mathbf{x}_{i'}$ using the following equation:

$$\hat{q}_{i'p}^R = \exp(\mathbf{x}_{i'}^T \hat{\mathbf{b}}_p) \tag{10}$$

This model is calibrated using at-site estimated quantiles q_{ip}^L. To ensure a good estimation quality, only sites with sufficiently long data series should be retained (such as more than $n_i = 30$ or $n_i = 40$ years), and these same sites are also retained for the validation of the model in a cross-validation (leave-one-out) procedure. For the remainder, let *S* denotes the number of sites with records

length exceeding a certain length ***l***. As mentioned above, the estimated at-site quantiles q^L present several sources of uncertainties, thus, they may induce significant errors in the modeled relationships quantiles/physiographical variables. In addition, the use of the log transformation, equation (9), may introduce additional biases to the model (Pandey and Nguyen 1999).

To address these limitations, we propose employing an adapted version of QR model to the RFA framework. The adaptation is related to a particular structure of RFA data. It can be seen that repeated measures from site to site is analogous to dose-response data commonly used in biostatistics (e.g. Harrell et al. 1996) or panel models (Hsiao 2014). However, within the latter, the model does not deal with quantiles. The conditional quantile of order ***p*** at a site ***i*** is written as:

$$q_{ip}^{R} = Q_p(Y_i \mid \mathbf{x}_i) = \mathbf{x}_i^T \mathbf{b}_p \tag{11}$$

where the parameters $\mathbf{b}_p$ are estimated using the following minimization problem:

$$\hat{\mathbf{b}}_p = \arg\min_{\mathbf{b}} \sum_{i=1}^{N} \left(\sum_{j=1}^{n_i} \rho_p \left(y_{ij} - \mathbf{x}_i^T \mathbf{b} \right) \right) \tag{12}$$

From (12), one of the main advantages of the QR model in RFA is that there is no need to perform at-site FA studies for each site to estimate at-site quantiles. As opposed to the traditional approach, a QR model is employed to directly establish the relationship between physio-meteorological variables **x** and observed annual maximum flood records Y**.** Hence, the whole available data set, even sites with very short records, can be employed in the calibration procedure. The use of the QR method would thus involve much more information than the traditional one. In addition, there is no need to employ a log transformation in the regression

model, unlike the classical regression approach which conducts to an additional bias. Figure 1 illustrates the steps involved in regional quantile estimation using both LR and QR approaches. One can identify differences concerning the calibration of the two considered approaches; when dealing with the conventional RFA, one has to carry out a FA at *each site* containing enough data records within the region of interest. This step could require an important time and experience since it includes for each site: i) the check of the basic assumption of FA including stationarity, independence and homogeneity; ii) the identification of the frequency distribution that fit the best each data series, iii) the estimation of the distribution parameters and iv) the estimation of the at-site quantile of order p (e.g. Chebana et al. 2013). These quantiles are then implemented in the LR model as the outputs or variables of interest. Using the proposed QR model, as illustrated in Figure 1, all observed data is directly inputted in the regression model to get the regional quantile without conducting at-site analysis.

3.2. Model quality assessment

One of the objectives of the present research is to evaluate the performance of the QR approach and to compare it to the classical LR model. All evaluation criteria commonly used for this purpose, namely the Nash criterion (NASH), the root mean squared error (RMSE), the relative RMSE (RRMSE), the mean bias (BIAS) and the relative mean bias (RBIAS), are established using at-site estimated quantiles. They consist in calculating the residual errors between the at-site *estimated* quantiles and the regional estimated ones. This approach considers estimated at-site quantiles as a perfect estimation. Indeed, the total error related to the regional LR model results from two main sources: a) the at-site estimation error (denoted $\boldsymbol{\varepsilon}$ in Figure 2) which is often not

considered in the assessment of regional modeling, and b) the regional error which is evaluated using classical criteria (Tasker and Moss 1979).

The proposed evaluation criterion in the current study, the Mean of the Piecewise Loss Function (MPLF), using the objective KF is expressed as follows:

$$MPLF\left(p\right)=\frac{10^3}{n}\sum_{i=1}^{N}\sum_{j=1}^{n_i}\rho_p\left(y_{ij}-\hat{q}_{ip}^{R}\right)\quad;\quad p\in(0,1)$$

(13)

where *n* denotes the total number of observations in all stations combined, $n=\sum_{i=1}^{N}n_i$.

The rationale of this criterion is to assess the performance of adopted models through the use of raw observed data. The lower the criterion value is, the more suitable is the considered model. As indicated in Koenker and Machado (1999), the use of the optimal value of the loss function as a goodness of fit measure is a very natural idea commonly used in the robust literature. Based on this argument, the MPLF criterion was calculated by summing up the values of KF computed at each site, and then standardized by the total number of observations at all sites. Thus, this calculation procedure will attribute a weight of $\frac{n_i}{n}$ to each site *i* depending on its number of observations, in other words, giving more importance to sites with long data series.

The concept behind this criterion as well as the classical ones is explained graphically in Figure 3. As indicated in this figure, these criteria (proposed and traditional ones) are calculated based on a cross validation procedure (e.g. Ouarda et al. 2001). It consists in removing temporarily each

site and considering it as an ungauged one. A new flood record value is thus estimated and the ability of each method is then evaluated. This figure illustrates the difference between classical evaluation indices and the proposed one. In reality, when using the classical approach, the error related to at-site estimation is not included in the evaluation procedure. Indeed, the at-site *estimated* quantiles are considered as references although they are not observed data. The quality of the at-site estimated quantiles depends strongly on the record lengths (Tasker and Moss 1979). Consequently, in order to ensure a minimum reliability, the RMSE criterion should be calculated over a subset **e** of sites with enough record lengths, for instance more than 30 years. Using the MPLF criterion, there is no need to reduce the dataset neither for the fitting nor for the evaluation step.

The basic concepts behind the proposed criterion are justified in a similar way as in Koenker and Machado (1999). In this study, the authors developed the R^1 statistic as an analog of the coefficient of determination R^2, which is a commonly used statistic to evaluate LR models, to be used for assessing QR model performance. It consists in using the piecewise quantile loss function instead of the sum of the squared error used in the traditional LR models. More theoretical details about this measure can be found in Koenker and Machado (1999) and Hao and Naiman (2007). In a cross validation procedure, the RMSE is often employed to assess the traditional LR model instead of R^2. In this respect, since R^1 was proposed as a natural analog of R^2, the proposed MPLF can be considered as an analog of the RMSE criterion.

Let us also define the Piecewise error (PE) for each site to be the natural analog of the absolute error:

$$PE_i(p) = \frac{1}{n_i}\sum_{j=1}^{n_i}\rho_p\left(y_{ij} - \hat{q}_{ip}^{R}\right) \quad ; \quad p \in (0,1) \text{ and } i = 1,...,N \qquad (14)$$

4. Case study and study design

The proposed procedure was applied to the dataset issued from the hydrometric station network of southern Quebec, provided by the Quebec Ministry of the Environment (Province of Quebec, Canada). One hundred fifty one stations, located between 45 and 55°N in the southern part of Quebec are selected. Two types of variables are considered: physio-meteorological variables and hydrological variables. The physio-meteorological variables are those used previously by Chokmani and Ouarda (2004) namely: the mean basin slope (PMBV) in %, the basin area (BV) in km^2, the proportion of the basin area covered by lakes (PLAC) in %, the annual mean total precipitation (PTMA) in mm and the annual mean degree-days over 0°C (DJBZ) in degree-day. Hydrological variables are at-site specific flood quantiles, denoted Q_{ST} in $m^3/s.km^2$, corresponding to return periods T= 10, 50 and 100 years. For each site, the most appropriate statistical distribution has been identified in order to estimate at-site quantiles for different return periods.

For the proposed approach, we use the same information as for the classical regional estimation methods (the flood record at gauged sites) but in a different way: we use the raw flow data available at each station for the period between 1900 and 2002, rather than using the processed quantile data at the gauged sites. Figure 4 represents the spatial variation of length of historical data records for each site ranging from 15 to 84 years.

The present study seeks to address the classical RFA gaps through: i) providing a direct quantile estimation without performing an at-site FA and ii) proposing an adopted evaluation criterion. This is performed in several steps:

- Apply and compare the considered models (QR and LR) using different criteria; the calibration and the application of both models are performed using the entire data.
- Take into account the at-site quantile estimation quality; modify the data used for the calibration step.
- Consider a more suitable case for which the LR performs well and the QR advantages are accounted for; the QR model built and assessed using the entire data / the LR model built using only sites with record length exceeding 30 years and evaluated using the entire data.
- Compare both models using the MPLF criterion; the concept of this criterion permits the model assessment using the entire data set.

5. Results

Results related to the application of both QR and LR using the whole dataset are initially reported. Then, the next step consists in investigating the effect of long local data series through a comparison of the results of the various models. Finally, comparison results of the two models based on the MPLF criterion are presented.

By considering the entire set of 151 sites, regional estimated quantiles using both LR and QR models are compared with those obtained from the at-site estimated quantiles. Obtained results are illustrated in Figure 5. It is important to recall that the at-site quantile estimates are considered as reference values, thus, regional estimation is as accurate as it is closer to the at-site estimation. A comparison of the two results reveals that the LR reproduces more adequately the at-site estimated quantiles than the QR model which generally overestimates them. Associated BIAS and RMSE values of each case are also reported. One can see that quantile estimates obtained from the LR model are less biased (i.e. smaller BIAS) and more accurate (i.e. smaller RMSE)

than QR quantiles. Conversely, the QR model was found to be more biased and least accurate. This finding is expected and can be explained by the definition of these criteria which are, by construction, based on the at-site estimated quantiles. Indeed, regardless of the quality of the at-site processed data, these criteria have been often used to measure the regional flood quantile estimation performance. Recall that the BIAS is related with the deviation from the true value while the RMSE is associated with the model accuracy. Thus, a good estimation quality is systematically related to long data series since the at-site estimation error, denoted $\boldsymbol{\varepsilon}$ in the illustration in Figure 2, will decrease. The longer the record data is, the more reliable the at-site estimation will be. This fact was also illustrated in Figure 5 using the colored map, where short data series are presented from gray to white and long data series are presented in black. It can be seen that the RMSE and BIAS are insensitive to the sites record lengths, meaning that both short and long records are weighted identically.

In order to quantify the at-site estimation error and to assess the record length effect on the estimation quality, we proceeded by Monte Carlo simulation (e.g. Arora and Singh 1989). For a selected site, with long data series, the appropriate distribution was identified and the quantile of interest (Q_{ST}^{Obs}) was calculated. Then, for each record length (from 1 to 100 years) 100 series were randomly generated from the identified distribution and the associate quantile was estimated (Q_{ST}^{est}). The examination of the obtained results reported in Figure 6 shows that, as pointed out above, the quadratic error (defined as $QE = (Q_{ST}^{Obs} - Q_{ST}^{est})^2$) decreases when the record length increases. Another interesting finding is that the RMSE values corresponding to the mean record length of the data set (32 years) (at-site RMSE) for Q_{S10}, Q_{S50} and Q_{S100} are respectively 8.77 10^{-5}, 1.46 10^{-4} and 2.25 10^{-4} $m^3/s.km^2$. Hence, an average quantification of the total at-site RMSE

may be obtained by multiplying the RMSE related to the mean record length 151 times (respectively 0.013, 0.022 and 0.034 $m^3/s.km^2$).

An important question that can be raised at this level is the effect of considering longer records, which systematically implies better at-site quantile estimation quality. In this respect, quantile estimation at ungauged sites was performed using different minimum record lengths $l \in \{15, 20, 30, 40, 50, 60\}$ linked to S sites ranging from 151 sites (associated to l =15 year, i.e., all sites) to only 14 sites, (Figure 7). The differences in the performance criterion RMSE using the LR and QR approaches are presented in Figure 8. It can be seen, from Figure 8-a and 8-b, that the LR method leads to a better performance in all considered cases. However, as pointed out previously, from a conceptual viewpoint, the consideration of such evaluation criterion is in favor of the LR model since both are based on the estimated values of flood quantiles. Thus, it would be advantageous to assess models with an objective tool such as the above-mentioned MPLF.

In Table 2 validation results of both LR and QR models using the proposed observation-based MPLF criterion are presented. It can be seen that QR outperforms the LR model. This finding remains the same when considering different record lengths for the LR calibration. Figure 8-c and 8-d summarize the results of the LR method and the QR method in terms of MPLF for Q_{S50} and Q_{S100}, respectively. It is obvious that, for different $l \in \{15, 20, 30, 40, 50, 60\}$, the QR model shows a better performance than the classical model (i.e. smaller MPLF). Note also that, for higher record lengths (meaning fewer considered sites for the calibration step) the LR performance decreases. On the other hand, since QR and the MPLF criterion do not depend on the at-site quantile estimation, the MPLF values associated to the QR approach are always constant.

One of the objectives of the present study was to tangibly assess (in reference to observed data) the performance of the two regional estimation approaches. To achieve this, comparison results using the MPLF criterion were further developed. For each site, the PE given in (14) was calculated and the differences between the LR and the QR approaches were highlighted in Figure 9 for Q_{S10}, Q_{S50} and Q_{S100}. This figure indicates that for long data series (deep black points) both models behave well, meaning that the errors are almost equal and small compared to those corresponding to short records (clear points). Furthermore, it can be seen that QR errors, especially for short record lengths, are smaller than LR errors, which is in accordance with our earlier findings. We have also identified some problematic sites associated to high PE for both models, namely sites with identification numbers: 030415, 073301, 076601 and 080104.

It is also of interest to verify whether the basin size has an effect on the performance criterion. Figure 10 illustrates the PE of Q_{S100} in all sites for the LR model (a) and the QR model (b). It was effectively noticed that the basin size seems to influence the performance criterion. Indeed, in both cases (LR and QR), the error was larger for smaller basins in comparison to larger ones. Note also that the QR errors are less scattered than the ones resulting from the classical model. This confirms the robustness of the QR model.

6. Discussion

This work was motivated by the weakness of the classical approach. Indeed, most traditional regression-based approaches are conceived to answer the question ‘’how to transfer estimated quantiles from gauged sites to an ungauged one?’’ Consequently, these classical approaches depend strongly on the quality of at-site quantile estimates which in turn depend on the choice of the distribution function, the parameter estimation method as well as the length of the data series.

To address these limitations, the proposed approach seeks an answer to the following question: ‘’why not estimate quantiles in an ungauged site by transferring information from observed hydrological data instead of using estimated quantiles?’’ The answer to this question resides in the principle of the traditional regression model. Because classical regression models give only the conditional mean of the response variable, traditional regression-based approaches in RFA are often performed using the processed quantile data at the gauged sites rather than using raw flow data. Evaluation of the at-site estimation error is often performed using simulations. In the present work, an attempt to assess the at-site error was elaborated using the Monte Carlo simulations. Presented results indicate that, somewhat surprisingly, for the mean record length, the at-site error is relatively high. To avoid training regression models with at-site estimated quantiles, the QR model is introduced in the present study in the RFA context. An evaluation criterion is also proposed based on raw data as reference to assess the models performance. This forms the main advantage of the MPLF criterion. The second advantage is that it includes information not only from long historic records, but also from short ones. Unlike the classical approach, this leads to retaining all available information.

From the results of this study, one may conclude to the good performance of the proposed criterion. Furthermore, when dealing with this criterion, each site has a weight proportionally on its data record length. Hence, the longer the data series is the greater the corresponding weight is. This is quite the opposite of the classical criteria, such as the RMSE, which are not sensitive to the data size characteristics, i.e., the site record length or the region size.

An alternative application of QR in RFA framework could be performed by generating data at ungauged sites. The idea consists on generating a set of quantiles given only physio-meteorological variables at a site. Indeed, the QR model can be used to estimate the conditional

distribution through generating quantiles randomly between 0 and 1 (Cannon 2011). The resulted sample is then compared to the observed one using non-parametric tests namely the two-sample Kolmogorov-Smirnov and the Kruskal-Wallis tests. This procedure was applied in the present work for some selected sites, supposed to be ungauged, with 5% significance level. Obtained results, in terms of p-values, suggest that for all chosen sites the observed and simulated samples come from the same continuous distribution. This finding may serve as a base for several practical applications within the RFA context.

Another important issue is related to the first step of the traditional RFA, namely the delineation of homogeneous regions DHR. This latter is a challenging step which consists in grouping sites within homogeneous regions, in other words, identifying groups of stations having similar hydrological, meteorological and physiographical behaviors. This would significantly improve quantile estimates at ungauged sites. Actually, when compared to the entire regions results, the DHR-based approaches provide the smallest standard error values (Gingras and Adamowski 1993). In the current work, we opted for two regional (one step) models to estimate regional flood quantiles at ungauged sites without identifying sites having a similar hydrological behavior to the target site. Indeed, the proposed methodology aimed only to compare the two regression models. It could be important to assess the true performance of the QR method in association with a number of commonly used DHR methodologies (e.g. Ouali et al. 2015).

Moreover, it should be noted that, as it is the case for the most of RFA models, the proposed approach is based on the so-called stationary assumption, meaning estimated quantiles for a given site will not vary significantly over time. An investigation of this assumption has been performed using the non-parametric Spearman test (Yue et al. 2002; Villarini et al. 2011). Application of this test to the maximum annual streamflow series for each gauged station, indicates that 3

stations over 151 are found to be non-stationary at a significance level of 1% (Table 3), as found by Kouider (2003). Given the small percentage of rejected stations (2%) and to maximize sources of information, these stations have been retained in this study as it was the case in previous studies (e.g. Chokmani and Ouarda (2004)). Note that with a 5 % significance level, results show that 20 stations are non-stationary. In this case, the proposed approach is also considered on the remaining stationary stations. The obtained results are similar to those with all stations and led to the same conclusions.

As a matter of fact, under changing climate conditions, flood extremes at a given site could be altered. Hence, we argue that the RFA models, such as the proposed approach, should be adapted to account for such a context in future work.

Further potential directions for future research would be to consider non-crossing constraints when dealing with the QR approach. Indeed, the crossing QR is a serious modelling problem which may leads to an invalid response distribution. To address this problem several solutions are proposed in the statistical literature. Indeed, depending on the goal of the modelling procedure two main approaches exists namely the direct approaches (e.g. Bondell et al. 2010; Liu and Wu 2011) whenever the focus is on estimating the entire parametric distribution, and the indirect approaches (e.g. Hall et al. 1999; Dette and Volgushev 2008; Chernozhukov et al. 2010) when the interest is to estimate particular quantiles. In the present work since the focus is on adequately estimate few quantile orders, an indirect estimation approach is required. This consists on determining the conditional cumulative distribution function and then inverting it to get the desirable quantile estimates.

7. Conclusions

QR is a tool often used in several fields such as economy, health, ecology and environmental studies including hydrology. Nevertheless, it remains unutilized in the specific hydrological RFA framework, even though the quantile is a very important quantity to estimate. It is the LR model that is commonly employed as an estimation model in RFA essentially because of its simplicity. In this work, we address the important issue of including at-site FA as a step within RFA and hence employing estimated at-site quantiles in RFA. The purpose of the present study is first to consider observed data directly in the RFA using the QR model, then to evaluate the estimation performance of the two regional models (LR and QR) through an objective criterion.

Initially, the efficiency of the QR model was investigated in comparison to the classical approach. The QR model was considered in the RFA in order to estimate flood quantiles at ungauged sites. At a second step the proposed criterion was applied within the same framework to ensure a fair model assessment. The methodology is validated on a real case study for different quantile orders. The developments are made using several datasets for each minimum record length l, and the performance criteria are calculated as a function of l. Then, the relevance of both models in terms of MPLF criterion was explored in each site and according to the basin area.

Overall, we can conclude that the proposed approach is a promising method for the estimation and evaluation of flood quantiles at-sites with short to medium length records.

Acknowledgments

Financial support for this study was graciously provided by the Natural Sciences and Engineering Research Council (NSERC) of Canada. The authors thank the ministry of the environment of Quebec for having provided the employed data sets.

References

Abrahart, R. and L. See (2007). "Neural network modelling of non-linear hydrological relationships." Hydrology and Earth System Sciences **11**(5): 1563-1579.

Alkenani, A. and K. Yu (2013). "Penalized Single-Index Quantile Regression." International Journal of Statistics and Probability **2**(3): p12.

Arnell, N. W. (1989). "Expected annual damages and uncertainties in flood frequency estimation." Journal of Water Resources Planning and Management **115**(1): 94-107.

Arora, K. and V. P. Singh (1989). "A comparative evaluation of the estimators of the log Pearson type (LP) 3 distribution." Journal of Hydrology **105**(1): 19-37.

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

Bondell, H. D., B. J. Reich and H. Wang (2010). "Noncrossing quantile regression curve estimation." Biometrika **97**(4): 825-838.

Brath, A., A. Castellarin, M. Franchini and G. Galeati (2001). "Estimating the index flood using indirect methods." Hydrological sciences journal **46**(3): 399-418.

Bro, R., N. D. Sidiropoulos and A. K. Smilde (2002). "Maximum likelihood fitting using ordinary least squares algorithms." Journal of Chemometrics **16**(8-10): 387-400.

Burn, D. H. (1990). "Evaluation of regional flood frequency analysis with a region of influence approach." Water Resources Research **26**(10): 2257-2265.

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

Castellarin, A., D. Burn and A. Brath (2001). "Assessing the effectiveness of hydrological similarity measures for flood frequency analysis." Journal of Hydrology **241**(3): 270-285.

Chamaillé-Jammes, S., H. Fritz and F. Murindagomo (2007). "Detecting climate changes of concern in highly variable environments: Quantile regressions reveal that droughts worsen in Hwange National Park, Zimbabwe." Journal of Arid Environments **71**(3): 321-326.

Chebana, F., C. Charron, T. Ouarda and B. Martel (2014). "Regional Frequency Analysis at Ungauged Sites with the Generalized Additive Model." Journal of Hydrometeorology **15**(6): 2418-2428.

Chebana, F. and T. Ouarda (2009). "Index flood–based multivariate regional frequency analysis." Water Resources Research **45**(10).

Chebana, F., T. B. Ouarda and T. C. Duong (2013). "Testing for multivariate trends in hydrologic frequency analysis." Journal of Hydrology **486**: 519-530.

Cheng, Y., J. G. De Gooijer and D. Zerom (2011). "Efficient estimation of an additive quantile regression model." Scandinavian Journal of Statistics **38**(1): 46-62.

Chernozhukov, V., I. Fernandez-Val and A. Galichon (2010). "Quantile and probability curves without crossing." Econometrica: 1093-1125.

Choi, W., R. Tareghian, J. Choi and C. s. Hwang (2014). "Geographically heterogeneous temporal trends of extreme precipitation in Wisconsin, USA during 1950–2006." International Journal of Climatology **34**(9): 2841-2852.

Chokmani, K. and T. B. M. J. Ouarda (2004). "Physiographical space-based kriging for regional flood frequency estimation at ungauged sites." Water Resources Research **40**(12).

Chokmani, K. and T. B. M. J. Ouarda (2004). "Physiographical space-based kriging for regional flood frequency estimation at ungauged sites." Water Resources Research **40**(12).

Coad, A. and R. Rao (2008). "Innovation and firm growth in high-tech sectors: A quantile regression approach." Research Policy **37**(4): 633-648.

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

Dette, H. and S. Volgushev (2008). "Non-crossing non-parametric estimates of quantile curves." Journal of the Royal Statistical Society: Series B (Statistical Methodology) **70**(3): 609-627.

Friederichs, P. and A. Hense (2007). "Statistical downscaling of extreme precipitation events using censored quantile regression." Monthly weather review **135**(6): 2365-2378.

Gebregziabher, M., C. Lynch, M. Mueller, G. Gilbert, C. Echols, Y. Zhao and L. Egede (2011). "Using quantile regression to investigate racial disparities in medication non-adherence." BMC medical research methodology **11**(1): 88.

Gingras, D. and K. Adamowski (1993). "Homogeneous region delineation based on annual flood generation mechanisms." Hydrological sciences journal **38**(2): 103-121.

Girard, C., T. B. Ouarda and B. Bobée (2004). "Étude du biais dans le modèle log-linéaire d'estimation régionale." Canadian Journal of Civil Engineering **31**(2): 361-368.

Guse, B., A. H. Thieken, A. Castellarin and B. Merz (2010). "Deriving probabilistic regional envelope curves with two pooling methods." Journal of Hydrology **380**(1): 14-26.

Haddad, K. and A. Rahman (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**: 142-161.

Hall, P., R. C. Wolff and Q. Yao (1999). "Methods for estimating a conditional distribution function." Journal of the American Statistical Association **94**(445): 154-163.

Hamed, K. and A. R. Rao (2010). Flood frequency analysis, CRC press.

Hao, L. and D. Q. Naiman (2007). Quantile regression, Sage.

Harrell, F., K. L. Lee and D. B. Mark (1996). "Tutorial in biostatistics multivariable prognostic models: issues in developing models, evaluating assumptions and adequacy, and measuring and reducing errors." Statistics in medicine **15**: 361-387.

Hartog, J., P. T. Pereira and J. A. Vieira (2001). "Changing returns to education in Portugal during the 1980s and early 1990s: OLS and quantile regression estimators." Applied Economics **33**(8): 1021-1037.

Holder, R. (1985). Multiple regression in hydrology, Institute of hydrology.

Hosking, J. R. M. and J. R. Wallis (2005). Regional frequency analysis: an approach based on L-moments, Cambridge University Press.

Hsiao, C. (2014). Analysis of panel data, Cambridge university press.

Hu, Y., R. B. Gramacy and H. Lian (2012). "Bayesian quantile regression for single-index models." Statistics and Computing: 1-18.

Javelle, P., T. B. Ouarda, M. Lang, B. Bobée, G. Galéa and J.-M. Grésillon (2002). "Development of regional flood-duration–frequency curves based on the index-flood method." Journal of Hydrology **258**(1): 249-259.

Johnston, J. and J. DiNardo (1972). "Econometric methods." New York **19**(7): 22.

Koenker, R. (2005). Quantile regression, Cambridge university press.

Koenker, R. (2011). "Additive models for quantile regression: Model selection and confidence bandaids." Brazilian Journal of Probability and Statistics **25**(3): 239-262.

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

Koenker, R. and K. Hallock (2001). "Quantile regression: An introduction." Journal of Economic Perspectives **15**(4): 43-56.

Koenker, R. and J. A. Machado (1999). "Goodness of fit and related inference processes for quantile regression." Journal of the American Statistical Association **94**(448): 1296-1310.

Kouider, A. (2003). Analyse fréquentielle locale des crues au Quebec, Université du Québec.

Kozubowski, T. J. and K. Podgórski (1999). "A class of asymmetric distributions." Actuarial Research Clearing House **1**: 113-134.

Lee, E. R., H. Noh and B. U. Park (2014). "Model selection via Bayesian information criterion for quantile regression models." Journal of the American Statistical Association **109**(505): 216-229.

Liu, Y. and Y. Wu (2011). "Simultaneous multiple non-crossing quantile regression estimation using kernel constraints." Journal of nonparametric statistics **23**(2): 415-437.

Melly, B. (2005). "Public-private sector wage differentials in Germany: Evidence from quantile regression." Empirical Economics **30**(2): 505-520.

Nelder, J. A. and R. Baker (1972). Generalized linear models, Wiley Online Library.

Ouali, D., F. Chebana and T. Ouarda (2015). "Non-linear canonical correlation analysis in regional frequency analysis." Stochastic Environmental Research and Risk Assessment: 1-14.

Ouarda, T. B. M. J. (2013). "Hydrological Frequency Analysis, Regional." Encyclopedia of Environmetrics: DOI:10.1002/9780470057339.vnn9780470057043.

Ouarda, T. B. M. J., J. Cunderlik, A. St-Hilaire, M. Barbet, P. Bruneau and B. Bobée (2006). "Data-based comparison of seasonality-based regional flood frequency methods." Journal of Hydrology **330**(1): 329-339.

Ouarda, T. B. M. J., C. Girard, G. S. Cavadias and B. Bobée (2001). "Regional flood frequency estimation with canonical correlation analysis." Journal of Hydrology **254**(1-4): 157-173.

Ouarda, T. B. M. J., A. St-Hilaire and B. Bobée (2008). "Synthèse des développements récents en analyse régionale des extrêmes hydrologiques." Revue des sciences de l'eau:/Journal of Water Science **21**(2): 219-232.

Palmen, L., W. Weeks and G. Kuczera (2011). "Regional flood frequency for Queensland using the quantile regression technique." Australian Journal of Water Resources **15**(1): 47.

Pandey, G. and V.-T.-V. Nguyen (1999). "A comparative study of regression based methods in regional flood frequency analysis." Journal of Hydrology **225**(1): 92-101.

Phien, H. N., B. K. Huong and P. D. Loi (1990). "Daily flow forecasting with regression analysis." Water S. A. **16**(3): 179-184.

Planque, B. and L. Buffaz (2008). "Quantile regression models for fish recruitment environment relationships: four case studies." Marine Ecology Progress Series **357**: 213-223.

Prinzio, M. D., A. Castellarin and E. Toth (2011). "Data-driven catchment classification: application to the pub problem." Hydrology and Earth System Sciences **15**(6): 1921-1935.

Reich, B. J. and L. B. Smith (2013). "Bayesian quantile regression for censored data." Biometrics **69**(3): 651-660.

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

Shu, C. and D. H. Burn (2004). "Artificial neural network ensembles and their application in pooled flood frequency analysis." Water Resources Research **40**(9).

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**(07).

Sousa, S., J. Pires, F. Martins, M. Pereira and M. Alvim-Ferraz (2009). "Potentialities of quantile regression to predict ozone concentrations." Environmetrics **20**(2): 147-158.

Sveinsson, O. G., D. C. Boes and J. D. Salas (2001). "Population index flood method for regional frequency analysis." Water Resources Research **37**(11): 2733-2748.

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

Tasker, G. D. and M. E. Moss (1979). "Analysis of Arizona flood data network for regional information." WATER RESOURCES RESEARCH **15**(6): 1791-1796.

Tokdar, S. and J. B. Kadane (2011). "Simultaneous linear quantile regression: A semiparametric bayesian approach." Bayesian Analysis **6**(4): 1-22.

Villarini, G., J. A. Smith, M. L. Baeck, R. Vitolo, D. B. Stephenson and W. F. Krajewski (2011). "On the frequency of heavy rainfall for the Midwest of the United States." Journal of Hydrology **400**(1): 103-120.

Wazneh, H., F. Chebana and T. Ouarda (2013). "Optimal depth-based regional frequency analysis." Hydrology and Earth System Sciences **17**(6): 2281-2296.

Wu, T. Z., K. Yu and Y. Yu (2010). "Single-index quantile regression." Journal of Multivariate Analysis **101**(7): 1607-1621.

Ying, Z., S. Jung and L. Wei (1995). "Survival analysis with median regression models." Journal of the American Statistical Association **90**(429): 178-184.

Yu, K. and R. A. Moyeed (2001). "Bayesian quantile regression." Statistics & Probability Letters **54**(4): 437-447.

Yue, S., P. Pilon and G. Cavadias (2002). "Power of the Mann–Kendall and Spearman's rho tests for detecting monotonic trends in hydrological series." Journal of Hydrology **259**(1): 254-271.

List of Tables

Table 1. LR vs QR model characteristics in RFA context 171
Table 2. MPLF values associated to QR and LR approaches 171
Table 3.Rejected stations at a significance level of 1% using Spearman test 171

Table 1. LR vs QR model characteristics in RFA context

LR model	QR model
- Estimates the conditional mean	- Estimates the conditional quantile
- Assumes : iid, gaussian errors	- Assumes: iid, Asymetric Laplacian errors
- Employs at-site *estimated* quantiles (aggregated series)	- Employs observed data (direct series)
- Requires at-site frequency analysis for each site in the region	- Does not perform any at-site frequency analysis
- Not robust to outliers	- Robust to outliers
- Affected by transformations (e.g. log-linear transformation induces bias)	- Invariant to any monotonic transformation
- Excludes sites with short record lengths	- Includes all available sites

Table 2. MPLF values associated to QR and LR approaches

	Q_{S10}		Q_{S50}		Q_{S100}	
	LR	QR	LR	QR	LR	QR
MPLF* ($m^3/s.km^2$)	16.07	**15.43**	6.62	**5.30**	4.65	**3.43**

Best results are in bold character.

Table 3. Rejected stations at a significance level of 1% using Spearman test

ID Station	**River**	**Test result**	**P-value**
030420	Aux brochets	Positive trend	7.06 E-5
072301	Moisie	Negative trend	1.75 E-4
060101	Petit Saguenay	Negative trend	9.00 E-3

List of figures

Figure 1. Steps involved in training the classical LR and the QR models. 173
Figure 2. Illustration of regional quantile estimation error related to the LR model and the QR model, compared to the 'true' quantile. 174
Figure 3. Procedure to evaluate the LR and QR models 175
Figure 4. Map showing the spatial variation of flood record length at gauged sites. 176
Figure 5. Scatter plots of at-site and regional estimated quantiles using the LR model (first column) and the QR model (the second column) for quantiles Q_{S10}, Q_{S50} and Q_{S100}. Both models are calibrated and evaluated over the entire data set. Points plotted in deep dark designate sites with long records. 177
Figure 6. Evaluation of the quadratic error related to the at-site flood quantile estimates for various record lengths. Simulations are performed using Weibull parameters of a fixed site belonging to the data set. For each record length 100 series were randomly generated using the same Weibull distribution. The quantile of each series is then estimated, and compared to the theoretical one (derived from the fixed distribution). 178
Figure 7. Bar plot of number of stations. Classes are defined to indicate the number of stations with records length exceeding a given minimum. 179
Figure 8. RMSE of the regional estimators of Q_{S50} (a) and Q_{S100} (b) as well as the MPLF of the regional estimators of Q_{S50} (c) and Q_{S100} (d) according to the length of regional data series. Both models are calibrated using sites with record length exceeding *l* years, except (c) and (d) where QR model was calibrated using the entire data set; QR and LR Validation is done using the entire data set. 180
Figure 9. QR Piecewise error (QRPE) vs LR Piecewise error (LRPE) associated to Q_{S10} (a), Q_{S50} (b) and Q_{S100} (c). Both models are validated over the entire data set. LR calibration is done using sites with record length exceeding 30 years; QR calibration is done over all sites. Points plotted in deep dark designate sites with long record length. 181
Figure 10. Piecewise error of regional quantiles estimations of Q_{S100} using the LR (a) and QR (b) models for various basin areas in the logarithmic scale. 182

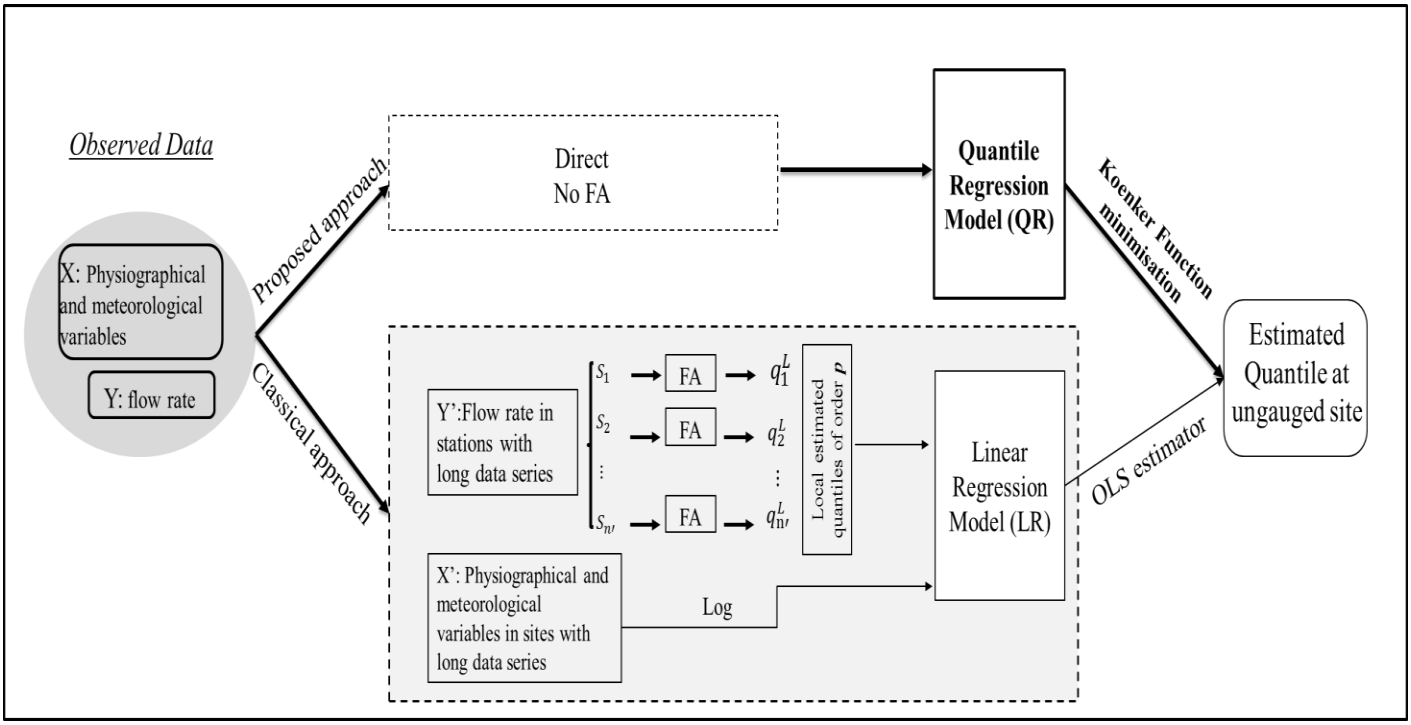


Figure 1. Steps involved in training the classical LR and the QR models.

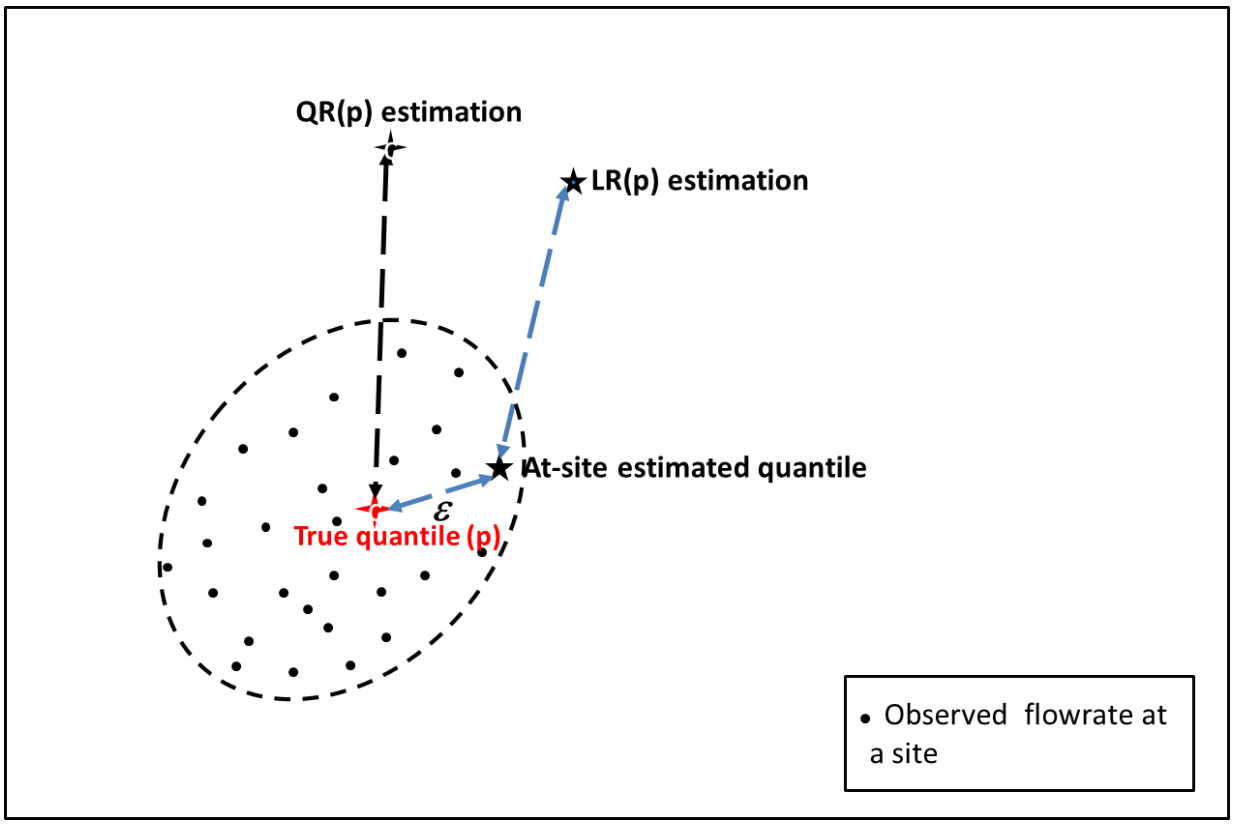


Figure 2. Illustration of regional quantile estimation error related to the LR model and the QR model, compared to the 'true' quantile.

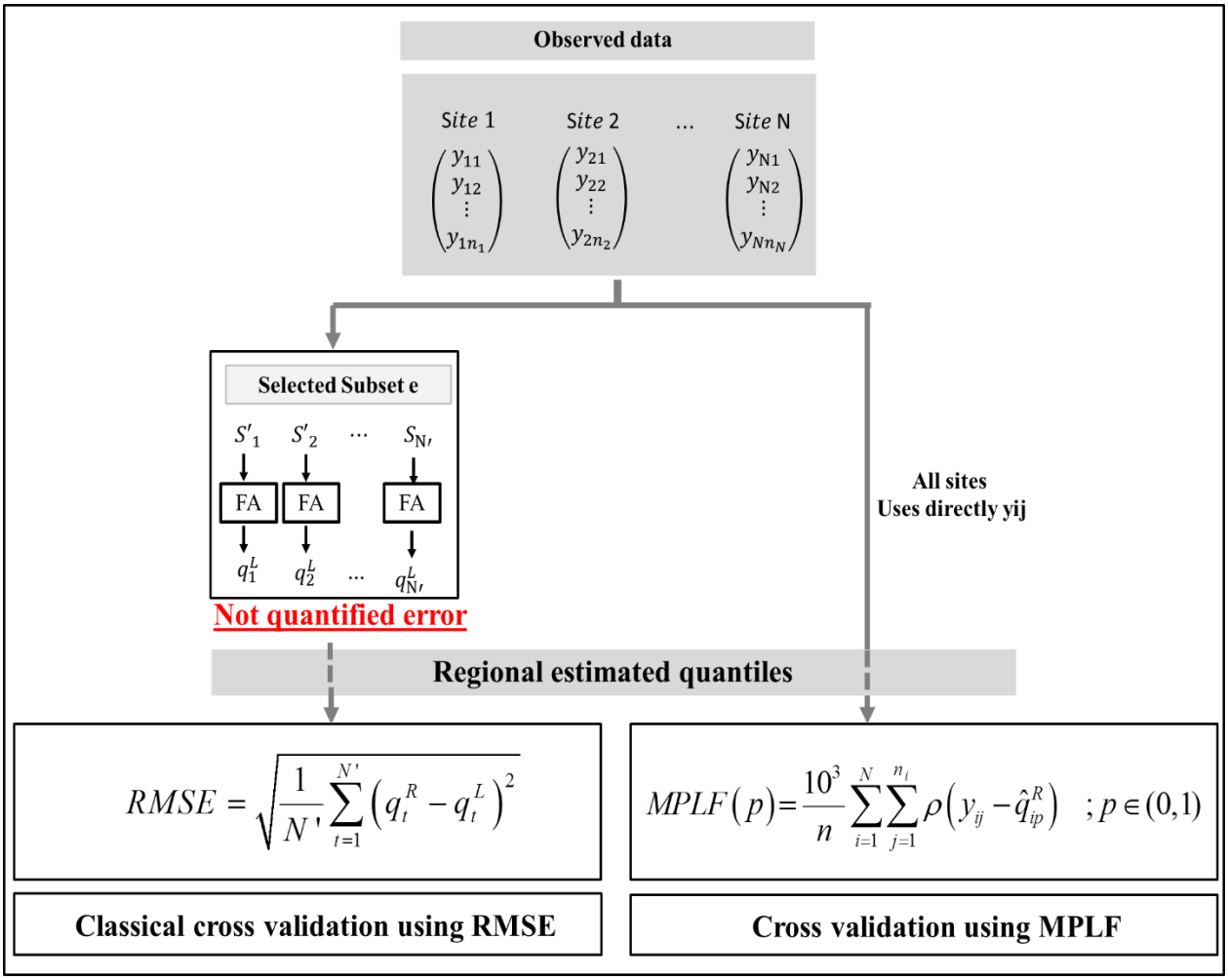


Figure 3. Procedure to evaluate the LR and QR models

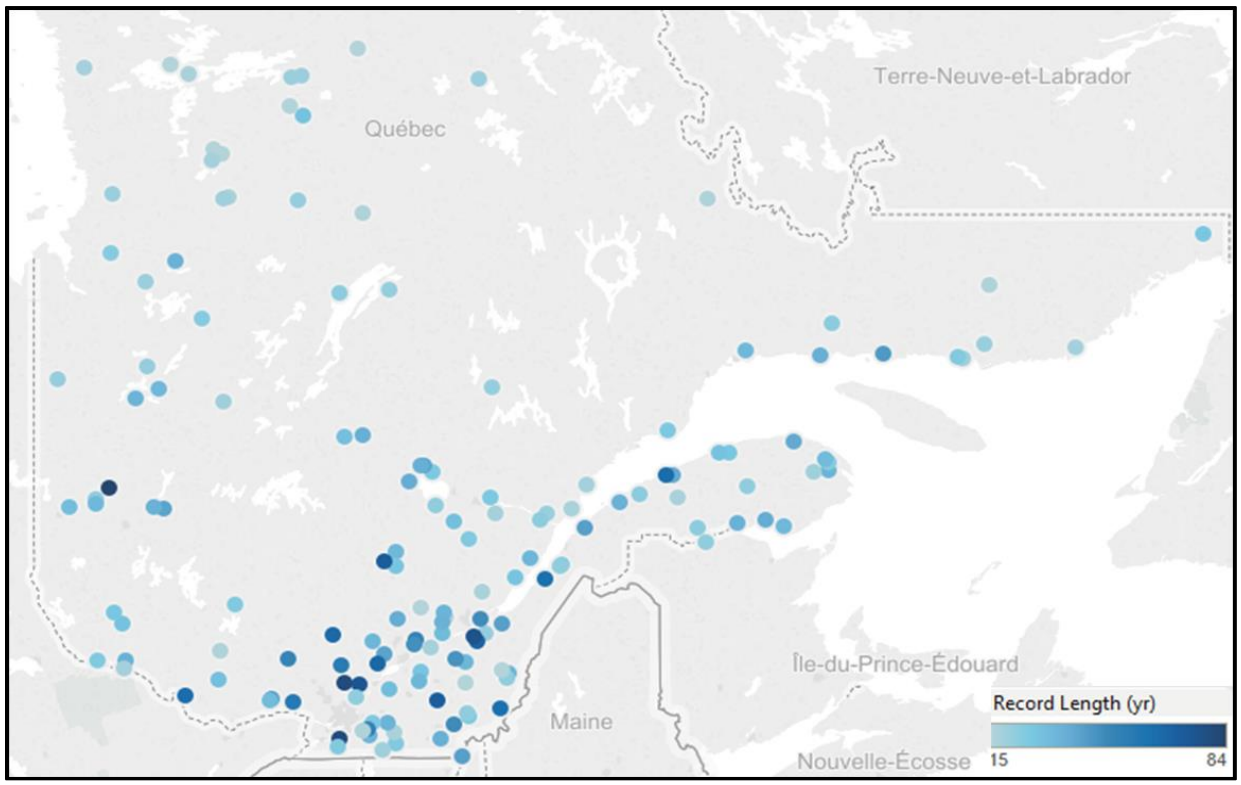


Figure 4. Map showing the spatial variation of flood record length at gauged sites.

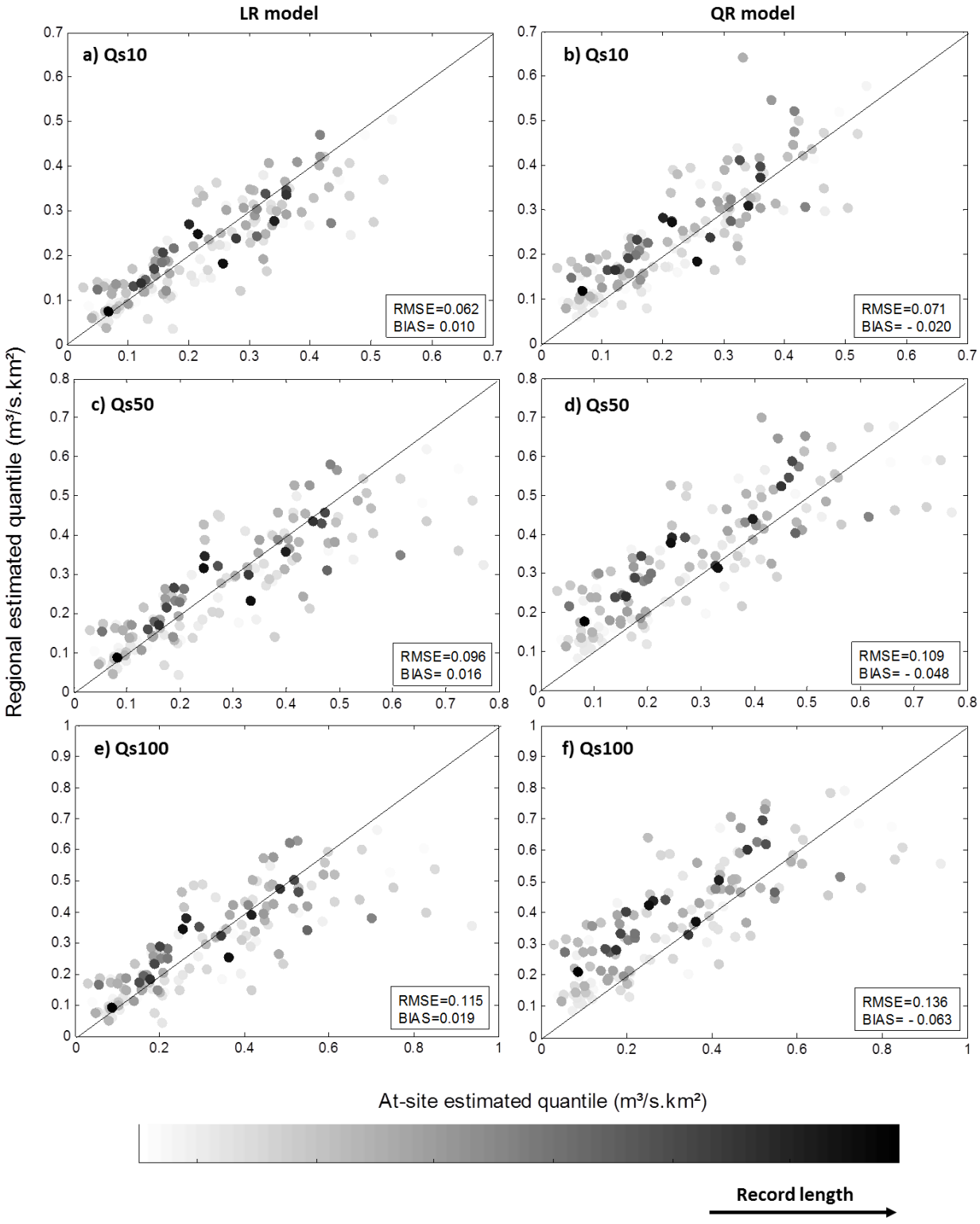


Figure 5. Scatter plots of at-site and regional estimated quantiles using the LR model (first column) and the QR model (the second column) for quantiles Q_{S10}, Q_{S50} and Q_{S100}. Both models are calibrated and evaluated over the entire data set. Points plotted in deep dark designate sites with long records.

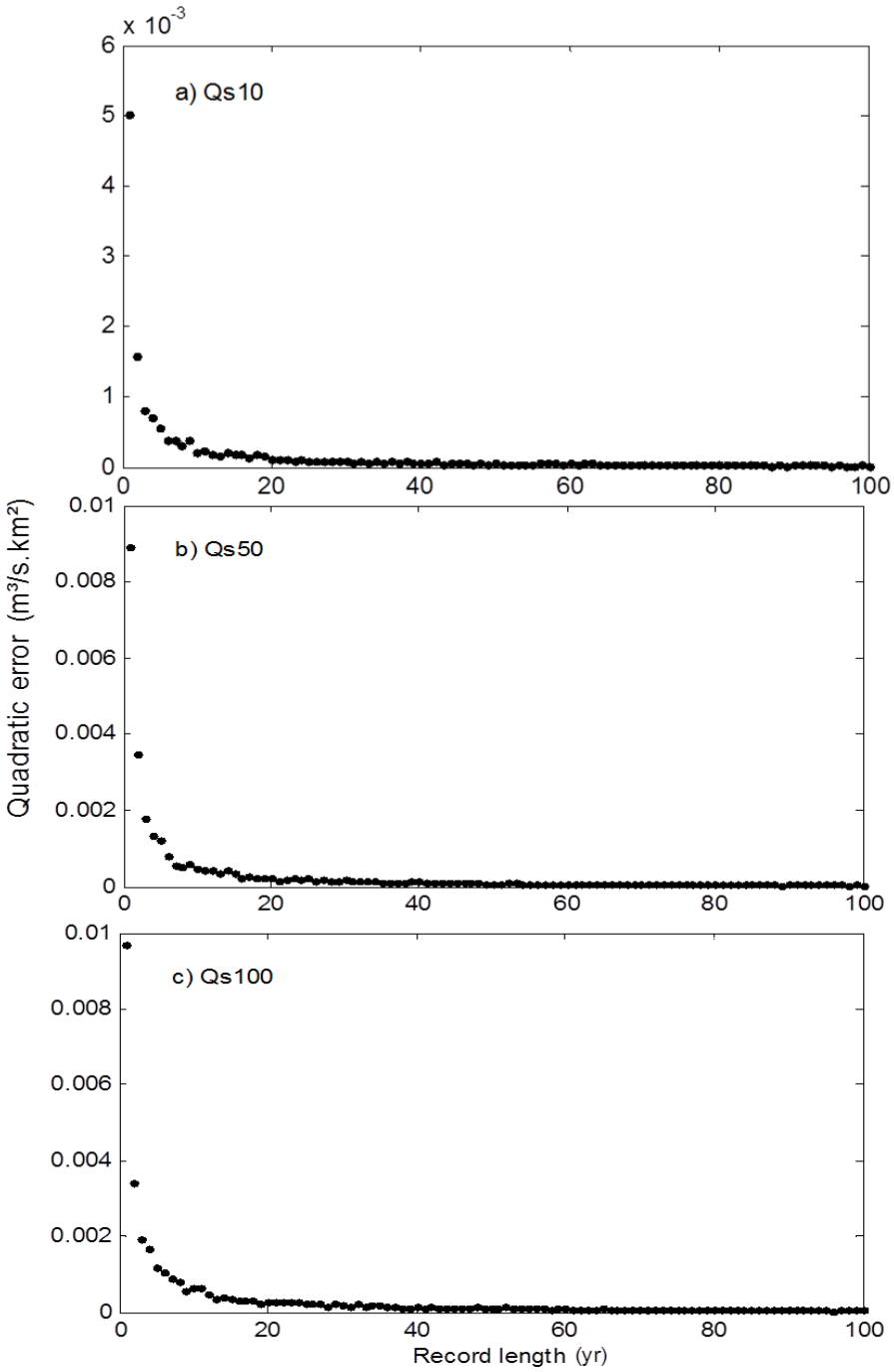


Figure 6. Evaluation of the quadratic error related to the at-site flood quantile estimates for various record lengths. Simulations are performed using Weibull parameters of a fixed site belonging to the data set. For each record length 100 series were randomly generated using the same Weibull distribution. The quantile of each series is then estimated, and compared to the theoretical one (derived from the fixed distribution).

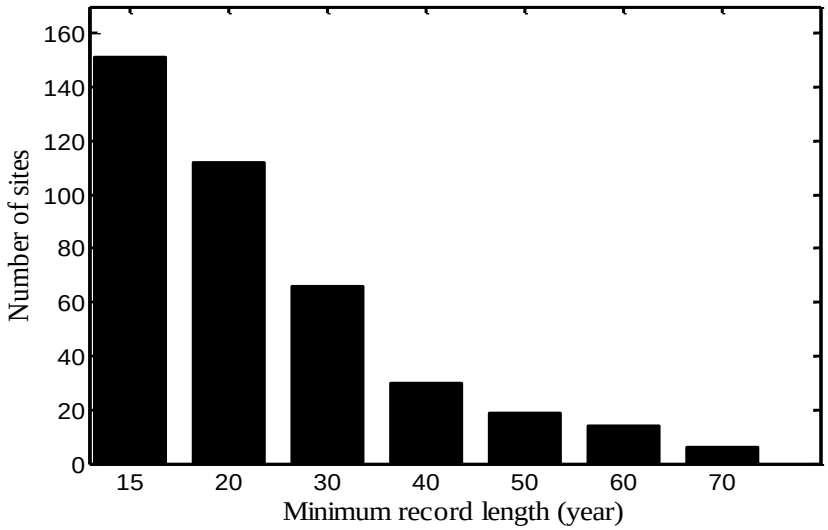


Figure 7. Bar plot of number of stations. Classes are defined to indicate the number of stations with records length exceeding a given minimum.

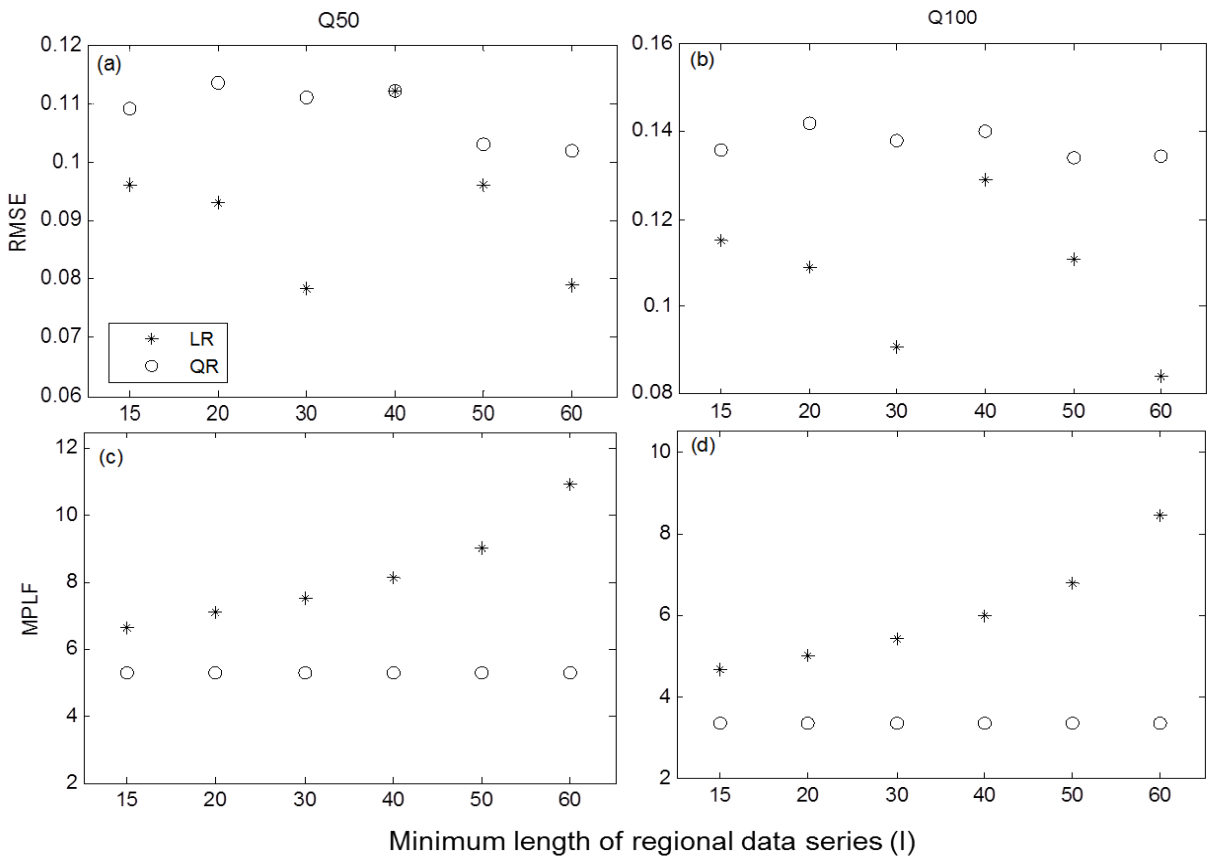


Figure 8. RMSE of the regional estimators of Q_{S50} (a) and Q_{S100} (b) as well as the MPLF of the regional estimators of Q_{S50} (c) and Q_{S100} (d) according to the length of regional data series. Both models are calibrated using sites with record length exceeding ***l*** years, except (c) and (d) where QR model was calibrated using the entire data set; QR and LR Validation is done using the entire data set.

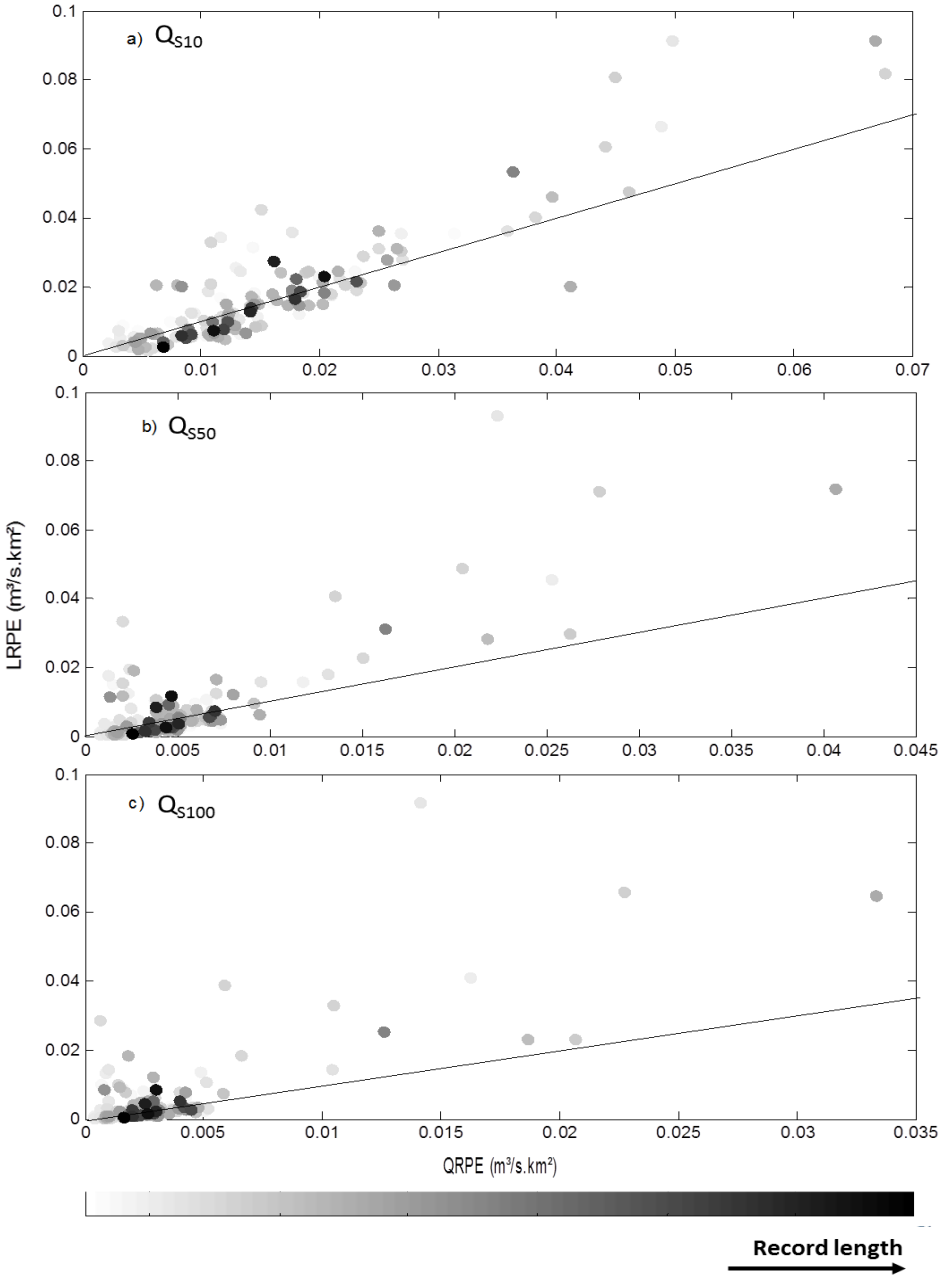


Figure 9. QR Piecewise error (QRPE) vs LR Piecewise error (LRPE) associated to Q_{S10} (a), Q_{S50} (b) and Q_{S100} (c). Both models are validated over the entire data set. LR calibration is done using sites with record length exceeding 30 years; QR calibration is done over all sites. Points plotted in deep dark designate sites with long record length.

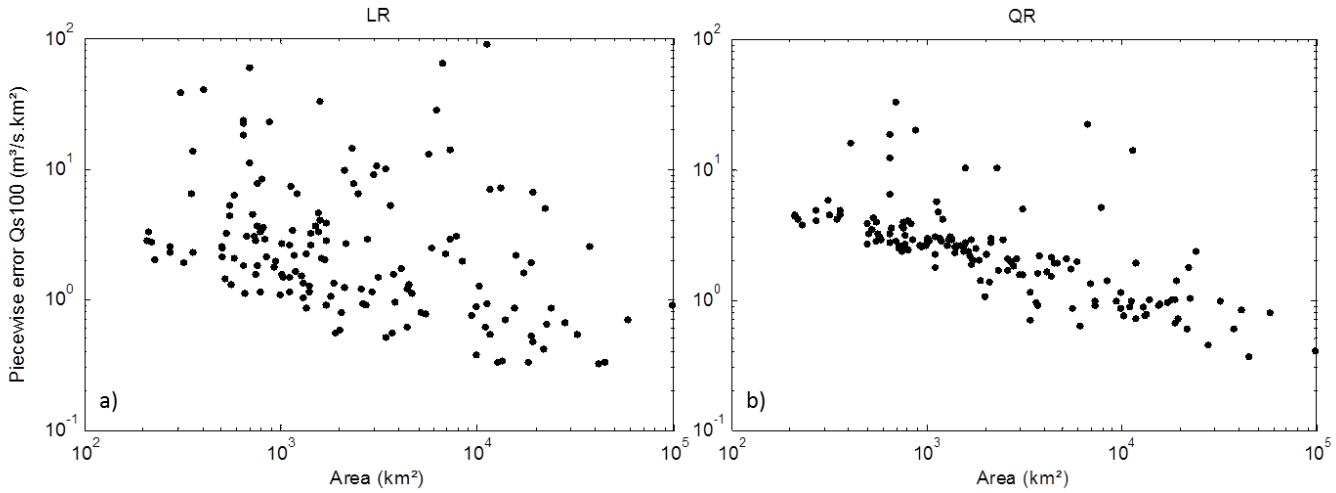


Figure 10. Piecewise error of regional quantiles estimations of Q_{S100} using the LR (a) and QR (b) models for various basin areas in the logarithmic scale.

CHAPITRE 5

HYDRO-CLIMATIC FREQUENCY ANALYSIS IN A CHANGING CLIMATE USING ADDITIVE QUANTILE REGRESSION

Hydro-climatic frequency analysis in a changing climate using additive quantile regression: an exploratory analysis

D. Ouali[1], F. Chebana[1], T.B.M.J. Ouarda[1,2]

[1]*Institut National de la Recherche Scientifique, Centre Eau Terre et Environnement,*

490, rue de la Couronne, Québec (Québec), G1K 9A9, Canada.

[2]*Institute Centre for Water Advanced Technology and Environmental Research*

P.O. Box 54224, Abu Dhabi, UAE

[*]**Corresponding author:** Tel: +1 (418) 654 2530#4477

Email: dhouha.ouali@ete.inrs.ca

2016-10-10

Abstract

The main aim of this study is to develop a new framework for hydrologic frequency analysis for a more realistic estimation of extreme events under a changing climate. In this respect, an exploratory analysis of quantile regression additive model (QRAM) abilities is performed through investigating the inter-annual stream flow variability in relation with climate teleconnections indices or climate variables. A comparison study is carried out between non-stationary models based on Generalized Extreme Value models (GEV), the non-stationary standard linear QR model and the considered QRAM model, using synthetic data sets and three real datasets from hydrologic stations located in California, Oregon and Québec. Owing to the flexibility of the additive model (AM), findings indicate that the QRAM, which is a flexible data driven tool, is able to capture any features present in the data.

Key-words: Non-stationary extreme events; flood frequency analysis, non-linear relationship, climate indices;

1. Introduction and literature review

During the last several years, extreme hydrological events such as droughts, floods and storms have attracted much attention as it affects more intensely human lives and society. Information on these extremes becomes a topical issue that requires increasingly sophisticated statistical tools to be characterized. Frequency analysis (FA) of hydro-meteorological variables is an important technique that aims to estimate the probability of occurrence of an extreme event. Estimation of a certain probability of occurrence of an extreme event is usually associated to its return period. This procedure is often performed through a selected probability distribution function used to characterize the occurrence frequency of a particular event [Hamed et Rao, 2010]. However, hypotheses that sustain this classical procedure are at odds with the scientific consensus that the climate is in continuous change.

As a matter of fact, several relevant researches applied for different areas in the world have identified significant changes for hydro-meteorological variables such as the flood flow rate [Cunderlik et Burn, 2002; Milly et al., 2002; Shankman et al., 2006; Cunderlik et Ouarda, 2009; Petrow et Merz, 2009; Ishak et al., 2013] and precipitation [Karl et Knight, 1998; Cheng et al., 2005; Hundecha et Bárdossy, 2005; Zhai et al., 2005; Pal et Al-Tabbaa, 2009; Allan et al., 2010; Chou et al., 2013; Jones et al., 2013].

To investigate the variation of hydro-meteorological variables in the past and its projection into the future, it is worthwhile to study trends in the extreme values. A continuously increasing number of FA studies that account for changing climatic conditions have been performed [Aissaoui-Fqayeh et al., 2009; Sugahara et al., 2009; Delgado et al., 2014; Condon et al., 2015; Mallakpour et Villarini, 2015]. For instance, in order to investigate the non-stationarity of the precipitation

patterns in California, El Adlouni et al. [2007] integrated time-varying covariates into the Generalized extreme value (GEV) distribution of the annual maximum precipitation. Several forms of non-stationarity have been considered in order to study the conditional distribution of the annual maximal precipitation as a function of the Southern Oscillation Index's variations. Villarini et al. [2009] developed a similar approach based on the Generalized Additive Models (GAM) for location, scale and shape parameters. It consists of modeling the time-dependent variation of annual maximum discharge (AMD) records by integrating several GAMs in the parameters of the conditional distributions. In the same regard, Cannon [2010] developed a neural network-GEV model capable of performing nonlinear and non-stationary analysis of hydro-meteorological time series. The motivation behind the use of these models (GEV, GAM and GEV with neural network) is that a parsimonious model is obtained using a statistical distribution, which in most cases can be justified based on some solid theoretical foundations such as the extreme value theory.

Despite significant advantages when dealing with the above mentioned non-stationary FA models, some critical points may be highlighted:

i) The inclusion of several steps such as exploratory data analysis, checking basic assumptions, identification of appropriate distributions, and estimation of the corresponding parameters. To be performed, each of these steps requires a selection among available methods that comes with some degree of subjectivity and uncertainty as well as time and effort;
ii) The fact of selecting a prior probability distribution. Indeed, the true probability distribution is unknown. When trying to identify the best distribution, various candidates arise and relatively complex analyses are performed over the central part as well as the upper and the lower tails.

iii) The limited-length of flood records is a well-known problem in the stationary FA of hydro-meteorological variables. Indeed, the data series lengths are generally insufficient to provide reliable estimates of flood quantile. In the non-stationary framework, this issue is much more problematic since in addition to the usual parameter estimation, the non-stationarity should be considered with the same insufficient amount of data;

iv) Depending on the selected time dependent parameter (generally the location and/or the scale parameter), its statistical distribution and the trend form several complex models can be defined;

v) Such statistical models (GEV-based models) generally consider a limited number of covariates, which reduce the sources of information;

vi) The difficulty to interpret results, since no direct link between the covariates and the response variable exists;

vii) As recognized by Madsen et al. [2013] there is a crucial need for developing more consistent FA models that account for the transient nature of a changing climate, i.e., developing a dynamic model that accounts for both stationary and non-stationary circumstances.

In summary, it seems that all the above limitations are mainly motivated by keeping the way of thinking in the classical and parametric framework. An alternative approach which provides a straightforward tool to estimate the conditional quantile of a considered variable is proposed by Koenker et Bassett [1978], namely, the Quantile regression model (QR). It offers the opportunity to obtain the entire response distribution without imposing a prior conditional distribution function. The QR contribution has been highlighted in several recent studies, for instance, to describe the effects of air pollution exposure on DNA methylation [Bind et al., 2015], to estimate

the value at risk in financial applications [Xiao et al., 2015], to study the relation between off-farm income and food costs [Mishra et al., 2015], to forecast electricity spot prices [Maciejowska et al., 2015] and to estimate daily and annual suspended sediment loads in rivers [Shiau et Chen, 2015].

A wide range of hydro-meteorological studies dealing with QR are also available in the literature [e.g Friederichs et Hense, 2007; Barbosa, 2008; Mazvimavi, 2010; Timofeev et Sterin, 2010; Ben Alaya et al., 2015; Shiau et Huang, 2015; Shiau et Lin, 2015]. Koenker et Schorfheide [1994] employed a QR model based on smoothing splines to study the global temperature change. In the same regard, Villarini et al. [2011] used a linear QR (LQR) model to investigate the presence of monotonic trends in the annual maximum daily rainfall distribution. In Jagger et Elsner [2009] and Elsner et al. [2008], the influence of some climate covariates on the intensity of tropical cyclones is studied; the QR model is compared with the Generalized Pareto Distribution. In the flood FA setting, relatively few studies dealing with QR have been performed, mainly, Sankarasubramanian et Lall [2003] in which authors seek to estimate the conditional flood quantile under a changing climate by comparing the LQR performances to that of a non-stationary Lognormal model. In the regional FA framework, Ouali et al. [2016] exploited the LQR potential to perform a one-step regional model allowing overcoming some limitations of the classical approaches.

Given on one hand the complexity involved in physical and environmental processes, and on the other hand the availability of high performant computational tools, it is reasonable to deal with more sophisticated statistical tools. An extension of the standard LQR method to the nonlinear setting is adopted in the current study through combining the QR with the additive model (AM). Indeed, the AM provides an efficient flexible tool to describe complex data relationships especially in the hydrological field [e.g. Campbell et Bates, 2001; Latraverse et al., 2002]. However, it might be the case that higher or lower quantile orders of the response variable depend on the covariates

quite differently from the center. Hence, given the QR skills, it would be of substantial interest to investigate the QRAM approach in the FA framework.

Further conceptual advantages of the QRAM arise such as: not requiring a prior distribution, not requiring basic assumptions, possibility to include several explanatory variables, providing one way of modeling both stationary and non-stationary cases, the ability to reproduce any shape of linear or nonlinear trend, producing direct and easy interpretable results. It also allows integrating the effects of potential meteorological variables that correspond to including more information in the model.

The main goal of this study is to explore the QR potential in estimating flood quantiles under a changing climate. In this regard, the LQR and QRAM approaches will be illustrated using both synthetic and real datasets.

This paper is structured as follows. In the next section we briefly describe the theoretical foundation behind the proposed approaches. In section 3, a Monte Carlo simulation analysis using synthetic data is performed to compare the performance of the considered models whereas real case studies are considered in section 4. Concluding remarks are presented in the last section.

2. Background

2.1. Non-stationary parametric models- GEV

Classical cumulative distribution functions, commonly used in hydro-meteorological FA, can be generalized to incorporate the concept of non-stationarity in the modeling process through integrating time-varying covariates in the distribution parameters. Such covariates could

incorporate, for instance, specific measures of land use/land cover change as well as some indices of low-frequency climatic modes.

Based on extreme value theory, th eannual maximum discharge (AMD) series often follows a GEV distribution. Combining the Gumbel, Frechet, and Weibull distributions, the GEV cumulative distribution function is given by:

$$F(q) = \exp\left\{-\left[1-k\left(\frac{q-\mu}{\alpha}\right)\right]^{1/k}\right\} \qquad k \neq 0$$
$$F(q) = \exp\left\{-\exp\left(-\frac{q-\mu}{\alpha}\right)\right\} \qquad k = 0 \tag{1}$$

where μ, *k* and α are location, shape and scale parameters respectively. In the case when **k**=0, the distribution matches the Gumbel distribution. To account for plausible non-stationarity of data series, each parameter, especially location and scale, is modelled as a function of one or more time-varying explanatory variables *U(t).* For instance, in the case of a linear dependence the location parameter can be written as:

$$\mu_t = \sum_i \beta_i U_i(t) \tag{2}$$

where β_i denote the regression parameters.

A number of non-stationary GEV models are considered for comparison purposes, depending on the time-varying parameter and the dependency structure:

- GEV_{00} (u, α , k): the classical stationary model with constant parameters;
- GEV_{10} (u_t, α , k): model with location parameter μ linearly dependent on one covariate;

- GEV_{01} (u_t, α , k): model with scale parameter α linearly dependent on one covariate;
- GEV_{11} (u_t, α_t , k): model where both location μ and scale α parameters depend linearly of the covariates;
- GEV_{20} (u_t, α , k): model where the location parameter μ is a quadratic function of the covariates;
- GEV_{21} (u_t, α , k): model where the location parameter μ is a quadratic function of the covariate and the scale parameter α is linearly dependent on the covariates.

As mentioned in the introduction, a major limitation of these parametric models resides in assuming a specific distribution that may change in future conditions. A second limitation is that they assume the conditional distribution parameters to be invariant except for location parameter and in some cases the scale. For example, it is usually presumed that there would be no change in the shape of the flow distribution over time [El Adlouni et al., 2007]. Clearly, it is of interest to ascertain whether this assumption is appropriate. Furthermore, using these models it can be difficult to interpret parameters physically with respect to issues of the climate's influence on the response variable. Additionally, when considering these approaches one assumes implicitly the non-stationarity which may not be valid at future-times.

An alternative is provided using the quantile regression approach, which allows direct quantile estimates, an easier interpretation of model coefficients (parameters), integrating naturally the non-stationary aspect and not assuming a fixed conditional distribution.

2.2. Quantile regression- based models

2.2.1. Linear quantile regression model (LQR)

Since the main goal of non-stationary FA is to provide quantile estimates at gauged sites according to the updated values of the time dependent covariates, it seems reasonable to proceed with a regression tool that directly estimates the flood quantile given a set of covariates. This can be achieved through the quantile regression (QR) approach. The latter is introduced by Koenker et Bassett [1978] as an extension of the median regression approach to directly predict any conditional quantile of the response variable y given a set of predictors $\mathbf{x}$. For a given quantile p (0<p<1), the linear QR (LQR) model is expressed by:

$$\hat{Q}_p(y \mid \mathbf{x}) = \mathbf{x}^T \mathbf{b}_p \tag{3}$$

where $\hat{Q}_p(y \mid \mathbf{x})$ is the quantile of the conditional distribution of y given $\mathbf{x}$, and $\mathbf{b}_p$ contains parameters of the linear quantile regression model. Regression parameters are estimated by solving the following minimization problem:

$$\hat{\mathbf{b}}_p = \arg\min_{\mathbf{b}} \sum_{i=1}^{n} \rho_p\left(y_i - \mathbf{x}_i^T \mathbf{b}\right) \tag{4}$$

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

$$\rho_{\mathrm{p}}(\mathrm{u}) = \begin{cases} \mathrm{u}(\mathrm{p}-1) & \text{if } \mathrm{u} < 0 \\ \mathrm{up} & \text{if } \mathrm{u} \geq 0 \end{cases} \qquad ;0 < \mathrm{p} < 1 \tag{5}$$

Unlike the classical FA, this approach does not require the same restrictive assumptions and presents a number of attractive statistical properties such as invariance with respect to any

monotonic transformation and robustness against outliers [Koenker et Bassett 1978]. More theoretical aspects can be found in Koenker [2005].

2.2.2. Quantile regression additive model (QRAM)

The natural complexity of the hydrological process arising for example from the watershed topography, the associated geological characteristics and meteorological variations, has been widely recognized and documented in the hydrological literature [e.g. Riad et Mania, 2004; Sivakumar et Singh, 2012]. In this regard, based on the same concept as for the LQR, more flexible QR-based approaches are proposed in the nonlinear setting assuming nonlinear transfer functions linking the predictors to the response variable. In this paper, we investigate a QR approach based on additive models (AMs) (QRAM) as a flexible estimation tool to model nonlinear data relationships.

As introduced by Hastie et Tibshirani [1986], in an Additive Model (AM) the modeled effect of explanatory variables $X=(X_i)_{i=1..n}$ on the response Y is generally described by non-parametric smoothing functions $f_i(X_i)$ for each explanatory variable X_i. The basic model formulation is explicitly given by:

$$Y=\eta+\sum_{i=1}^{n} f_i(X_i)+\varepsilon \tag{6}$$

where f_i is a smooth function of explanatory variable X_i such that:

$$f_i(x)=\sum_{j=1}^{q}\beta_{ij}b_{ij}(x) \tag{7}$$

where b_{ij} are basis functions and β_{ij} are parameters to be estimated.

Typically, smooth functions can take both parametric and nonparametric forms. As a matter of fact, several types of smoothing functions designed to represent the non-linear relationships between X and Y exist, such as smoothing by parametric regression, mobile average, kernel smoothing and spline smoothing. Noteworthy that spline models are the most commonly used approach to smoothly fit data [Wahba, 1990]. Overall, the ability to consider non-parametric fitting provides the potential for AM to better describe regression relationships.

Hereafter, we provide a brief description of the theoretical foundation behind QRAM. For a more detailed discussion, the reader is pointed to Koenker [2011]. The nonlinear QRAM for a quantile order $p \in (0,1)$ has the following general form:

$$Q_p(y \mid \mathbf{x}) = \mathbf{x}^T \mathbf{b}_p + \sum_{i=1}^{m} f_i(z_i) \tag{8}$$

Instead of fitting a linear quantile function using expression in (4), using the QRAM a quantile of order *p* can be estimated through resolving the following optimisation problem:

$$\min_{(b_p, f)} \sum_{i=1}^{n} \rho_p \left(y_i - x_i^T b_p - \sum_j f_j(z_{ij}) \right) + \lambda_0 \left\| b_p \right\|_1 + \sum_{j=1}^{m} \lambda_j \int \left\| \nabla^2 f_j(s) ds \right\| \tag{9}$$

where $\rho_p(.)$ is the usual check function, λ_i are the smoothing parameters, $\nabla^2 f_i$ denotes the Hessian of the function f_i with absolutely continuous gradient, and $\| . \|$ denotes the usual Hilbert-Schmidt norm. As can be perceived, the last term in (9) is a penalisation term controlling the smoothing shape. As mentioned in Koenker [2011] the total variation penalty has been adopted during the concept of this model.

A solution to the scheme in (9) that greatly reduces computational cost is provided using piecewise linear with knots at the observed values z_i .

As in any regressive model, in the presence of multiple covariates a challenging task is to avoid redundant information and reduce the model dimensionality during the variables selection. This issue of selecting significant covariates has been addressed in Koenker [2011] using the LASSO approach. For further details the reader is referred to Koenker [2011].

Regarding the QRAM, it will be possible to investigate the response characteristics beyond its central part especially when investigating non-linear relationships in a non-stationary framework. In the FA setting, the large class of models derived from AMs combined to the LQR benefits confer several convincing advantages to the QRAM model. When compared to the classical non-stationary FA, this latter presents the following advantages: i) it does not require nor a prior distribution neither basic assumptions, ii) it is easily interpretable since it directly describes the observed data conditional on some chosen co-variates, iii) it offers the flexibility to fit the real relationship between variables due to a high sensitivity to changes in the evolution of co-variates, iv) it is insensitive to outlier observations, v) it offers the possibility to include several explanatory variables and vi) it provides one way for modeling both stationary or non-stationary cases.

3. Monte Carlo simulation data

In this section, synthetic data issued from Monte Carlo simulation experiments are used to compare the QRAM and the classical non-stationary GEV models.

3.1. Experimental design

In order to describe changes in flood regimes over time, the classical non-stationary FA approach incorporates climate indices as covariates in the model parameters. The aim of this study is to

address the non-stationary issue relating to extreme flooding events using some promising approaches such as the QR. This is initially performed using Monte Carlo simulations.

Several models have been considered when performing Monte Carlo simulations involving classical parametric models (GEV_{00}, GEV_{10}, GEV_{01}, GEV_{11}, GEV_{20}, and GEV_{21}) as well as the QR-based models (LQR and QRAM). The concept consists in estimating particular conditional quantiles of N=1000 randomly generated samples, using the considered models. This replication number has been chosen in a way to ensure stability of simulation results. For simplicity, only the case where the location parameter of the GEV distribution depends on one single covariate ***t*** representing time has been considered in this paper.

Consider the annual maximum discharge (AMD), Y_t, that arises from a GEV distribution. Then, two simple test cases are considered here: for the first test case, we assume a linear dependency structure for the location parameter, $Y_t \sim GEV_{10}(\mu_t, \alpha, k)$. Note that parameters of the non-stationary parametric models are estimated through the Generalized Maximum Likelihood method [El Adlouni et al., 2007].

Hence, the distribution parameters are fixed to vary as follows: $\mu_t = 0.1\ t + 5; \alpha = 1; k = 0.1$.

For the second synthetic case, we assume a quadratic dependency structure between the location parameter and the covariate, $Y_t \sim GEV_{20}(\mu_t, \alpha, k)$. Then, parameters are fixed as $\mu_t = 0.1\ t^2 + 5, \alpha = 1, k = 0.1$.

Note that these cases are undertaken to investigate the ability of all considered models, in particular the QR-based ones, to adequately estimate non-stationary quantiles of synthetic data.

Using the above parameters, N=1000 realizations of Y_t are generated with a record length of 50. Quantiles are conditionally estimated at particular values of the covariate, principally the median

value. Then, a comparison study is carried out based on evaluation criteria computed for the QR, QRAM and all parametric models, for each quantile of non-exceedance probability p=0.50, 0.90 and 0.99:

Relative root mean square error:
$$RRMSE = \sqrt{\frac{1}{N}\sum_{i=1}^{N}\left(\frac{Q_{ip} - \hat{Q}_{ip}}{Q_{ip}}\right)^2} \qquad (10)$$

Relative bias:
$$RBIAS = \frac{1}{N}\sum_{i=1}^{N}\left(\frac{Q_{it} - \hat{Q}_{ip}}{Q_{ip}}\right) \qquad (11)$$

where Q_{ip} denotes the empirical p-quantile associated to the initial samples generated from GEV_{10} or GEV_{20}; $\hat{Q}_{ip}$ is the estimated p-quantiles using the fitted model.

3.2. Results of Monte Carlo simulations

- GEV_{10} test case

As noted above, the GEV_{10} model assumes a linear dependency structure between the location parameter and a covariate. An example of data series randomly generated using the GEV_{10} model is illustrated in Figure 1-a. Averaged values of RBIAS and RRMSE over the 1000 realizations are given in Table 1 for each of the six GEV models and the two QR-based models. Obtained results show that, in terms of RRMSE, the QRAM is the most appropriate model for non-exceedance probabilities p=0.5, 0.9 and 0.99. Indeed, RRMSE values of the QRAM model are the smallest values compared to other models (for instance the RRMSE is around 12.60% for QRAM and 14.4% for the GEV_{10} model for quantile order 0.99). On the other side, using the RBIAS criterion it can be seen that the GEV_{10} is the less biased model. Indeed, QRAM produced relatively high RBIAS values especially for higher probabilities. This is somewhat expected especially using the QR tool. Indeed, the bias in the LQR is a recognized problem which is due to sampling [Buchinsky,

1995]. When dealing with the QRAM the question of bias may arise from the shrinkage term in the estimation procedure (9) [Koenker, 2011].

- GEV_{20} test case

Using the GEV_{20} test case, the location parameter is a quadratic function of a covariate and the scale parameter is constant. Example of data series randomly generated using the GEV_{20} model is illustrated in Figure 1-b. Results associated to this model are presented in Table 1 for quantiles of non-exceedance probabilities p=0.5, 0.9 and 0.99. As shown in the table, the QRAM outperforms all other considered models for all quantile orders. Indeed, for the 0.99 non-exceedance probability the RRMSE value goes from 31.76% using the GEV_{20} to only 23.35 % using QRAM, a relative reduction of around 26%.

Compared to the first test case results, it has been found that the non-linear QRAM model performs better in the second case test. This could be explained by the non-linear dependency structure assumed when generating data by the GEV_{20}.

4. Applications

The motivation behind these applications arose from an effort to investigate the conditional distribution of the AMD as a function of standard climate indices as well as meteorological variables.

4.1. Case studies and initial exploratory analysis

To illustrate the proposed QR-based approaches, we first analyze AMD data records from two U.S. Geological Survey (USGS) stations, namely the Arroyo Seco and Bear Creek at Medford. Figure 2-a illustrates the geographic locations of these two basins. A third application is also considered to investigate the impact of introducing meteorological variables in the QR-based models. In this case, data from Dartmouth gauging station is considered (Figure **2**-b).

4.1.1. Datasets

- Arroyo Seco: It is a sub-watershed of the Los Angeles River Watershed in California with a drainage area of 41.4 km^2. Its record station (with the USGS number 11098000) is located at latitude 34°13'20" and longitude 118°10'36". It should be emphasized that the hydrological pattern in this site is characterized by large stream flow volumes. Indeed, the steep slope which characterizes the basin morphology associated to the heavy rainfalls caused by Pacific storms has produced several catastrophic floods in the past. The analysed gaging record of AMD used in the current study extends from 1949 to 1999.

Given the geographical location of this station, the hydrological pattern could be strongly affected by the El Niño or La Niña phenomena. A measure of the intensity of these phenomena is provided using the Southern Oscillation Index (SOI). Hence, for both case studies, the SOI index was taken to be the covariate over the same record period. The SOI is calculated using monthly mean sea level pressure anomalies at Tahiti and Darwin. Sustained positive values, meaning high atmospheric pressure in Tahiti and low atmospheric pressure in Darwin, indicate La Niña episodes. In the counterpart, negative SOI values, meaning low atmospheric pressure in Tahiti and high atmospheric pressure in Darwin, indicate El Niño episodes [e.g. Allan et al., 1991].

- Bear Creek station at Medford: The Bear Creek station, maintained by the USGS Oregon Water Science Center (with the USGS number 14357500) is located within Jackson County in southwestern Oregon just north of the California border, at 42.32º N latitude and 122.86º W longitude. This site is a tributary to the Rogue River with a drainage area of 748.50 km^2. Urbanization has strongly affected the Bear Creek watershed's hydrologic response through increasing the sedimentation amounts and decreasing water level from water withdrawals [Vogt,

2002]. Given its geographical location, the SOI index is also taken to be the covariate for the current case study. The analysed AMD records over the period between 1916 and 2015 are used.

- Dartmouth River in Gasépsie: This station is located on the Dartmouth River near the Ruisseau du Pas de Dame in the Gaspésie region of the province of Quebec (Canada) with a drainage area of 626 km^2 [Masselot et al., 2016]. Hydrological data (AMD) are provided from the hydrometric gauging station with federal reference number 01BH005 for the period between 1981 and 2012.

Given the geographical location of this station on the Atlantic side, the North Atlantic Oscillation (NAO) index is considered as a covariate. This index is a large scale phenomenon resulting from alternation of two atmospheric mass in the North Atlantic Ocean, with centers of action located near the Icelandic low and the Azores high. This dipole system may induce considerable effects over the eastern United States and Canada side. The NAO index used in this application is obtained using the sea level pressure anomalies between Lisbon, Portugal and Stykkisholmur/Reykjavik, Iceland [Hurrell, 1996].

In order to maximize the explained response variance, the mean annual temperature (T) is also included. Indeed, the mean annual temperature could be a relevant variable to model the maximal streamflow temporal variability. Considered data are derived from the meteorological station with identification number 7052605, very close to the hydrometric station, over the same record period as the AMD.

4.1.2. Exploratory analysis

The time series of the AMD recorded in the three considered stations are provided in

Figure 3 where a simple linear regression was used to describe the possible linear trends. Hence, a possible monotonic and increasing trend can be perceived in all cases. The significance of these trends was then investigated through applying the Spearman test on the data of the three cases. The associated results, reported in Table 2, indicate that the AMD series recorded at the Arroyo Seco and Dartmouth sites were found to be stationary, at both 5 % and 10 % significance levels. However, the AMD series of Bear Creek station shows a statistically significant trend for both significance levels. Accordingly, using classical approaches one may perform a non-stationary analysis only in the Bear Creek station. However, it should be noted that the Spearman test provides information about only the trend in the central part of the distribution [Villarini et al., 2011]. Since the focus in the current study is on higher quantiles, it is of interest to examine not only the overall trend (through the median or the mean) but especially extreme value trends (through higher quantiles). This issue has been addressed using the LQR approach to examine the trend of flood quantile as a function of time. A visual investigation of Figure 4, which represents several quantile curves, shows a significant trend especially for high quantile orders for all considered cases.

Further investigation has been performed by applying the LQR approach at each site for high quantile orders. Then, based on the t-test at 5% significance level, it was found that the slopes are significantly different from zero which indicates the existence of a significant trend. Hence, it is worthwhile to account for non-stationarity of extreme values for all the considered sites.

In Figure 5, the relationship between AMD and SOI is illustrated in Arroyo Seco and Bear Creek stations. Accordingly, one can observe a negative correlation between these two variables with Pearson's correlation coefficient $\zeta = -0.40$ and $\zeta = -0.19$ respectively. In fact, one can note that

negative SOI values correspond to high values of AMD and conversely. This hydrological pattern is explained by the heavy precipitations occurring during El Niño years at the studied regions.

Additionally, relations between AMD, NAO and mean annual temperature over the Dartmouth station are displayed in Figure 5c and d. Accordingly, positive correlations can be interpreted through the patterns displayed on scatterplots with Pearson's correlation coefficients: $\zeta(AMD, NAO) = 0.37$ and $\zeta(AMD, T) = 0.48$. Typically, positive NAO values indicate strengthening of both Icelandic low and Azores high resulting drier air masses over the Northern Canada, and conversely. This can be easily perceived from Figure 5c where the lowest AMD values are recorded during positive NAO values.

4.2. Results

In this section, the general behavior of the QR-based models is explored and analyzed. Therefore, the three case studies are considered with the associated results illustrated hereafter.

4.2.1. Arroyo Seco watershed

Each of the above mentioned non-stationary GEV models, as well as LQR and QRAM models, is used to study the conditional distribution of the AMD according to the SOI index. Figure 6 presents the conditional quantiles of non-exceedance probabilities 0.90 and 0.99 estimated using the GEV_{00}, GEV_{01}, GEV_{20}, LQR and QRAM models for various SOI values.

Investigation of the estimated quantiles with the considered non-stationary models shows that the SOI index describes satisfactorily the temporal variability of floods recorded in Arroyo Seco. Inspection of Figure 6-a indicates that for negative SOI values, corresponding to El Niño episodes

and consequently extreme AMD values, the QRAM model leads to the highest quantile values. Note that for this range, both the GEV_{20} and the LQR models behave similarly and lead to estimates in the midst of the others. On the other hand, for positive SOI values meaning during La Niña, both GEV_{01} and GEV_{20} models have similar profiles providing the highest estimates after the stationary GEV_{00} model. Note that the QRAM estimates during La Niña vary slightly for different SOI values. This deeply pronounced behavior indicates the ability of QRAM to distinguish between the two phenomena (La Niña and El Niño). Indeed, in contrast to El Niño episodes, during La Nina episodes no heavy precipitations and hence no high discharges are recorded. Note also that the QR-based models provide the lowest estimates values during La Niña. These findings are numerically indicated in Table 3 where the maximum and minimum SOI values are -3.16 and 2.07 respectively. Indeed, for the lowest SOI value, estimation of 0.90 quantile obtained for instance by GEV_{20} model is around 167 m^3/s versus 197 m^3/s using QRAM.

From Figure 6-b, it appears likewise that QRAM provides a similar profile of the 0.99 flood quantile as for the 0.90 one, meaning high variations during El Niño versus slight variations during La Niña. Note that obtained GEV_{01} and GEV_{20} estimates are significantly higher than that of QRAM. Estimated 0.99 flood quantile by each of the considered models is also summarized in Table 3. Figure 7 shows the estimated effect of SOI on AMD variability for non-exceedance probabilities 0.90 and 0.99 using the QRAM model. Clearly, the SOI index has a negative effect on AMD values. This effect appears to be non-linear for the 0.90 non-exceedance probability and linear for the 0.99 non-exceedance probability.

4.2.2. Bear Creek watershed

Unlike the previous case study, derived data from Bear Creek station was divided into calibration and validation subsets due to its long record period (Figure 8). Results of the application of the considered models over both calibration (around 66% of observed data) and validation periods (about 33% of observed data) for the non-exceeding probability 0.90 are shown in Figure 9. The obtained results from modelling flood frequency under non-stationary conditions using classical models and the SOI index show that this latter is not a very illustrative covariate. Explicitly, results for the calibration period show that the non-stationary models are not sensitive to the SOI variations, especially the GEV01 model. Also, it is interesting to underline that the LQR model result expressed a slight dependence on the SOI variation, especially during the validation step. On the other side, it can be seen that the non-stationary QRAM result indicates that the flood magnitude experienced significant variability. This supports the idea that the QRAM captures more adequately the dispersion of stream flow values, and shows the effect of climate indices on floods magnitude. These findings are observed for both calibration and validation steps.

4.2.3. Dartmouth watershed

The last application is considered to explain more adequately the stream flow variation through integrating new explanatory variables. We first begin by analysing results of application of the considered models for the non-exceeding probability 0.90 as function of only the NAO index. Several conclusions can be drawn from Figure 10 concerning the behaviour of each model. Actually, it can be seen that for a high level of activity of the NAO (highly positive or negative NAO values) both the GEV_{01} and the GEV_{20} models estimate poorly the observed streamflow values. When using the QR-based models, it can be seen that they do well with slightly better

behaviour for the QRAM. Unlike GEV models which do not show a significant trend, the QR-based models show an overall trend explained in part by the NAO pattern over the studied region. Indeed, as discussed in Anctil et Coulibaly [2004], the NAO may have a strong signature over the province of Quebec inducing high discharge levels for negative NAO values and low levels for positive NAO values. This feature is verified using both the LQR and the QRAM models. In Figure 11 the effect of each explanatory variable on the AMD variability is displayed using the QRAM. One can note the complexity of the hydrological system, and perceive the necessity to consider both the additive and the non-linear aspects.

For a better understanding of the streamflow variability, an additional analysis has been performed for the 0.5 non-exceeding probability integrating both the NAO index and the mean annual temperature. Estimation of the median form of the AMD time series using QR models is illustrated in Figure 12a. Adopted QR models consist in integrating only the NAO as explanatory variable on one hand, and both NAO and T as explanatory variables on the other hand. Obtained results indicate clearly the superiority of the second model to represent the AMD temporal variability. In Figure 12b the same application was conducted using the QRAM. AMD time series using the GEV20 model is also shown. The main finding that can be drawn from this figure is the overall ability of the QRAM, with the two considered explanatory variables, to reproduce the AMD variability when compared to the other models: LQR models, GEV_{20} model and QRAM (with only NAO) model.

5. Conclusions

Temporal variability of streamflows is seen to be connected to many factors particularly the meteorological ones. This ascertainment would suggest that the stationarity assumption of flood

data series can hardly be met in reality, especially for extreme values. In these concerns, the non-stationarity issue is addressed in this analysis through adopting some advanced statistical tools. In the current study, the QRAM model was investigated in the non-stationary framework to estimate extreme quantiles conditional on some covariate variations. In this respect, synthetic case studies were first performed. Faced to the different issues and inconveniences related to the classical FA framework, especially based on parametric distributions, the proposed QRAM was found to perform well when modeling extreme values especially when dealing with a non-linear dependency structure. Indeed, The QRAM has a number of conceptual advantages and can be seen as a different framework to deal with FA in the presence (or not) of non-stationarity in a nonlinear (or linear) shape effect.

This additive approach was then illustrated using annual maximum discharge data from three watersheds from which two are situated in the USA and one is located in Quebec, Canada. Given the geographical location of the considered sites, a great impact of the atmospheric circulation variability may be felt on these regions, affecting notably the precipitations amounts, temperatures and consequently discharges variability. Accordingly, an incorporation of teleconnections indices describing the regional climate variability is found to be useful for understanding changes in local flood frequency and for an adequate forecasting of the flood risk. In these concerns, it was of interest to consider the SOI and the NAO indices as relevant climatic patterns to describe the flood data variability in the considered sites.

Obtained results revealed that the QRAM is more flexible since it does not require the stationarity assumption, does not impose a prior selection of a probability distribution and tend to better reproduce the flood temporal variability.

Results have also shown the feasibility of incorporating not only the large-scale climate indices as covariates but also some meteorological variables such as the mean annual temperature. A direct implication of this is to enable a better description of changes in flood regimes over time. However, an interesting fact to note is the importance to account for other factors that highly influence the modelling of flood frequency such as the impact of anthropogenic activities. This may allow a more precise description of the data variability especially using the proposed additive approach.

Acknowledgments

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

References

Aissaoui-Fqayeh, I., S. El-Adlouni, T. B. M. J. Ouarda et A. St-Hilaire (2009). "Non-stationary lognormal model development anf comparison with non-stationary GEV model." Hydrological sciences journal **54**(6): 1141-1156.

Allan, R. J., N. Nicholls, P. D. Jones et I. J. Butterworth (1991). "A further extension of the Tahiti-Darwin SOI, early ENSO events and Darwin pressure." Journal of Climate **4**(7): 743-749.

Allan, R. P., B. J. Soden, V. O. John, W. Ingram et P. Good (2010). "Current changes in tropical precipitation." Environmental Research Letters **5**(2): 025205.

Anctil, F. et P. Coulibaly (2004). "Wavelet analysis of the interannual variability in southern Québec streamflow." Journal of Climate **17**(1): 163-173.

Barbosa, S. M. (2008). "Quantile trends in Baltic sea level." Geophysical Research Letters **35**(22).

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

Bind, M., B. A. Coull, A. Peters, A. A. Baccarelli, L. Tarantini, L. Cantone, P. S. Vokonas, P. Koutrakis et J. D. Schwartz (2015). "Beyond the Mean: Quantile Regression to Explore the Association of Air Pollution with Gene-Specific Methylation in the Normative Aging Study." Environmental health perspectives.

Buchinsky, M. (1995). "Estimating the asymptotic covariance matrix for quantile regression models a Monte Carlo study." Journal of Econometrics **68**(2): 303-338.

Campbell, E. P. et B. C. Bates (2001). "Regionalization of rainfall-runoff model parameters using Markov chain Monte Carlo samples." WATER RESOURCES RESEARCH **37**(3): 731-739.

Cannon, A. J. (2010). "A flexible nonlinear modelling framework for nonstationary generalized extreme value analysis in hydroclimatology." Hydrological processes **24**(6): 673-685.

Cheng, Y., U. Lohmann, J. Zhang, Y. Luo, Z. Liu et G. Lesins (2005). "Contribution of changes in sea surface temperature and aerosol loading to the decreasing precipitation trend in southern China." Journal of climate **18**(9): 1381-1390.

Chou, C., J. C. Chiang, C.-W. Lan, C.-H. Chung, Y.-C. Liao et C.-J. Lee (2013). "Increase in the range between wet and dry season precipitation." Nature Geoscience **6**(4): 263-267.

Condon, L., S. Gangopadhyay et T. Pruitt (2015). "Climate change and non-stationary flood risk for the upper Truckee River basin." Hydrology and Earth System Sciences **19**(1): 159-175.

Cunderlik, J. M. et D. H. Burn (2002). "Local and regional trends in monthly maximum flows in southern British Columbia." Canadian Water Resources Journal **27**(2): 191-212.

Cunderlik, J. M. et T. B. M. J. Ouarda (2009). "Trends in the timing and magnitude of floods in Canada." Journal of Hydrology **375**(3): 471-480.

Delgado, J., B. Merz et H. Apel (2014). "Projecting flood hazard under climate change: an alternative approach to model chains." Natural Hazards and Earth System Science **14**(6): 1579-1589.

El Adlouni, S., T. B. M. J. Ouarda, X. Zhang, R. Roy et B. Bobée (2007). "Generalized maximum likelihood estimators for the nonstationary generalized extreme value model." WATER RESOURCES RESEARCH **43**(3).

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

Friederichs, P. et A. Hense (2007). "Statistical downscaling of extreme precipitation events using censored quantile regression." Monthly weather review **135**(6): 2365-2378.

Hamed, K. et A. R. Rao (2010). Flood frequency analysis, CRC press.

Hastie, T. et R. Tibshirani (1986). "Generalized additive models." Statistical science: 297-310.

Hundecha, Y. et A. Bárdossy (2005). "Trends in daily precipitation and temperature extremes across western Germany in the second half of the 20th century." International Journal of Climatology **25**(9): 1189-1202.

Hurrell, J. (1996). "Decadal trends in the North Atlantic Oscillation: regional temperatures and precipitation." Oceanographic Literature Review **2**(43): 116.

Ishak, E., A. Rahman, S. Westra, A. Sharma et G. Kuczera (2013). "Evaluating the non-stationarity of Australian annual maximum flood." Journal of Hydrology **494**: 134-145.

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

Jones, M. R., H. J. Fowler, C. G. Kilsby et S. Blenkinsop (2013). "An assessment of changes in seasonal and annual extreme rainfall in the UK between 1961 and 2009." International Journal of Climatology **33**(5): 1178-1194.

Karl, T. R. et R. W. Knight (1998). "Secular trends of precipitation amount, frequency, and intensity in the United States." Bulletin of the American Meteorological society **79**(2): 231-241.

Koenker, R. (2005). Quantile regression, Cambridge university press.

Koenker, R. (2011). "Additive models for quantile regression: model selection and confidence bandaids." Brazilian Journal of Probability and Statistics **25**(3): 239-262.

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

Koenker, R. et F. Schorfheide (1994). "Quantile spline models for global temperature change." Climatic Change **28**(4): 395-404.

Latraverse, M., P. F. Rasmussen et B. Bobée (2002). "Regional estimation of flood quantiles: Parametric versus nonparametric regression models." WATER RESOURCES RESEARCH **38**(6).

Maciejowska, K., J. Nowotarski et R. Weron (2015). "Probabilistic forecasting of electricity spot prices using Factor Quantile Regression Averaging." International Journal of Forecasting.

Madsen, H., D. Lawrence, M. Lang, M. Martinkova et T. Kjeldsen (2013). "A review of applied methods in Europe for flood-frequency analysis in a changing environment."

Mallakpour, I. et G. Villarini (2015). "The changing nature of flooding across the central United States." Nature Climate Change **5**(3): 250-254.

Masselot, P., S. Dabo-Niang, F. Chebana et T. B. M. J. Ouarda (2016). "Streamflow forecasting using functional regression." Hydrology (538): 754-766.

Mazvimavi, D. (2010). "Investigating changes over time of annual rainfall in Zimbabwe." Hydrology and Earth System Sciences **14**(12): 2671-2679.

Milly, P., R. Wetherald, K. Dunne et T. Delworth (2002). "Increasing risk of great floods in a changing climate." Nature **415**(6871): 514-517.

Mishra, A. K., K. A. Mottaleb et S. Mohanty (2015). "Impact of off-farm income on food expenditures in rural Bangladesh: an unconditional quantile regression approach." Agricultural Economics **46**(2): 139-148.

Ouali, D., F. Chebana et T. B. M. J. Ouarda (2016). "Quantile regression in regional frequency analysis: a better exploitation of the available information." Journal of Hydrometeorology.

Pal, I. et A. Al-Tabbaa (2009). "Trends in seasonal precipitation extremes–An indicator of 'climate change'in Kerala, India." Journal of Hydrology **367**(1): 62-69.

Petrow, T. et B. Merz (2009). "Trends in flood magnitude, frequency and seasonality in Germany in the period 1951–2002." Journal of Hydrology **371**(1): 129-141.

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

Riad, S. et J. Mania (2004). "Rainfall-Runoff Model Using an Artificial Neural Network Approach." Mathematical and Computer Modelling **40**: 839-846.

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

Shankman, D., B. D. Keim et J. Song (2006). "Flood frequency in China's Poyang Lake region: trends and teleconnections." International Journal of Climatology **26**(9): 1255-1266.

Shiau, J.-T. et T.-J. Chen (2015). "Quantile regression-based probabilistic estimation scheme for daily and annual suspended sediment loads." Water Resources Management **29**(8): 2805-2818.

Shiau, J.-T. et W.-H. Huang (2015). "Detecting distributional changes of annual rainfall indices in Taiwan using quantile regression." Journal of Hydro-environment Research **9**(3): 368-380.

Shiau, J.-T. et J.-W. Lin (2015). "Clustering Quantile Regression-Based Drought Trends in Taiwan." Water Resources Management: 1-17.

Sivakumar, B. et V. Singh (2012). "Hydrologic system complexity and nonlinear dynamic concepts for a catchment classification framework." Hydrology and Earth System Sciences **16**(11): 4119-4131.

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

Timofeev, A. et A. Sterin (2010). "Using the quantile regression method to analyze changes in climate characteristics." Russian Meteorology and Hydrology **35**(5): 310-319.

Villarini, G., J. A. Smith, M. L. Baeck, R. Vitolo, D. B. Stephenson et W. F. Krajewski (2011). "On the frequency of heavy rainfall for the Midwest of the United States." Journal of Hydrology **400**(1): 103-120.

Villarini, G., J. A. Smith, F. Serinaldi, J. Bales, P. D. Bates et W. F. Krajewski (2009). "Flood frequency analysis for nonstationary annual peak records in an urban drainage basin." Advances in Water Resources **32**(8): 1255-1266.

Vogt, J. (2002). "Upper Rogue smolt trapping project, 2002." Oregon Department of Fish and Wildlife, Rogue Fish District, Central Point, OR.

Wahba, G. (1990). Spline models for observational data. Philadelphia, SIAM.

Xiao, Z., H. Guo et M. S. Lam (2015). Quantile Regression and Value at Risk. Handbook of Financial Econometrics and Statistics, Springer**:** 1143-1167.

Zhai, P., X. Zhang, H. Wan et X. Pan (2005). "Trends in total precipitation and frequency of daily precipitation extremes over China." Journal of climate **18**(7): 1096-1108.

List of Tables

Table 1. RBIAS and RRMSE of quantiles estimated by non-stationary parametric models and QR-based models generated using the GEV_{10} and the GEV_{20} models for N=1000 216
Table 2. Results of Spearman test for the studied regions 217
Table 3. Flood quantile estimations for p=0.90 and 0.99 conditional on the minimum and maximum SOI values 217

Table 1. RBIAS and RRMSE of quantiles estimated by non-stationary parametric models and QR-based models generated using the GEV_{10} and the GEV_{20} models for N=1000

	GEV_{10}		**GEV_{20}**	
	RRMSE (%)	RBIAS (%)	RRMSE (%)	RBIAS (%)
	p=0.50			
Gev_{00}	3.67	-2.42	3.50	-0.45
Gev_{01}	9.53	-0.45	3.58	-0.48
Gev_{10}	2.37	**-0.02**	8.59	-1.54
Gev_{11}	2.40	-0.07	8.59	-1.54
Gev_{20}	3.03	0.28	4.48	-0.53
Gev_{21}	3.32	0.40	4.49	-0.52
LQR	2.65	-0.35	4.07	-0.67
QRAM	**2.30**	-0.42	**3.13**	**-0.20**
	p=0.90			
Gev_{00}	10.14	-9.00	8.03	-4.62
Gev_{01}	24.73	-9.72	8.51	-3.82
Gev_{10}	4.48	0.46	10.93	-3.88
Gev_{11}	4.49	0.174	10.90	-3.85
Gev_{20}	4.72	0.82	9.36	-3.07
Gev_{21}	4.80	1.09	9.38	-3.91
LQR	5.70	**0.16**	9.85	-3.97
QRAM	**4.36**	1.32	**5.78**	**-0.36**
	p=0.99			
Gev_{00}	12.97	-2.09	31.75	-18.16
Gev_{01}	86.47	-18.97	37.40	-21.74
Gev_{10}	14.40	**-0.76**	41.44	-24.79
Gev_{11}	14.83	-1.83	36.46	-21.75
Gev_{20}	14.87	-1.26	31.76	-20.72
Gev_{21}	15.54	-1.40	35.68	-21.40
LQR	13.58	6.65	28.78	**-6.86**
QRAM	**12.60**	6.13	**23.35**	-14.31

Table 2. Results of Spearman test for the studied regions

Station	5% significance level	10% significance level	P-value
Arroyo Seco	No trend	No trend	0.259
Bear Creek	Positive trend	Positive trend	0.001
Dartmouth	No trend	No trend	0.368

Table 3. Flood quantile estimations for p=0.90 and 0.99 conditional on the minimum and maximum SOI values

	SOI	
	Min = -3.16	Max = 2.07
	p=0.90	
GEV0	82	82
GEV01	101	56
GEV10	144	41
GEV11	143	41
GEV20	167	59
GEV21	167	59
LQR	168	11
QRAM	198	11
	p=0.99	
GEV0	710	710
GEV01	733	385
GEV10	772	670
GEV11	767	673
GEV20	796	687
GEV21	791	691
LQR	478	11
QRAM	460	11

List of Figures

Figure 1. Series generated by: (a) the GEV_{10} model and (b) the GEV_{20} model. 219
Figure 2. Geographic locations of Arroyo Seco and Bear Creek stations (a) and Dartmouth station (b). 221
Figure 3. Observed annual maximal discharges in Arroyo Seco (a) and Bear Creek (b) stations. 221
Figure 4. Fitted quantile curves using LQR for p= 0.5, 0.8, 0.9, 0.99: (a) Arroyo Seco, (b) Bear Creek and (c) Dartmouth River. 222
Figure 5. Scatterplots of observed AMD and SOI values for Arroyo Seco (a) and Bear Creek (b) watersheds. Scatterplots of observed AMD and NAO values (c), AMD and the mean annual temperature (d) for the Dartmouth River. 223
Figure 6. Quantile estimations of: (a) the 0.90 non-exceeding probability and (b) the 0.99 non-exceedance probability conditional upon values of the SOI, obtained using the QRAM, LQR, GEV_{20}, GEV_{01}, GEV_{0} models, Arroyo Seco. 224
Figure 7. Effects of the SOI index on floods variability on Arroyo Seco basin, for non-exceeding probabilities of 0.90 and 0.99. 224
Figure 8. SOI recorded values over the calibration and validation periods, Bear Creek. 225
Figure 9. Quantile estimations of the 0.90 non-exceeding probability conditional upon values of the SOI on the Bear Creek station, obtained using the QRAM, LQR, GEV_{20}, GEV_{01}, GEV_{00} models for the calibration (a) and the validation (b) periods. 225
Figure 10. Quantile estimations of the 0.90 non-exceeding probability conditional upon values of the NAO on the Dartmouth station, obtained using the QRAM, LQR, GEV_{20}, GEV_{01}, GEV_{00} models. 226
Figure 11. Effects of NAO and mean annual temperature on flood variability on Dartmouth station using the QRAM, for the 0.90 non-exceeding probability. 226
Figure 12. AMD time series and estimated median form using LQR models (with and without T) (a) and QRAM (with and without T) and GEV_{20} (b) 227

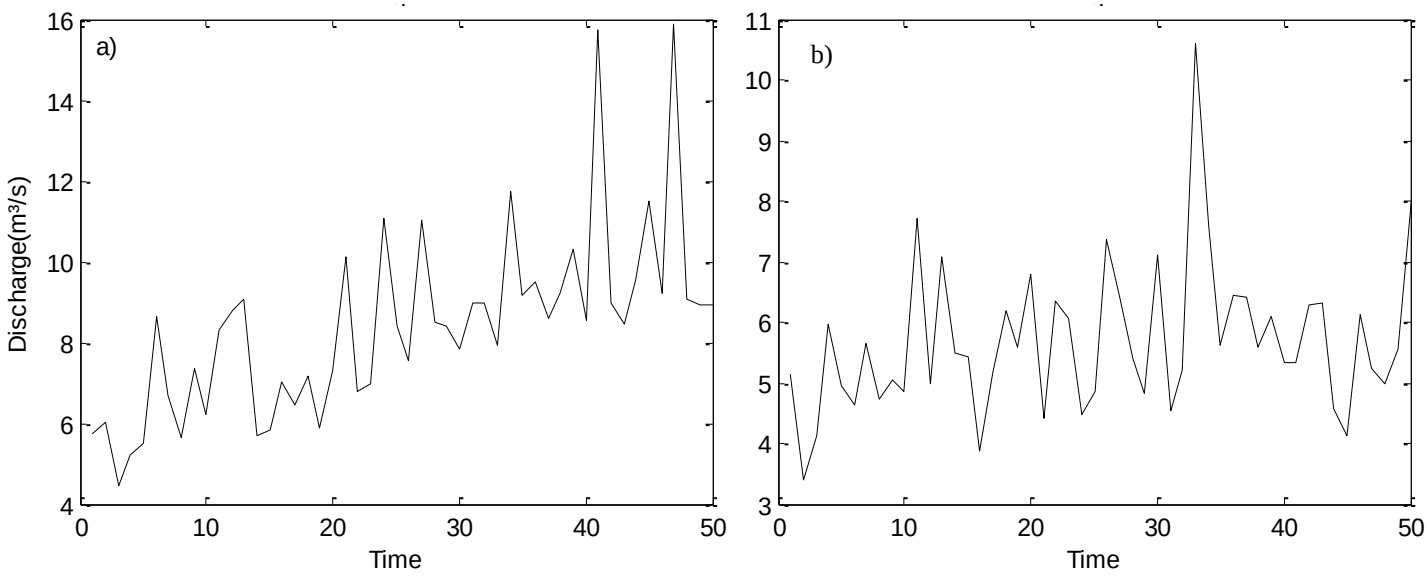
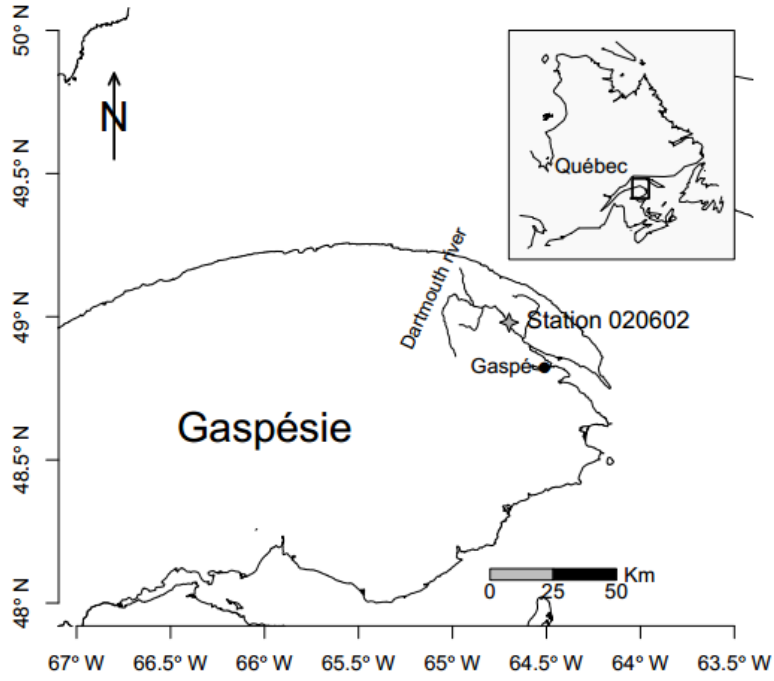


Figure 1. Series generated by: (a) the GEV_{10} model and (b) the GEV_{20} model.


Oregon
Bear Creek
Pacific Ocean
California
Arroyo Seco
United States of America

a)


50° N
49.5° N
49° N
48.5° N
48° N
N
Québec
Dartmouth river
Station 020602
Gaspé
Gaspésie
0
25
50
Km
67° W
66.5° W
66° W
65.5° W
65° W
64.5° W
64° W
63.5° W

b)

Figure 2. Geographic locations of Arroyo Seco and Bear Creek stations (a) and Dartmouth station (b).

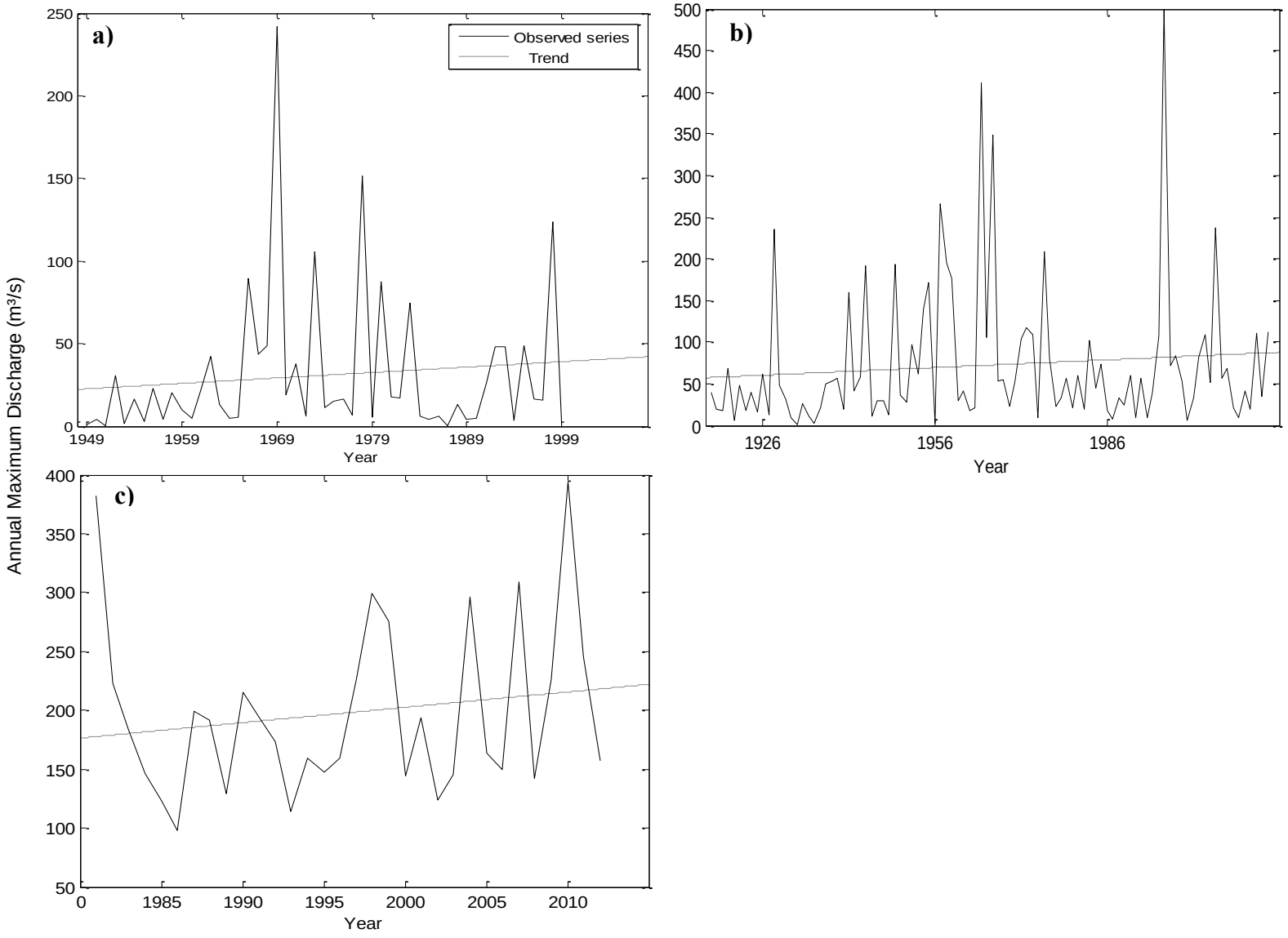


Figure 3. Observed annual maximal discharges in Arroyo Seco (a) and Bear Creek (b) stations.

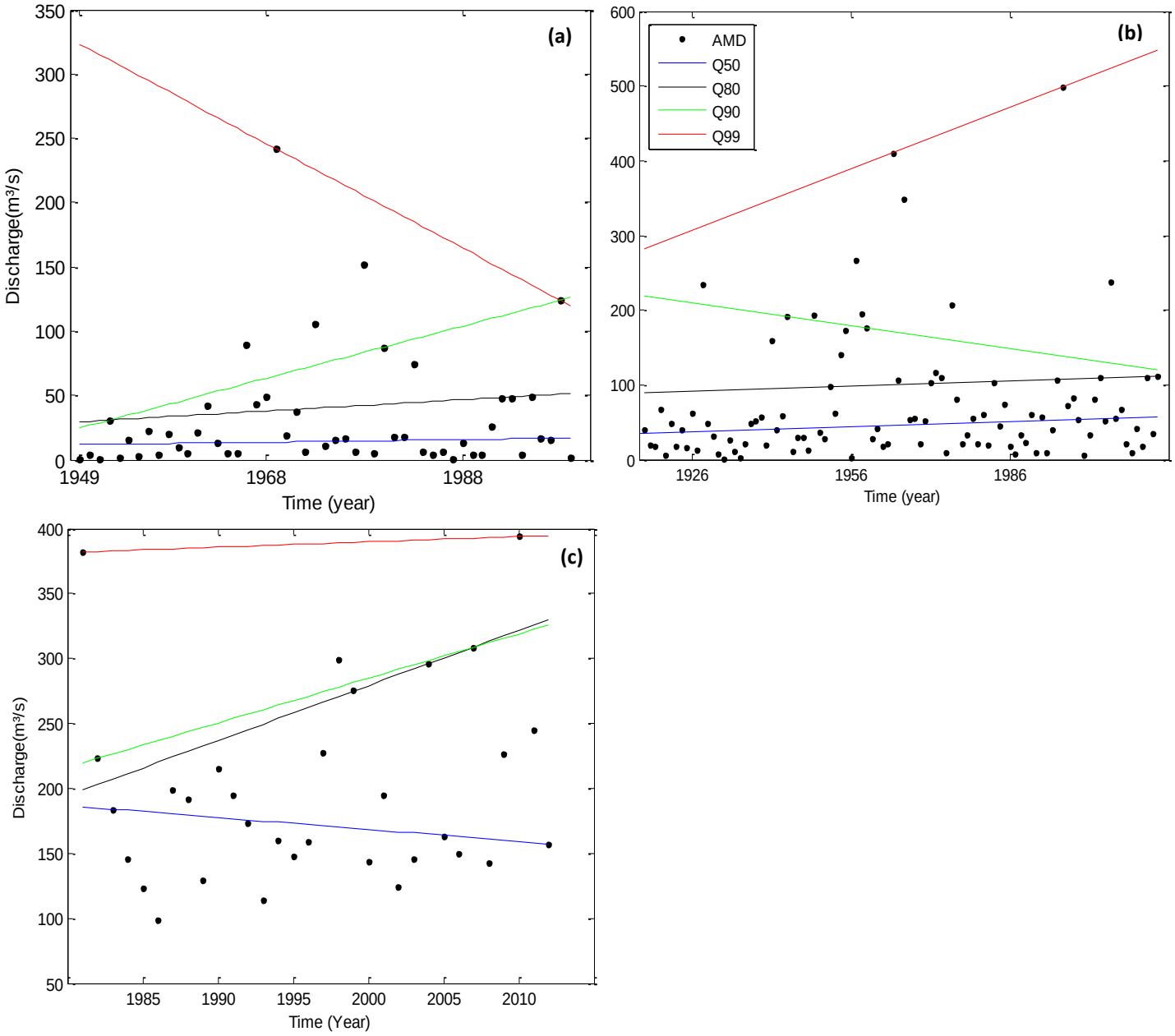


Figure 4. Fitted quantile curves using LQR for p= 0.5, 0.8, 0.9, 0.99: (a) Arroyo Seco, (b) Bear Creek and (c) Dartmouth River.

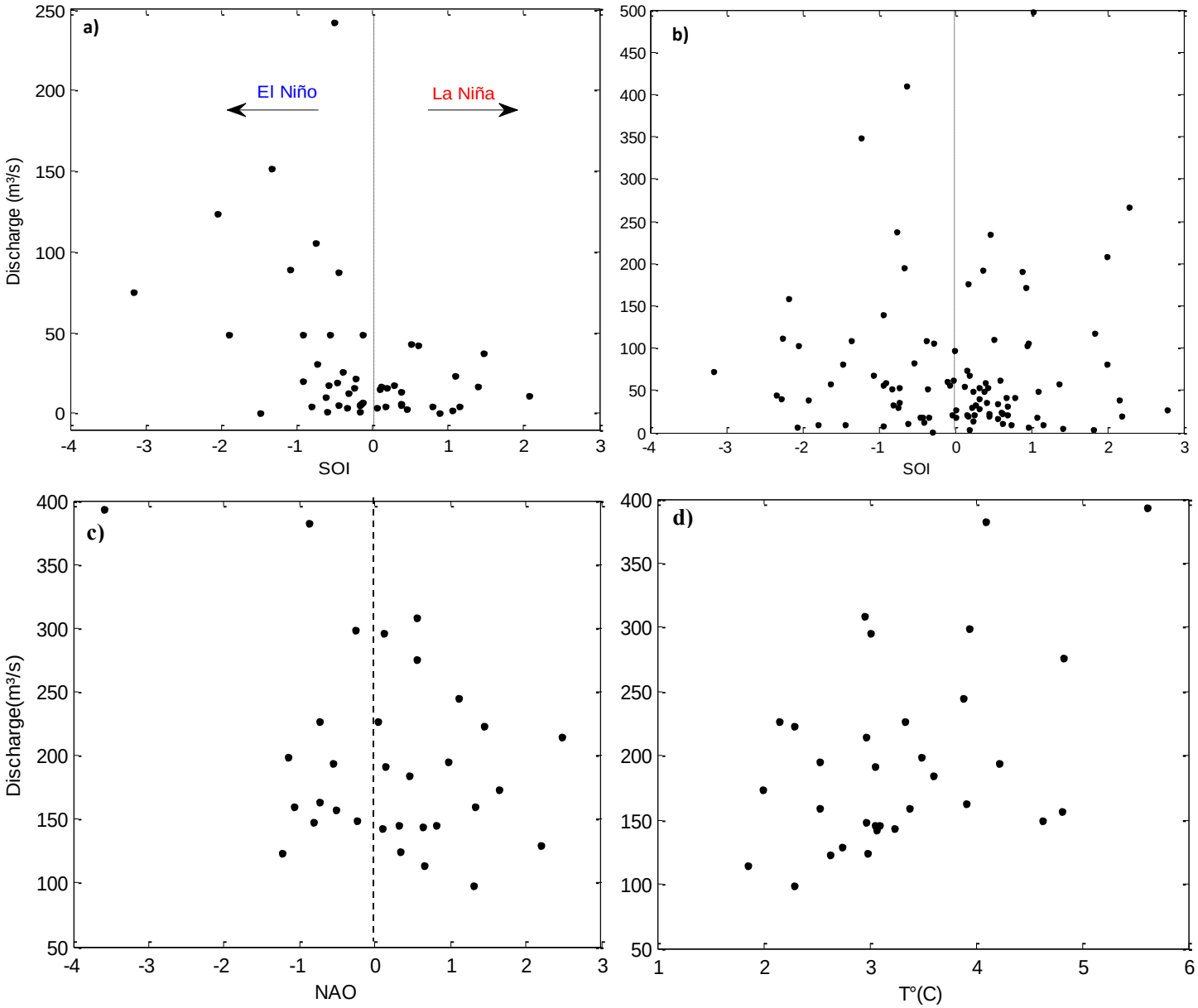


Figure 5. Scatterplots of observed AMD and SOI values for Arroyo Seco (a) and Bear Creek (b) watersheds. Scatterplots of observed AMD and NAO values (c), AMD and the mean annual temperature (d) for the Dartmouth River.

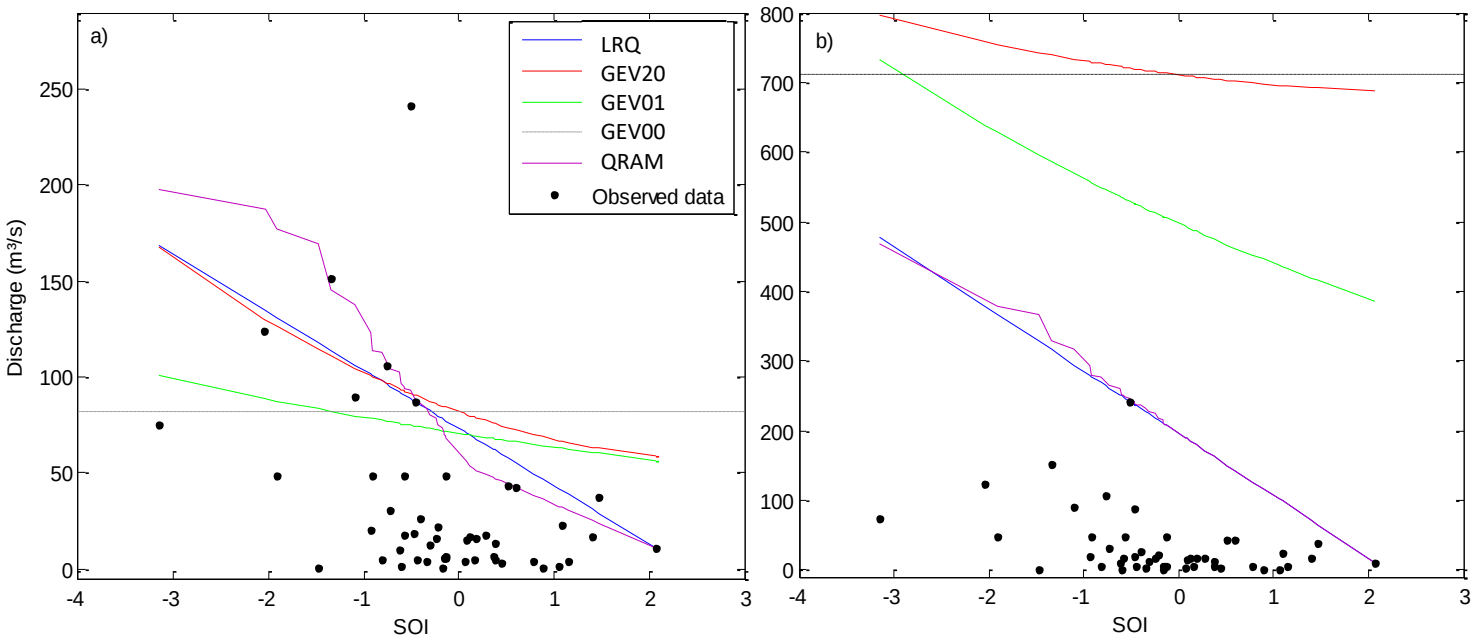


Figure 6. Quantile estimations of: (a) the 0.90 non-exceeding probability and (b) the 0.99 non-exceedance probability conditional upon values of the SOI, obtained using the QRAM, LQR, GEV_{20}, GEV_{01}, GEV_0 models, Arroyo Seco.

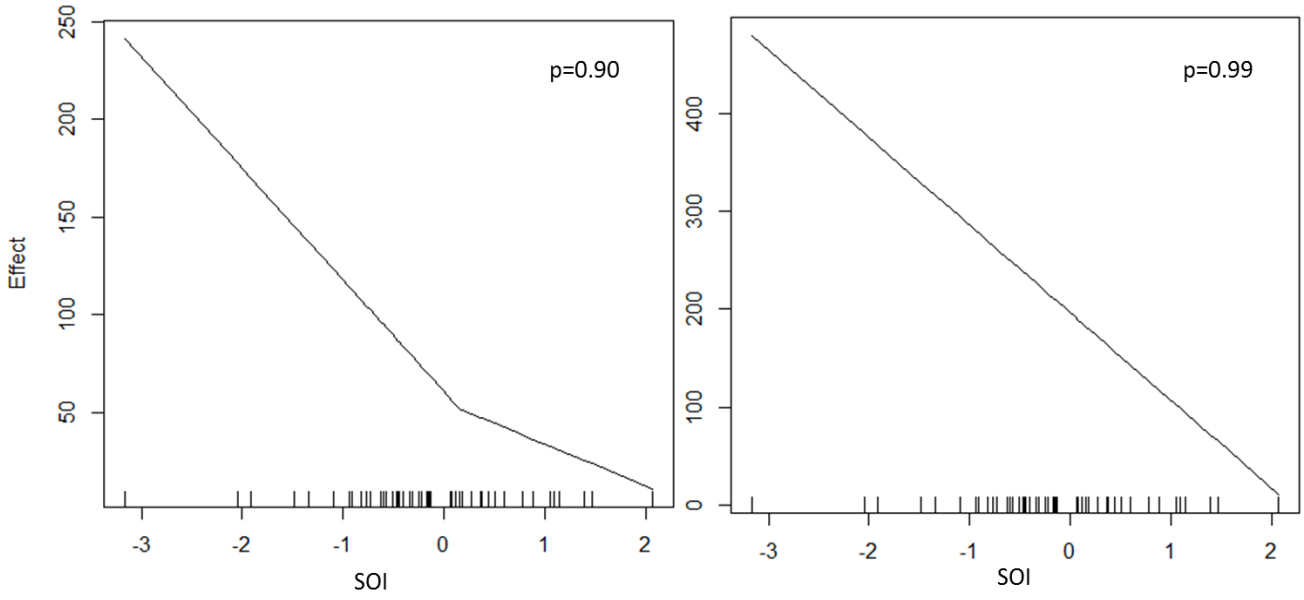


Figure 7. Effects of the SOI index on floods variability on Arroyo Seco basin, for non-exceeding probabilities of 0.90 and 0.99.

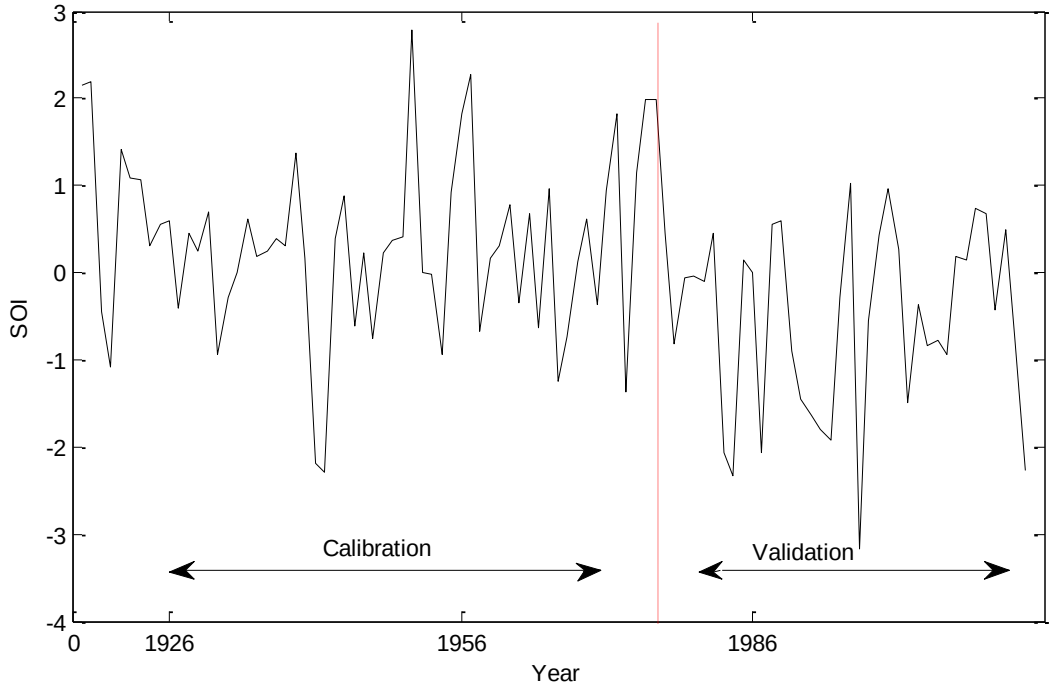


Figure 8. SOI recorded values over the calibration and validation periods, Bear Creek.

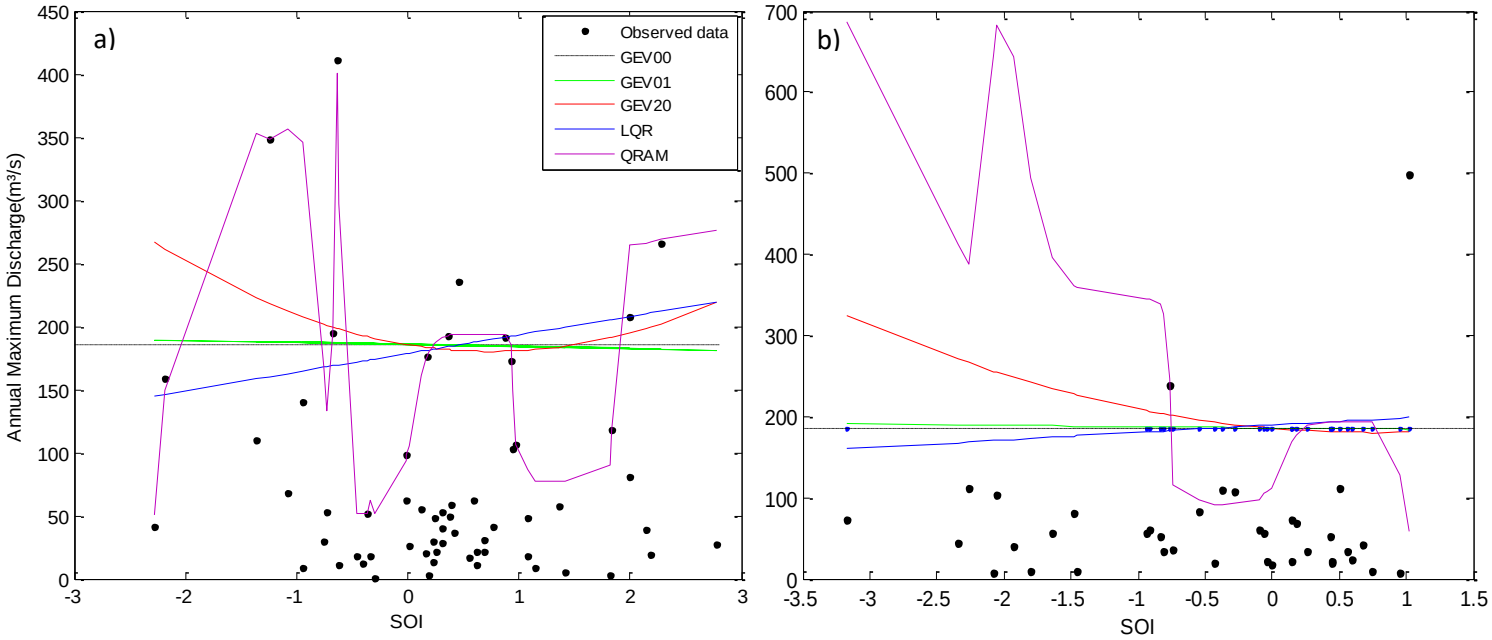


Figure 9. Quantile estimations of the 0.90 non-exceeding probability conditional upon values of the SOI on the Bear Creek station, obtained using the QRAM, LQR, GEV_{20}, GEV_{01}, GEV_{00} models for the calibration (a) and the validation (b) periods.

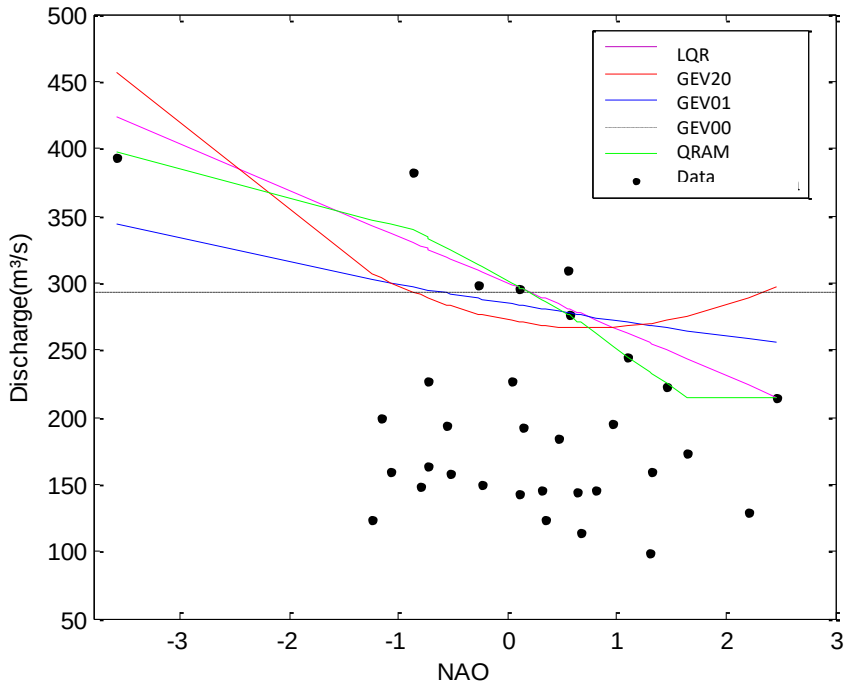


Figure 10. Quantile estimations of the 0.90 non-exceeding probability conditional upon values of the NAO on the Dartmouth station, obtained using the QRAM, LQR, GEV_{20}, GEV_{01}, GEV_{00} models.

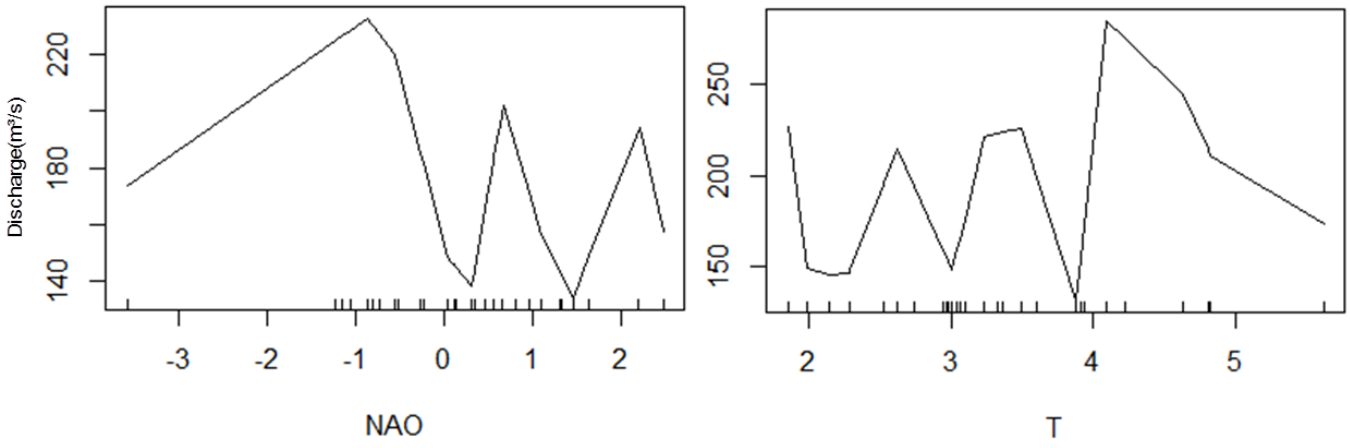


Figure 11. Effects of NAO and mean annual temperature on flood variability on Dartmouth station using the QRAM, for the 0.90 non-exceeding probability.

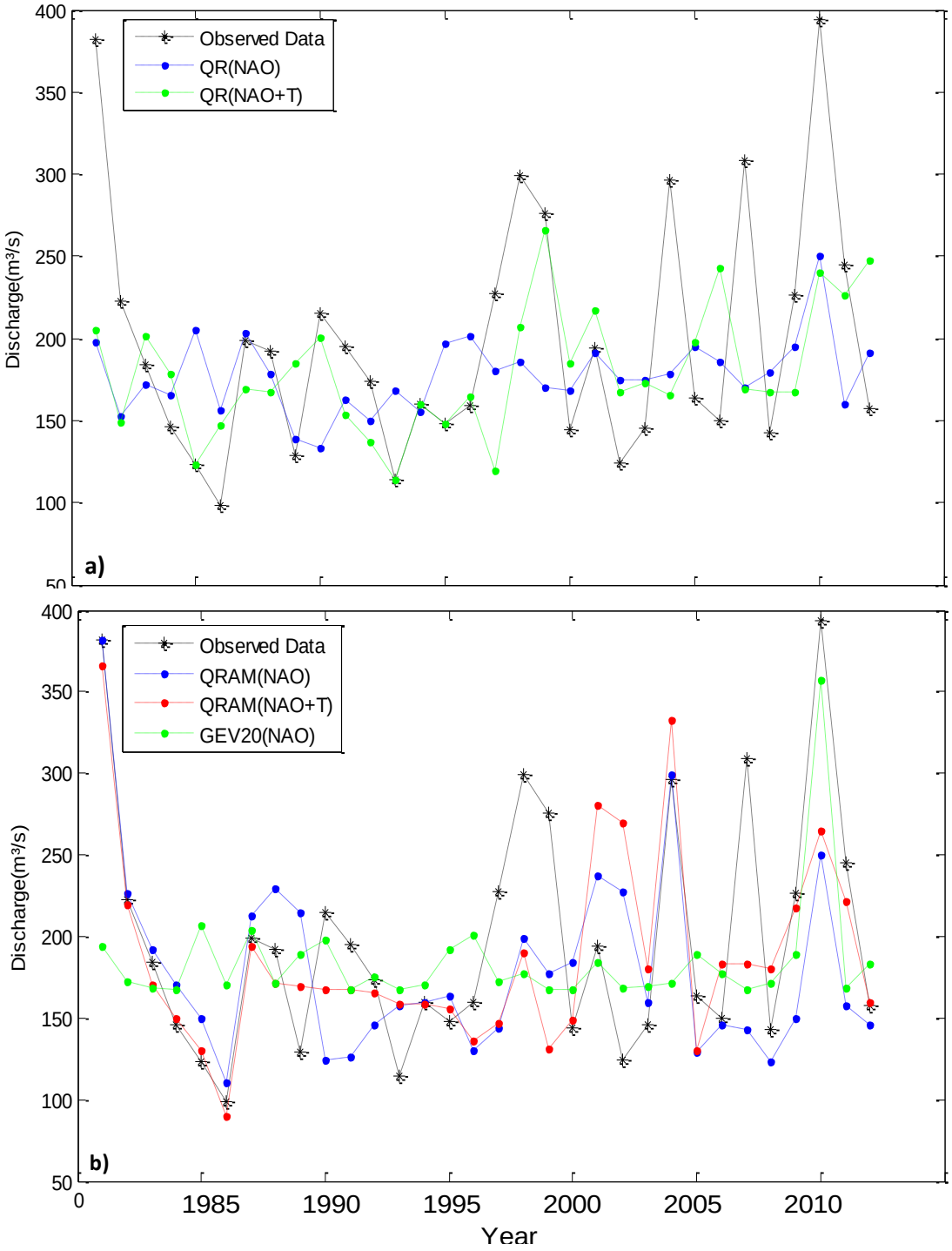


Figure 12. AMD time series and estimated median form using LQR models (with and without T) (a) and QRAM (with and without T) and GEV_{20} (b)