Université du Québec

Institut Nationale de la Recherche Scientifique

Eau, Terre et Environnement

Approches flexibles et optimales en analyse fréquentielle régionale des crues en se basant sur les fonctions de profondeur

Préparé par

Hussein Wazneh

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

Jury d'évaluation :

Examinateur externe	Professeur Amir Aghakouchak Université de Californie, Irvine
Examinateur externe	Professeur Jan Adamowski Université McGill
Examinateur interne	Professeur André St-Hilaire INRS-ETE
Codirecteur	Professeur Taha B.M.J. Ouarda INRS-ETE
Directeur	Professeur Fateh Chebana INRS-ETE

Février 2015

Je dédie cette thèse à ma mère, mes frères et mes sœurs, pour leur soutien malgré la distance; à ma femme Katia et ma fille Lea, pour leur patience et leur encouragement dont elles ont fait preuve pendant toute la durée de cette thèse.

Remerciements

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

Professeur Fateh Chebana, mon directeur de thèse, pour avoir encadré ce travail. Merci d'avoir cru en moi et de m'avoir fait confiance depuis le début, et ce, jusqu'à la fin de ce doctorat. Merci aussi pour tout ce que tu as apporté à ces travaux, tant grâce à ton recul scientifique et tes précieux conseils, que pour m'avoir appris la rigueur de la progression et de l'analyse scientifique. Merci de m'avoir accompagné dans le monde de la recherche. Merci aussi pour nos nombreuses discussions, parfois scientifiques, parfois moins. En attendant que nos chemins se recroisent peut être un jour, je te souhaite le meilleur, tant d'un point de vue professionnel que personnel.

Professeur Taha B.M.J. Ouarda, mon codirecteur de thèse, pour sa confiance et la liberté qu'il m'a accordées pour mener cette étude. Les discussions fructueuses que nous avons eues lors de nos rencontres et ses conseils ont été pour moi une source de motivation. Lors de notre dernière discussion, tu m'as dit '*Je crois que tu es prêt pour t'envoler de tes propres ailes*', donc voilà ce jour arrive, j'espère être toujours à la hauteur de tes attentes.

Le Conseil de Recherche en Sciences Naturelles et Génie du Canada pour avoir financé ma thèse et les membres du groupe de recherche en hydroclimatologie statistique, professeurs et étudiants, pour leur coopération et leur aide.

À la mémoire de *Ali Wazneh*

Préface

Cette thèse présente les travaux de recherche menés au cours de mes études doctorales. La structure de la présente thèse suit la structure standard des thèses par articles de l'INRS-ETE. La première partie de la thèse comporte une synthèse générale des travaux effectués. Cette synthèse a pour objectif de survoler la méthodologie adoptée et les principaux résultats obtenus au cours de la thèse. La deuxième partie contient quatre articles comme des chapitres, publiés (3) et soumis (1) à des revues internationales.

Articles et contribution des auteurs

[1] **Wazneh, H.,** Chebana, F., et Ouarda, T. B. M. J. (2015): Identification of hydrological neighborhood using statistical depth function (Soumis).

[2**] Wazneh, H.,** Chebana, F., et Ouarda, T. B. M. J. (2014): Delineation of homogeneous region for regional frequency analysis using statistical depth function, *Journal of Hyd.*, 521, 232-244, doi:10.1016/j.hydrol.2014.11.068.

[3] **Wazneh, H.,** Chebana, F., et Ouarda, T. B. M. J. (2013): Optimal depth-based regional frequency analysis, *Hydrol. Earth Syst. Sci.*, 17, 2281-2296, doi:10.5194/hess-17-2281-2013.

[4] **Wazneh, H.,** Chebana, F., et Ouarda, T. B. M. J. (2013)*:* Depth-based regional index-flood model, *Water Resour. Res.*, 49, 7957–7972, doi:10.1002/2013WR013523.

Dans le premier article, H. Wazneh a présenté une nouvelle technique basée sur les fonctions de profondeur pour déterminer les régions hydrologiques homogènes de type voisinage. F. Chebana et T. B. M. J. Ouarda ont commenté et révisé la version finale du manuscrit.

Dans le deuxième article, H. Wazneh a proposé une nouvelle approche pour délimiter les régions homogènes de type géographiquement non contiguës, pour l'analyse fréquentielle régionale (AFR). Cette approche est basée sur la notion statistique des fonctions de profondeur. Tout au long de ce travail, F. Chebana et T. B. M. J. Ouarda ont donné de précieux conseils et suggestions et ils ont révisé la version finale du manuscrit.

Dans le troisième article, H. Wazneh a présenté un nouveau modèle régional d'indice de crue pour l'estimation des évènements hydrologiques extrêmes. F. Chebana et T. B. M. J. Ouarda ont discuté l'aspect hydrologique de ce modèle et ils ont révisé la version finale du manuscrit.

Dans le quatrième article, H. Wazneh a reformulé le modèle de régression régional afin d'optimiser sa performance et réduire sa complexité. F. Chebana et T. B. M. J. Ouarda ont commenté et révisé la version finale du manuscrit.

Résumé

L'estimation adéquate des phénomènes hydrologiques extrêmes est primordiale en raison des risques importants associés à une compréhension insuffisante de ces phénomènes. Cette estimation est obtenue avec une analyse fréquentielle hydrologique pour un site jaugé. Toutefois, pour diverses raisons, on est souvent amené à produire des estimations dans des sites non jaugés. Dans cette situation, on fait appel à une procédure de régionalisation. Elle comporte deux principales étapes, la délimitation des régions hydrologiquement homogènes et l'estimation régionale. Les approches de régionalisation disponibles dans la littérature présentent certaines contraintes et limitations. Par exemple, les approches traditionnelles de délimitation des régions non contiguës où voisinages sont basées sur des mesures non robustes, et les modèles d'estimation régionale ne sont pas flexibles et avec des performances pas nécessairement optimales.

Dans cette thèse, nous proposons de nouvelles approches robustes, flexibles et optimales pour les deux étapes de l'analyse régionale des crues. Ces nouvelles approches sont mises au point en introduisant des nouvelles notions et quantités statistiques dans les approches classiques couramment utilisées dans la littérature. Par construction, les approches traditionnelles représentent des cas spéciaux de ces nouvelles approches.

En ce qui concerne la délimitation des régions homogènes, nous proposons deux nouvelles méthodes basées sur les fonctions de profondeur. Ces fonctions évaluent la dissimilarité entre le site cible et les sites jaugés d'une région. Les résultats issus de ces nouvelles méthodes sont indépendants des échelles et des distributions des variables physiographiques, ce qui n'est pas le cas des méthodes traditionnelles. En plus, nous montrons que les méthodes

proposées conduisent à des régions plus homogènes, et produisent des estimations des quantiles moins biaisées que celles obtenues par les approches traditionnelles.

En ce qui concerne l'estimation régionale, nous proposons deux modèles régionaux flexibles et optimaux. Ces deux modèles sont fondés respectivement sur le modèle d'indice de crue et le modèle de régression multiple. En y incluant les fonctions de profondeur, les modèles proposés sont plus représentatifs des phénomènes hydrologiques. La flexibilité de ces modèles est obtenue en introduisant les fonctions de poids dans l'estimation de leurs paramètres. Cette flexibilité a permis également d'optimiser la performance des modèles. Nous montrons que pour l'estimation des quantiles des crues, ces nouveaux modèles surpassent très nettement les approches traditionnelles. Ce résultat est d'autant plus vrai lorsque les périodes de retour deviennent importantes.

Mots-clés : Analyse fréquentielle régionale, fonctions de profondeur, fonctions de poids, algorithme d'optimisation, région d'influence, classification hiérarchique, analyse canonique des corrélations, indice de crue, modèle de régression multiple, quantile, période de retour, sites non-jaugés.

Table des matières

CHAPITRE 1 : SYNTHÈSE

1. Introduction

L'eau, une des plus importantes ressources naturelles, doit être protégée de manière à garantir de façon durable un équilibre entre les besoins et son utilisation. Le développement durable des activités humaines s'appuie, en particulier, sur une gestion intégrée des eaux. Une gestion efficace et durable des eaux ne se limite pas à garantir une quantité et une qualité suffisantes pour les demandes humaines (eau potable, industrielle, irrigation) et pour les besoins des milieux naturels; mais également, cette gestion doit prendre en compte la manifestation des événements extrêmes, tels que les étiages et les crues.

Pour protéger la population avec leurs habitations des événements extrêmes, plusieurs mesures peuvent être prises. À titre d'exemple, des ouvrages hydrauliques (barrages, digues, ponts) peuvent être construits et des plans de préventions des risques peuvent être établis. Le dimensionnement des ouvrages hydrauliques, l'établissement des plans de prévention sont fondés sur les prévisions des débits d'une rivière ainsi que sur l'estimation de l'amplitude et la fréquence de débits extrêmes.

L'analyse fréquentielle (AF) des variables hydrologiques est une approche couramment utilisée pour obtenir des estimations des événements extrêmes. Quand l'information hydrologique est disponible au site d'intérêt et est de bonne qualité, l'AF peut être envisagée et peut donner des résultats efficaces. Dans ce cas on parle de l'AF locale. L'objectif principal de l'AF des variables hydrologiques est de relier l'amplitude des événements extrêmes à leur fréquence d'occurrence à travers des distributions statistiques [*Khaliq et al.*, 2005; *Naulet et al.*,

2005]. Dans le cadre de cette thèse, on s'intéresse particulièrement à l'estimation de l'amplitude des crues, comme variables hydrologiques extrêmes.

On dispose de trois différents modèles d'AF permettant d'effectuer l'estimation des amplitudes de crue [*Cunnane*, 1987]. Ces modèles sont basés sur: (1) des séries de débits maximums annuels (DMA), (2) des séries de dépassements (séries partielles) et (3) des séries temporelles à haute fréquence. Dans le cadre de ce travail nous intéressons aux méthodes d'estimation basées sur l'utilisation des DMA. Dans ce cas, on considère une crue par année hydrologique correspondant à la valeur maximale du débit (pointe de crue). Les DMA constitue une variable aléatoire, notée X, ayant une fonction de distribution F ainsi qu'une fonction de densité de probabilité f. Selon cette approche, l'amplitude du débit de la crue Q_T associée à une période de retour T, qui correspond à une probabilité de non-dépassement prédéfinie t, est donnée par :

$$Q_T = F^{-1}(t) = F^{-1}\left(1-\frac{1}{T}\right) \qquad (1)$$

avec $F^{-1}(.)$ est la fonction de répartition inverse (ou fonction quantile) de la variable aléatoire X.

De manière générale, l'AF des crues consiste à ajuster une distribution F en utilisant les mesures de DMA disponibles au site d'intérêt. Plusieurs distributions ont été proposées pour l'analyse locale des crues, comme la loi log-normale (LN); la distribution des valeurs extrêmes généralisée (GEV) et la loi log-Pearson type 3 (LP3). Notons qu'une crue de période de retour de T années correspond au débit de la crue qui est dépassée en moyenne une fois toutes les T années.

En raison des grandes étendues territoriales et du coût associé à l'installation et au maintien de stations de mesure, il arrive fréquemment qu'on doit produire des estimations de Q_T

dans des sites non jaugés où l'on ne dispose pas d'information hydrologique. Pour tenter de contrer ce problème, les approches de régionalisation sont nécessaires et utilisées afin de transférer l'information disponible dans les sites jaugés vers le site cible non jaugé ou partiellement jaugé [e.g., *Burn*, 1990; *Dalrymple*, 1960]. De manière générale, la détermination des régions hydrologiquement homogènes et l'estimation régionale sont les deux étapes principales pour effectuer une AF régionale (AFR).

Le but de la détermination des régions homogènes (DRH) est de regrouper les sites ayant un comportement semblable [*Chebana et Ouarda*, 2008]. Plus précisément, cette homogénéité doit tenir compte des caractéristiques physiographiques, hydrologiques et météorologiques des bassins versants. Les régions homogènes peuvent être géographiquement contiguës, géographiquement non contiguës ou de type voisinage [*Ouarda et al.*, 2001]. Toutefois, les régions non contiguës et de type voisinage sont recommandées dans la littérature hydrologique pour les estimations des crues [*GREHYS*, 1996]. Pour définir les régions non contiguës, la classification ascendante hiérarchique (CAH) est la principale approche utilisée en AFR [*Hosking et Wallis*, 1997]. Par ailleurs, pour les régions de type voisinage, Burn [1990] a proposé une approche nommée région d'influence (ROI) où une distance euclidienne a été utilisée pour mesurer la similarité entre le site cible et les sites jaugés. Cavadias [1990] a proposé une deuxième approche de type voisinage en utilisant l'analyse canonique de corrélations (ACC). Les trois approches CAH, ROI et ACC couramment utilisées dans la littérature en AFR et considérées dans cette thèse sont présentées brièvement dans la section suivante.

L'estimation régionale (ER), deuxième étape d'une AFR, fait référence au transfert d'information des sites jaugés vers un site cible non jaugé ou partiellement jaugé, à l'intérieur

d'une même région homogène. De nombreuses procédures et techniques d'estimation régionale des quantiles des crues ont été proposées et appliquées dans plusieurs régions du monde [*Burn*, 1990; *Chebana et Ouarda*, 2008; *Ouarda et al.*, 2001]. À titre d'exemple, Chokmani et Ouarda [2004] ont proposé une approche basée sur le krigeage ordinaire pour interpoler les quantiles de crue dans l'espace physiographique créé à l'aide de l'ACC. Shu et Burn [2004] ont introduit les réseaux de neurones artificiels (RNA) dans le modèle d'estimation régionale. Par la suite, la logique floue et les RNA sont également utilisées dans plusieurs modèles régionaux des crues (p. ex., Shu et Ouarda [2007; 2008]). Pour une présentation récente et complète des variantes et des combinaisons de ces modèles, ainsi que d'autres modèles, le lecteur est invité à consulter Ouarda [2013]. Notons que malgré les nombreux modèles d'estimation régionale qui existent, les deux modèles indice de crue et régression régionale sont les plus couramment utilisés dans la littérature hydrologique [e.g., *Haddad et Rahman*, 2012; *Hosking et Wallis*, 1997]. GREHYS [1996] comparent ces deux modèles régionaux. Cette comparaison indique que généralement, il n'est pas possible de déterminer le meilleur (absolu) modèle d'estimation. Toutefois, pour une région caractérisée par un degré d'homogénéité élevé, le modèle d'indice de crue est préféré, sinon un modèle de régression régionale peut représenter une meilleure alternative. Dans le modèle de la régression régionale des quantiles [*Benson*, 1962], on suppose une relation log-linéaire entre les quantiles des crues (comme variable dépendante) et les variables physiographiques et climatologiques (comme variables explicatives) des différents sites d'une région. Ce modèle est simple, rapide et permet d'utiliser différentes distributions pour les différents sites. Tandis qu'avec le modèle d'indice de crue [*Dalrymple*, 1960], on fait l'hypothèse que les données de pointes de crues aux différents sites d'une région homogène suivent la même

distribution statistique à un facteur d'échelle près. Une description détaillée de ces deux modèles (c.-à-d. indice de crue et régression régionale) utilisés dans cette thèse se trouve dans la section suivante.

La présente synthèse est organisée comme suit. La section 2 présente une brève revue de littérature sur les méthodes de régionalisation couramment utilisées dans les études hydrologiques. La section 3 présente la problématique, les objectifs et l'originalité du projet de recherche. Les outils statistiques utilisés pour réaliser les objectifs de la thèse sont présentés dans la section 4. La section 5 présente la méthodologie adoptée dans la thèse. Les principaux résultats obtenus sont présentés dans la section 6. Enfin, la conclusion et les perspectives de recherche sont présentées dans la section 7.

2. Méthodes traditionnelles de régionalisations

Cette section contient la formulation des méthodes de régionalisation des débits de crue couramment utilisées dans la littérature hydrologique. Elle est divisée en deux sous-sections selon les deux étapes de l'AFR. La première est consacrée à la description des méthodes utilisées pour déterminer les régions homogènes. La deuxième montre les techniques utilisées pour produire une estimation régionale pour des sites non-jaugés.

2.1. Méthodes de détermination des régions homogènes

Le but principal de la DRH est de regrouper les sites ayant un comportement hydrologique similaire. La DRH est donc très importante, puisqu'elle pourra avoir des

répercussions significatives sur les phases subséquentes de la régionalisation. Tout d'abord, l'ACC et la méthode des ROI seront présentées, puis suivra la CAH.

2.1.1. Analyse canonique de corrélation (ACC)

L'ACC est une méthode d'analyse statistique multivariée qui permet de décrire la relation de dépendance existant entre deux ensembles de variables aléatoires. Une description générale de cette méthode peut être trouvée dans Ouarda et al. [2001] dans le contexte d'AFR.

Soient $X=\left(X_1,..,X_i,..,X_r\right)$; $X_i=\left(X_i^1,...,X_i^N\right)'$ et $Y=\left(Y_1,..,Y_i,..,Y_s\right)$; $Y_i=\left(Y_i^1,..,Y_i^N\right)'$ les deux matrices contiennent respectivement les *r* et les *s* variables physio-météorologiques et hydrologiques des *N* bassins versants. La matrice de covariance entre *X* et *Y*, sous forme de blocs, est donnée par :

$$\Sigma=\mathrm{cov}\begin{pmatrix}X\\Y\end{pmatrix}=\begin{bmatrix}\Sigma_X & \Sigma_{XY}\\ \Sigma_{YX} & \Sigma_Y\end{bmatrix} \qquad (2)$$

L'ACC consiste en la détermination de deux ensembles de combinaisons linéaires *V* et *W*, respectivement pour *X* et *Y*. Ces combinaisons sont appelées variables canoniques, et sont données par:

$$\begin{aligned}V&=a'X\\W&=b'Y\end{aligned} \qquad (3)$$

où *a* et *b* sont choisis de telle sorte qu'ils maximisent $\rho=\dfrac{a'\,\Sigma_{XY}b}{\sqrt{a'\,\Sigma_X ab'\,\Sigma_Y b}}$ sous la contrainte que $Var\left(V\right)=Var\left(W\right)=1.$

Soient $W^m = \left(W_1^m, ..., W_p^m\right)'$ et $V^m = \left(V_1^m, ..., V_p^m\right)'$ respectivement les deux vecteurs des variables canoniques hydrologiques et physio-météorologiques d'un site m $\left(m = 1, .., N\right)$, avec p est le rang de la matrice Σ_{XY}. Notons que pour un site cible non jaugé l la variable canonique physio-météorologique V^l du site l est connue tandis que sa variable hydrologique W^l est inconnue. En supposant que le vecteur $\begin{pmatrix} W \\ V \end{pmatrix}$ est 2p-normalement distribué, le voisinage hydrologique pour le site cible l en utilisant la méthode ACC, pour un niveau de confiance de 100(1-α)%, est donné par :

$$\left\{ \text{site } k \in \{1, .., N\};\ d^2_{I_p - \Lambda'\Lambda}\left(W^k, \Lambda V^l\right) = \left(W^k - \Lambda V^l\right)' \left(I_p - \Lambda'\Lambda\right)^{-1} \left(W^k - \Lambda V^l\right) < \chi^2_{\alpha,p} \right\} \qquad (4)$$

où I_p est la matrice identité de dimension p, $\Lambda = \text{diag}(\lambda_1, ..., \lambda_p)$ avec $\lambda_i = \text{cor}(V_i, W_i), i = 1, ..., p$, $V_i = \left(V_i^1, ..., V_i^N\right)$ et $W_i = \left(W_i^1, ..., W_i^N\right)$, d^2 est la distance de Mahalanobis, et $\chi^2_{\alpha,p}$ est le (1-α) quantile associé à la distribution du Chi-2 avec p degrés de liberté.

Le voisinage (4) d'un site cible, en utilisant la méthode ACC, décrit l'intérieur d'une région ellipsoïdale de coefficient α. La valeur de ce dernier α peut varier entre [0,1], dont la valeur α = 0 implique l'inclusion de tout l'espace hydrologique et α = 1 seulement un point de l'espace. La valeur de α doit être telle que l'ellipse peut contenir le plus des sites similaires, mais aussi éviter de considérer les sites ayant une faible similarité hydrologique. La valeur optimale de α est calculée en optimisant un critère qui quantifie l'erreur du modèle [*Ouarda et al.*, 2001] (p. ex., le biais relatif 'RB', l'erreur relative quadratique moyenne 'RRMSE').

2.1.2. Région d'influence (ROI)

Cette méthode de détermination des régions homogènes a été proposée pour la régionalisation des crues dans les années 80 [*Acreman et Sinclair*, 1986]. Burn [1990] a désigné l'application de cette méthode par le terme « région d'influence (ROI) ». Dans cette méthode, comme dans la méthode d'ACC, chaque site peut être considéré comme le centre d'une région formée des sites dont les caractéristiques de crue du basin sont similaires.

Pour identifier le voisinage d'un site cible (non-jaugé), Burn [1990] utilise une distance euclidienne dans un espace multidimensionnel, dont les axes sont les variables physio-météorologique des bassins. Plus formellement, soit N le nombre total des sites jaugés dans la région et $X^m = \left(X_1^m, .., X_r^m\right)$ le vecteur qui contient les r variables physio-météorologiques (attributs) du site m $\left(m = 1, .., N\right)$. Le voisinage hydrologique pour un site cible l en utilisant la méthode ROI est donné par :

$$\left\{ \text{site } k \in \{1, .., N\}; D\left(X^k, X^l\right) = \left[\sum_{s=1}^{r}\left(X_s^k - X_s^l\right)^2\right]^{1/2} \leq \delta \right\} \qquad (5)$$

où D est la distance euclidienne et δ est un seuil choisi. Généralement, une standardisation de la matrice X qui contient les N vecteurs des variables physio-météorologiques est nécessaire afin d'enlever l'effet des unités de ces variables. La valeur de δ représente un compromis entre la quantité d'information (nombre de sites voisins) et l'homogénéité hydrologique du voisinage. Zrinji et Burn [1994] proposent un test d'homogénéité pour déterminer le nombre optimal de sites voisins dans la ROI.

2.1.3. Classification ascendante hiérarchique (CAH)

Généralement la CAH est utilisée pour constituer des groupes d'individus similaires sur la base de leur description par un ensemble de variables quantitatives. Elle consiste à agréger progressivement les individus selon leur ressemblance, mesurée à l'aide d'un indice de similarité ou de dissimilarité. L'algorithme commence par rassembler les couples d'individus qui se ressemblent le plus, puis à agréger progressivement les autres individus ou groupes d'individus en fonction de leur ressemblance, jusqu'à ce que la totalité des individus ne forme plus qu'un seul groupe. La CAH produit un dendrogramme (arbre binaire de classification), dont sa racine correspond à la classe regroupant l'ensemble des individus.

Il existe de nombreuses mesures de ressemblances (similarités ou dissimilarités), et plusieurs méthodes pour recalculer la ressemblance lorsque l'algorithme forme des groupes (critères d'agrégations). Toutefois, la méthode de Ward [1963] est la plus couramment employée en AFR. Elle utilise la distance euclidienne pour calculer la similarité entre les sites et l'inertie (variance) comme critère d'agrégation. Cette méthode est utilisée dans cette thèse, car elle tend à découper une région non homogène en des sous-régions homogènes et de tailles équilibrées, ce qui est approprié en AFR [*Hosking et Wallis*, 1997].

2.2. Méthodes d'estimation régionale

Dans cette section les deux approches de modélisation régionale couramment utilisées en AFR sont présentées, à savoir le modèle de l'indice de crue et celui de la régression log-linéaire.

2.2.1. Indice de crue

La méthode de l'indice de crue (index-flood) a été introduite par Dalrymple [1960]. Cette méthode fait l'hypothèse que dans une région homogène, toutes les données locales normées par un indicateur position centrale (e.g. moyenne, médiane) ont la même distribution, ce qui permet de définir une distribution régionale.

Plus formellement, dans une région homogène, en utilisant le modèle de l'indice de crue, le quantile de crue correspond à une probabilité de non-dépassement *t* (i.e. période de retour $T = 1/(1-t)$) est estimé par :

$$Q_l(t) = \mu_l q_t\left(\theta^R\right) \quad ; 0 < t < 1 \text{ et } \theta^R = \left(\theta_1^R, .., \theta_S^R\right) \qquad (6)$$

où *l* est l'indice du site cible, μ_l est la quantité indice de crue (indicateur de position centrale), $q_t(.) = F^{-1}(t,.)$ est la courbe de croissance (inverse de la fonction de répartition) et θ^R est le vecteur de paramètres régionaux à *S* composantes. Le nombre de composantes *S* est généralement égal à 2 ou 3 puisque la majorité des distributions utilisées en AFR ont 2 ou 3 paramètres (p. ex. paramètres de position et d'échelle pour la distribution LN; paramètres de position, d'échelle et de forme pour la distribution GEV).

L'estimation des quantiles de crue pour un site cible *l* nécessite la sélection de $q_t(.)$ et l'estimation de μ_l et de θ^R. Pour sélectionner $q_t(.)$, le diagramme des *L*-moments régionaux est utilisé [*Vogel et al.*, 1993]. Cependant, l'indice de crue μ_l est généralement estimé par une régression sur les variables physio-météorologiques [*Grover et al.*, 2002]. Finalement, les

paramètres régionaux sont estimés par des combinaisons linéaires des paramètres des distributions locales normées. En effet, le $s^{\text{ème}}$ paramètre régional est estimé par :

$$\hat{\theta}_s^R = \sum_{h=1}^{N'} \omega_h \hat{\theta}_s^h, \quad s = 1,...,S \tag{7}$$

où $\hat{\theta}_s^h$ est le $s^{\text{ème}}$ paramètre obtenu à partir de la distribution locale du site *h*, ω_h est le poids associé au site *h* pour $h = 1,...,N'$ et N' est le nombre total des sites dans la région homogène. Notons que tous les sites de la région homogène ont la même distribution locale, et cette dernière est identique à celle associée à $q_t^{-1}(.)$. Dans la littérature, trois différents types de poids ω_h ont été proposés pour l'estimation des paramètres régionaux. En effet, Hosking et Wallis [1997] ont proposé le poids proportionnel défini par (désignée par "PW-index-flood") :

$$\omega_h = \frac{n_h}{n}; \quad h = 1,..,N' \tag{8}$$

avec n_h est la longueur d'enregistrement du site *h* et $n = n_1 + ... + n_{N'}$ est la longueur totale d'enregistrement. Si la région n'est pas parfaitement homogène, il est possible que le poids proportionnel (8) conduit à surestimer ou à sous-estimer le vecteur des paramètres régionaux. Par conséquent, le poids uniforme (UW-index-flood) est proposé comme une alternative [*Jin et Stedinger*, 1989; *Minghui et Stedinger*, 1989]. Dans le cas de UW-index-flood, le vecteur des paramètres régionaux est égal à la moyenne simple des vecteurs des paramètres locaux :

$$\omega_h = \frac{1}{N'}; \quad h = 1,..., N' \tag{9}$$

Un autre choix qui réduit les poids attribués à des sites ayant une longue période d'enregistrement a été proposé par Stedinger et al. [1992], noté RW-index-flood :

$$\omega_h = \frac{n_h R/(n_h + R)}{\sum_{h=1}^{N'} n_h R/(n_h + R)}; \quad h = 1,..,N' \tag{10}$$

où *R* est une constante positive généralement considérée comme $R \approx 25$. Toutefois, la valeur optimale de *R* dépend de l'hétérogénéité de la région et de la taille de l'échantillon. Il est intéressant d'utiliser ce poids dans les régions où certains sites ont des enregistrements beaucoup plus longs que les autres

2.2.2. Méthode de régression régionale

La méthode de la régression régionale permet d'établir une relation directe entre un quantile de crue Q_T correspondant à une période de retour *T* et les *r* variables explicatives physio-météorologique $X_1,...,X_r$ des bassins versants. Cette méthode a l'avantage d'être simple, rapide et de permettre d'utiliser des différentes distributions dans chacun des sites d'une même région. De plus, la méthode n'est pas sensible à l'hétérogénéité qui peut exister dans la région considérée. En pratique, le modèle s'exprime sous la forme de puissance suivante [*Pandey et Nguyen*, 1999] :

$$Q_T = \beta_0 X_1^{\beta_1} X_2^{\beta_2} ... X_r^{\beta_r} e \tag{11}$$

où $\beta_0, \beta_1,..., \beta_r$ sont des paramètres à estimer et *e* est l'erreur du modèle. L'estimation de cette fonction non linéaire est généralement effectuée en utilisant une transformation logarithmique de façon à obtenir le modèle classique de régression linéaire multiple (RM) suivant :

$$\log(Q_T) = \beta_0 + \beta_1 \log(X_1) + .. + \beta_r \log(X_r) + \varepsilon \tag{12}$$

En supposant une telle relation régionale, il est alors possible d'estimer les paramètres du modèle et de faire le choix des variables explicatives. Les paramètres du modèle de RM (12) sont

généralement estimés par la méthode de moindres carrés ordinaires (MCO) [*Thomas et Benson*, 1970]. Cette estimation est faite en supposant que l'erreur résiduelle associée aux observations individuelles doit être indépendant et de variance constante égale à Γ :

$$E(\varepsilon) = 0 \quad \text{et} \quad Var(\varepsilon) = \Gamma \tag{13}$$

Soit $X = [1, X_1, ..., X_r]$ et $Y = [QT_1, .., QT_s]$ les deux matrices contiennent les *r* variables physio-météorologiques et les *s* variables hydrologiques des *N* bassins versants. Le vecteur des paramètres $\beta = (\beta_0, ..., \beta_r)$ peut être estimé par :

$$\begin{aligned} \hat{\beta} &= \arg\min_{\beta} (logY - logX\beta)' \Omega (logY - logX\beta) \\ &= ((\log X)' \Omega \log X)^{-1} (\log X)' \Omega \log Y \end{aligned} \tag{14}$$

où $\Omega = \text{diag}(w_1, ..., w_N)$ avec $w_i = \begin{cases} 1 \text{ si le site } i \text{ appartient au voisinage (région) du site cible } l \\ 0 \text{ sinon} \end{cases}$

La matrice Γ est estimée par :

$$\hat{\Gamma} = \frac{\left(\log Y - \log X\hat{\beta}\right)' \left(\log Y - \log X\hat{\beta}\right)}{N - r - 1} \tag{15}$$

3. Problématiques et objectifs de recherche

Selon les différentes combinaisons des méthodes de DRH et les techniques d'ER, différentes méthodologies de régionalisations des débits de crue sont obtenues. Ces méthodes de régionalisations ont des limitations et des contraintes. En effet, les méthodes de DRH ont un ou plusieurs des désavantages suivants : (1) elles ne sont pas invariantes pour les transformations affines, p. ex. ROI et CAH, donc leurs résultats dépendent des échelles des attributs, (2) elles ne

sont pas robustes vis-à-vis des données aberrantes, p. ex. ROI et ACC [*Neykov et al.*, 2007], donc les régions obtenues sont affectées par quelques sites particuliers, p. ex. sites ayant petit ou grand bassin versant, (3) elles reposent sur l'intervention de l'utilisateur pour la sélection des nombres des régions homogènes, p. ex. CAH, donc leurs résultats reposent sur l'intervention de l'utilisateur, et (4) elles supposent la normalité des attributs, p. ex. ACC et CAH [*Leclerc et Ouarda*, 2007], donc leurs applications dépendent de la validation des hypothèses. En plus, les méthodes d'ER ne tiennent pas compte, de façon appropriée, de la similarité hydrologique entre les sites jaugés et le site cible, donc les sites jaugés ont la même contribution dans l'estimation régionale. De plus, ces méthodes ne sont pas nécessairement flexibles ou/et optimales, donc ces méthodes ne sont pas capables à s'adapter à des différentes situations.

Le principal objectif de cette recherche est de développer des méthodes de régionalisation basée sur les fonctions de profondeur. On vise à proposer des méthodes représentatives, robustes, flexibles, invariantes par transformation affine, qui prennent en compte de façon appropriée la similarité hydrologique entre le site cible et les sites jaugés, avec des performances optimales et n'exigent pas d'intervention de l'utilisateur. Tous ces avantages peuvent être atteints principalement à cause des bonnes propriétés des fonctions de profondeur couplée avec les fonctions de poids et les algorithmes d'optimisation (voir sections 4 et 5). En termes d'application, on considère des débits de crue. Afin d'atteindre l'objectif principal, ce dernier se décompose en quatre objectifs spécifiques qui sont les suivants :

A. Proposer une nouvelle approche robuste et invariante pour identifier les régions de type *voisinage*. L'approche proposée est basée sur une nouvelle mesure géométrique de dissimilarité utilisant la fonction de profondeur. L'application de cette nouvelle approche

de voisinage n'exige pas la normalité des attributs tout en tenant compte de leurs propres distributions;

B. Similaire à A mais pour les régions de type *non contiguës* où l'approche proposée est basée sur un algorithme itératif avec des sous-régions homogènes selon leurs valeurs de profondeur des sites. L'algorithme proposé automatise d'une manière objective le choix du nombre de sous-régions homogènes;
C. Proposer un nouveau modèle régionale *d'indice de crue* flexible pour l'estimation régionale des quantiles de crue. Ce nouveau modèle tient compte de la similarité hydrologique entre les sites de la région avec leurs profondeurs et utilise un poids optimal pour obtenir les paramètres de la fonction de croissance;
D. Optimiser la performance du modèle de la *régression* régionale des quantiles de crue qui utilise les profondeurs et automatiser le choix du poids attribué à chaque site jaugé dans l'estimation du quantile d'un site cible non-jaugé.

La Figure 1 montre les liens entre les différentes étapes et approches, ainsi elle résume les inconvénients des approches classiques et les objectifs de cette thèse.

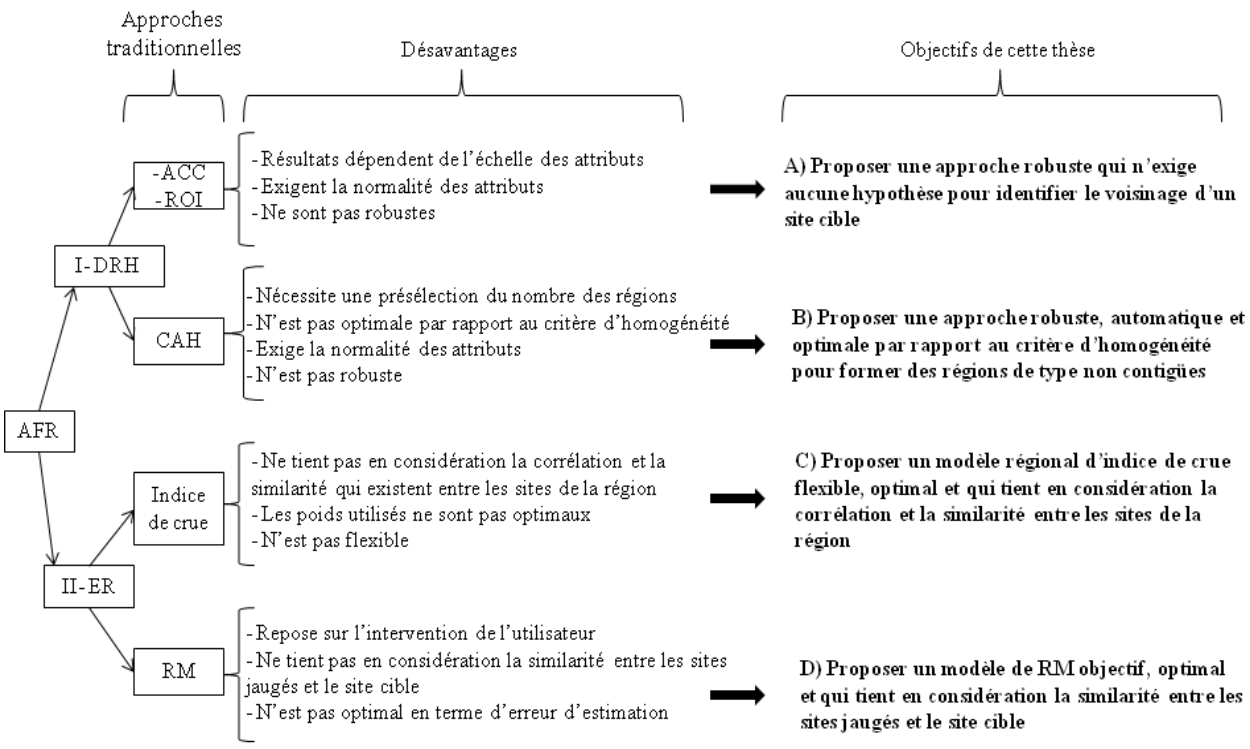


Figure 1. Approches traditionnelles en AFR et leurs inconvénients, ainsi que les objectifs de cette recherche.

Les sujets traités dans le cadre de cette thèse sont tous nouveaux en AFR des variables hydrologiques. En effet, les deux approches de délimitations des régions homogènes proposées sont robustes et basées sur la notion statistique de la fonction de profondeur. Malgré leur disponibilité en statistique, les fonctions de profondeurs n'ont jamais été introduites dans l'étape de délimitation des régions homogènes en AFR. D'autre part, les deux modèles d'estimation régionale proposés prennent en considération la similarité entre les sites jaugés et le site cible de la région, et produisent des performances optimales tout en évitant l'intervention de l'utilisateur.

4. Outils statistiques

Cette section présente d'abord les notions et les outils statistiques utilisés dans cette thèse, incluant les fonctions de profondeur et les fonctions de poids, pour introduire la robustesse, l'invariance, la flexibilité, la similarité et l'optimalité dans les deux étapes d'AFR.

4.1. Fonctions de profondeur

Ordonner une série de données est une opération très utile en statistique. Cela permet entre autres d'établir les quantiles, les mesures de localisation ou de dispersion comme l'étendue, les moyennes tronquées et les points médians. Ordonner des données devient plus difficile lorsque l'on passe en dimension supérieure à 1. C'est pour cette raison que les fonctions de profondeur ont été introduites en statistique dans les années 70 et ont par la suite connu un important développement et une très grande variété d'applications.

Si F est une distribution sur $\Re^d$ $(d \geq 1)$, alors une fonction de profondeur par rapport à la distribution F est une fonction $D_F : \Re^d \to \Re$ qui a tendance à prendre de grandes valeurs si son argument est près du "centre" de la distribution et à prendre de petites valeurs sinon [*Tukey*, 1975]. Selon Zuo et Serfling [2000], une fonction de profondeur devrait posséder quatre propriétés désirables :

i. Invariance par transformation affine : la profondeur d'un point $x \in \Re^d$ ne doit pas dépendre du système de coordonnées ou bien des échelles des mesures;

ii. Maximalité au centre : si F est une distribution qui possède un centre unique, alors D_F doit atteindre sa valeur maximale en ce centre;

iii. Monotonie à partir d'un point de profondeur maximale : si D_F atteint son maximum en $x \in \Re^d$ et si x_0 s'éloigne de *x* sur une demi-droite dans n'importe quelle direction, alors $D_F(x_0)$ doit être non-croissante sur cette demi-droite;

iv. Valeur nulle atteinte à l'infini : D_F doit tendre vers 0 lorsque $\|x\| \to \infty$.

Dans le cadre d'AFR, le caractère multivarié se présente dans plusieurs situations, comme par exemple, des vecteurs qui contiennent les *p* attributs des sites de la région, des quantiles de différentes périodes de retour (QT_1,..,QT_s), ou des vecteurs des paramètres des distributions ajustées pour chacun des sites. Les propriétés de ces fonctions contribuent à surmonter les limitations et les contraintes des méthodes classiques de régionalisation des débits de crues. Par exemple, (i) est utile pour éliminer l'effet d'échelle dans les deux étapes de régionalisation. En plus, si on suppose que le centre, où la fonction de profondeur est maximale, représente le site cible, la propriété (ii) est donc utile pour ordonner les sites jaugés selon leur similarité avec ce site cible. Enfin, la contribution des sites diminue et s'annule en l'éloignant du site cible grâce aux propriétés (iii) et (iv). Une description détaillée de la façon dont les fonctions profondeurs sont utilisées dans cette thèse est présentée dans la section 5.

Plusieurs fonctions de profondeur ont été présentées dans la littérature. Dans les travaux de cette thèse, trois fonctions de profondeur sont utilisées. En effet, la fonction de profondeur Simpliciale '*SD*' [*Liu*, 1990] est utilisée pour identifier les régions de type voisinage (objectif A). Cette fonction est utilisée car elle peut être transformée comme une mesure de dissimilarité (voir chapitre 2). La profondeur Spatiale '*SPD*' est utilisée quant à elle pour identifier les régions de type non contiguës (objectif B). Cette fonction ne prend pas la valeur 0, donc elle est utile pour

classer les sites dans des régions homogènes (voir chapitre 3). Finalement, pour les deux modèles d'estimation régionale proposés (objectifs C et D), la fonction de Mahalanobis '*MHD*' [*Mahalanobis,* 1936] est utilisée. Cette fonction est utilisée car elle tient compte de la variabilité des données des différents sites à travers la matrice de dispersion (voir chapitres 4 et 5).

4.2. Fonctions de poids

Une fonction de poids est un outil mathématique utilisé lors d'une estimation pour mieux contrôler l''influence des données (des observations, des sites, des paramètres) sur le résultat. Ainsi les méthodes proposées sont flexibles et peuvent être optimisées. Dans cette thèse, deux classes des fonctions de poids, Gompertz [*Gompertz*, 1825] et logistique [*Verhulst*, 1838], sont introduites dans l'étape d'estimation régionale.

La fonction de Gompertz φ_G est donnée par :

$$\varphi_G(x) = c\exp\left\{-ae^{-bx}\right\} \quad a,b,c>0\ ;\ x\in\Re \tag{16}$$

avec *c* est la limite supérieure, *a* et *b* sont deux coefficients qui permettent respectivement de translater et de modifier la forme de la courbe. La fonction logistique, notée par $\varphi_{\log istic}$, est donnée par :

$$\varphi_{\text{logistic}}(x) = \frac{c}{1+ae^{-bx}} \qquad a,b,c>0; x\in\Re \tag{17}$$

Les coefficients *a*, *b* et *c* jouent les mêmes rôles que dans φ_G.

Les deux classes des fonctions φ_G et $\varphi_{\log istic}$ ont une forme sigmoïdale (en S) avec trois phases. Cette forme est utile pour surmonter les limitations des méthodes d'estimation régionale couramment utilisées en hydrologie. En effet, les trois phases permettent d'inclure

progressivement les sites jaugés dans une région homogène pour estimer les variables hydrologiques d'un site cible. La phase de démarrage (proche de zéro) est utile pour pondérer les sites jaugés qui ont une faible contribution dans l'estimation, c.-à-d. les sites qui sont un peu similaires au site cible. L'inverse est valable pour la phase finale (proche de 1). Entre ces deux phases, les fonctions de poids ont une phase de croissance qui peut être utile pour pondérer les sites dont la similarité est 'intermédiaire' avec le site cible.

4.3. Algorithme d'optimisation

Par construction, les modèles proposés pour l'estimation des variables hydrologiques d'un site cible dépendent de plusieurs facteurs et en particulier de la fonction de poids φ. Par conséquent, la performance de ces modèles dépend du choix de φ. Pour éviter toute subjectivité dans le choix de φ, une étape d'optimisation est introduite. Cette étape automatise le choix des coefficients de φ par rapport à un critère quantifiant la performance de modèle d'estimation. La forme complexe et non explicite des critères de performance à optimiser (par exemple RB, ou bien RRMSE) nécessite l'utilisation des algorithmes d'optimisation d'ordre zéro [*Torczon*, 2000]. Ces algorithmes sont appropriés lorsque la fonction objective n'est pas différentiable ou le gradient n'est pas disponible et ont l'avantage d'être robustes, simples en termes de programmation, et utiles pour résoudre des problèmes d'optimisation non linéaire avec ou sans contraintes. Ces algorithmes sont décrits dans les articles associés et présentés dans la partie 2 de la thèse.

5. Approches et modèles proposés

Dans cette section on présente les nouvelles approches de régionalisations proposées. Elle est divisée en deux sous-sections selon les deux étapes de l'AFR. La première est consacrée à la description des méthodes proposées pour déterminer les régions homogènes. La deuxième montre les techniques proposées pour produire une estimation régionale dans des sites non jaugés.

5.1. Détermination des régions homogènes

Dans cette section, deux approches robustes pour déterminer les régions homogènes sont présentées. Ces deux approches permettent respectivement de définir des régions de type voisinages et de type non contiguës.

5.1.1. Identification des voisinages en utilisant les fonctions de profondeur

Les principales méthodes disponibles pour identifier le voisinage hydrologique pour un site cible sont basées sur des distances. Plus précisément, l'ACC est basée sur la distance de Mahalanobis, tandis que la ROI est basée sur la distance Euclidienne. Les résultats de ces deux méthodes peuvent être affectés par différents facteurs, y compris la distribution et l'échelle des variables physio-météorologiques [*Leclerc et Ouarda*, 2007; *Lin et Chen*, 2006] et la présence de valeurs aberrantes [*Neykov et al.*, 2007]. Afin de réduire ou éliminer l'impact négatif de ces éléments, une nouvelle méthode pour identifier le voisinage hydrologique d'un site cible basé sur la fonction de profondeur est présentée dans cette section.

La profondeur simpliciale de $x \in R^d$ est définie comme étant la probabilité que x appartient à un simplexe, dont les sommets, sont d+1 variables aléatoires indépendantes [*Liu*, 1990]. En suivant le même raisonnement, on définit la 'similarité simpliciale' entre deux points x

et *y* comme étant la probabilité que ces deux points appartiennent à un même simplexe. Plus formellement, soit *x* et *y* dans $\Re^d$ $(d \geq 1)$ et *F* une fonction de répartition, la similarité simpliciale entre *x* et *y* par rapport à *F* est définie par [*López et Romo*, 2010] :

$$SS(x, y, F) = P\left(x, y \in S\left[X_1, .., X_{d+1}\right]\right) \tag{18}$$

avec $X_1, ..., X_{d+1}$ sont des observations indépendantes de *F* et $S\left[X_1, ..., X_{d+1}\right]$ est un simplexe de sommets $X_1, ..., X_{d+1}$. La version empirique de cette fonction est notée par $SS_n(x, y)$.

La similarité simpliciale est symétrique, robuste, affine invariante, nulle à l'infini, et tient en compte la géométrie des données [*Zuo et Serfling*, 2000]. Ces propriétés peuvent réduire ou éliminer les inconvénients des méthodes traditionnelles de détermination de voisinage hydrologique. Afin d'être compatible avec les méthodes traditionnelles de voisinage hydrologique (ACC et ROI), la similarité simpliciale devrait être transformée en une mesure de dissimilarité. Ainsi, une fonction logarithmique est appliquée, et la mesure de dissimilitude suivante est utilisée pour définir le voisinage d'un site cible en utilisant l'approche de voisinage basée sur la quantité :

$$DSS_n(x, y) = -\log\left(SS_n(x, y)\right) \tag{19}$$

Supposons que *N* est le nombre des sites jaugés dans une région, le voisinage d'un site cible *l* en utilisant la méthode de voisinage basé sur *DSS* est défini comme :

$$\left\{ \text{site } k \in \{1, .., N\};\ DSS_N\left(U^k, U^l\right) \leq \tau \right\} \tag{20}$$

avec U^k est le vecteur des caractéristiques physio-météorologiques d'un site *k* et τ est une valeur seuil.

Dans cette thèse, pour définir le vecteur U^k, l'espace physio-météorologique et l'espace canonique physio-météorologique sont utilisés. Donc, si $U^k = X^k = \left(X_1^k, .., X_r^k\right)$ le vecteur de r variables physio-météorologiques du site k tel que dans (5), la méthode de voisinage basée sur *DSS* est notée par ROI-Depth et si $U^k = V^k = \left(V_1^k, ..., V_p^k\right)'$ les p variables canoniques physio-météorologiques du site k tels que dans (4), la méthode de voisinage basée sur *DSS* est notée par ACC-Depth.

Plus de détails concernant la détermination des voisinages en utilisant les fonctions de profondeurs sont présentés dans le chapitre 2 qui correspond à l'article Wazneh et al., [2014a].

5.1.2. Classification ascendante hiérarchique robuste pour AFR en utilisant les fonctions de profondeur

La CAH est utilisée en AFR pour décomposer une région non homogène en K sous-régions homogènes fixes. Les sites de ces sous-régions homogènes sont généralement géographiquement non contiguës. En AFR, la CAH souffre de deux principaux inconvénients. Tout d'abord, elle est basée sur des distances et utilisent des statistiques non robustes ce qui conduit à des résultats sensibles aux valeurs aberrantes [*Ilorme et Griffis*, 2013; *Jörnsten*, 2004]. D'autre part, elle nécessite une présélection du nombre de sous-régions homogène K, ce qui rend l'étape de délimitation subjective et dépend du choix de l'utilisateur. Afin de réduire ou éliminer l'impact négatif de ces éléments, une nouvelle approche de classification basée sur la profondeur spatiale est présentée dans cette thèse. Cette approche est notée par 'D-clustering' dans le reste de ce document. La profondeur spatiale est utilisée dans cette approche pour deux raisons. Premièrement, cette fonction de profondeur ne prend pas la valeur zéro, donc elle est utile pour

classifier les sites dans des sous-régions. Deuxièmement, cette fonction est simple en terme de calcul en grande dimension comme dans le cas des variables de classification.

Pour trouver les *K* sous-régions homogènes en utilisant l'approche D-clustering, les trois étapes suivantes sont utilisées :

- Initialisation : Utiliser une méthode traditionnelle de classification, comme la CAH, pour former les *K* sous-régions initiales et comme valeur initiale on prend *K*=2;
- Modification : Modifier la position initiale des sites dans les sous-régions en utilisant la fonction de profondeur;
- Homogénéité : Tester l'homogénéité des sous-régions obtenues après l'étape de modification. Si les sous-régions obtenues ne sont pas homogènes, revenez à l'étape d'initialisation avec $K = K + 1$.

Supposons que *N* est le nombre total des sites dans la région, $Y^m = \left(Y_1^l, ..., Y_{n_m}^m\right), m = 1, .., N;$ est le vecteur de longueur n_m qui contient les données hydrologiques de site *m* et $X^m = \left(X_1^m, \ldots, X_r^m\right)$ est le vecteur qui contient les *r* attributs du site *m* c.-à-d. les variables de classification telles que les coordonnées géographiques ou la superficie du bassin versant. La première étape de la procédure consiste à utiliser une méthode classique de CAH (par exemple, la méthode de Ward) pour former les sous-régions initiales. Par construction, l'approche D-clustering doit commencer par *K=2* sous-régions initiales. Les sous-régions initiales ne sont pas nécessairement homogènes. Soit $i(k)$ l'ensemble qui contient les identifiants des sites de la sous-région k, $k = 1, .., K;$ c.-à-d. $i(k) = \{\text{site m} \in \text{sous-région } k; m = 1, .., N\}$, et $I(k)$

l'ensemble qui contient les attributs des sites de sous-région *k* c.-à-d. $I(k)=\{X^m; m\in i(k); m=1,..,N\}$.

Dans la seconde étape de l'approche D-clustering, nous vérifions et modifions les sous-régions initiales obtenues dans la première étape. En effet, dans cette étape, chaque site dans l'ensemble de données doit être affecté à la sous-région qui maximise sa profondeur [*Jörnsten*, 2004]. Pour ce faire, dans cette étude la profondeur spatiale est utilisée. Plus précisément, la profondeur spatiale de chaque vecteur d'attributs, $X^m; m\in i(k)$, est calculée par rapport à la sous-région initiale $I(k)$ à laquelle il appartient. Dans le cadre de cette thèse, la profondeur de X^m par rapport à $I(k)$ est appelée profondeur 'intra sous-région' et notée par $D_m^w = SPD(X^m, I(k))$. Notons que D_m^w quantifie la centralité du site *m* par rapport à sa propre sous-région. Ensuite, on calcule la profondeur 'inter sous-région' du site *m*, définie comme étant la profondeur maximale de X^m par rapport aux sous-régions à laquelle il n'appartient pas $D_m^b = \max_{\substack{y=1,..,K \\ y\neq k}} \left[SPD(X^m, I(y)) \right]$. Ainsi, un site *m* est bien classé dans sa sous-région *k* si $D_m^w > D_m^b$.

Par conséquent, pour chaque site *m* on définit la profondeur 'déviance' $\mathrm{DeD}_m = D_m^w - D_m^b$ comme étant la différence entre D_m^w et D_m^b. Comme résultat, dans la seconde étape de l'approche D-clustering, chaque site *m* est réaffecté à la sous-région qui maximise son DeD_m. Un site *m* est bien classé dans une sous-région si $\mathrm{DeD_m}$ est positif, un $\mathrm{DeD_m}$ proche de 0 signifie que le site *m* se situe entre deux sous-régions et une valeur négative de $\mathrm{DeD_m}$ signifie que le site *m* est mal placé et doit changer sa sous-région.

Après la modification de la position initiale des sites en fonction de leurs valeurs de DeD, dans la dernière étape de l'approche D-clustering, nous testons l'homogénéité des sous-régions formées. Pour cette raison, de nombreux tests ont été proposés dans la littérature hydrologique [e.g., *Hosking et Wallis*, 1997; *Lu et Stedinger*, 1992; *Viglione et al.*, 2007a]. Dans cette thèse, la mesure *H* de Hosking et Wallis est utilisée pour tester et estimer le degré d'hétérogénéité dans les sous-régions. L'expression du *H* est donnée au chapitre 3. Si les sous-régions obtenues ne sont pas homogènes, nous répétons les deux premières étapes de l'approche en augmentant *K* à *K*+1.

Une description plus détaillée de cette approche est présentée au chapitre 3 qui correspond à l'article Wazneh et al., [2014b].

5.2. Approches flexibles et optimales pour l'estimation régionale

Dans cette section, deux approches flexibles et optimales pour l'estimation régionale sont présentées. Ces deux approches sont basées sur le modèle de l'indice de crue et celui de la régression multiple qui sont deux modèles d'estimation régionale couramment utilisés en AFR.

5.2.1. Modèle d'indice de crue optimal basé sur les fonctions de profondeur

Le modèle d'indice de crue (6) est utilisé pour estimer l'amplitude de crue dans des sites non jaugés où on ne dispose pas des informations hydrologiques. La performance de ce modèle dépend de plusieurs facteurs, en particulier l'estimation de vecteur des paramètres régionaux. Généralement, le vecteur des paramètres régionaux est estimé par des combinaisons linéaires pondérées des paramètres des sites jaugés (7), c.-à-d. des paramètres locaux.

Tel que indiqué ci-dessus, pour ce modèle, trois types de poids ont été proposés dans la littérature hydrologique, à savoir PW (8), UW (9) et RW (10). Ces poids sont basés uniquement

sur les longueurs d'enregistrements, et ils ne contiennent aucune mesure objective de similarité. Ces poids ne prennent pas en considération suffisamment d'informations contenues dans les séries, telle que la redondance qui peut exister entre les sites, la similarité et la corrélation entre les sites jaugés de la région, et la similarité entre les sites jaugés et le site cible. En plus, en termes de performance du modèle, ces poids ne conduisent pas nécessairement à des résultats optimaux.

Dans ce travail, on propose un nouveau schéma de poids qui contribue à surmonter les problématiques mentionnées des pondérations classiques et permet de maximiser la performance du modèle d'indice de crue. L'approche proposée est basée principalement sur l'utilisation des fonctions de profondeur et elle est appelée DW-index-flood. Le modèle vise d'abord à généraliser le modèle d'indice de crue classique en termes de représentativité et de flexibilité. En plus, la performance du modèle proposé est optimisée par rapport à la fonction de poids. Notons que dans DW-index-flood, on utilise une nouvelle procédure itérative pour estimer le vecteur des paramètres régionaux de la courbe de croissance θ^R (7) mais la quantité indice de crue μ_l est estimée par les approches traditionnelles (approche de régression ou spatiale). Le modèle proposé pour estimer θ^R (7) est composé d'un certain nombre d'ingrédients. En premier lieu, une fonction de profondeur est utilisée pour introduire et/ou évaluer la similarité hydrologique. En introduisant une fonction de poids, le modèle devient plus général. Afin d'améliorer l'estimateur de θ^R, une procédure itérative est considérée. Enfin, un algorithme d'optimisation est utilisé afin d'obtenir la fonction de poids optimale selon un critère de performance du modèle. Notez que dans cette étude, la similarité hydrologique est évaluée en comparant les valeurs des paramètres de la

distribution ajustée pour chaque site jaugé (par exemple position, échelle et forme dans le cas de distribution GEV).

Dans le DW-index-flood, les poids ne sont pas liés seulement et directement aux longueurs d'enregistrements comme dans le cas de (8), (9) et (10). Ils sont définis grâce à une fonction de profondeur et une fonction de poids croissante et continue. En effet, la profondeur de Mahalanobis de chaque vecteur des paramètres locaux $\hat{\theta}^h = \left(\hat{\theta}_1^h, ..., \hat{\theta}_S^h\right)$ d'un site *h*, est calculée par rapport à l'ensemble $\Theta = \left\{\hat{\theta}^1, .., \hat{\theta}^h, .., \hat{\theta}^{N'}\right\}$ formé par les N' vecteurs des paramètres locaux des sites jaugés de la sous-région homogène. Ainsi, la profondeur de chaque site *h* par rapport à Θ est notée par $MHD\left(\hat{\theta}^h; \Theta\right)$. Ces valeurs de profondeur sont obtenues en supposant que le centre de l'ensemble Θ (c.-à-d. le point plus profond) est le vecteur des paramètres régionaux θ^R. Pour améliorer la précision de l'estimation de θ^R, l'approche proposée utilise une procédure itérative où à chaque itération les poids sont mise à jour par rapport à θ^R estimé à l'itération précédente. Notons que dans ce travail, la fonction de profondeur de Mahalanobis est utilisée pour ses propriétés adéquates. Par exemple, cette fonction tient compte de la variabilité des données des différents sites à travers la matrice de dispersion et ainsi permettant de considérer des sous-régions ayant une large corrélation croisée entre les sites, ce qui n'est pas le cas du modèle traditionnel.

Afin de contrôler et amplifier les valeurs de profondeur, une fonction de poids φ continue et croissante est appliquée sur les valeurs de profondeur. Cela permet à l'utilisateur, par exemple, sur la base d'autres informations disponibles ou d'expérience précédente, d'augmenter ou de

réduire le poids attribué à chaque site en respectant l'ordre des sites, puisque la fonction de poids est croissante. Cette fonction de poids rend le modèle d'indice de crue plus flexible et général. En outre, la fonction de poids est utile, car la fonction de profondeur permet uniquement d'ordonner les sites, mais puisque les valeurs de profondeur sont très proches, donc ces valeurs sont incapables de quantifier les contributions des sites jaugées dans l'estimation régionale. Formellement, le poids DW proposé pour l'estimation des paramètres régionaux (7) est donc défini par :

$$\omega_h = \varphi\left[MHD\left(\hat{\theta}^h, \Theta\right)\right];\ \ h = 1,..., N' \qquad (21)$$

Dans le contexte de ce travail, les deux fonctions de poids Gompertz (16) et logistique (17) ont été utilisées.

Pour plus de détails concernant la méthode d'estimation de la courbe de croissance dans le modèle DW-index-flood ainsi que l'algorithme et la procédure itérative utilisée pour estimer le vecteur des paramètres régionaux, le lecteur est invité à consulté le chapitre 4 ou l'article associé de Wazneh et al.[2013a].

5.2.2. Modèle de régression optimal basé sur les fonctions profondeurs

Le modèle de RM (12) est généralement utilisé pour estimer les variables hydrologiques dans des sites non jaugés où on ne dispose pas des informations hydrologiques. Un élément clé de ce modèle est l'estimation de son vecteur des paramètres β (14). L'estimateur des MCO de β, utilisé dans l'approche classique (14), fait face à deux désavantages. Premièrement, il ne tient pas en considération la similarité qui peut exister entre le site cible et les sites jaugés de son voisinage, puisque tous les sites jaugés du voisinage ont la même contribution dans l'estimation.

Deuxièmement, les résultats issus de cet estimateur ne sont pas nécessairement optimaux. Afin de réduire ou éliminer l'impact négatif de ces désavantages, une nouvelle approche itérative pour estimer le vecteur des paramètres β, ainsi que pour estimer les variables hydrologiques d'un site cible, est proposée. Cette approche est notée par DBRM (Depth-Based Regression Multiple).

En bref, l'approche DBRM est basée sur les éléments suivants. Une fonction de profondeur est introduite pour ordonner les sites en se basant sur la similarité hydrologique entre le site cible et les sites jaugés. Une fonction de poids φ est employée afin d'amplifier et contrôler les valeurs de profondeur ainsi que pour rendre le modèle flexible et général. Afin d'améliorer la précision d'estimation, une procédure itérative est considérée. Enfin, un algorithme d'optimisation est utilisé afin d'automatiser objectivement le choix de la fonction de poids.

Plus précisément, en utilisant DBRM le vecteur des paramètres β est estimé itérativement par la méthode de moindres carrés pondérés (MCP) :

$$\hat{\beta}_{MHD,\varphi} = \left((\log X)'\Omega_{MHD,\varphi}\log X\right)^{-1}(\log X)'\Omega_{MHD,\varphi}\log Y \qquad (22)$$

Le point crucial dans cette méthode est la définition non usuelle de la pondération. La matrice de pondération Ω utilisée dans cette méthode est calculée en appliquant la fonction de poids φ sur les valeurs de profondeur des variables hydrologiques :

$$\Omega_{MHD,\varphi} = \mathrm{diag}\left(\varphi\left(MHD(Y_1;Y)\right),...,\varphi\left(MHD(Y_N;Y)\right)\right) \qquad (23)$$

où $Y = \{Y_1,...,Y_N\}$ est l'ensemble formé par les variables hydrologiques du N sites du voisinage, $MHD(Y_i;Y)$ est la profondeur de Mahalanobis appliquée aux variables hydrologiques du site i. La profondeur de Mahalanobis est utilisée dans cette approche car elle prend en compte de la variabilité des données hydrologiques des différents sites à travers la matrice de dispersion, donc

cette profondeur rend le modèle applicable dans des régions ayant une large corrélation croisée entre les sites.

L'estimateur de β (22) dépend en particulier du choix de la fonction de poids φ. Afin d'automatiser objectivement le choix de φ, un algorithme d'optimisation est utilisé dans DBRM. Comme fonction objective de l'optimisation, les critères de performance du modèle sont considérés, par exemple RB ou RRMSE. L'algorithme complet de l'approche DBRM est présenté dans le chapitre 5 ou l'article associé [*Wazneh et al.*, 2013b].

6. Applications et principaux résultats

Dans cette section on présente les cas d'étude et les résultats d'application des approches proposées et présentées dans la section précédente.

6.1. Identification des voisinages hydrologique avec fonction de profondeur

L'approche d'identification des voisinages hydrologiques en utilisant les fonctions de profondeur est appliquée sur un ensemble de 151 sites situés dans la partie sud de la province de Québec, Canada. Cette approche est appliquée dans deux différents espaces : l'espace physiographique (ROI-Depth) et l'espace canonique (ACC-Depth). Chaque site dans l'ensemble de données a plus de 15 années d'observations et ses séries des crues maximums annuelles historiques sont homogènes, stationnaires et indépendantes. Les superficies des bassins versants drainés par ces stations variaient entre 200 km^2 et 100 000 km^2. Plus d'informations sur ces stations, comme la localisation géographique, sont présentées au chapitre 2.

La sélection des variables physio-météorologiques est basée sur l'étude de Chokmani et Ouarda [2004]. Les variables sélectionnées sont : la superficie du bassin versant (AREA), la

pente moyenne du bassin versant (MBS), la fraction de la superficie couverte par des lacs (FAL), les précipitations annuelles moyennes totales (AMP), le nombre moyen des jours de plus de 0 C (AMD).

Un des objectifs de cette application est d'évaluer et de comparer la performance des approches traditionnelles et celles proposées pour identifier le voisinage d'un site cible. Puisque l'étape de DRH affecte les résultats d'estimation, deux catégories de critères de performance sont utilisées pour cette application: i) les critères qui permettent d'évaluer les approches d'identification des régions homogènes et ii) les critères qui permettent d'évaluer les modèles d'estimation régionale.

Pour évaluer et comparer les voisinages, la mesure d'hétérogénéité *H* [*Hosking et Wallis*, 1993] est utilisée. Cette mesure est basée sur les L-moments. Comme mentionné dans Hosking et Wallis [1997], *H* quantifie le degré d'hétérogénéité d'une région. Par conséquent, la meilleure approche est celle qui identifie des régions avec de faibles valeurs de *H.* Dans cette thèse, pour déterminer la meilleure approche, chaque site *l* dans la région est considéré comme un site cible en le retirant temporairement de la région pour ensuite identifier son voisinage en utilisant les différentes approches. Ensuite, à travers tous les sites de la région, on calcule la fréquence de *H* dans des intervalles prédéfinis. Les résultats correspondants sont résumés dans le Tableau 1. Ce tableau montre que la plupart des voisinages obtenus par la méthode ROI sont 'hétérogènes'. Ce constat l'est moins avec ACC. Cependant, plus de la moitié des voisinages obtenus par les méthodes basées sur la fonction de profondeur, ACC-Depth ou ROI-Depth, sont 'homogènes' ou 'possiblement homogènes'. D'une manière générale, ce tableau montre que les voisinages

obtenus par les méthodes basées sur la fonction de profondeur sont plus homogènes que ceux obtenus par les méthodes traditionnelles.

Tableau 1. Fréquence de *H* dans les différents intervalles avec les différentes approches de délimitation

Intervalles	Méthode de délimitation			
	ACC	ROI	ACC-Depth	ROI-Depth
H<1 ‘Homogène’	31	2	56	43
1<H<2 ‘Possiblement homogène’	16	6	22	34
H>2 ‘Hétérogène’	104	143	73	74

En terme d’estimation, la meilleure approche de voisinage est celle qui minimise l’erreur de prédiction de $Q(T)$, le quantile de crue correspond à une période de retour *T*, à travers tous les sites de la région. Pour quantifier cette erreur, la procédure de ré-échantillonnage jackknife est utilisée [*Chernick*, 2012]. Les résultats de cette procédure, en termes d’erreurs d’estimation des quantiles de période de retour 10 et 100 ans sont résumés dans le Tableau 2. Notons que la méthode d’estimation RM a été utilisée pour les différentes méthodes de voisinage. Ce tableau montre que, dans les deux différents espaces, les méthodes de voisinage basées sur la fonction de profondeur conduisent à des estimations plus efficaces que ceux obtenus à l'aide des méthodes traditionnelles. Cela est en partie dû au degré d’hétérogénéité des voisinages identifié par ces méthodes (Tableau 1). Pour le critère RRMSE, l'amélioration est plus importante en utilisant ROI-Depth que ACC-Depth. Ce résultat est expliqué par le fait que certaines informations sont perdues par la procédure de réduction de dimension ACC. Ceci n'est pas le cas de ROI-Depth où la dissimilarité simpliciale est appliquée aux variables physio-météorologique originales.

Tableau 2. Résultats d'estimation des quantiles avec les différentes approches de délimitation.

	Méthode de délimitation	QS10		QS100	
		RRMSE(%)	RB(%)	RRMSE(%)	RB(%)
Approches classiques	ACC	44.62	-7.54	51.84	-8.14
	ROI	43.10	**-2.13**	49.09	**-5.13**
Approches basée sur la fonction de profondeur	ACC-Depth	37.42	-2.74	43.87	-5.35
	ROI-Depth	**34.50**	-5.07	**41.92**	-6.50

Les meilleurs résultats sont en caractères gras.

Afin de visualiser le voisinage d'un site cible en utilisant les différentes méthodes de voisinage, on suppose que le site numéro 97 soit un site cible (c.à.d. les variables hydrologiques de ce site sont inconnues) et il doit être estimé en utilisant les 150 sites jaugés. Ce site est sélectionné puisque ces voisinages définis par les approches classiques sont très hétérogènes. La Figure 2 illustre les voisinages obtenus par l'ACC et l'ACC-Depth. Les voisinages pour le site cible 97 en utilisant ROI et ROI-Depth ne sont pas illustrés sur cette figure puisque ces deux méthodes sont calculées sur un espace de dimension 5 (nombre des variables physio-météorologiques). On remarque que le voisinage obtenu par l'approche ACC contient des sites qui sont loins du site cible dans l'espace hydro-physiographique (Figure 2a), tel que le cas des sites 1, 4, 109, 120 et 122. Ces sites sont en plus géographiquement loins du site cible (Figure 3). En testant la normalité du vecteur $\begin{pmatrix} W \\ V \end{pmatrix}$ qui contient les variables canoniques hydro-physiographiques, on trouve que l'hypothèse nulle de normalité est rejetée. Donc, l'hypothèse de base derrière l'approche ACC n'est pas respectée dans cette étude, ce qui pourrait expliquer l'inclusion de certains sites 'loins' dans le voisinage d'un site cible. Toutefois, ces sites ne sont pas inclus dans le voisinage obtenu par l'approche ACC-Depth (Figure 2c). En effet, en utilisant

l'approche ACC-Depth, le voisinage est formé sans aucune hypothèse sur les variables d'origines et/ou canoniques, ce qui explique la supériorité de cette approche dans l'étape d'estimation. En plus, nous observons que le voisinage formé par l'ACC traditionnelle décrit l'intérieur d'une région elliptique (Figure 2b). Cependant, pour l'ACC-Depth, le voisinage a une forme triangulaire (Figure 2d), ce qui est plus flexible en termes des angles et d'arrêtes. Cela est dû à la forme triangulaire des simplexes qui définissent la similarité simpliciale utilisée dans cette approche.

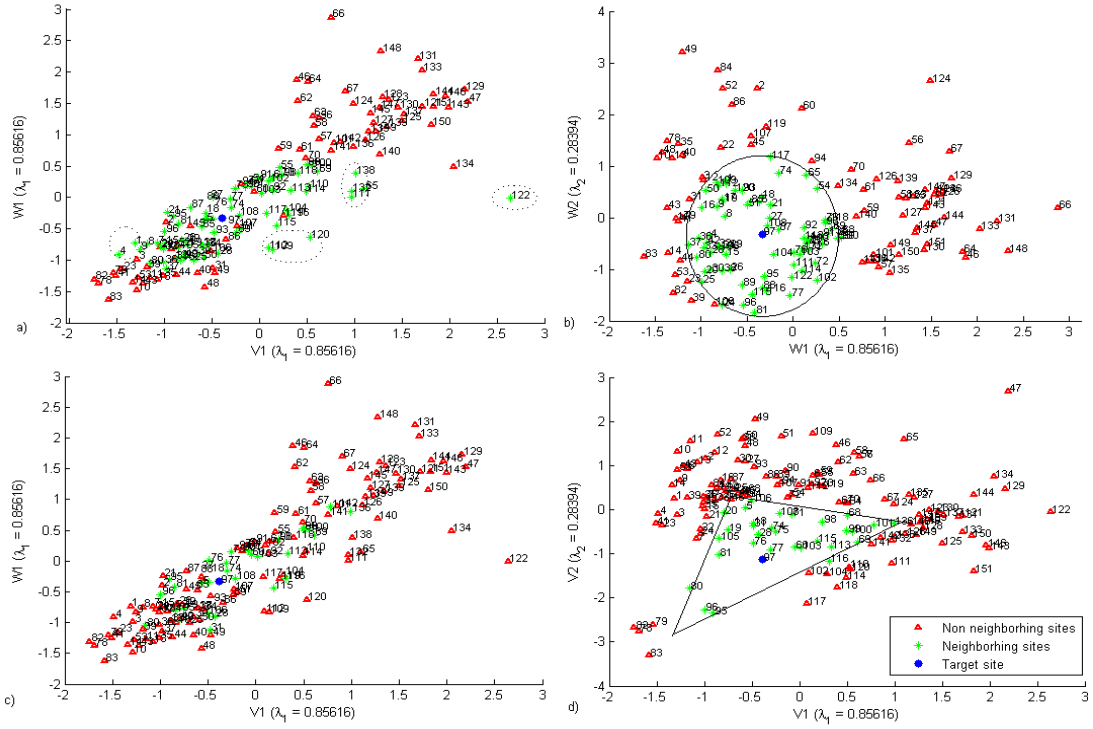


Figure 2. Le voisinage du site cible 97en utilisant (a) ACC dans l'espace canonique (V1, W1) (b) ACC dans l'espace canonique (W1, W2). (c) ACC-Depth dans l'espace canonique (V1, W1) et (d) ACC-Depth dans l'espace canonique (V1, V2).

Dans ce paragraphe, la sensibilité des approches proposées pour une transformation affine est étudiée. Plus précisément, le voisinage du site cible 97 est déterminé par les deux approches ROI et ROI-Depth en utilisant (a) les variables physio-météorologiques originales et (b) une version normalisée de ces variables. La Figure 3 illustre les résultats obtenus de cette application. On remarque qu'en utilisant la méthode ROI-Depth (Figure 3a), le même voisinage est obtenu en utilisant les variables originales ou les variables transformées. Ceci est dû au fait que la similarité simpliciale utilisée dans cette approche est affine invariante. Toutefois, pour la méthode ROI, le voisinage change en fonction de la transformation utilisée (Figure 3b et 3c). Les mêmes résultats sont obtenus en comparant le voisinage obtenu par ACC et celui par ACC-Depth (voir chapitre 2). Par conséquent, l'approche de voisinage basée sur la fonction de profondeur a l'avantage d'être invariante pour une transformation affine, ceci n'est pas le cas des approches traditionnelles.

En conclusion, les résultats ci-dessus montrent que l'approche d'identification de voisinage basée sur les fonctions de profondeur conduit à des estimations plus efficaces, en termes de RRMSE, que celles obtenues en utilisant les méthodes traditionnelles de voisinage. L'approche proposée rend l'étape de DRH affine invariante et ne nécessite aucune hypothèse sur les variables d'attributs. En plus, elle conduit à des voisinages plus homogènes que ceux obtenus à l'aide des approches traditionnelles.

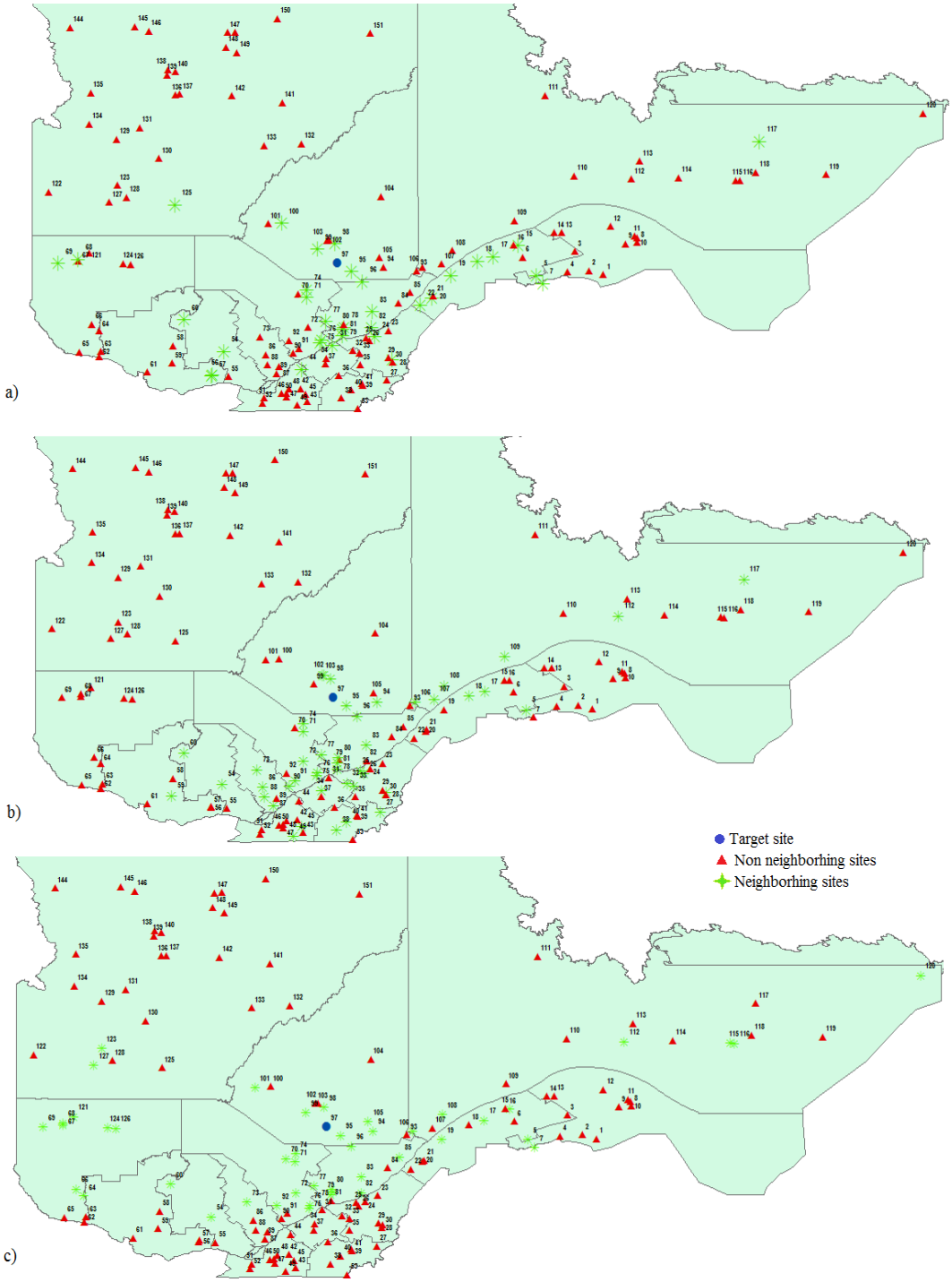


Figure 3. Localisation géographique du voisinage du site cible 97 obtenue par l'approche (a) ROI-Depth en utilisant les variables physiographiques originales ou normalisées (b) ROI en utilisant les variables normalisées et (c) ROI en utilisant les variables originales.

6.2. Classification robuste pour AFR en utilisant les fonctions profondeurs

L'approche D-clustering est appliquée à un ensemble de sites situés au nord-ouest de l'Italie. Ensuite, elle est comparée à celle obtenue en utilisant la méthode de Ward, une méthode classique de CAH. Les débits maximums annuels de 47 sites jaugés sont disponibles avec des longueurs d'enregistrements (années d'observation) variant entre 6 et 65 ans. Neuf de ces sites jaugés ont une longueur d'enregistrement moins de 15 ans et ils ne sont pas considérés dans cette étude. Les surfaces des bassins versants drainés par ces stations variaient entre 22 km^2 et 8024 km^2 avec des précipitations annuelles moyennes entre 501 et 2380 mm.

Les variables de classification utilisées dans cette application pour délimiter les sous-régions homogènes sont extraits de l'étude de Viglione et al. [2007b]. Les variables physio-météorologiques utilisées sont les coordonnées des sites dans le système UTM (X_{bar} et Y_{bar}) et l'élévation moyenne du bassin versants (H_m). Ces variables ont été également sélectionnées dans une étude précédente sur le même ensemble de données [*Viglione,* 2010]. .

L'algorithme de calcul pour l'approche D-clustering départ avec K = 2 sous-régions initiales. Cependant, les sous-régions obtenues dans ce cas ne sont pas homogènes. Par conséquent, l'algorithme redémarre avec 3 sous-régions initiales, soit $i(1), i(2)$ et $i(3)$. Dans la deuxième étape de l'approche D-clustering, nous calculons D_m^w, D_m^b et DeD_m de X^m, avec $X^m = \left(X_{\mathrm{bar}}, Y_{\mathrm{bar}}, H_m\right)^m$ est le vecteur formé par les valeurs des variables de classification du site $m \in i(k), k = 1,2,3$. Ensuite, les sites mal placés dans leurs sous-régions (c. à-dire les sites qui ont une valeur négative de DeD) sont identifiés. Le Tableau 3 montre que dans l'étape initiale, il y a 9 sites mal placés. Ces sites sont déplacés vers des sous-régions qui maximisent leurs valeurs

de DeD (Tableau 3). On teste l'homogénéité des nouvelles sous-régions obtenues après le déplacement de ces 9 sites. Le Tableau 4 présente la mesure de l'hétérogénéité *H* de ces trois sous-régions. Les valeurs de *H* indiquent que ces nouvelles sous-régions peuvent être considérées comme homogènes. Par conséquent, l'algorithme s'arrête à cette itération. La localisation géographique des sites des sous-régions finales obtenues par l'approche D-clustering est présentée dans la Figure 4a. Les sites de la sous-région 1 sont situés dans le nord de la région étudiée. Tandis que les sites des sous-régions 2 et 3 sont situés respectivement à l'ouest et au sud de la région étudiée.

Par la suite la délimitation des sous-régions homogènes en utilisant la méthode de Ward est présentée. Le nombre de sous-régions est choisi selon le critère d'hétérogénéité *H*. Par conséquent, quatre sous-régions sont identifiées par la méthode de Ward. La localisation géographique des sites de chaque sous-région est présentée dans la Figure 4b. Nous remarquons que les sous-régions sont géographiquement contiguës. En particulier, la plupart des sites de la sous-région 1 sont situés dans la région du Val d'Aoste. Les sites de la sous-région 2 sont situés dans le nord de la région du Piémont. Cependant, les sites de sous-régions 3 et 4 sont situés respectivement dans le sud et le sud-ouest de la région du Piémont. Les valeurs de *H* (Tableau 4) indiquent que les sous-régions 3 et 4 sont ' homogènes ' (*H* <1). Toutefois, les sous-régions 1 et 2 sont ' possiblement homogènes ' ($1 \leq H < 2$).

Tableau 3. Sites ayant des valeurs négatives de "DeD". Ces sites sont déplacés de leurs sous-régions initiales en utilisant l'approche D-clustering.

Numéro de site	Sous-région initiale	*DeD*	Sous-région finale
6	1	-0.10	3
11	1	-0.12	2
12	1	-0.21	2
14	2	-0.15	3
17	2	-0.13	3
20	2	-0.17	3
23	2	-0.21	3
29	2	-0.15	3
32	2	-0.13	3

Tableau 4. La mesure d'hétérogénéité *H* obtenue par D-clustering et Ward.

Méthode de délimitation	Sous-régions	Nombre des sites	H
D-clustering	1	14	1.03
	2	9	-0.60
	3	15	0.01
Ward	1	10	1.93
	2	7	1.96
	3	13	-0.30
	4	8	0.50

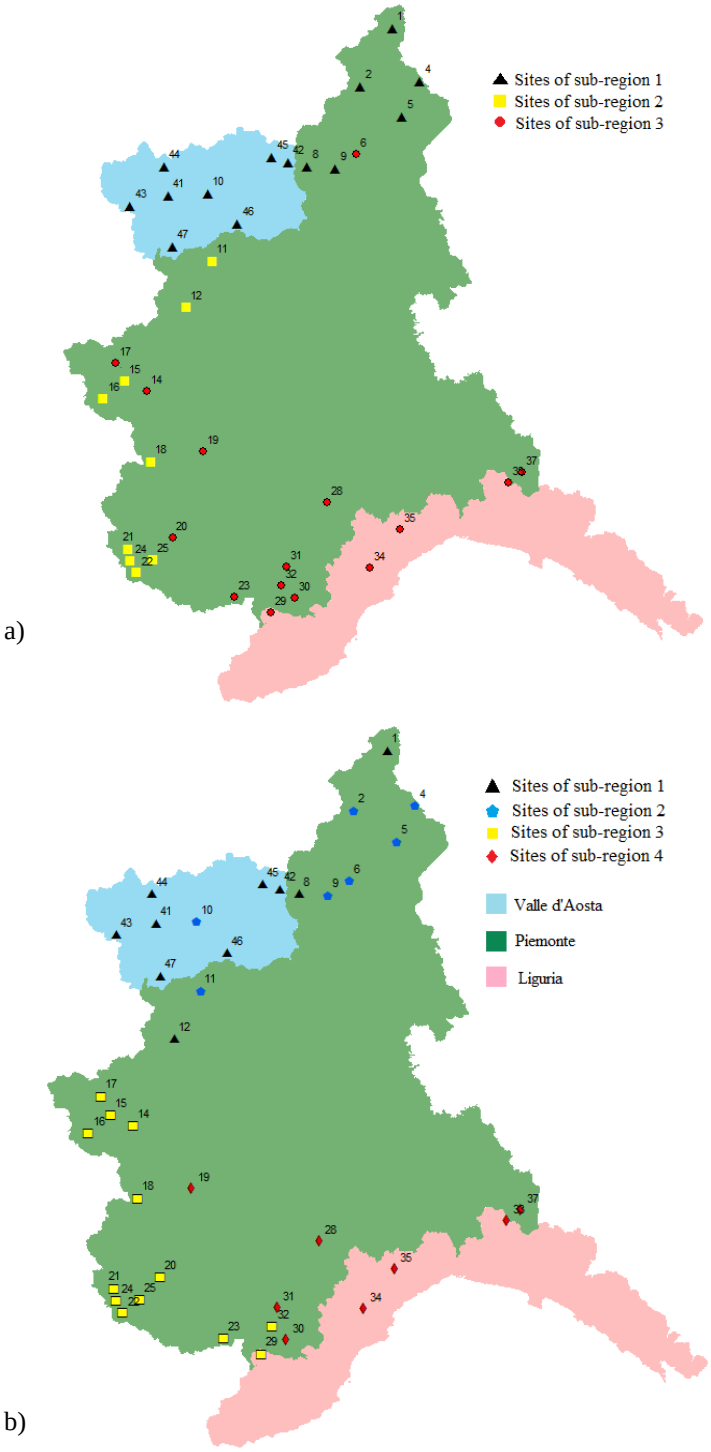


Figure 4. a) Localisation géographique des sites des trois sous-régions homogènes formés en utilisant l'approche D-clustering. b) Localisation géographique des sites des quatre sous-régions homogènes formés en utilisant la méthode Ward.

Le Tableau 4 montre que la méthode D-clustering conduit à des sous-régions avec des faibles valeurs de *H* qui sont les trois homogènes. Ce qui signifie que les sous-régions obtenues avec D-clustering sont plus homogènes que celles obtenues en utilisant la méthode Ward. En termes de nombre des sites, les sous-régions obtenues en utilisant D-clustering sont plus adaptées pour l'estimation régionale que celles obtenues par la méthode de Ward (Tableau 4).

Puisque l'étape de délimitation des sous-régions homogènes affecte les résultats d'estimation régionale, la performance de D-clustering et Ward sont comparées en termes d'estimation des crues. Pour réaliser ces estimations, le modèle d'indice de crue a été utilisé. Cette comparaison est basée sur le RB et le RRMSE des quantiles de crues pour les probabilités de nondépassement t = 0.9, 0.99, 0.995 et 0.999. Les résultats liés aux différentes approches de délimitation sont résumés dans le Tableau 5. Ce tableau montre que l'approche D-clustering conduit à des estimations plus efficaces en termes de RB et RRMSE que l'approche de Ward, en particulier pour la grande valeur de *t*. Comme indiqué précédemment, les sous-régions obtenues par l'approche D-clustering sont plus grandes et plus homogène que celles formées par Ward, ceci pourrait être une des raisons de cette amélioration.

Tableau 5. Résultats d'estimation en % des quantiles de crue en utilisant le modèle d'indice de crue avec les différentes approches de délimitation.

Méthode de délimitation	0.9		0.99		0.995		0.999	
	RB	RRMSE	RB	RRMSE	RB	RRMSE	RB	RRMSE
Ward	**0.21**	4.43	0.23	10.70	**0.18**	12.70	**-0.05**	17.33
D-clustering	-0.78	**3.25**	**0.18**	**9.05**	0.32	**9.40**	-0.12	**11.54**

Les meilleurs résultats sont en caractères gras

En conclusion, les résultats obtenus à partir de l'application de l'approche D-clustering sur un ensemble de sites dans le Nord-Ouest de l'Italie montrent que l'approche proposée conduit

à des sous-régions plus homogènes et plus grandes que celles obtenues en utilisant une méthode traditionnelle CAH de Ward. Les estimations obtenues en utilisant l'approche D-clustering sont plus efficaces en termes de RB et RRMSE que celles obtenues avec la méthode de Ward. L'approche proposée rend l'étape de DRH plus pratique et ne nécessite pas l'intervention de l'utilisateur. L'utilisateur n'a qu'à sélectionner un critère, tel que celui de l'hétérogénéité *H* utilisée dans cette application, pour obtenir les sous-régions optimales par rapport à ce critère.

6.3. Modèle d'indice de crue optimal basé sur les fonctions profondeurs

L'approche DW-index-flood est appliquée à un ensemble de 50 sites situés dans l'île de Sicile (Italie). Les débits maximums annuels mesurés par SIRI (Sisteme Informative Regionale Idrologico) sont disponibles avec des longueurs d'enregistrements variant entre 10 et 65 ans. Les conditions d'application de l'AF, c'est-à-dire l'homogénéité, la stationnarité et l'indépendance, ont été testées sur les données historiques de ces stations dans plusieurs études [*Cannarozzo et al.*, 2009; *Noto et La Loggia*, 2009]. Les superficies des bassins versants variaient entre10 km^2 et 2000 km^2. La région d'étude a été divisée en trois sous-régions homogènes [*Wazneh et al.*, 2013a]. Les emplacements géographiques des sites de ces sous-régions sont représentés sur la Figure 5.

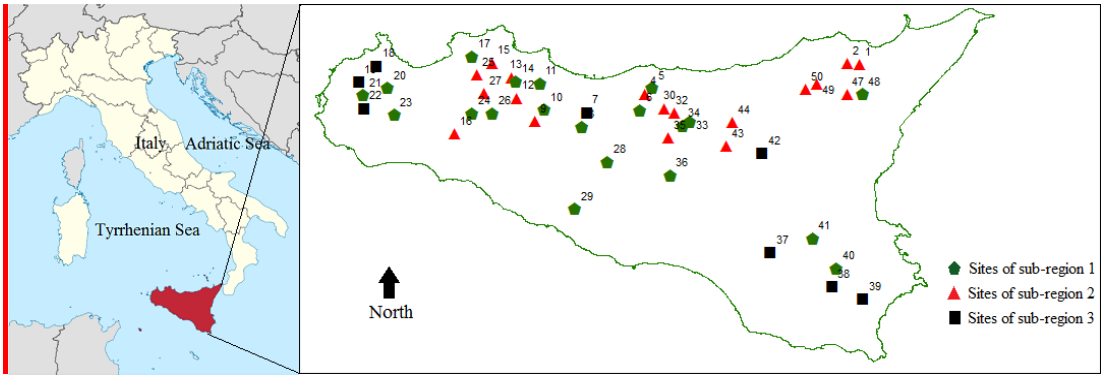


Figure 5. Localisation géographique des sites étudiés dans l'ile de Sicile.

Dans cette section l'approche d'estimation régionale DW-index-flood est appliquée aux trois sous-régions homogènes situées dans l'île de Sicile (Figure 5). Ensuite, les résultats de cette application sont comparés à ceux obtenus par les approches traditionnelles (PW (8), UW (9) et RW (10)) utilisées dans la littérature hydrologique.

L'application de l'approche DW-index-flood nécessite la détermination des courbes de croissance pour ces sous-régions. Dans cette étude, quatre distributions de probabilités fréquemment utilisées en AFR sont considérées comme candidates pour les courbes de croissance, à savoir la distribution des valeurs extrêmes généralisée (GEV), la loi de Pearson type 3 (PE3), la loi logistique généralisée (GLO), et la loi de Pareto généralisée (GPD). Pour sélectionner la distribution appropriée pour chaque sous-région, le diagramme des L-moments est utilisé (voir chapitre 3). Ainsi, la GEV est choisie pour les trois sous-régions. Cette distribution a trois paramètres ξ, α et κ (respectivement position, échelle et forme). Selon la notation (7), le vecteur des paramètres régionaux est $\theta^R = \left(\kappa^R, \alpha^R, \xi^R\right)$ et donc le nombre de composantes S=3.

Tableau 6 montre les résultats de l'évaluation de la performance des différentes approches d'estimation. La constante positive R pour l'approche RW-index-flood (10) est fixée à 25,

comme il a été suggéré par Stedinger et al. [1992]. Pour DW-index-flood, la fonction de poids de Gompertz (16) est utilisée. La fonction de poids optimale est obtenue en utilisant la procédure proposée dans le chapitre 2 (ou [*Wazneh et al.*, 2013a]). L'optimisation est effectuée par rapport aux deux critères RB et RRMSE. Comme prévu et indiqué dans un certain nombre d'études, l'estimation régionale est généralement plus précise pour des faibles probabilités de non-dépassement (c.-à-d. petites périodes de retour). Le Tableau 6 montre que le DW-index-flood avec φ optimale conduit à des estimations plus efficaces en termes de RB et RRMSE que celles obtenues par les autres approches. L'amélioration est plus importante dans la sous-région 3 que dans les deux autres. En effet, une corrélation croisée significative existe entre les sites de cette sous-région, ce qui pourrait être la raison des valeurs élevées du RRMSE des approches classiques et ainsi de l'amélioration importante de l'approche proposée. Donc le DW-index-flood fonctionne mieux pour les régions avec corrélation croisée entre les sites, ce qui n'est pas le cas des approches traditionnelles [Hosking et Wallis, 1997-P.8]. Ceci est expliqué par le fait que la profondeur de Mahalanobis utilisée dans DW-index-flood prend en considération la variabilité à travers les paramètres locaux des sites. Les mêmes résultats sont obtenus en utilisant la fonction de poids logistique $\varphi_{\log istic}$.

Tableau 6. Résultats d'estimation en% en utilisant le modèle d'indice de crue avec les diverses poids.

		Probabilité de non-dépassement *p*											
		0.9	0.99	0.999	0.9	0.99	0.999	0.9	0.99	0.999	0.9	0.99	0.999
		UW			PW			RW			DW		
		$\omega_h = \frac{1}{N'}$			$\omega_h = \frac{n_h}{n}$			$\omega_h = \frac{\frac{n_h R}{n_h + R}}{\sum_{h=1}^{N'} \frac{n_h R}{n_h + R}}$			φ_G		
Sous-région 1	RB	-0.8	-1.8	-9.0	0.2	-0.9	-8.4	-0.3	-1.6	-9.2	**0.1**	**0.4**	**0.8**
	RRMSE	6.5	25	55.0	6.2	24.8	53.3	6.2	25.1	54.1	**5.8**	**23.8**	**45.7**
Sous-région 2	RB	1.9	1.0	-0.9	0.2	-1.4	-0.9	1.9	-1.3	-6.0	**0.1**	**0.4**	**0.7**
	RRMSE	8.7	28	47.5	7.3	25.0	46.7	7.4	25.0	46.2	**6.9**	**23.0**	**40.9**
Sous-région 3	RB	-2.4	-5.5	-21.1	-2.0	-6.3	-22.8	-3.2	-5.5	-19.4	**-0.3**	**0.1**	**0.5**
	RRMSE	16.6	17	65.7	16.5	16.1	63.4	17	15.8	61.2	**13**	**11.3**	**34.6**

Les meilleurs résultats pour chaque sous-région sont en caractère gras.

La fonction de poids optimale pour chaque sous-région, pour l'estimation de $q_{0.995}(.)$, est présentée dans la Figure 6. On remarque que pour les deux sous-régions 1 et 2, la fonction de poids optimale (φ_G) a une forme S avec trois phases. Une phase de démarrage qui pondère faiblement (proche de zéro) les sites jaugés qui ont une faible contribution dans l'estimation. L'inverse est valable pour la phase finale (proche de 1). Entre ces deux phases, il y a une phase de croissance qui représente les sites jaugés avec une contribution 'intermédiaire'. Par contre, la fonction de poids optimale de la sous-région 3 a la forme d'une ligne quasi-horizontale. Ce qui signifie que tous les sites jaugés de cette sous-région ont presque la même contribution dans l'estimation de quantile d'un site cible. Ceci est dû à la forte corrélation positive entre les sites de

cette sous-région. Donc, en plus d'améliorer la performance du modèle, l'approche DW-index-flood reconnaît la corrélation entre les sites à partir de la forme de la fonction de poids.

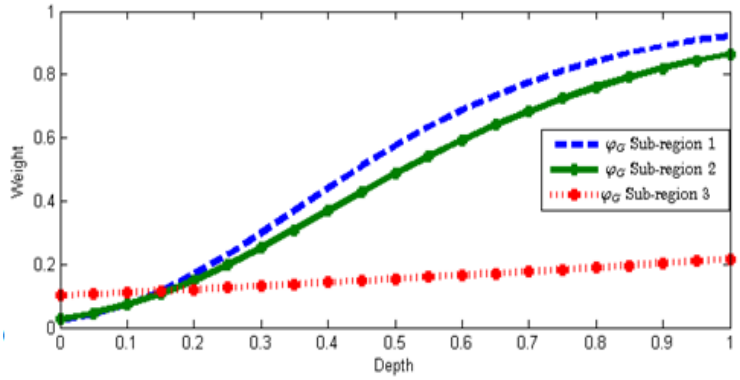


Figure 6. Fonction de poids optimale pour l'estimation de $q_{0.995}(.)$ par l'approche DW-index-flood pour les trois sous-régions et en utilisant φ_G .

La Figure 7 illustre les différents éléments de l'application de DW-index-flood à la sous-région 1. En effet, cette Figure illustre la profondeur et le poids attribués à chaque site jaugé dans l'estimation de vecteur des paramètres régionaux. On remarque que les sites 14, 20 et 28 sont situés proches du centre de l'espace des paramètres (Figure 7a). Ces sites ont des valeurs élevées de profondeur et donc ils ont une forte contribution (poids) dans l'estimation de la courbe de croissance régionale (Figure 7b). Toutefois, l'inverse est vrai pour les sites 8 et 29.

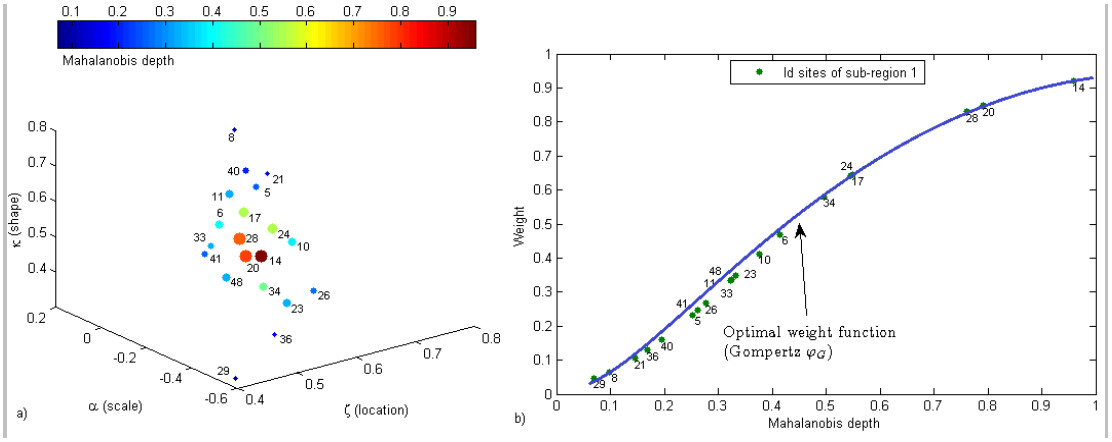


Figure 7. a) La profondeur de Mahalanobis des sites jaugés de la sous-région 1. La taille de cercle qui représente le site est proportionnelle à sa valeur de profondeur. b) Les pondérations allouées pour chaque site de la sous-région 1 dans l'estimation du vecteur des paramètres régionaux.

Les résultats obtenus à partir de l'application du modèle DW-index flood montrent sa supériorité en termes de performances. L'étude des trois sous-régions montre une association entre la corrélation entre les sites d'une sous-région et la forme de la fonction de poids optimale. L'approche proposée est capable d'identifier la corrélation croisée dans la région et fournit une amélioration significative des performances.

6.4. Modèle de régression optimal basé sur les fonctions profondeurs

L'approche DBRM est appliquée sur trois régions situées dans les états d'Arkansas et du Texas (États-Unis) et dans la partie sud de la province du Québec (Canada). L'ensemble de données de la région du Québec est composé de 151 sites. Toutefois, celle des régions d'Arkansas et du Texas contient respectivement 204 et 69 sites. L'objectif de cette application est d'estimer les quantiles de crue pour une période de retour de 10 et 100 ans pour la région du Québec et pour une période de retour de 10 et 50 ans pour Arkansas et Texas.

Les résultats de l'application de l'approche d'estimation régionale DBRM sont comparés à ceux obtenus par l'approche de régression classique (14) et ils sont résumés dans le Tableau 7. Pour l'approche DBRM, la fonction de poids de Gompertz est utilisée dans cette application. L'optimisation de l'approche DBRM est faite par rapport au critère RRMSE. On remarque que, pour une région et une méthode données, l'estimation régionale de quantile de crue est plus précise pour des petites périodes de retour. Le Tableau 7 montre que pour toutes les régions, DBRM conduit à des estimations plus précises en termes de RB et de RRMSE que celles obtenues en utilisant l'approche traditionnelle. Les mêmes résultats sont obtenus en utilisant la fonction de poids logistique $\varphi_{\log istic}$.

Tableau 7. Résultats d'estimation en% pour les trois régions en utilisant les diverses approches.

	Région											
	Sud Québec				Arkansas				Texas			
	Q10		QS100		QS10		QS50		QS10		QS50	
Approche d'estimation	RB	RRMSE	RB	RRMSE	RB	RRMSE	RB	RRMSE	RB	RRMSE	RB	RRMSE
RM	-7.5	44.6	-8.1	51.8	-7.8	48.1	-9.3	59.5	-1.2	42.3	-7.4	57.4
DBRM	**-3.5**	**38.7**	**-2.2**	**44.5**	**-6.0**	**41.5**	**-6.3**	**47.7**	**-1.0**	**36.8**	**-6.0**	**50.7**

Les meilleurs résultats pour chaque région sont en caractère gras

Afin de visualiser l'influence des sites jaugés dans l'estimation régionale d'un site cible en utilisant les deux approches traditionnelle et DBRM, supposons que le site numéro 25 du Texas est un site cible et il doit être estimé en utilisant les 68 sites jaugés de cette région. La Figure 8 illustre les coefficients de pondération attribués à chaque site dans l'espace canonique hydrologique (W1, W2). Nous observons qu'avec l'approche classique, la fonction de poids ne

prend que deux valeurs, 1 à l'intérieur du voisinage et 0 sinon (Figure 8a). Cependant, l'influence d'un site jaugé dans l'estimation régionale en utilisant DBRM est proportionnelle à la similarité hydrologique entre ces deux sites. En effet, la fonction de poids prend la forme de cloche dans l'espace 3D (Figure 8b).

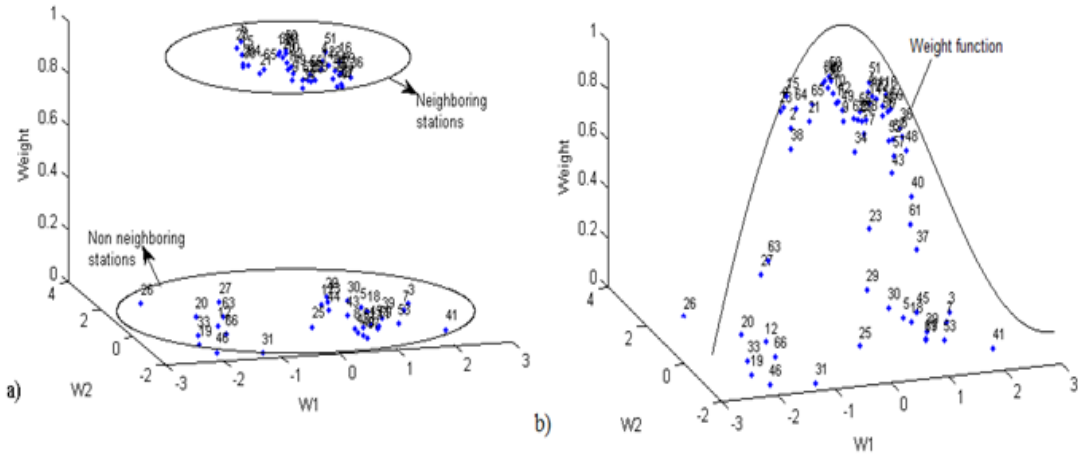


Figure 8. Les pondérations attribuées aux sites jaugés dans l'estimation de la variable hydrologique du site numéro 25 en utilisant a) le modèle de régression classique, et b) l'approche DBRM.

Le Tableau 8 présente la comparaison entre DBRM et les différentes méthodes d'estimation qui ont été appliquées dans la région sud Québec, comme l'approche Ensemble RNA-ACC (ERNA-ACC) [*Shu et Ouarda*, 2007] et l'approche krigeage-ACC [*Chokmani et Ouarda*, 2004]. Les résultats montrent généralement la supériorité de l'approche DBRM en termes de RB et RRMSE, sauf une très légère différence de 1% dans le RRMSE de QS10 avec ERNA-ACC. Cela pourrait être lié aux approximations numériques dans les algorithmes de calcul.

Tableau 8. Résultats d'estimation en% avec les approches disponibles pour la région du Québec et leurs références.

Approach	Reference	QS10		QS100	
		RB (%)	RRMSE (%)	RB (%)	RRMSE (%)
Régression non linéaire (RNL)	Shu and Ouarda [2008]	-9	61	-12	70
RM	Tableau7 ci-dessus	-7	44	-8	52
Krigeage-ACC	Chokmani and Ouarda [2004]	-20	66	-27	86
Adaptive Neuro-Fuzzy Inference Systems (ANFIS)	Shu and Ouarda[2008]	-8	57	-14	64
Réseaux de Neurones Artificiels (RNA)	Shu and Ouarda [2008]	-8	53	-10	60
Single RNA-ACC (SRNA-ACC)	Shu and Ouarda [2007]	-5	38	-4	46
Ensemble RNA (RNA)	Shu and Ouarda [2007]	-7	44	-10	60
Ensemble RNA-ACC (ERNA-CCA)	Shu and Ouarda [2007]	-5	**37**	-6	45
DBRM	Tableau7 ci-dessus	**-3**	38	**-2**	**44**

Les meilleurs résultats sont en caractère gras

En résumé, les résultats ci-dessus montrent la supériorité de l'approche DBRM en terme de performance. En effet, cette approche introduit la similarité hydrologique entre le site cible et les sites jaugés dans l'estimation de son quantile. Cette approche est flexible, générale et optimale en terme de performance.

7. Conclusions générales et perspectives

7.1. Conclusions générales

L'AFR est un outil statistique pour prédire les événements extrêmes dans des sites où on ne dispose pas d'informations hydrologiques. Cette procédure comporte deux principales étapes, la délimitation des régions homogènes et l'estimation régionale. Les méthodes traditionnelles utilisées pour ces deux étapes souffrent des plusieurs inconvénients et limitations. Durant notre travail de recherche, nous avons proposé de nouvelles approches dans le cadre de l'AFR des crues. Notre travail a porté sur le développement de nouvelles approches statistiques dans le but d'avoir des approches robustes, invariantes aux transformations affines, flexibles et conduisant à des performances optimales. Ceci correspond à l'objectif général de la présente thèse. Pour atteindre cet objectif, on a réalisé quatre études qui constitueront les quatre chapitres de la thèse. La Figure 9 présente une brève illustration des quatre chapitres de la thèse et montre comment ils sont reliés entre eux et avec la structure méthodologique de l'AFR. L'ingrédient essentiel et commun entre toutes ces approches est les fonctions de profondeur.

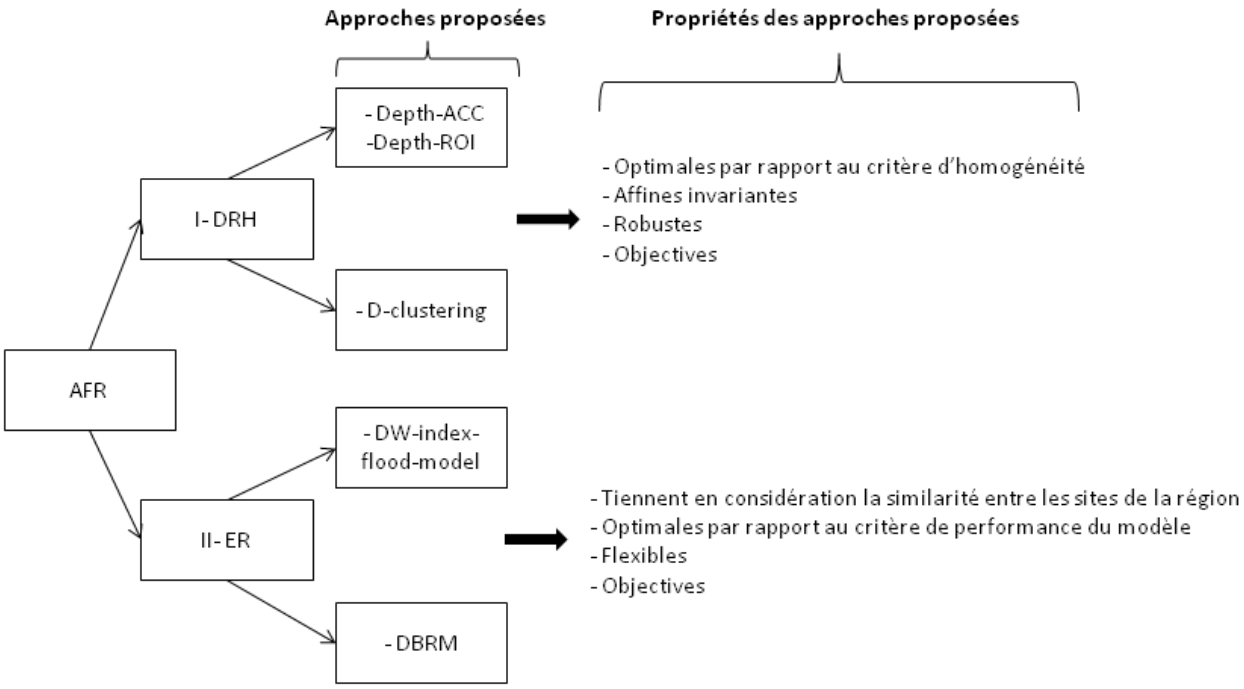


Figure 9. Les approches proposées au cours de cette thèse et leurs principales propriétés.

Dans un premier temps, nous avons proposé une nouvelle méthode pour identifier le voisinage d'un site cible. Cette méthode rend l'étape de délimitation robuste, affine invariante et indépendante de la distribution des variables. La méthode proposée a été appliquée sur un ensemble des sites de la partie sud de la province de Québec (Canada). Les résultats montrent que le voisinage d'un site cible obtenu en utilisant la méthode proposée est plus homogène que ceux obtenus en utilisant les approches traditionnelles. En plus, la méthode proposée conduit à des estimations plus efficaces que celles utilisant les méthodes traditionnelles (ROI et ACC).

Ensuite, pour identifier les régions homogènes de type non contiguës, nous avons proposé une nouvelle approche robuste et objective. Les régions homogènes sont formées en attribuant chaque site à la région qui maximise sa valeur de profondeur. L'approche proposée rend l'étape de délimitation pratique et ne nécessite pas l'intervention subjective de l'utilisateur. L'utilisateur

n'a qu'à sélectionner un critère pour obtenir les régions homogènes par rapport à ce critère. Les résultats obtenus à partir d'un ensemble des sites dans le Nord-Ouest de l'Italie montrent que l'approche proposée conduit à des régions plus homogènes que celles en utilisant l'approche classique de Ward.

Partant du constat que le modèle d'indice de crue utilise l'information disponible au sein d'une région de manière rationnelle, nous avons visé à construire une nouvelle approche utilisant l'information d'une manière sensée. Ceci a conduit à développer un modèle d'estimation optimale. Dans ce modèle, nous avons introduit la similarité hydrologique entre les bassins d'une région homogène dans l'étape d'estimation. Cette nouvelle approche a été appliquée sur les données des crues de la Sicile, Italie. Nous avons obtenu des estimations meilleures, en termes d'erreur quadratique moyenne, que celles issues par des approches classiques.

Enfin, nous avons proposé une procédure d'estimation afin d'optimiser la performance du modèle régional de régression des crues. Cette procédure d'estimation a été appliquée sur trois différentes régions en Amérique du Nord. Les résultats montrent la flexibilité et la généralité de la procédure proposée. La procédure d'optimisation de l'approche, non seulement améliore la performance, mais aussi rend l'approche automatisée et moins subjective quant au choix des poids.

7.2. Limitations et perspectives

Même si les approches proposées sont supérieures aux approches traditionnelles en termes de flexibilité, de représentativité, de généralité, et de l'optimalité, elles aussi peuvent être

améliorées. En effet, ces approches sont composées de plusieurs étapes et comprennent plusieurs outils, donc elles peuvent être assez lourdes en termes de temps de calcul.

Dans les travaux effectués, les pluparts des éléments liés aux approches proposées sont traités. Cependant, d'autres méritent d'être développés dans des efforts futurs. Par exemple:

- Une première direction des travaux futurs pourrait être une étude de sensibilité approfondie de l'impact de différents facteurs qui peuvent affecter la performance et l'efficacité des approches proposées. Ces facteurs comprennent le choix de la fonction de profondeur dans les différentes approches proposées, le choix des variables d'attributs dans les approches proposées pour l'étape de DRH, la méthode d'estimation des paramètres dans les approches proposées pour l'étape d'ER.
- Une autre direction est d'appliquer les approches proposées à des situations particulières. Ces situations comprennent, par exemple, des régions avec corrélation entre les sites et/ou avec faible niveau d'homogénéité.
- Dans cette thèse, pour évaluer les méthodes d'ER proposées, les deux critères couramment utilisés dans la littérature hydrologique, RB et RRMSE, ont été utilisés. La précision de ces deux critères dépend de la qualité des estimations locales (at-site) réalisées. En outre, pour quantifier le degré d'homogénéité des régions obtenues, l'indice *H* de Hosking et Wallis a été utilisé. Cet indice n'est pas borné et peut prendre des valeurs négatives. Une perspective sur ce point serait d'approfondir la recherche en proposant des critères pour évaluer les méthodes d'estimation qui ne dépendent pas des estimations locales et ainsi réduire les étapes de l'étude et les incertitudes de l'estimation. En plus, il serait utile de proposer une nouvelle statistique qui quantifie le degré d'homogénéité des

régions et qui prend des valeurs interprétables, par exemple dans [0, 1]. Ces critères et cette statistique doivent être validés et testés à l'aide d'études par simulations.

- Dans cette thèse le vecteur des paramètres β dans le modèle de RM est estimé par la méthode de moindres carrés pondérés (MCP). Enfin, il serait sans doute intéressant d'utiliser d'autres méthodes plus sophistiquées, comme la méthode de moindres carrés généralisée où/et les méthodes robustes pour la régression, pour estimer le vecteur des paramètres. Ces méthodes permettent d'inclure la dépendance qui peut exister entre les sites de la région et de réduire l'influence des sites aberrants dans l'estimation de vecteur des paramètres. Dans un premier temps, il serait intéressant de considérer une comparaison entre ces méthodes et celles proposées dans la thèse. De plus, il est possible de les améliorer et les rendent plus avantageuses en leur intégrant les fonctions de profondeur.

Références

Acreman, M., and C. Sinclair (1986), Classification of drainage basins according to their physical characteristics; an application for flood frequency analysis in Scotland, *Journal of Hydrology*, *84*(3-4), 365-380.

Benson, M. A. (1962), Factors influencing the occurrence of floods in a humid region of diverse terrain.

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

Cannarozzo, M., L. V. Noto, F. Viola, and G. La Loggia (2009), Annual runoff regional frequency analysis in Sicily, *Physics and Chemistry of the Earth*, *34*(10-12), 679-687.

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

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

Chernick, M. R. (2012), The jackknife: A resampling method with connections to the bootstrap, *Wiley Interdisciplinary Reviews: Computational Statistics*, *4*(2), 224-226.

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), 1-13.

Cunnane, C. (1987), Review of statistical models for flood frequency estimation, ed., Dordrecht, The Netherlands, D. Reidel Publishing Co., 1987, Section 1981, p.1949-1995.

Dalrymple, T. (1960), Flood frequency methods, *Water Supply Paper No. 1543 A*.

Gompertz, B. (1825), On the nature of the function expressive of the law of human mortality, and on a new mode of determining the value of life contingencies, *Philos. Trans. R. Soc. Lond.*, *115*, 513-585.

GREHYS (1996), Inter-comparison of regional flood frequency procedures for Canadian rivers, *Journal of Hydrology*, *186*(1–4), 85-103.

Grover, P. L., D. H. Burn, and J. M. Cunderlik (2002), A comparison of index flood estimation procedures for ungauged catchments, *Canadian Journal of Civil Engineering*, *29*(5), 734-741.

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*, *430-431*, 142-161.

Hosking, J. R. M., and J. R. Wallis (1993), Some statistics useful in regional frequency analysis, *Water Resour. Res.*, *29*(2), 271-281.

Hosking, J. R. M., and J. R. Wallis (1997), *Regional Frequency Analysis: An Approach Based on L-Moments*, Cambridge University Press.

Ilorme, F., and V. W. Griffis (2013), A novel procedure for delineation of hydrologically homogeneous regions and the classification of ungauged sites for design flood estimation, *Journal of Hydrology*, *492*, 151-162.

Jin, M., and J. R. Stedinger (1989), Flood frequency analysis with regional and historical information, *Water Resources Research 25*(5), 925-936.

Jörnsten, R. (2004), Clustering and classification based on the L1 data depth, *Journal of Multivariate Analysis*, *90*(1 SPEC. ISS.), 67-89.

Khaliq, M. N., A. St-Hilaire, T. B. M. J. Ouarda, and B. Bobée (2005), Frequency analysis and temporal pattern of occurrences of southern Quebec heatwaves, *International Journal of Climatology*, *25*(4), 485-504.
Leclerc, M., and T. B. M. J. Ouarda (2007), Non-stationary regional flood frequency analysis at ungauged sites, *Journal of hydrology*, *343*(3), 254-265.
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*(1-4), 1-9.
Liu, Y. (1990), On a Notion of Data Depth Based on Random Simplices, 405-414.
López, Á., and J. Romo (2010), Simplicial similarity and its application to hierarchical clustering, Universidad Carlos III, Departamento de Estadística y Econometría.
Lu, L.-H., and J. R. Stedinger (1992), Sampling variance of normalized GEV/PWM quantile estimators and a regional homogeneity test, *Journal of Hydrology*, *138*(1–2), 223-245.
Mahalanobis, P. C. (1936), On the generalized distance in statistics, *Calcutta Statist. Assoc. Bull.*, *14*, 9.
Minghui, J., and J. R. Stedinger (1989), Flood frequency analysis with regional and historical information, *Water Resources Research*, *25*(5), 925-936.
Naulet, R., M. Lang, T. B. M. J. Ouarda, D. Coeur, B. Bobée, A. Recking, and D. Moussay (2005), Flood frequency analysis on the Ardèche river using French documentary sources from the last two centuries, *Journal of Hydrology*, *313*(1-2), 58-78.
Neykov, N. M., P. N. Neytchev, P. H. A. J. M. Van Gelder, and V. K. Todorov (2007), Robust detection of discordant sites in regional frequency analysis, *Water Resources Research*, *43*(6), W06417.
Noto, L. V., and G. La Loggia (2009), Use of L-moments approach for regional flood frequency analysis in Sicily, Italy, *Water Resources Management*, *23*(11), 2207-2229.
Ouarda, T. B. M. J. (2013), Hydrological Frequency Analysis, Regional, in *Encyclopedia of Environmetrics*, edited, John Wiley & Sons, Ltd.
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.
Pandey, G. R., and V. T. V. Nguyen (1999), A comparative study of regression based methods in regional flood frequency analysis, *Journal of Hydrology*, *225*(1-2), 92-101.
Shu, C., and D. H. Burn (2004), Artificial neural network ensembles and their application in pooled flood frequency analysis, *Water Resources Research*, *40*(9), W09301.
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*(7), W07438.
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*(1–2), 31-43.
Stedinger, J. R., R. M. Vogel, and E. Foufoula-Georgiou (1992), *Frequency analysis of extreme events* 18.11-18.66 pp., McGraw-Hill, New York.
Thomas, D. M., and M. A. Benson (1970), *Generalization of streamflow characteristics from drainage-basin characteristics*, US Government Printing Office.
Torczon, V. (2000), On the Convergence of Pattern Search Algorithms, *SIAM Journal on Optimization*, *7*(1), 1-25.

Tukey, J. W. (1975), Mathematics and the picturing of data, *Proceedings of the International Congress of Mathematicians*, *2*, 523-531.
Verhulst, P. F. (1838), *Notice sur la loi que la population pursuit dans son accroissement*.
Viglione, A. (2010), Non-supervised Regional Frequency Analysis, *http://cran.r-project.org/web/packages/nsRFA/index.html*.
Viglione, A., F. Laio, and P. Claps (2007a), A comparison of homogeneity tests for regional frequency analysis, *Water Resources Research*, *43*(3), W03428.
Viglione, A., P. Claps, and F. Laio (2007b), Mean annual runoff estimation in north-western Italy,, in *Water Resources Assessment and Management Under Water Scarcity Scenarios*, edited by L. Loggia, CDSU Publication, Milan.
Vogel, R. M., T. A. McMahon, and F. H. S. Chiew (1993), Floodflow frequency model selection in Australia, *Journal of Hydrology*, *146*(C), 421-449.
Ward Jr, J. H. (1963), Hierarchical grouping to optimize an objective function, *Journal of the American Statistical Association*, *58*(301), 236-244.
Wazneh, H., F. Chebana, and T. B. M. J. Ouarda (2013a), Depth-based regional index-flood model, *Water Resources Research*, *49*(12), 7957-7972.
Wazneh, H., F. Chebana, and T. B. M. J. Ouarda (2013b), Optimal depth-based regional frequency analysis, *Hydrology and Earth System Sciences*, *17*(6), 2281-2296.
Wazneh, H., F. Chebana, and T. B. M. J. Ouarda (2014a), Identification of hydrological neighborhoods using statistical depth function, *Submitted*.
Wazneh, H., F. Chebana, and T. B. M. J. Ouarda (2014b), Delineation of homogeneous regions for regional frequency analysis using statistical depth function, *journal of hydrology*, *Accepted*.
Zrinji, Z., and D. H. Burn (1994), Flood frequency analysis for ungauged sites using a region of influence approach, *Journal of Hydrology*, *153*(1–4), 1-21.
Zuo, Y., and R. Serfling (2000), General notions of statistical depth function, *Annals of Statistics*, 461-482.

LISTE DES TABLEAUX

LISTE DES FIGURES

LISTE DES NOTATIONS

Symbole	Définition
AF	Analyse fréquentielle
T	Période de retour
DMA	Débits maximums annuels
Q_T	Amplitude du débit de la crue
t	Probabilité de dépassement
LN	Log normal
GEV	Valeurs extrêmes généralisées
LP3	Log-Pearson type 3
AFR	Analyse fréquentielle régionale
DRH	Détermination des régions homogènes
CAH	Classification ascendante hiérarchique
ROI	Région d'influence
ACC	Analyse canonique de corrélations
ER	Estimation régionale
N	Nombre total des bassins versants dans la région
$X = (X_1,.,X_i,.,X_r)$	Matrice contient les *r* variables physio-météorologiques
$X_i = (X_i^1,...,X_i^N)'$	Vecteur contient les valeurs du *i* ème variable physio-météorologiques pour les *N* sites
$Y = (Y_1,.,Y_i,.,Y_s)$	Matrice contient les *s* variables hydrologiques
$Y_i = (Y_i^1,..,Y_i^N)'$	Vecteur contient les valeurs du *i* ème variable hydrologique pour les *N* sites
Σ_{XY}	Matrice de covariance entre *X* et *Y*
Σ_X	Matrice de variance de *X*
Σ_Y	Matrice de variance de *Y*
p	Rang de la matrice Σ_{XY}.
$W^m = (W_1^m,...,W_p^m)'$	Vecteurs des variables canoniques hydrologiques pour le site *m*
$V^m = (V_1^m,...,V_p^m)'$	Vecteurs des variables canoniques Physio-météorologiques pour le site *m*
$W_i = (W_i^1,...,W_i^N)$	Vecteur contient les valeurs du *i* ème variable canonique hydrologique pour les *N* sites
$V_i = (V_i^1,...,V_i^N)$	Vecteur contient les valeurs du *i* ème variable canonique Physio-météorologique pour les *N* sites
l	Indice du site cible

d^2	Distance de Mahalanobis
$\chi^2_{\alpha,p}$	(1-α) quantile associé à la distribution du khi-deux avec p degrés de liberté
$X^m = \left(X_1^m, .., X_r^m\right)$	Vecteur contient les r variables physio-météorologiques du site m
D	Distance euclidienne
δ	Seuil
$Q_l(t)$	Quantile de crue correspond à une probabilité de non-dépassement t
μ_l	Quantité indice de crue associée au site l
$q_t(.)$	Courbe de croissance
$\theta^R = \left(\theta_1^R, .., \theta_S^R\right)$	Vecteur des paramètres régionaux
$\hat{\theta}_s^R$	$s^{\text{ème}}$ paramètre régional
$\hat{\theta}_s^h$	$s^{\text{ème}}$ paramètre obtenu à partir de la distribution locale du site h
ω_h	Poids associé au site h
N'	Nombre des sites dans la région homogène
n_h	Longueur d'enregistrement du site h
PW-index-flood"	Indice de crue avec le poids proportionnel
UW-index-flood	Indice de crue avec le poids uniforme
RW-index-flood	Indice de crue avec le R-poids
DW-index-flood	Indice de crue basée sur la fonction de profondeur
RM	Modèle de régression multiple
$\beta = \left(\beta_0, ..., \beta_r\right)$	Vecteur des paramètres du modèle de régression
$\Omega = \text{diag}(w_1, ..., w_N)$	Matrice de pondération
ε	Erreur du modèle de régression
MCO	Moindres carrés ordinaires
Γ	Variance de l'erreur du modèle
MHD	La profondeur de Mahalanobis
SPD	Profondeur spatiale
SD	Profondeur simpliciale
φ_G	Fonction de Gompertz
$\varphi_{\text{log}istic}$	Fonction logistique
H	Mesure d'hétérogénéité
RB	Biais relatif
RRMSE	Erreur relative quadratique moyenne
SS	Similarité simpliciale

DSS	Dissimilarité simpliciale
ACC-Depth	Méthode de voisinage fondée de l'ACC basé sur la fonction de profondeur
ROI-Depth	Méthode de voisinage fondée de ROI basé sur la fonction de profondeur
K	Nombre de sous régions homogènes
D-clustering	Approche de classification basée sur la fonction de profondeur
$i(k)$	Ensemble qui contient les numéros des sites de la sous-région *k*
$I(k)$	Ensemble qui contient les attributs des sites de sous-région *k*
D_m^w	Profondeur 'intra sous-région' du site *m*
D_m^b	Profondeur 'inter sous-région' du site *m*
DeD_m	Profondeur 'Déviance' du site *m*
DBRM	Modèle de régression multiple basée sur la fonction de profondeur
MCP	Moindres carrées pondérées
AREA	Superficie du bassin versant
MBS	Pente moyenne du bassin versant
FAL	Fraction de la superficie couverte par des lacs
AMP	Précipitations annuelles moyennes totales
AMD	Nombre moyen des jours de plus de 0 C
X_{bar} et Y_{bar}	Coordonnées des sites dans le système UTM
H_m	Élévation moyenne du bassin versants

CHAPITRE 2: IDENTIFICATION DES VOISINAGES HYDROLOGIQUES EN UTILISANT LES FONCTIONS DE PROFONDEUR

Identification of hydrological neighborhoods using statistical depth function

H. Wazneh*[1], F. Chebana[1] and T.B.M.J. Ouarda[2]

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

*__Corresponding author:__ Tel: +1 (418) 654 2530#4468
Email: hussein.wazneh@ete.inrs.ca

December 13th 2014

(TO BE SUBMITTED)

Abstract

Regional frequency analysis (RFA) aims to estimate extreme hydrological events at sites where little or no hydrological data are available. The delineation of homogeneous regions and the regional estimation are the two main steps of a RFA procedure. Hydrological neighborhood is one of the common approaches employed for the delineation step. Traditional methods proposed for building hydrological neighborhoods are mainly based on a distance metric. These methods have some limitations. They are not robust against outliers, they are not affine invariant and require that site characteristics be normally distributed. To overcome these limitations, the present study aims to propose new robust methods to identify neighborhood of a target site. The proposed methods are based on the statistical notion of depth function. A data set from the southern part of the province of Quebec (Canada) is used to compare the proposed methods with traditional ones. The obtained results indicate that the depth-based methods lead to neighborhoods which are more homogeneous in terms of *H* heterogeneity measure, and are more efficient for quantile estimation in terms of relative bias and relative root mean square error, than those obtained by the traditional methods.

Keywords: Ungauged basins, Neighborhood, Target site, Regional Frequency Analysis, Statistical depth function, Canonical Correlation Analysis, Region of Influence.

1. Introduction

Extreme hydrological events have important and different consequences on human society. These extreme events are rare and their record lengths are generally short at the desired site. Consequently, statistical inference is difficult in such sites. To overcome this problem, information must be transferred from other sites that are hydrologically similar to the target one. The estimation of extreme events, such as floods, at sites where little or no data are available is the main aim of regional frequency analysis (RFA). The latter comprises two main steps: delineation of homogeneous regions and regional estimation [e.g., *De Michele and Rooso*, 2002; *Ouarda et al.*, 2008; *Zaman et al.*, 2012].

Homogenous regions can be defined as geographically contiguous regions, geographically noncontiguous regions, or as hydrological neighborhoods. The use of hydrological neighborhoods was recommended in the literature for different parts of the world. In the neighborhood category each target site is assumed to have its own region. Two main neighborhood methods are proposed in the literature: the region of influence (ROI) [*Burn*, 1990]; and the canonical correlation analysis (CCA) [*Ouarda et al.*, 2001].

Generally, a neighborhood of a target-site is composed of a set of hydrologically similar gauged sites [*Chokmani and Ouarda*, 2004]. Therefore, the determination of the neighborhood involves the computation of a similarity measure between the target site and the gauged ones. To this end, the available methods in the literature, i.e. ROI and CCA, use a simple distance metric. More precisely, ROI uses the weighted Euclidian distance within a multidimensional space defined by the physio-meteorological variables, such as the catchment centroids or the basin area [*Burn*, 1990; *Gaál et al.*, 2008]. However, CCA uses the Mahalanobis distance within the

multidimensional space defined by the canonical physio-meteorological and hydrological variables [*Ribeiro-Corréa et al.*, 1995].

The available neighborhood methods suffer from three principal drawbacks. First, they are not robust against outliers in the data set, i.e. they exhibit a high sensitivity to the presence of very different sites, and consequently, the neighborhood of a target site can be strongly affected by some particular sites such as sites with very small or/and large drainage area [*Neykov et al.*, 2007]. Second, these methods are not affine invariant which make the neighborhood of a target site depend on the coordinate system, and in particular, on the scales of the site characteristics [*Chebana and Ouarda*, 2008]. Third, these traditional methods require the normality of site characteristics [*Leclerc and Ouarda*, 2007].

These elements motivate the development of new delineation methods that overcome some of these drawbacks. These methods developed in the present study are based on the notion of depth function [*Tukey*, 1975]. More precisely, to define a neighborhood of a target site, the proposed methods use a new similarity measure, derived from the statistical depth notion. This similarity measure has several desirable properties including robustness, affine invariance, and taking into account the shape of the data set (see section 3 below). To define the space where the proposed similarity measure is computed, two approaches are used in this study including: the physio-meteorological space and the canonical space.

Statistical depth functions have been introduced in the estimation step of RFA [*Chebana and Ouarda*, 2008; *Wazneh et al.*, 2013a; b]. Recently, Wazneh el al., [2014] introduced the notion of depth functions in the delineation step of RFA. More precisely, the authors have introduced depth function in the formation of homogeneous noncontiguous regions (i.e. in the clustering approach). They showed that the depth-based approach leads to regions which are more homogeneous and more efficient for flood quantile estimation than those obtained by the

traditional one. However in the present study, depth functions are introduced in the formation of hydrological neighborhood, which is the difference between this study and the previously one.

The present paper is structured as follows. Section 2 gathers the different elements of the background necessary to introduce the proposed neighborhood methods. Section 3 describes the proposed methods. In section 4, the proposed depth neighborhood methods are applied and compared to the traditional methods based on a data set from the hydrometric station network of the province of Quebec (Canada). The last section is devoted to the conclusions of this work.

2. Background

In this section, the background materials required to introduce and apply the identification of hydrological neighborhoods using statistical depth function are briefly presented. This section contains a number of basic notions as well as a brief reminder of the multiple regression model and the traditional neighborhood methods.

2.1. Simplicial depth function

The absence of a natural order to classify multivariate data led to the introduction of the depth functions (Tukey, 1975). In the present work the simplicial depth function is used. This particular case of depth functions is used because it can easily be transformed to a similarity measure (see section 3 c). A detailed description of the way the simplicial depth is used to introduce the similarity between data sets of different sites and to define the neighborhood of a target site is presented below in section 3.

The simplicial depth of $x \in R^d$ is defined as the probability that the point x belongs to a random simplex whose vertices are a (d+1) independent random variables with distribution F [*Liu*, 1990]. More formally, the depth of x is defined by:

$$SD(x, F) = P\left[x \in S\left[X_1, ..., X_{d+1}\right]\right] \quad (1)$$

where $X_1,...,X_{d+1}$ are random variables from F and $S\left[X_1,...,X_{d+1}\right]$ is the simplex with vertices $X_1,...,X_{d+1}$. The sample version of this depth function consists in computing the proportion of simplex containing the point x; i.e., given a random sample $x_1,...,x_n$ from F with $n > d$, the depth value of x is estimated by:

$$SD_n(x) = \binom{n}{d+1}^{-1} \sum_{1 \le i_1 < ... < i_{d+1} \le n} 1_{S\left[x_{i_1},....,x_{i_{d+1}}\right]}(x) \qquad (2)$$

where $1_A(x) = \begin{cases} 1 \text{ if } x \in A \\ 0 \text{ if } x \notin A \end{cases}$. Figure 1 illustrates the way that simplicial depth is computed in the two dimensional space. In fact, the depth of the point x is the number of simplex (i.e. triangle in R^2) containing this point divided by the total number of simplex that can be formed using the data set. Therefore the depth of x in Figure 1 is 3/10.

2.2. Traditional neighborhood methods

A brief description of the two main traditional neighborhood methods is presented herein.

Canonical correlation analysis (CCA)

Let $X = \left(X_1,.,X_i,.,X_r\right)$ with $X_i = \left(X_i^1,...,X_i^N\right)'$ and $Y = \left(Y_1,.,Y_i,.,Y_s\right)$ with $Y_i = \left(Y_i^1,..,Y_i^N\right)'$ are respectively the r physio-meteorological (e.g., basin area, annual mean total precipitation) and the s hydrological variables (e.g., flood quantiles) of the N gauged sites. CCA provides two sets of basis vectors highly correlated (called canonical variables), one for X and the other for Y. More precisely, let V and W be linear combinations of X and Y respectively, i.e.,

$$V = a'X \text{ and } W = b'Y \qquad (3)$$

Let Σ be the covariance matrix of the variables X and Y;,and ρ the correlation between V and W.The goal of the CCA is to find two vectors a and b maximizing ρ subject to the constraint that

W and *V* must have unit variance. Once the first pair of canonical variables is obtained, other pairs can be similarly obtained in the uncorrelated directions to the previous.

Let $W^n = \left(W_1^n, ..., W_p^n\right)'$ and $V^n = \left(V_1^n, ..., V_p^n\right)'$ be respectively the first canonical hydrological and physio-meteorological variables of site n $\left(n = 1, .., N\right)$, where p is the rank of the covariance matrix Σ. Assume that the vector $\begin{pmatrix} W \\ V \end{pmatrix}$ is 2*p*-normally distributed, the hydrological neighborhood of a target site *l*, using the CCA method at 100(1-α)% confidence level is given by:

$$N_{ACC}^l = \left\{ \text{site } k \in \{1, .., N\};\ d^2_{I_p - \Lambda'\Lambda}\left(W^k, \Lambda V^l\right) = \left(W^k - \Lambda V^l\right)' \left(I_p - \Lambda'\Lambda\right)^{-1} \left(W^k - \Lambda V^l\right) < \chi^2_{\alpha,p} \right\} \quad (4)$$

where I_p is the $p \times p$ identity matrix, $\Lambda = \text{diag}(\lambda_1, ..., \lambda_p)$, $\lambda_i = \text{corr}(V_i, W_i), i = 1, ..., p$ with $V_i = \left(V_i^1, ..., V_i^N\right)$ and $W_i = \left(W_i^1, ..., W_i^N\right)$, d^2 is the Mahalanobis distance, and $\chi^2_{\alpha,p}$ is the (1-α) quantile associated to the chi-squared distribution with *p* degrees of freedom. This neighborhood describes the interior of an ellipsoidal region. For more details concerning CCA application in RFA, the reader is referred to Ouarda et al. [2001].

Region of influence

The ROI method was introduced by Burn [1990]. As in CCA method, the target-site is considered as the center of its own neighborhood. The identification of a neighborhood for a given target-site is based on an Euclidian distance in multidimensional physio-meteorological space [*Tasker et al.*, 1996].

More formally, suppose that $X^n = \left(X_1^n, .., X_r^n\right)$ be the vector of *r* physio-meteorological variables of site *n*. Let *N* be the number of gauged sites in the studied region, the neighborhood of an ungauged site *l*, i.e. the vector of physio-meteorological variable X^l is known, using ROI method is defined as:

$$N_{ROI}^{l} = \left\{ \text{site } k \in \{1,..,N\};\ D\left(X^{k}, X^{l}\right) = \left[\sum_{s=1}^{r}\left(X_{s}^{k} - X_{s}^{l}\right)^{2}\right]^{1/2} \le \delta \right\} \qquad (5)$$

where $D\left(X^{k}, X^{l}\right)$ is the Euclidean distance between the target site *l* and the gauged one *k* ; and δ is the threshold value. As the physio-meteorological variable $X^{n} = \left(X_{1}^{n},..,X_{r}^{n}\right)$ may have substantially different magnitudes, such as in the case of *X*=(basin area, mean basin slope, annual mean total precipitation), a standardization of the matrix *X*, containing the *N* vector of physio-meteorological variables, before calculating the dissimilarity measure defined in (5) is generally applied.

2.3. Multiple regression model

In RFA, the neighborhood of a target site is used to estimate its flood quantiles. In this study, the multiple regression (MR) model is used to estimate the flood quantile Q_T corresponding to a return period *T*. A brief description of the MR model in RFA is presented herein.

The MR model has the advantage to be simple, fast, and not requiring the same distribution for hydrological data at each site within the region. It is assumed that the relationship between Q_T , as the hydrological variable, and the basin characteristics A_1, A_2, . . A_r takes the form [*Pandey and Nguyen*, 1999; *Tasker et al.*, 1996]:

$$\log\left(Q_T\right) = \beta_0 + \beta_1 \log\left(A_1\right) + \beta_2 \log\left(A_2\right)... + \beta_r \log\left(A\right) + e \qquad (6)$$

where $\beta = \left(\beta_0,....,\beta_r\right)$ is the vector of parameters and *e* is the model error. The vector of parameters β can be estimated using the weighted least squares estimation method [Rencher 2002].

2.4. Performance criteria

An objective of the present work is to assess and compare the performance of neighborhood methods used in RFA. Since the delineation step affects the results of estimation of flood quantiles, two categories of performance criteria are used in this study: i) criterion that assess the neighborhood methods and ii) criteria that assess the estimation results.

To assess and compare the neighborhood methods, the *H* heterogeneity measure defined bellow is used. As mentioned in Hosking and Wallis [1997], *H* quantifies the degree of heterogeneity of a neighborhood. For a target site *l*, the heterogeneity measure H_l corresponding to its neighborhood is computed:

$$H_l = \frac{\left(V_l - \mu_l^V\right)}{\sigma_l^V} \text{ such that } V_l = \frac{\sum_{i \in N^l} w_i \left(L_{cv}^i - \bar{L}_{cv}\right)^2}{\sum_{i \in N^l} w_i} \qquad (7)$$

where N^l represent the set of sites included in the neighborhood of site *l*, w_i is the sample size at site *i*, L_{cv}^i and $\bar{L}_{cv}$ are respectively the *L*-coefficient of variation of site *i* and the average one across all sites in N^l, and μ_l^V and σ_l^V are the mean and the standard deviation of the simulated neighborhoods respectively. For more details concerning the way that L_{cv}^i, $\bar{L}_{cv}$, μ_l^V and σ_l^V are computed, the reader is referred to Hosking and Wallis [1997]. The best neighborhood method will be the one that leads to the lowest H_l value.

As mentioned above, in the RFA framework, the neighborhood of a target site *l* is used to estimate its flood quantiles Q_T^l corresponding to a return period *T*. Therefore, the best neighborhood method is the one that minimizes the prediction error of Q_T over all sites in the data set. In this study, to quantify the prediction error, the following criteria are used:

$$RB_T = \frac{1}{N}\sum_{l=1}^{N} RE\left(\hat{Q}_T^l\right) \quad \text{with} \quad RE\left(\hat{Q}_T^l\right) = \frac{Q_T^l - \hat{Q}_T^l}{Q_T^l} \tag{8}$$

$$RRMSE_T = \sqrt{\frac{1}{N}\sum_{l=1}^{N}\left[RE\left(\hat{Q}_T^l\right)\right]^2} \tag{9}$$

where RB_T, $RRMSE_T$ and $RE\left(\hat{Q}_T^l\right)$ are respectively the relative bias, the relative root mean square error, and the relative quantile error of site *l*; corresponding to the return period *T*, Q_T^l is the at-site quantile estimation and $\hat{Q}_T^l$ is the regional estimation by the regression model, using a given neighborhood method.

3. Depth-based neighborhood method

The neighborhood methods proposed in the literature to identify homogeneous regions, i.e. CCA and ROI, are based on a distance dissimilarity measure. More precisely, CCA (4) is based on the Mahalanobis distance, whereas ROI (5) is based on the Euclidean distance. The results of these methods can be affected by several elements including the distributions of sites characteristics [*Leclerc and Ouarda*, 2007; *Lin and Chen*, 2006], the presence of outliers [*Hadi*, 1992; *Neykov et al.*, 2007] and the scale of variables [*Castellarin et al.*, 2001]. To reduce or eliminate the negative impact of these elements, a new similarity function based on the concept of statistical depth function is presented herein.

The simplicial depth function of a point $x \in R^d$ is based on the membership of this point to random simplices with $d + 1$ vertices, where d is the dimension of the space of x. This function (1) measures the probability that a random simplex contains the point for which depth is calculated. In the same reasoning, Lopez and Romo [2010] defined the simplicial similarity between two points x and y as the probability that a random simplex contains both points. The two examples in Figure 2 show that the proposed idea is an appropriate way to measure

similarity. It can be observed that, when *x* and *y* are close (Figure 2a), two of the five sample triangles drawn contain both points at the same time, whereas if the points are far (Figure 2b), none of these five triangles contain both of them at the same time.

More formally, given a *d*-dimensional distribution function *F* and the points *x* and *y* in R^d, the simplicial similarity between *x* and *y* with respect to *F* is defined as:

$$SS(x, y, F) = P\left(x, y \in S\left[X_1, .., X_{d+1}\right]\right) \tag{10}$$

where $X_1, ..., X_{d+1}$ are a random sample from *F* and $S\left[X_1, ..., X_{d+1}\right]$ is the simplex with vertices $X_1, ..., X_{d+1}$. The sample version of this similarity function consists in computing the proportion of simplices containing the points *x* and *y* together; i.e., given a random sample $x_1, ..., x_n$ of *F*, the simplicial similarity between *x* and *y* is:

$$SS_n(x, y) = \binom{n}{d+1}^{-1} \sum_{1 \le i_1 < ... < i_{d+1} \le n} 1_{S\left[x_{i_1}, ..., x_{i_{d+1}}\right]}(x, y) \tag{11}$$

The most important proprieties of SS_n are:

1. Given a point *x*, the maximum similarity value between *x* and any other *y* is equal to the similarity between *x* and itself and is equal to the simplicial depth of *x* i.e. $SS_n(x, x) = \max_y \left[SS_n(x, y)\right] = SD_n(x)$;
2. Given any points *x* and *y*, the similarity between *x* and any point in the segment from *x* to *y* is greater than or equal to the similarity between points *x* and *y*;
3. The similarity between points *x* and *y* tends to zero when point *y* is far from *x*;
4. Given two points *x* and *y*, the distribution function *F* and an affine transformation *T(.)*, the similarity between points *T*(*x*) and *T*(*y*) with respect to the transformed distribution function is equal to the similarity between points *x* and *y* with respect to *F*.

5. Simplicial similarity is symmetric i.e. $SS_n(x,y) = SS_n(y,x)$.

As simplicial depth, simplicial similarity is robust, distribution free (i.e. no assumption about the distribution of the data), affine invariant, and takes into account the shape of the data [*Zuo and Serfling*, 2000]. These proprieties can reduce and eliminate the drawbacks of traditional neighborhood methods. To make values interpretable, the simplicial similarity measure (11) is normalized as:

$$NSS_n(x,y) = \frac{SS_n(x,y)}{\sqrt{SD_n(x).SD_n(y)}} \quad (12)$$

where *SD(.)* is the simplicial depth function defined in (2). Traditionnal neighborhood methods, such as the case in (4) and (5), are defined using a dissimilarity measure (distance). In order to be compatible with these methods, the normalized simplicial similarity should be transformed into dissimilarity measure. To this end, a logarithm function, which can help spread out the magnitude of NSS, is applied. As result, the following dissimilarity measure is used in this study to define the neighborhood of a target site using depth-based neighborhood approach:

$$DSS_n(x,y) = -\log\left(NSS_n(x,y)\right) \quad (13)$$

Since $NSS_n(x,y)$ takes values in the interval [0, 1], $DSS_n(x,y)$ takes positive values.

The neighborhood of a target site *l* using depth-based neighborhood method is defined as:

$$N^l_{DSS} = \left\{ \text{site } k \in \{1,..,N\};\ DSS_N\left(U^k, U^l\right) \le \tau \right\} \quad (14)$$

where U is the vector of site characteristics and τ is a threshold value.

In this study, to define the vector *U*, two different spaces are used: either the physiographical space or the canonical space. If $U^l = X^l = \left(X^l_1,..,X^l_r\right)$ the original vector of *r* physio-meteorological variables of site *l* such as in (5), the depth-based neighborhood method is denoted

by ROI-Depth; if $U^l = V^l = \left(V_1^l, ..., V_p^l\right)'$ the p canonical physio-meteorological variables of site l such as in (4), the depth-based neighborhood method is denoted by CCA-Depth.

Like CCA and ROI, the depth-based neighborhood method requires the choice of a threshold value τ that is as a cut-off point for the dissimilarity measure. All stations with a dissimilarity measure higher than the threshold τ are excluded from the neighborhood of target site. Note that determining the optimal value of threshold τ is equivalent to determining the optimal number of sites in the neighborhood $n^{opt.}$ that are close to the target site (i.e. have the lowest values of DSS_N). To this end, in this study, a "forward" pooling procedure is used [*Gaál and Kyselý*, 2009]. Starting with the target site l which represents a single-site pooling group at the very beginning of the procedure, the next closest site (i.e. the site j with the next lowest value of $DSS_N\left(U^l, U^j\right)$, j=1, ..., N) is appended to the existing neighborhood in each turn, and a pre-selected criterion quantifying the performances of the model (such as RB (8) or RRMSE (9)) is computed. The optimal $n^{opt.}$ number of sites in the neighborhood is the one that optimizes the performance of model. This same procedure is used in this study to determine the values of α and δ respectively for the CCA (4) and the ROI (5) methods [Ouarda el al., 2001].

4. Applications

In this section, the delineation of a neighbourhood using the depth-based and the traditional approaches are applied and compared on a real world dataset.

4.1. Dataset

The considered dataset comes from the hydrometric station network of the southern part of the province of Quebec (Canada). Annual maximum peak flow data of 151 stations are available with record lengths ranging from 15 to 84 years. These stations are located between 45° N and 55° N

in the southern part of Quebec, Canada. The area of these catchments is larger than 208 km^2 but less than 96600 km^2. The geographical location of these stations is shown in Figure 3.

An at-site frequency analysis was carried out at each station of the data base by [*Kouider et al.*, 2002]. Data were tested for homogeneity and stationarity, and appropriate statistical distributions were fitted to data in order to estimate at-site flood quantiles corresponding to several return periods. In this study, we focus on 10 and 100 years return periods. Eaton el al., [2002] indicated that in order to investigate the underlying physical behavior of drainage systems, scale effects must be eliminated from data. Consequently, we use specific quantiles (flood quantiles standardized by the basin area), noted by QS_{10} and QS_{100} (in m^3/km^2s),.

The selected physio-meteorological variables used in this study are used in the study of Wazneh el al. [2013a]. It includes: the basin area (AREA) in km^2, the mean basin slope (MBS) in %, the fraction of the basin area covered with lakes (FAL) in %, the annual mean total precipitation (AMP) in mm and the annual mean degree days over 0°C (AMD) in degree-day.

4.2. Results

In order to determine the optimal number of sites, for the different neighbourhood methods, that must be used in the estimation of flood quantile for a target site in the considered dataset. Figure 4 illustrates the variation of the two performance criteria RB and RRMSE, obtained by the MR model, and using ROI, ROI-Depth and CCA-Depth neighborhood methods, as a function of number of sites in the neighborhood (i.e. size of neighborhood). By construction, the size of a neighborhood can vary between 0, empty neighborhood, and 151, the whole region. However, in this study, the size of neighborhood varies between 10 and 80 to ensure that the neighborhood of sites contains sufficient stations to allow the estimation by the MR model, but not too large to maintain neighborhood homogeneity. Figure 4 indicates that, for a given criterion i.e. RB or

RRMSE, the optimal size of neighborhood $n^{opt.}$ is almost the same for the three neighborhood methods (i.e. ROI, ROI-Depth and CCA-Depth) and for the two return periods (T = 10 and 100 yrs), even though this is not a general result. However, the value of $n^{opt.}$ depends on the selected criterion [*Ouarda et al.*, 2001]. In this study, for these three methods, $n^{opt.}$ = 35 sites with respect to the RRMSE criterion, whereas $n^{opt.}$ = 50 sites for the RB criterion. Note that for the CCA method, as shown in a number of studies that used this dataset, the optimal number of sites in the neighborhood with respect to RB and RRMSE criteria is 76 [e.g. *Chebana et al.*, 2014; *Chokmani and Ouarda*, 2004; *Wazneh et al.*, 2013a]. This optimal size of neighborhood corresponds to $\alpha = 0.25$ in (4).

The results of the two performance criteria RB and RRMSE, related to the different neighborhood methods, and using the size of neighborhood $n^{opt.}$ that optimizes the RRMSE criterion, are summarized in Table 1. This table shows that the depth-based neighborhood methods lead to more precise estimates than those obtained using traditional CCA and ROI methods. In fact, the neighborhoods obtained using depth-based methods are more homogeneous and more representative than those obtained using traditional methods (see below), which could be the reason of this improvement. The improvement is more significant using ROI-Depth than CCA-Depth. This result can be explained by the fact that some information is lost by the CCA dimension reduction which is not the case of ROI-Depth where the simplicial dissimilarity is applied to the original set that contains 5 physio-meteorological variables.

Figure 5 shows the scatter of the at-site specific quantiles versus the regional estimate ones for the QS_{100} using the different neighborhood methods. In this Figure, each site i (i=1,..,151) is temporarily excluded from the data set and the corresponding regional quantile $\hat{Q}S_{100}^{i}$ is estimated by the MR model and using its neighborhood defined by the various methods. By

fitting a straight line with a slope equal to 1:1 between QS_{100} and $\hat{Q}S_{100}$, it is seen that for some sites, the regional quantiles estimated, using the neighborhood obtained by depth methods, are closer to the at-site quantiles (i.e. more uniformly distributed around 1:1 line) than those obtained using traditional CCA and ROI. The values of coefficient of determination R^2 show that the linearity between the at-site specific quantiles and the regional estimates one is more respected in the depth-neighborhood methods than in the traditional methods. Besides, we remark that traditional neighborhood methods lead to underestimate sites with high quantile values, i.e. sites that have *QS* values greater than 0.8 m^3/km^2s, such as the case of sites 10, 48, 49 and 78 (Figure 5). These high quantile values are associated to sites with small basin area, less than 500 km^2. However, the depth-based neighborhood methods, appear to produce more precise values for these sites. This result can be explained by the fact that traditional neighborhood approaches are highly related by the values of site characteristics, so these delineation approaches are not effective in the case of specific target sites such as sites with small basin area. However, the depth neighborhood methods does not depend directly on the values of sites characteristics, but it depends on the position of target sites in the hydrological and the physiographical space. The effectiveness of depth-based delineation methods in the case of specific target sites can also explain the results obtained in Table 1.

To compare the relative errors (8) of flood quantiles estimates obtained by different neighborhood methods, Figure 6 illustrates these errors with respect to the logarithm of basin area. It is generally observed that relative errors obtained using depth neighborhood methods are lower than those obtained using traditional neighborhood methods. We also observe that generally relative errors are randomly distributed around zero except for some sites such as numbers 46, 64, 66, 148. These sites have large negative relative errors using any of various

delineation methods. Chokmani and Ouarda [2004] found that for these four sites, the catchment areas are underestimated, which can be the cause of these larges relative errors. The exclusion of these four sites can improve the overall results. Note that, sites 10, 48, 49 and 78 have proportionally small relative errors because their high at-site quantile values.

The homogeneity of 151 neighborhoods, related to the temporarily excluded sites, are quantified using the *H* heterogeneity measures defined in (7). Table 2 shows the frequency of *H* in each predefined class intervals and using the different neighborhood methods. We remark that most neighborhoods obtained by the ROI method are 'heterogeneous' (H>2). However, more than half of neighborhoods obtained by the depth-based methods (CCA-Depth or ROI-Depth) are 'homogeneous' or 'possibly homogeneous'. Generally, Table 2 shows that neighborhoods obtained using depth-based methods are more homogeneous than those obtained using traditional methods, which could be the reason of the improvement in the estimation step (Table 1). Note that, the estimation of flood quantile using MR model does not require the assumptions of neighborhood homogeneity, which can justify the use of this model in the estimation step in this study.

In order to visualize the neighborhoods of a target site using the various methods, assume that site number 97 is a target site and has to be estimated using the remaining 150 gauged sites. Figure 7 illustrates the neighborhoods using CCA and CCA-Depth methods. The neighborhoods region for the target site using ROI and ROI-Depth are not illustrated in this Figure since these two methods are computed in a 5 dimensional space (number of physio-meteorological variables). By observing the neighborhood of the target site 97, we remark that using CCA method, some sites relatively far from the target site in the canonical hydro-physiographical space (W1, V1), such as sites 109, 120, 122, 4 and 1, are included in its neighborhood (Figure 7a). These sites are also geographically far from the target site (Figure 3). By testing the

normality of the vector $\begin{pmatrix} W \\ V \end{pmatrix}$ that contains the canonical hydro-physiographical variables, we find that the null hypothesis is rejected. Therefore, the basic assumption of normality behind the CCA neighborhood method is not respected in this study, which can be the reason that some relatively far sites are include in the neighborhood using CCA method. However, these far sites are not included in the neighborhood using Depth-CCA method (Figure 7b). This is because the neighborhood using depth-based methods is computed without any assumption. That can also explain the superiority of depth neighborhood methods in the estimation step (Table 1). Furthermore, we observe that the neighborhood using traditional CCA describes in the canonical hydrological space the interior of an ellipsoidal region (Figure 7b) as formed in Ouarda et al., [2001]. However, for the CCA-Depth, the neighborhoods in the canonical physiographical space present a triangular shape. This is due to the angular shape of the simplices that define the simplicial similarity (11) used in this method.

The sensitivity of the neighborhood approaches to an affine transformation is investigated. More precisely, the neighborhoods of the target site 97 are computed using the original and the normalized version of the physio-meteorological variables, as well as using classical and depth-based approaches. The results of this application are summarized in Figure 8. We remark that using ROI-Depth method, the same sites are included in the neighborhood obtained by the original end the transformed physio-meteorological variables (Figure 8a). That is because the simplicial similarity used in depth approach is affine invariant. However for the ROI method (Figure 8b and Figure 8c), the sites included in the neighborhood change depending on the used transformation. As a result, the depth-based neighborhood approach has the advantage to be affine invariant and does not require that site characteristics to be normally distributed. Note

that, the same results are obtained using CCA and CCA-Depth methods (results are not presented due to space limitation).

The homogeneity of neighborhood of the target site 97 obtained by the different methods is summarized in Table 3. The neighborhood using depth-based methods can be considered as "possibly homogeneous". However, the neighborhoods obtained using traditional methods are "heterogeneous" (H>2). Table 3 shows that, for the target site 97, the neighborhoods obtained using depth-based method are more homogeneous than those obtained using traditional methods.

In this paper, the depth-based neighborhood approach combined with the log-linear estimation model is compared with the most popular RFA approaches, which is the CCA and the ROI approaches. In order to widen the comparison, it is interested to compare it with different approaches already applied to the studied region. Table 4 summarizes the approaches previously applied to the region of Quebec with those proposed in this study. The results indicate that the CCA-Depth and the ROI-Depth lead to more efficient estimates, especially in term of RRMSE, than those obtained using traditional approaches.

5. Conclusion

In the present paper, a new method is proposed and applied to identify the neighbourhood of a target site in RFA. The proposed depth-based neighbourhood method is based on the statistical notion of depth function. This method employs depth functions to compute the similarity between the target site and the gauged ones. Therefore, aside from leading to the best estimation results, the proposed method allows the delineation step of RFA procedure to be robust, affine invariance and distribution free.

The obtained results from a set of sites in the southern part of Quebec (Canada) show that the neighbourhood of a target site using depth method is more homogeneous than those obtained

using traditional CCA and ROI. Additionally, it leads to more efficient estimates in terms of RB and RRMSE than those obtained using traditional method.

Reference

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

Castellarin, A., D. H. Burn, and A. Brath (2001), Assessing the effectiveness of hydrological similarity measures for flood frequency analysis, *Journal of Hydrology*, *241*(3-4), 270-285.

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

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*.

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), 1-13.

De Michele, C., and R. Rooso (2002), A multi-level approach to flood frequency regionalisation, *Hydrology and Earth System Sciences*, *6*(2), 185-194.

Eaton, B., M. Church, and D. Ham (2002), Scaling and regionalization of flood flows in British Columbia, Canada, *Hydrological Processes*, *16*(16), 3245-3263.

Gaál, L., and J. Kyselý (2009), Comparison of region-of-influence methods for estimating high quantiles of precipitation in a dense dataset in the Czech Republic, *Hydrol. Earth Syst. Sci.*, *13*(11), 2203-2219.

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

Hadi, A. S. (1992), Identifying multiple outliers in multivariate data, *Journal of the Royal Statistical Society. Series B (Methodological)*, 761-771.

Kouider, A., H. Gingras, T. B. M. J. Ouarda, Z. Ristic-Rudolf, and B. Bobée (2002), Analyse fréquentielle locale et régionale et cartographie des crues au Québec,*Rep. R-627-el*, INRS-ETE, Canada.

Leclerc, M., and T. B. Ouarda (2007), Non-stationary regional flood frequency analysis at ungauged sites, *Journal of hydrology*, *343*(3), 254-265.

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*(1-4), 1-9.

Liu, R. Y. (1990), On a Notion of Data Depth Based on Random Simplices, 405-414.

López, Á., and J. Romo (2010), Simplicial similarity and its application to hierarchical clustering, Universidad Carlos III, Departamento de Estadística y Econometría.

Neykov, N. M., P. N. Neytchev, P. H. A. J. M. Van Gelder, and V. K. Todorov (2007), Robust detection of discordant sites in regional frequency analysis, *Water Resources Research*, *43*(6), W06417.

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., K. M. 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*(1–2), 40-58.

Pandey, G. R., and V. T. V. Nguyen (1999), A comparative study of regression based methods in regional flood frequency analysis, *Journal of Hydrology*, *225*(1-2), 92-101.

Ribeiro-Corréa, J., G. S. Cavadias, B. Clément, and J. Rousselle (1995), Identification of hydrological neighborhoods using canonical correlation analysis, *Journal of Hydrology*, *173*(1–4), 71-89.

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*(7).

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*(1-2), 31-43.

Tasker, G. D., S. A. Hodge, and C. S. Barks (1996), Region of influence regression for estimating the 50-year flood at ungaged sites, *Water Resources Bulletin*, *32*(1), 163-170.

Tukey, J. W. (1975), Mathematics and the picturing of data, paper presented at Proceedings of the international congress of mathematicians.

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

Wazneh, H., F. Chebana, and T. B. M. J. Ouarda (2013b), Depth-based regional index-flood model, *Water Resources Research*, *49*(12), 7957-7972.

Wazneh, H., F. Chebana, and T. B. M. J. Ouarda (2014), Delineation of homogeneous regions for regional frequency analysis using statistical depth function, *journal of hydrology, Accepted*.

Zaman, M. A., A. Rahman, and K. Haddad (2012), Regional flood frequency analysis in arid regions: A case study for Australia, *Journal of Hydrology*, *475*, 74-83.

Zuo, Y., and R. Serfling (2000), General notions of statistical depth function, *Annals of Statistics*, 461-482.

Table 1. Quantile estimation results with the various delineation methods and using the MR estimation model.

	Delineation method	Optimal coefficients	QS10		QS100	
			RRMSE(%)	RB(%)	RRMSE(%)	RB(%)
Classical approaches	CCA	$n^{opt.}$=76	44.62	-7.54	51.84	-8.14
	ROI	$n^{opt.}$=35	43.10	***-2.13***	49.09	***-5.13***
Depth-based approaches	CCA-Depth	$n^{opt.}$=35	*37.42*	-2.74	43.87	-5.35
	ROI-Depth	$n^{opt.}$=35	**34.50**	-5.07	**41.92**	-6.50

Best result is in bold character, the second result is in italic.

Table 2. The frequency table of *H* obtained by the different neighborhood methods

	Delimitation methods			
Class intervals	CCA	ROI	CCA-Depth	ROI-Depth
H<1 (Homogeneous)	31	2	56	43
1<H<2 (Possibly homogeneous)	16	6	22	34
H>2 (Heterogeneous)	104	143	73	74

Table 3. Heterogeneity measures *H* of neighborhoods for the target site number 97 obtained by the various delineation methods

Heterogeneity measure	Delineation method			
	CCA	ROI	CCA-Depth	ROI-Depth
H	5.56	8.87	1.43	1.53

Table 4. Quantile estimation result for Quebec with available approaches and their references

Approach	Reference	QS10		QS100	
		RB (%)	RRMSE (%)	RB (%)	RRMSE (%)
Linear regression (LR)	Wazneh et al., [2013a]	-9	55	-11	64
Nonlinear regression (NLR)	Shu and Ouarda [2008]	-9	61	-12	70
NLR with regionalisation approach	Shu and Ouarda [2008]	-19	67	-24	79
CCA	Table 1 above	-7	44	-8	52
ROI	Table 1 above	**-2.1**	43.1	-5.1	49.0
Kriging-CCA space	Chokmani and Ouarda [2004]	-20	66	-27	86
Kriging-Principal Component Analysis space	Chokmani and Ouarda [2004]	-16	51	-23	70
Adaptive Neuro-Fuzzy Inference Systems (ANFIS)	Shu and Ouarda [2008]	-8	57	-14	64
Artificial Neural Networks (ANN)	Shu and Ouarda [2008]	-8	53	-10	60
Single ANN-CCA (SANN-CCA)	Shu and Ouarda [2007]	-5	38	*-4*	46
Ensemble ANN (EANN)	Shu and Ouarda [2007]	-7	44	-10	60
Ensemble ANN-CCA (EANN-CCA)	Shu and Ouarda [2007]	-5	37	-6	45
CCA-Depth	Table 1 above	*-2.7*	*37.4*	-5.3	*43.8*
ROI-Depth	Table 1 above	-5.0	**34.50**	-6.5	**41.9**

Best result is in bold character, the second result is in italic

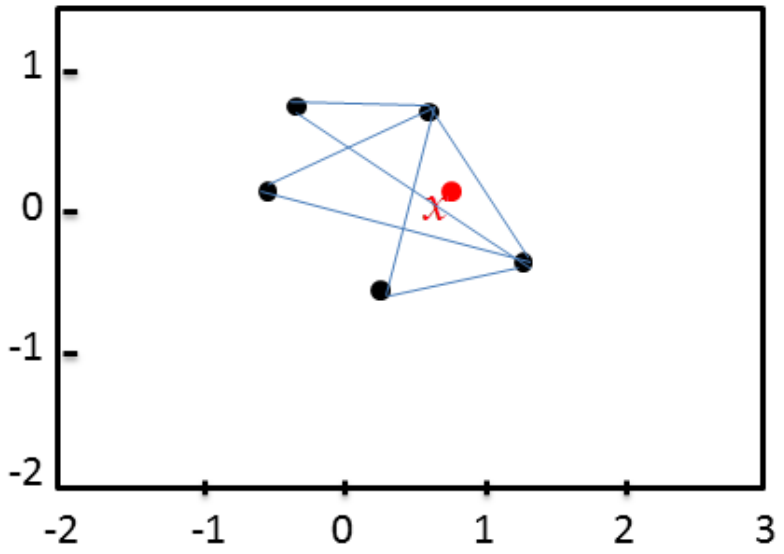


Figure 1. Illustration of the way that simplicial depth of point x in R^2 is computed.

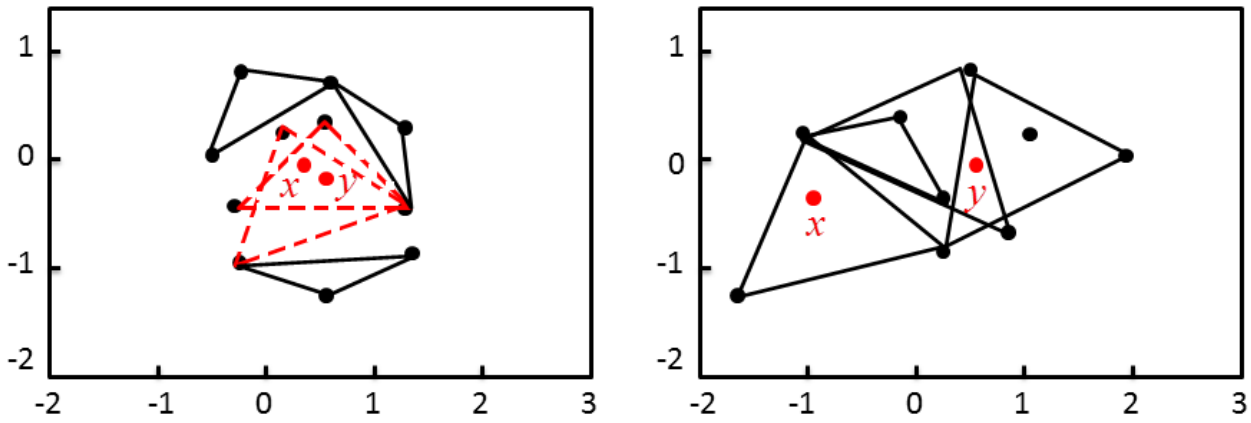


Figure 2. Examples of random simplices for close and distant points.

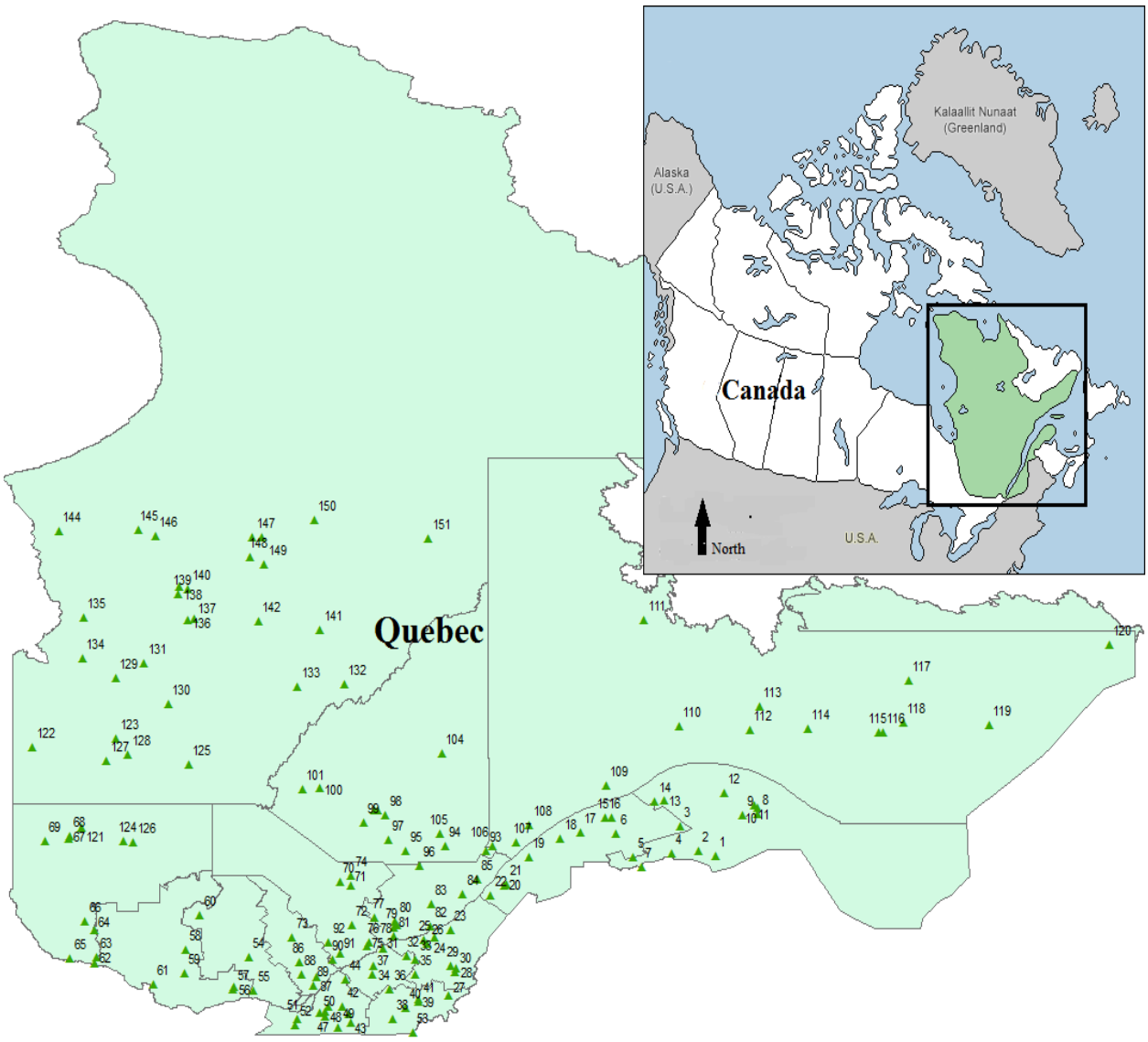


Figure 3. Geographical location of the studied sites in the southern part of the province of Quebec, Canada.

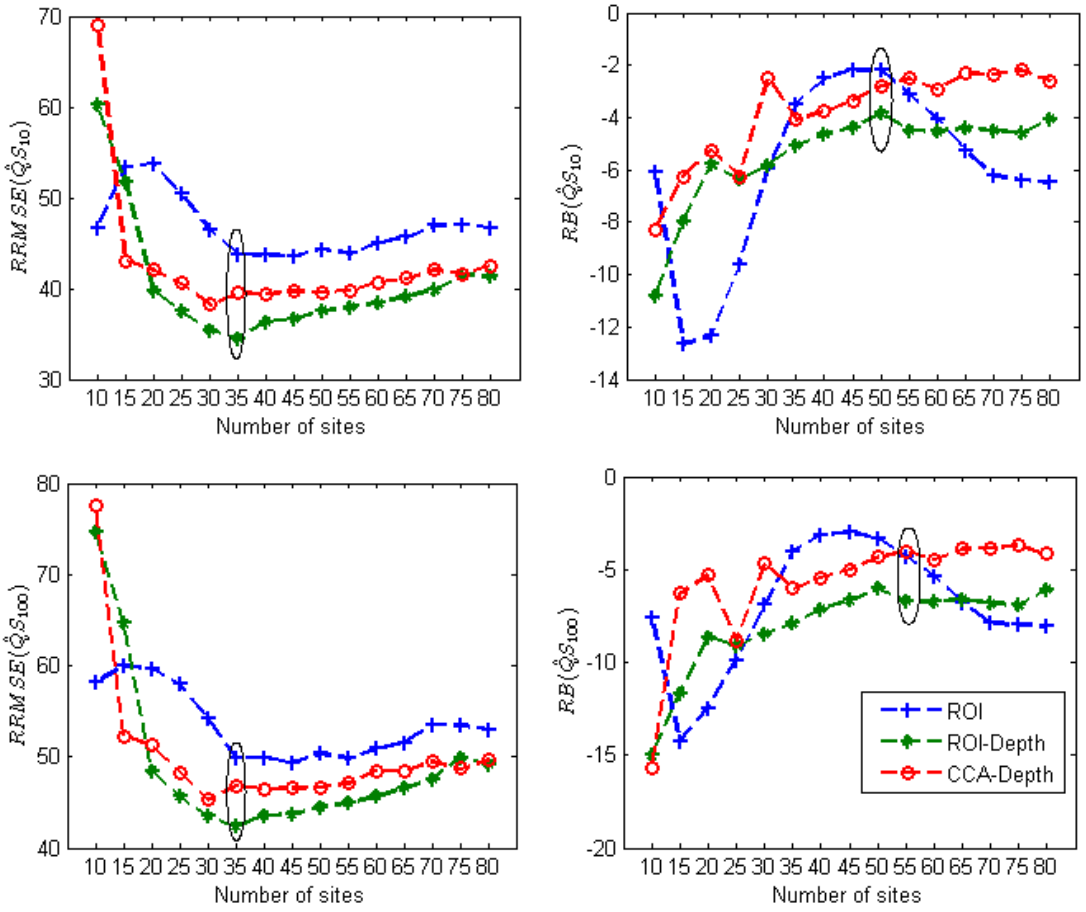


Figure 4. Variation of criteria RB and RRMSE obtained by the various neighborhood methods as function of the number of site in the neighborhood. For the CCA method, as shown in a number of studies, the optimal number of sites with respect RB and RRMSE criteria is 76.

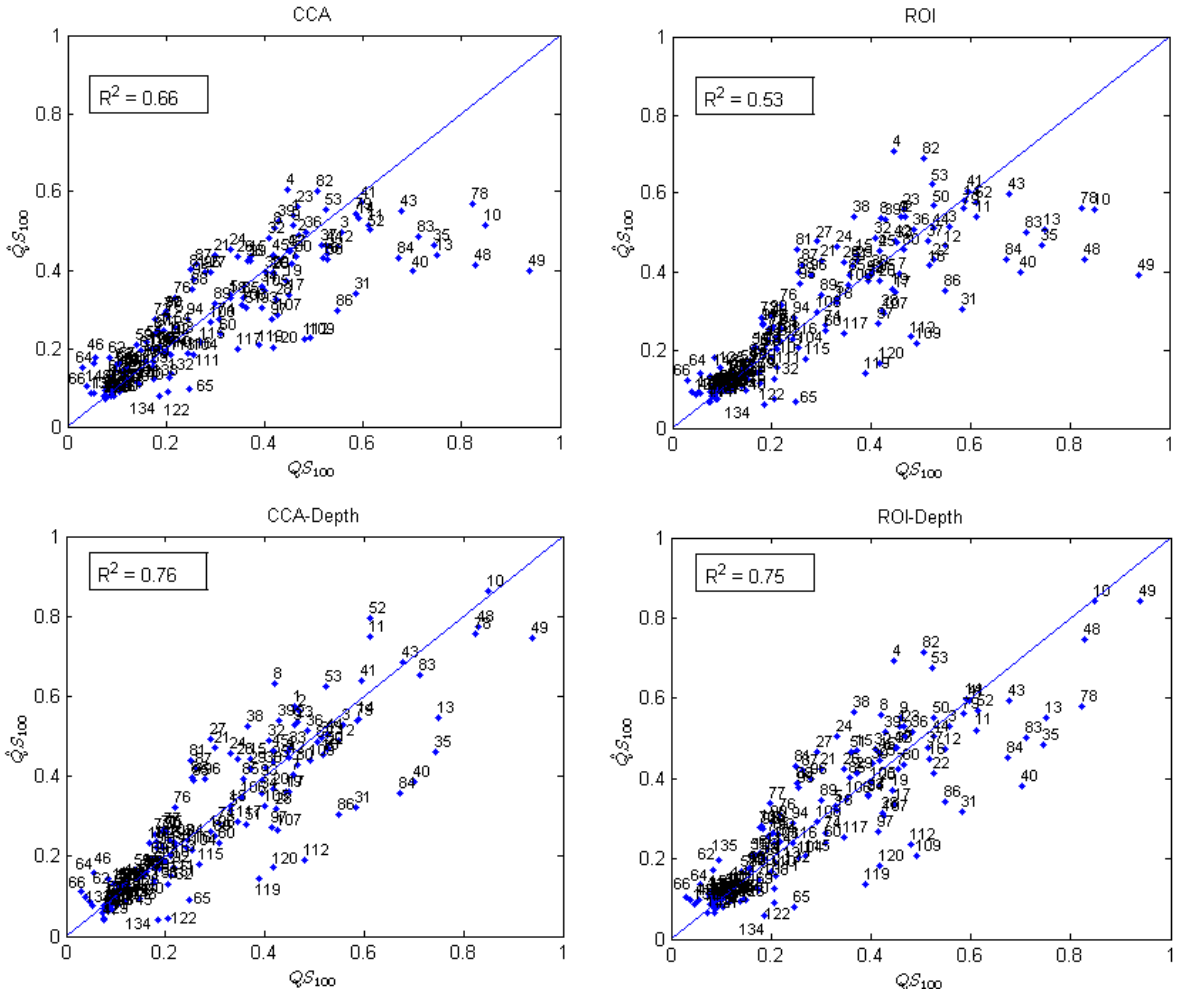


Figure 5. Scatter diagram of at-site quantiles (specific) versus the regional estimates ones for the 100 yr return period. R^2 represent the coefficient of determination.

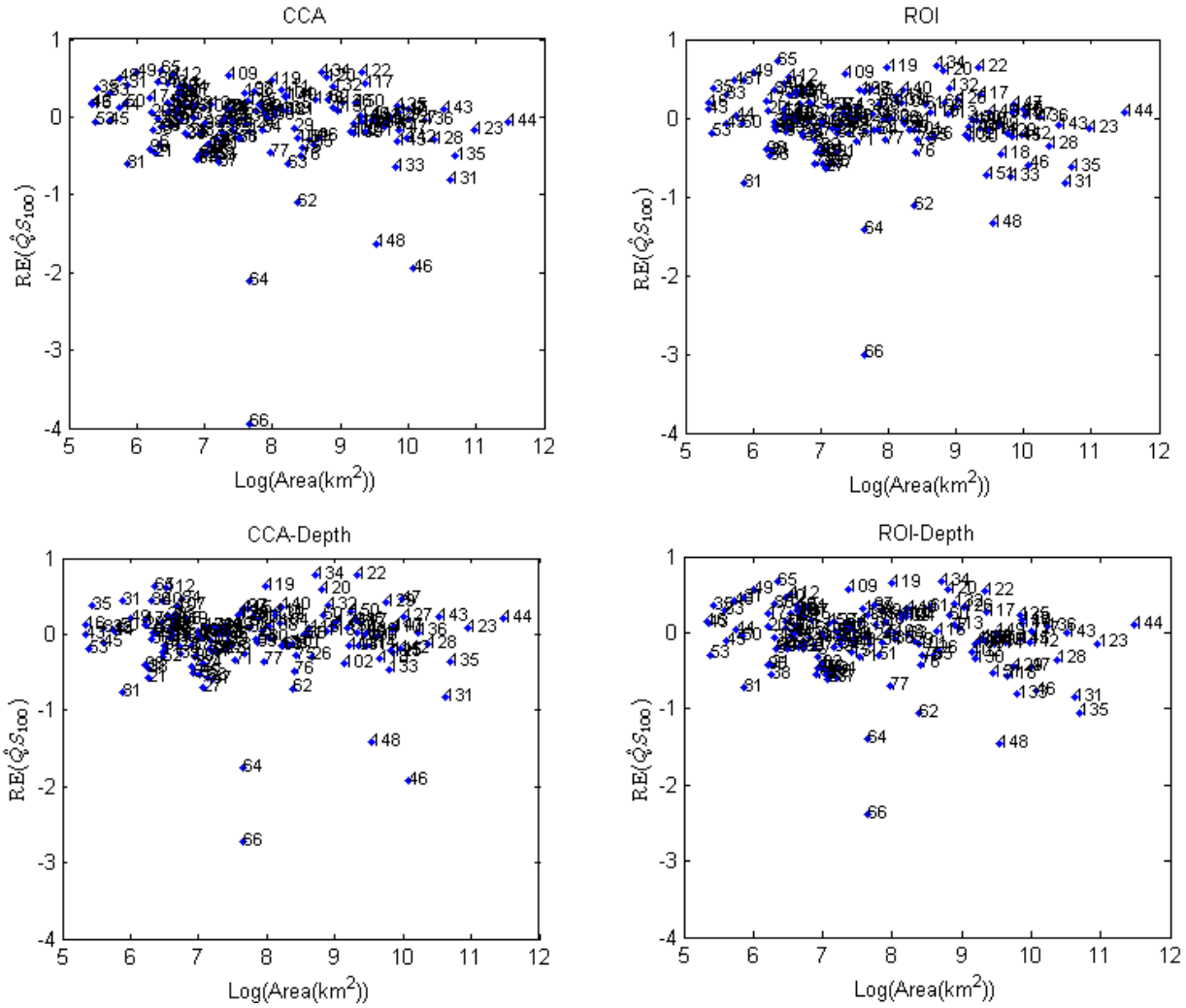


Figure 6. Relative quantile errors for the 100 yr return period using various neighborhood methods

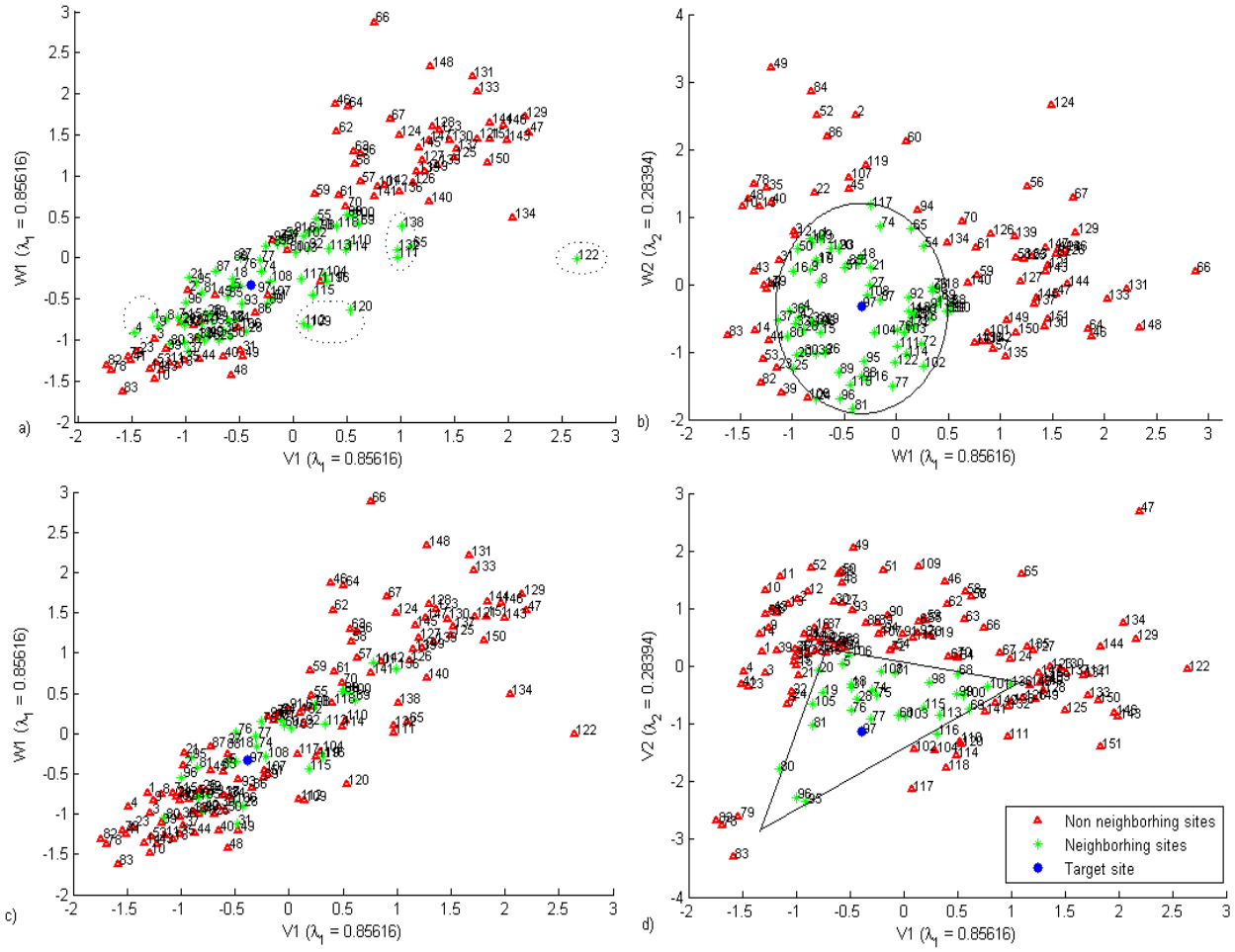


Figure 7. The neighborhood of the target site number 97 using (a) CCA and (b) CCA-Depth methods.

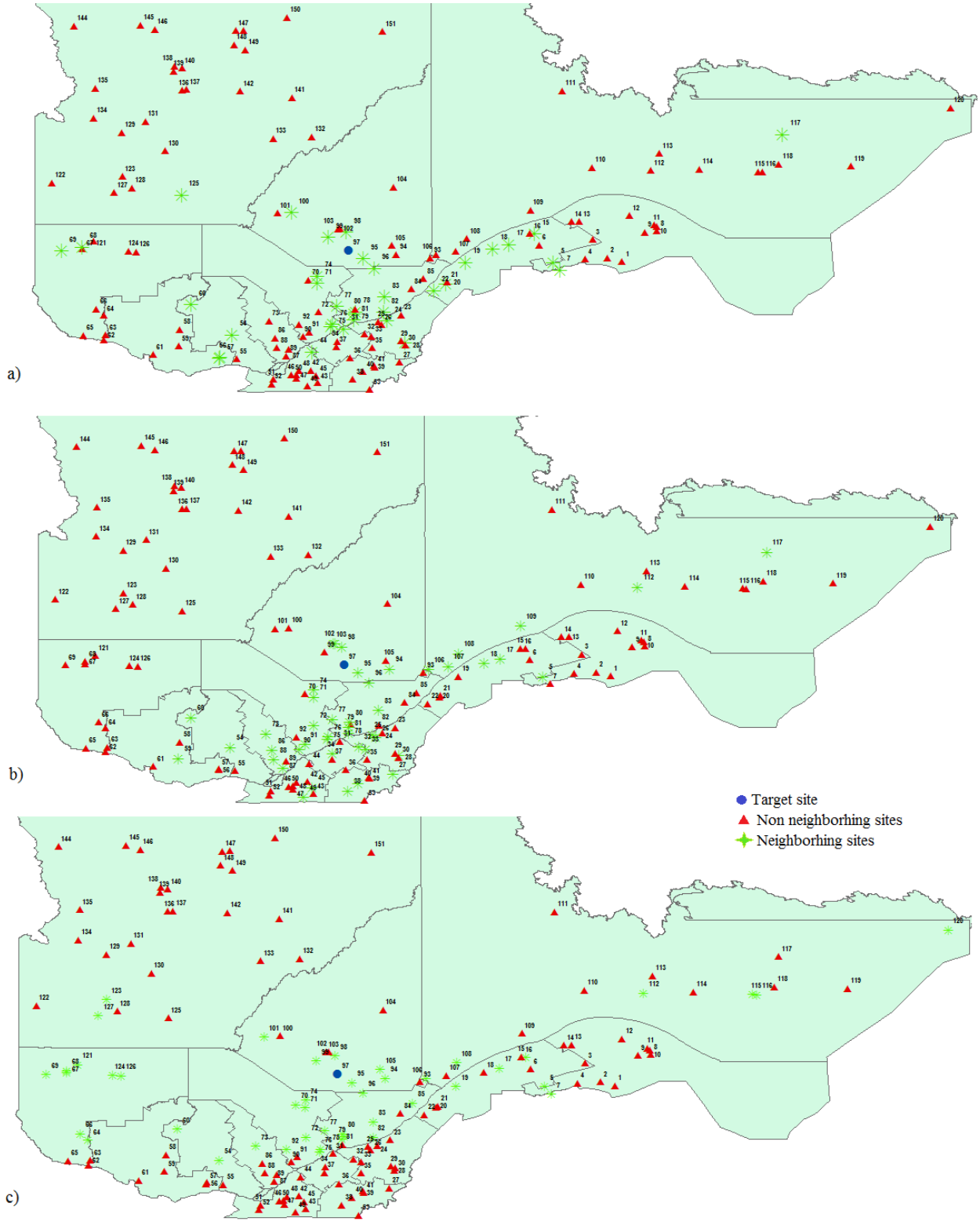


Figure 8. The geographical location of sites in the neighbourhood of the target site 97 using a) ROI-Depth with original or normalized physiographical variables b) ROI with normalized variables and c) ROI with original variables**.**

CHAPITRE 3: DÉLIMITATION DES RÉGIONS NON CONTIGUËS POUR L'ANALYSE FRÉQUENTIELLE RÉGIONALE EN UTILISANT LES FONCTIONS DE PROFONDEUR

Delineation of homogeneous regions for regional frequency analysis using statistical depth function

H. Wazneh*[1], F. Chebana[1] and T.B.M.J. Ouarda[2]

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

*__Corresponding author:__ Tel: +1 (418) 654 2530#4468
Email: hussein.wazneh@ete.inrs.ca

Accepted 24th November 2014

(Journal of Hydrology)

Abstract

The aim of regional frequency analysis (RFA) is to estimate extreme hydrological events at sites where little or no hydrological data are available. The delineation of sub-regions and the regional estimation within these sub-regions are the two main steps of RFA. As currently practiced, the delineation step is unrobust and subjective. To overcome these limitations, the present paper aims to propose a new robust approach for delineating homogeneous sub-regions. The proposed approach is objective and based on the concept of depth function. A data set from three geographical regions in the North-West of Italy is used to apply and compare the proposed approach with a traditional one. Results indicate that the proposed depth-based approach leads to more homogeneous sub-regions in terms of *H* heterogeneity measure, and leading to more efficient quantile estimations in terms of relative bias and relative root mean square error, than those obtained by the traditional approach.

Keywords: Ungauged basins, Homogeneous sub-region, Regional frequency analysis, Statistical depth function, Clustering analysis.

1. Introduction

Hydrological data records are often short and are not always available in the desired site. Consequently, at-site frequency analysis is not always accurate or even possible at the sites of interest. Regional frequency analysis (RFA) is used to estimate extreme hydrological events at sites where little or no hydrological data are available [e.g., Basu and Srinivas, 2014; Haddad et al., 2014; Ouarda et al., 2000; Reed et al., 1999]. Transfer of the available information from gauged sites, within a homogeneous region, to the target (ungauged) site is the main idea behind the RFA. The two main steps of RFA are (i) the identification of homogeneous hydrological sub-regions and (ii) the regional estimation within these sub-regions [e.g., Wazneh et al., 2013b].

The identification of sub-regions has received important consideration in hydrology, but no common methodology has been developed. Thus, various methods of defining sub-regions can be found in the literature, leading to geographically contiguous regions, geographically non-contiguous regions, or hydrological neighborhoods [Ouarda et al., 2001]. Using non-contiguous regions was recommended in the literature [e.g., Haddad and Rahman, 2012; Ouarda et al., 2008]. Cluster analysis based on site characteristics is one of the most practical methods used to define the non-contiguous regions [Hosking and Wallis, 1997].

Cluster analysis groups sites based on a distance (measure) reflecting the similarity among a set of attributes of the gauging site (e.g., basin area, latitude) [Rao and Srinivas, 2006a]. The identified regions highly depend on the choice of the set of attributes [Castellarin et al., 2001; Oudin et al., 2010]. Generally, the set of attributes used for RFA approaches includes: (i) physiographic catchment characteristics such as drainage area, average basin slope, main stream slope, stream length [e.g., Acreman and Sinclair, 1986]; (ii) geographical location attributes such as latitude, longitude and altitude of catchment centroid [e.g., Burn and Goel, 2000]; (iii)

measures of basin response time such as basin lag or time-to-peak [e.g., Potter and Faulkner, 1987]; (iv) meteorological factors such as storm direction, mean annual rainfall, precipitation intensities [e.g., Chebana and Ouarda, 2008]; and (v) at-site flood statistics such as L-moments or other statistical measures calculated from the available flow series [e.g., Wazneh et al., 2013a]. A combination of two or more of the above variables may also constitute an attribute in a cluster analysis [e.g., Bargaoui et al., 1998; Nathan and McMahon, 1990].

Several clustering algorithms are available in the statistical literature [Johnson and Wichern, 2002]. Clustering algorithms used for the delineation of sub-regions in RFA can be broadly classified into two categories: hierarchical and partitional clustering [Rao and Srinivas, 2006b]. The hierarchical category includes single linkage, complete linkage, average linkage and Ward method [e.g., Baeriswyl and Rebetez, 1997; Bhaskar and O'Connor, 1989; Chiang et al., 2002]. The partitional category includes k-means and fuzzy c-means [e.g., Bargaoui et al., 1998; Rao and Srinivas, 2003]. Ward hierarchical method is the most commonly used in RFA. In fact, this method tends to delineate sub-regions approximately equivalent in size, and is thus considered more convenient in the context of regionalizing flood data [Hosking and Wallis, 1997].

In the context of RFA, the application of the above clustering approaches in the delineation step faces two drawbacks. First, these approaches are based on distance measures (e.g., Ward or linkage) and/or use non robust statistics (e.g., k-means), and making the delineation results sensitive to noise and to outliers [Ilorme and Griffis, 2013; Jörnsten, 2004]. Second, some of these approaches require a preselection of the number of sub-regions (e.g., k-means), which makes the delineation step subjective and depends on the user choice.

The aim of the present work is to identify sub-regions for RFA with a particular focus on the formation of sub-regions that can be used for estimating extreme flow quantiles for ungauged

sites. More precisely, in this study, a new robust approach for delineation of hydrological sub-regions based on the notion of data depth [Tukey, 1975] is presented and applied. The proposed approach uses, as a starting step, a traditional approach (such as Ward method) to form initial sub-regions. Then, the sites of the initial sub-regions are redistributed in a manner that maximizes their depth values (see section 3). The proposed approach determines objectively the number of homogeneous sub-regions using a preselected criterion such as the *H* heterogeneity measure [Hosking and Wallis, 1997].

Wazneh el al., [2013a] introduced statistical depth function in regional estimation through the index-flood model. To delineate the homogeneous sub-regions, they used traditional Ward method. They showed that employing depth function in the estimation step provides improved results. However, as argued by the authors, it is more appropriate to consider depth function in the delineation step as well as in the estimation which ensures compatibility between the two main RFA steps (delineation and estimation) and could lead to better results. Therefore, it is necessary to develop a depth-based regional delineation.

Several depth functions are available in the literature and have been used for generalizing many univariate statistical methods to the multivariate set-up [e.g., Chen et al., 2009; Donoho and Gasko]. In particular, Jörnsten [2004] introduced the notion of depth functions in the clustering approach. In this study, the author defined the maximum depth clustering, where an observation is assigned to the cluster within it has the maximum depth value. From a simulation study, the author shows that the depth clustering approach is robust to noise and outliers. Jörnsten shows that employing depth functions improves clustering accuracy. After Jörnsten [2004], statistical depth functions are employed in a number of clustering and classification approaches [e.g., Dutta and Ghosh, 2012; Ghosh and Chaudhuri, 2005].

This paper is organized as follows: section 2 assembles the various background elements needed to introduce the proposed approach and compare it with the traditional Ward approach. The proposed approach in its general form is described in Section 3. Its application in a real world data set is presented in Section 4. The last section is devoted to the conclusions of this work.

2. Background

This section briefly presents the background material required to introduce and apply the delineation of sub-regions using depth functions. In addition, this section includes several basic concepts as well as a summary of the main steps of the index-flood model used for regional estimation model in order to quantify the performance of the proposed approach.

2.1. Ward method

In this section, the Ward method, commonly used to delineate homogeneous sub-regions in RFA analysis, is presented. This method is considered in this paper for comparison purposes.

Ward method [Ward Jr, 1963] is a hierarchical algorithm which initially begins with each site serving as its own sub-region. The algorithm successively merges sub-regions using an analysis of variance approach in which the similarity among sites in a sub-region is measured in terms of the Error Sum of Squares (*ESS*). More formally, for a sub-region *r* containing *s* sites, where the flood regime is represented by *p* attributes $X = \left(X_1, X_2, .., X_p\right)$, the *ESS* is given by:

$$ESS_r = \sum_{j=1}^{s} \left(X_j - \bar{X}_r\right)' \left(X_j - \bar{X}_r\right) \quad (1)$$

where $X_j = \left(X_{j1}, X_{j2}, .., X_{jp}\right)$ is a vector of the attributes at site *j*, and $\bar{X}_r$ is a vector of the means of the attributes within the sub-region *r*. At each step, ESS_r is computed for the hypothetical merger of any two sub-regions, and the actual mergers chosen to occur are those

which minimize the increase in the total *ESS* across all sub-regions. A dendrogram is commonly used to illustrate the mergers made at successive levels, where the vertical axis represents the value of the *ESS*.

2.2. Spatial depth function

Data depth is a quantitative measure of how central (or deep) a point is with respect to a data set or a distribution in a multivariate framework. This gives us a central outward ordering of multivariate data points and gives rise to new ways to quantify the many complex multivariate features of the underlying multivariate distribution [Li et al., 2012; Liu et al., 1999]. The depth functions were first introduced by Tukey [1975]. They are employed in many research fields including water sciences [e.g., Bárdossy and Singh, 2008; Bárdossy and Singh, 2011; Chebana and Ouarda, 2008; Kraußе and Cullmann, 2012; Kraußе et al., 2012]. For a given cumulative distribution function F on $\Re^d$ $(d \geq 1)$, a depth function is any non-negative bounded function which has a number of convenient properties. These properties fit well the RFA requirements and constraints [Chebana and Ouarda, 2011].

Several types of depth functions have been developed e.g., half space, projection, simplicial, spatial and Mahalanobis depth functions. Several depth functions, practical in two dimensions, become impractical where the dimensionality increases e.g., half space [Hugg et al., 2006]. The spatial depth function that was used in this study because of its convenient properties. First, spatial depth is non-zero outside the convex hull of the data set and therefore can be used to clustering sites of the region (see section 3.1). Second, it is fast and easy to compute in any dimension, contributing to its application for the study of large high-dimensional data sets, such clustering studies.

The spatial depth of point *x* is the amount of probability mass needed at *x* to make it the multivariate median (spatial median) of the data. Formally, the spatial depth of point *x* on $\Re^d$ $(d \geq 1)$ is:

$$SPD(x, F) = 1 - \left\| E_F \left[S(x - X) \right] \right\| \tag{2}$$

where *X* is a random variable with a distribution *F*, $S(u) = \frac{u}{\|u\|}$ is the spatial sign function where $S(0) = 0$ and $\|.\|$ is the Euclidean norm. The empirical version of the *SPD* of *x* with respect to the sample $X = \{x_1, ..., x_n;\ x_i \text{ in } \Re^d\}$ is defined by replacing *F* by a suitable empirical function $\hat{F}_X$. Therefore, the empirical version is:

$$SPD\left(x, \hat{F}_X\right) = 1 - \frac{1}{n} \left\| \sum_{i=1}^{n} \frac{x - x_i}{\|x - x_i\|} \right\|, \quad x \in \Re^d \tag{3}$$

In the following, the notation $SPD\left(x, \hat{F}_X\right)$ is replaced by $SPD(x, X)$. Figure 1 illustrates, on a simple example, the way that the spatial depth is computed in the bivariate case (i.e. $d = 2$) of y_1 and y_2 respectively with respect to a sample $X = \{x_1, ., x_7\}$. Let $e_i^s = S(y_s - x_i) = \frac{y_s - x_i}{\|y_s - x_i\|}$ be the unit vector from y_s to x_i $(s = 1, 2 \text{ and } i = 1, .., 7)$. For $s = 1$, y_1 is located deep inside the cloud of *x*'s, summing up e_i^1 will result in a vector with a norm close to 0, since unit vectors have different directions and they cancel each other out. Therefore, the spatial depth of y_1 is approaching 1. For *s*=2, y_2 is outside the data cloud, the sum of e_i^2 has a large norm, thus the depth is approaching 0. The point where the spatial depth reaches its maximum value 1 is called the spatial median. The spatial median represents the geometric center of the data.

2.3. Index-flood model

In RFA homogeneous sub-regions are used to estimate $Q_{i_0}(t)$ the flood quantiles of an ungauged site i_0 corresponding to a nonexceedance probability t. In this study, the index-flood model [Chebana and Ouarda, 2009; Hosking and Wallis, 1997] is used to estimate flood quantiles $Q(t)$. The main steps of the index-flood model are briefly described in the following:

Compute the sample L-moments

For a random variable X, the $(r+1)^{\text{th}}$ L-moment is defined by:

$$\lambda_{r+1} = \sum_{v=0}^{r} p^*_{r,v}\beta_v \quad \text{where} \quad p^*_{r,v} = (-1)^{r-v}\binom{r}{v}\binom{r+v}{v} \quad \text{and} \quad \beta_v = E\left\{X\left[F(X)\right]^v\right\} \tag{4}$$

where F is the cumulative distribution function of X. In the index-flood model, L-moment ratios (LMRs) are employed where L_{cv} is equal to $\tau = \lambda_2/\lambda_1$ while the other LMRs are given by:

$$\tau_r = \frac{\lambda_r}{\lambda_2} \; ; r = 3,4 \; (L_{skew} \text{ for } r = 3 \text{ and } L_{kur} \text{ for } r = 4) \tag{5}$$

Initial screening of data

For a data set from N sites, the discordancy measure D_i for site i is defined by:

$$D_i = \frac{N}{3}(u_i - \bar{u})^{'} S^{-1}(u_i - \bar{u}) \tag{6}$$

where $u_i = \left[L^i_{cv}\ L^i_{skew}\ L^i_{kur}\right]'$, $\bar{u} = \frac{1}{N}\sum_{i=1}^{N} u_i$ and $S = \sum_{i=1}^{N}(u_i - \bar{u})(u_i - \bar{u})'$. A site i is considered discordant if D_i is larger than 3, but it is advised to examine sites with the largest D_i values.

Heterogeneity measure

The homogeneity of a region is measured by the following H statistic [Hosking and Wallis, 1997]:

$$H = \frac{\left(V - \mu_V\right)}{\sigma_V} \text{ such that } V = \frac{\sum_{i=1}^{N} n_i \left(L_{cv}^i - \bar{L}_{cv}\right)^2}{\sum_{i=1}^{N} n_i} \quad (7)$$

where n_i is the sample size at site *i, N is* the number of sites within the region, L_{cv}^i and $\bar{L}_{cv}$ are the L_{cv} at site *i* and the average regional one respectively, and μ_V and σ_V are the mean and the standard deviation of the simulated regions respectively. Regions are considered as "acceptably homogenous" if $H < 1$, "possibly homogenous" if $1 \le H < 2$ and "definitely heterogeneous" if $H \ge 2$.

Regional growth curve

This step consists in selecting the appropriate distribution (growth curve) corresponding to each identified homogeneous sub-region. To this end, for each homogeneous sub-region, a set of candidate distributions are fitted. A candidate distribution is selected if the following goodness-of-fit statistic $\left|Z^{DIST}\right|$ is less than the test threshold 1.64 which corresponds to the 90% normal quantile [Hosking and Wallis 1997]:

$$Z^{DIST} = \frac{\left(\bar{L}_{kur} - L_{kur}^{DIST}\right)}{\sigma_{kur}} \quad (8)$$

where L_{kur}^{DIST} is the L_{kur} of the candidate distribution, $\bar{L}_{kur}$ is the regional average and σ_{kur} is the standard variation obtained from appropriately simulated regions.

Estimate flood quantiles at the target-site

For a given homogenous sub-region, the index-flood model estimates the flood quantile corresponding to a nonexceedance probability *t*, i.e. return period $T = 1/(1-t)$, through the expression:

$$Q_{i_0}(t) = \mu_{i_0} q_t\left(\theta^R\right) \quad ; 0 < t < 1 \text{ and } \hat{\theta}_l^R = \frac{\sum_{h=1}^{N'} n_h \hat{\theta}_l^h}{\sum_{h=1}^{N'} n_h}, \quad l = 1,..L \tag{9}$$

where i_0 is the identifier of the target site, μ_{i_0} is the scale factor called the index flood, $q_t(.)$ is the growth curve function, $\hat{\theta}_l^h$ is the l^{th} parameter obtained from the at-site standardized distribution at the h^{th} gauged site, θ^R is the vector with L components of the regional growth curve parameters and N' is the number of sites in the homogenous sub-region that contain the site i_0. Note that, for gauged sites, the index flood μ_{i_0} is estimated by a location measure such as the mean or the median. In this study, we consider the sample mean of the observed data (sample mean) for each gauged site. However, in the case of an ungauged site, μ_{i_0} is estimated using a multi-regressive model that relates the index flood to the catchment characteristics [Viglione et al., 2007].

The weight given in (9) to estimate the regional growth curve parameters θ^R is based solely on the record lengths. However, recently Wazneh el al., [2013a] proposed a new iterative weighting scheme to the index-flood model where the vector of regional growth curve parameters θ^R is estimated by:

$$\left(\hat{\theta}^R\right)_D = \sum_{h=1}^{N'} \omega_h \theta_l^h, \quad l = 1,..., L \text{ and } \omega_h = \varphi\left[MHD\left(\hat{\theta}^h; \Theta\right)\right]; \; h = 1,..., N' \tag{10}$$

where φ is an increasing weight function, *MHD* is the Mahalanobis depth function and Θ is the set formed by the N' vectors of at-site parameters of N' gauged sites in the homogeneous sub-region. This approach is also considered in this study for a comparison proposes and it is called

Depth-based index-flood model. A detailed presentation of Depth-based index-flood model can be founded in [Wazneh et al., 2013a].

3. Approach development

In this section the proposed approach in its general form is described. An algorithm is presented to facilitate the approach presentation.

The aim of the proposed approach is to delineate the non-homogeneous region into *K* homogeneous regions (*K* is unknown) using the spatial depth function. This approach will be denoted by D-clustering, for Depth clustering, in the remainder of the paper. If the whole region is homogeneous, there is no need to use the proposed approach. In order to find homogeneous sub-regions using D-clustering, the following three main steps are used:

- Initialization: use a traditional clustering approach to form the initial *K* sub-regions. Since the aim of the proposed approach is to delineate homogeneous sub-regions from the whole region which is assumed non-homogeneous, then *K* must be greater than 1 (in this study we start with *K=2* see Computation Algorithm below);
- Modifying: Modify the initial position of sites into the sub-regions using the depth function;
- Homogeneity: Test the homogeneity of the final sub-regions obtained after the modification step. If the obtained sub-regions are not homogeneous go to the initial step with *K=K+1*.

The description of each step and the computation algorithm are given below.

3.1. Description of the various steps of the approach

Assume that *N* is the number of gauged sites in the data set after removing discordant sites, $Y_r = \left(Y_{r1}, ..., Y_{rn_r}\right), r = 1, .., N;$ is the at-site hydrological data at site *r* with record length n_r (e.g., the annual maximum peak flood) and $X_r = \left(X_{r1}, \ldots, X_{rp}\right)$ are the *p* selected attributes that

represent the flood regime for this site i.e. the clustering variables such as the catchment centroids or the basin area. In the first step of the approach, we use a traditional method (e.g., Ward method) to form the initial sub-regions. By construction the D-clustering approach must started with *K=2* initial sub-regions. The initial sub-regions are not necessarily homogeneous. Let $i(k)$ be the set formed by the Id number of sites in the sub-region *k*, $k=1,..,K;$ i.e. $i(k)=\{\text{site } r \in \text{sub-region } k; r=1,..,N\}$, and let $I(k)$ be the set formed by the values of attributes of sites in the sub-region *k* i.e. $I(k)=\{X_r; r \in i(k); r=1,..,N\}$.

In the second step of the D-clustering approach, we check and modify the initial sub-regions obtained in the first step. In fact, in this step, each site in the data set is assigned to the sub-region within it has the maximum depth value [Jörnsten, 2004]. To this end, the spatial depth function (3) is used. More precisely, the spatial depth of the site characteristics X_r, $r \in i(k)$, is computed with respect to the initial sub-region $I(k)$ to which it belongs. In the context of this paper, the depth of X_r with respect to $I(k)$ is denoted as the within sub-region depth $D_r^w = SPD(X_r, I(k))$. Note that D_r^w quantifies how central the site *r* is with respect to its sub-region. Then, we compute the between sub-regions depth of site *r*, D_r^b, defined as the maximum depth of X_r with respect to the sub-region to which it does not belong i.e. $D_r^b = \max_{\substack{l=1,..,K \\ l \neq k}} \left[SPD(X_r, I(l)) \right]$. From the definition of D_r^w and D_r^b, we remark that site *r* is well classified in its sub-regions *k* if $D_r^w > D_r^b$. Therefore, for each site *r* in the data set, we define the deviance depth DeD_r as the difference between the corresponding within and between depths i.e. $\text{DeD}_r = D_r^w - D_r^b$. As a result, in the second step of the D-clustering approach, each site *r* is reassigned to the sub-region that maximizes DeD_r.

The DeD_r ranges between -1 and 1 since D_r^w and D_r^b ranges between 0 and 1. A site *r* is well classified in the sub-region if DeD_r is positive, a small DeD_r (around 0) means that the site *r* lies between two sub-regions, and a negative DeD_r means that site *r* is placed in the wrong sub-region. Note that the deviance depth DeD_r defined in this study is similar to the silhouette criterion [Rousseeuw, 1987] commonly used in the traditional clustering algorithm (defined below). An advantage of DeD is its independent of the scale of clustering variables (because the spatial depth function is affine invariant) which is not the case of silhouette [Ding et al., 2007; Jörnsten, 2004].

For more clarity of the second step of the D-clustering approach, consider as an example the case of *p=2* with variables the basin area (AREA) and the mean elevation (H_m) where for each site *r* $X_r = \left(AREA_r, \left(H_m \right)_r \right)$. Suppose also that the dataset is clustered into three initial sub-regions, and $I(k)$, $k = 1,\ldots,3$ are the corresponding attribute sets (see Figure 2). The within depths of sites 1 and 2 are respectively $D_1^w = SPD\left(X_1, I\left(1\right) \right)$ and $D_2^w = SPD\left(X_2, I\left(1\right) \right)$ because $X_1, X_2 \in I(1)$. Assume that $SPD\left(X_1, I\left(2\right) \right) > SPD\left(X_1, I\left(1\right) \right) > SPD\left(X_1, I\left(3\right) \right)$ and $SPD\left(X_2, I\left(3\right) \right) > SPD\left(X_2, I\left(1\right) \right) > SPD\left(X_2, I\left(2\right) \right)$. In this case the between depth of site 1 and 2 are respectively $D_1^b = SPD\left(X_1, I\left(2\right) \right)$ and $D_2^b = SPD\left(X_2, I\left(3\right) \right)$. Based on the previous assumptions, in the second step of the D-clustering algorithm, these two sites must modify their sub-regions. In fact, site 1 must be assigned to sub-region 2 and site 2 must be assigned to sub-region 3.

After modifying the initial position of sites according to their DeD values, in the last step of the D-clustering approach, we test the homogeneity of the final formed sub-regions. If the obtained sub-regions are not homogeneous, we repeat the first two steps with $K = K + 1$.

3.2. Computation Algorithm

To identify homogeneous sub-regions using the D-clustering approach, we propose to use the following algorithm where the main steps are summarized in Figure 3.

1. Form the initial *K* sub-regions. This is usually achieved by partitioning the gauged sites of the data set into disjoint groups using traditional clustering approach. Let $i(k) = \{\text{site } r \in \text{sub-region } k;\ k = 1,..,K\}$ and $I(k) = \{X_r; r \in i(k)\}$.
2. Modify the initial position of sites into the sub-regions. This is achieved by:
   2.1. Computing D_r^w, D_r^b and DeD_r of X_r with respect to $I(1),...,I(K)$, for $r = 1,..,N$ (i.e. for all sites in the data set);
   2.2. Identifying the set of sites *S* that are placed in the wrong sub-regions, $S = \{\text{site } r : \text{DeD}_r < 0\}$;
   2.3. Reassigning each site in *S* to the sub-region that maximizes its DeD. The new sub-regions are $\tilde{i}(1),...,\tilde{i}(K)$;
3. Test the homogeneity of the new formed sub-regions $\tilde{i}(1),...,\tilde{i}(K)$, using a commonly used homogeneity test.
   3.1.If the sub-regions are not homogeneous, go to step 1 with $K = K + 1$.
   3.2.If the sub-regions are homogeneous, end the algorithm;

The proposed algorithm requires the specification of an initial number *K* of sub-regions (step 1). By construction, and to be sure that the number of sites in the final sub-regions is large enough to

ensure the applicability of a regional estimation, this algorithm must be started with $K = 2$ as the minimum possible number of sub-regions. Then, the optimal *K* is determined by the algorithm based on the preselected homogeneity criterion (step 3).

A possible option to form the initial sub-regions (step 1) is to use one of the traditional clustering approaches commonly used in RFA (e.g., k-means or Ward). In this study, we use Ward method presented above in section 2.1.

The assessment of regional homogeneity is a critical point in this approach. Many homogeneity tests have been proposed in the hydrological literature, including Lu and Stedinger [1992], Hosking and Wallis [1993; 1997] and Viglione et al., [2007b]. In this study, the *H* heterogeneity measure (7) is used to evaluate the degree of heterogeneity in the sub-regions and to assess whether the sites might reasonably be treated as homogeneous regions. This homogeneity measure is the most commonly used in RFA and especially in the index-flood model framework. Note that, this measure does not affect the validity of the proposed approach since the procedure is general and any other heterogeneity measure can be considered.

If a particular region is subdivided several times without improving the statistical homogeneity for the obtained sub-regions i.e. $H>2$ for at least one of the formed sub-regions, then the region can be considered heterogeneous and should be kept without sub-division. In this case, the index-flood estimation model is not appropriate since this model is based on the homogeneity of the sub-regions. As an alternative, one can consider the multiple regression model where the homogeneity assumption is not required [Chebana and Ouarda 2008].

3.3. Performance criteria

One of the objectives of this study is to assess and compare the performance of the Ward and the D-clustering approach. Since the delineation step affects the results of estimation of flood

quantiles, two categories of performance criteria are used in this study: i) criteria that assess the clustering approaches and ii) criteria that assess the estimation results.

To assess and compare the clustering approaches, the *H* heterogeneity measure (7) and the Silhouette coefficient (defined below) are used. As mentioned in Hosking and Wallis [1997], *H* quantifies the degree of heterogeneity of a given sub-region. Therefore, the best clustering approach will be the one that leads to sub-regions with low *H* values. The notion of silhouette as a measure of clustering quality was first introduced by Rousseeuw [1987]. For each site *r* in the sub-region *k* i.e. $r \in i(k)$, let a_r be the average Euclidean distance between clustering variables of site *r* and clustering variables of sites in its sub-region i.e. $a_r = \delta\left(r, i\left(k\right)\right)$ where $\delta\left(r, i\left(k\right)\right) = \frac{1}{\#i\left(k\right)} \sum_{j \in i(k)} d\left(X_r, X_j\right)$ and *d*(.,.) denotes the Euclidean distance; and b_r be the average distance between clustering variables of site *r* and clustering variables of sites in its nearest neighboring sub-region i.e. $b_r = \min_{\substack{l=1,..K \\ l \neq k}} \left[\delta\left(r, i\left(l\right)\right)\right]$. The silhouette of site *r* then defined as:

$$sil_r = \frac{b_r - a_r}{\max\left\{a_r, b_r\right\}} \tag{11}$$

sil_r ranges between -1 and 1. A high sil_r indicates that site *r* is well-matched to its own sub-region, and poorly-matched to neighboring sub-regions. The silhouette of sub-region *k* (SIL_k) is defined as the average silhouette over its member sites i.e.:

$$SIL_k = \frac{\sum_{r \in i(k)} sil_r}{\#i\left(k\right)} \tag{12}$$

The global silhouette of the clustering approach (*GSIL*) is defined as the average silhouette over all sub-regions in the data set i.e.:

$$GSIL = \frac{\sum_{k=1}^{K} SIL_k}{K} \tag{13}$$

where K is the number of homogeneous sub-regions. Similar to *sil* and *SIL*, $GSIL$ ranges between -1 and 1. A clustering approach with high $GSIL$ value indicates that corresponding clusters are well-separated. Therefore, the best clustering approach will be the one that leads to sub-regions with high $GSIL$ values.

As mentioned above the delineation step affects the results of the estimation of flood quantiles. Therefore, the best approach for the delineation of homogeneous sub-regions is the one that minimizes the prediction error of $Q(t)$, the flood quantiles corresponding to a nonexceedance probability t over all sites in the data set. In this study, to quantify this error, a jackknife resampling procedure is used [e.g., Chernick, 2012; Shu and Ouarda, 2007]. It consists in considering each site i as an ungauged one by removing it temporarily from the data set and calculating its flood quantile $\hat{Q}_i(t)$ using index-flood model and according to its sub-regions obtained by the two different approaches (Ward and D-clustering). Then, over all sites in the data set, for each approach, we consider the relative bias (RB) and the relative root mean square error (RRMSE) given respectively by:

$$RB = \frac{1}{N}\sum_{i=1}^{N} \frac{Q_i(t) - \hat{Q}_i(t)}{Q_i(t)} \times 100\% \tag{14}$$

$$RRMSE = \sqrt{\frac{1}{N}\sum_{1}^{N}\left(\frac{Q_i(t)-\hat{Q}_i(t)}{Q_i(t)}\right)^2} \times 100\% \qquad (15)$$

where $Q_i(t)$ is the local flood quantile corresponding to a nonexceedance probability *t* for the site *i*, $\hat{Q}_i(t) = \mu_i q_t^{Sub_i}(\theta^R)$ is the regional flood quantile estimated by the index-flood model, $q_t^{Sub_i}(.)$ is the growth curve corresponding to the sub-region that contains the site *i*, θ^R is the regional parameter of that sub-region and *N* is the total number of sites in the whole region. The local (at-site) quantile $Q_i(t)$ is estimated using local frequency analysis. The way that the local analysis is made for this data set is already presented in the R-package "nsRFA" [Viglione 2010]. Note that, in this study, we focus on the ungauged basins problem, assuming that the observations at the gauged basins are sufficient to correctly estimate local quantiles corresponding to high return periods, and to be used as a reference for regional model performance evaluation. To this end, each site should have a long record of measured floods and its historical data must be homogeneous, stationary and independent. For the present data set, this is achieved by removing the stations that contain less than 15 years of record.

In this study, two different depth functions are used, one in the delineation step and the other one in the estimation step. The spatial depth is considered in the delineations step since it does not take a zero value as discussed above. However, for the estimation step, the Mahalanobis is chosen because it takes into account the variability of at-site parameters across all gauged sites of the sub-region through the variance-covariance matrix. The estimation results were improved because the accuracy of this matrix increases at each iteration [Wazneh et al., 2013a]. Note that

the adoption of the same depth function for both steps of RFA was also considered. However, this option led to inconsistent results or a lack of convergence.

4. Application

In this section, the delineation of homogeneous sub-regions using the D-clustering and the Ward approaches are applied and compared on a real world data set. In addition, their effect in the performance of the estimation step is considered by index-flood (9) and Depth-based index-flood (10) models. Consequently, the four combinations (delineation and estimation) defined below are considered:

- Ward and index-flood;
- Ward and Depth-based index-flood;
- D-clustering and index-flood;
- D-clustering and Depth-based index-flood.

4.1. Case study and clustering variables

The methodology was applied to Piemonte, Valle d'Aosta and Liguria, three geographical regions in the North-West of Italy (Figure 4). Annual maximum peak flood data of 47 stream flow gauging sites measured by the SIMN (Servizio Idrografico Mareografico Nazionale) are available with record lengths ranging from 6 (DoraRhemes Pelaud, Id site 39) to 65 (Ticino Miorina, Id site 4) years (Figure 5). Nine of the gauged sites have a record length less than 15 years. These nine sites are removed from the study of Viglione [2010] because of their record length. These sites are also removed from the analysis in the present study. As a result, 38 sites are selected for this study (Table 1). The area of the selected catchments is larger than 22 km^2 (RioPiz Pietraporzio, Id site 24) and less than 8024 km^2 (Tanaro Montecastello, Id site 28) and their mean annual streamflow varies from 501 to 2380 mm.

The clustering variables used to delineate homogeneous sub-regions are extracted from the study of Viglione et al. [2007a]. In total, 14 physiographical and meteorological variables are available for this set of gauged sites. However, some of these variables are correlated and are not usable as clustering variables. Therefore, it is necessary to select a subset of these 14 predictors. Viglione [2010] suggested to use the following three catchment characteristics as clustering variables: the coordinates in the UTM system of the catchment centroids (X_{bar} and Y_{bar}) and the mean elevation (H_m) of the drainage basin. These clustering variables are used in this study.

4.2. Results

The delineation of homogeneous sub-regions using the Ward method is first presented, since they can serve for the initial step of the proposed approach, then compared to those obtained using the D-clustering approach.

By considering the entire set of 38 gauged sites as a single region, the discordancy statistic D_i (6) is consistently lower than the critical value 3 (Table 1). As a result, all gauging sites are suitable for use in the present study. The corresponding heterogeneity measure H (7) is $H = 7.8 > 2$ and hence the entire region cannot be considered as homogeneous. Therefore, smaller sub-regions should be identified.

Ward method is applied to identify homogeneous sub-regions on the basis of the three characteristics described above. The number of regions is selected to ensure low H values. Hence, four different hydrometric sub-regions are identified as shown in Figure 6a. The geographical locations of the sites of these sub-regions are shown in Figure 6b. We observe that these sub-regions are almost geographically contiguous. In particular, most sites of sub-region 1 are located in the Valle d'Aosta region. The sites of sub-region 2 are located in the North of the Piemonte

region. However, the sites of sub-regions 3 and 4 are located respectively in the South Western and the Southern areas of the Piemonte region.

The discordancy statistics and the heterogeneity measure are evaluated for each one of the identified sub-regions. Within the latter, there are no discordant stations (result not presented due to space limitations). Table 3 gives the heterogeneity measure *H* of the four Ward sub-regions. The values of *H* indicate that sub-regions 3 and 4 can be considered as homogeneous (*H* < 1). However, sub-regions 1 and 2 are "possibly homogenous" ($1 \le H < 2$). Note that, the negative value of *H* (but close to 0) corresponding to sub-region 3 indicates that the data of this region have a dispersion less than the amount we expect for a homogeneous region, i.e. a positive correlation exists between the sites of the region [Hosking and Wallis, 1997].

The proposed D-clustering approach was applied to define homogeneous sub-regions on the basis of the same clustering variables described above. As mentioned in section 3, by construction, the D-clustering algorithm start with *K*=2 initial sub-regions. However, this first iteration is not presented in this paper since the obtained sub-regions are not homogeneous (step 3). Therefore, the algorithm restarts with *K*=3 using Ward method (Figure 6a). Following above notation, $i(1)$, $i(2)$ and $i(3)$ are the initial sets that contain the Id number of sites of three sub-regions (see Figure 6a). Note that, two of the three initial sub-regions are not homogeneous (result not presented due to space limitations). As a result, the initial sub-regions are not suitable for RFA and a reallocation of sites into the sub-regions is necessary. After the initialization step, in the second step of the D-clustering approach, we calculate D_r^w, D_r^b and DeD_r (see section 3) of X_r, where $X_r = (X_{\mathrm{bar}}, Y_{\mathrm{bar}}, H_m)_r$ is the vector of clustering variables for site $r \in i(k), k = 1,2,3$. Then, the sites wrongly placed in their sub-regions are determined i.e. sites with negative DeD

values. The corresponding results are summarized in Table 2. This table shows that 9 sites are placed in wrong sub-regions. These sites are moved from their initial sub-regions to the sub-regions that maximize their DeD value (Table 2). After the moving step, in the third step of the D-clustering approach the homogeneity for the new obtained sub-regions formed after moving the 9 sites is tested. Table 3 gives the heterogeneity measure *H* of the three obtained sub-regions. The values of *H* indicate that these new sub-regions can be considered as homogeneous. Therefore, the algorithm stops in this iteration. As a result, the D-clustering approach suggests, according to *H* criterion, three homogeneous sub-regions. The geographical locations of the sites of the final sub-regions are shown in Figure 7. Sites of sub-region 1 are located in the North of the studied regions. The sites of sub-region 2 and 3 are located respectively in the West and the South of the studied regions.

One of the objectives of this study is to compare the performance of the Ward and D-clustering approaches used for the delineation of homogeneous sub-regions. To this end, the *H* heterogeneity measure (7) and the global silhouette *GSIL* (13) are computed (Table 3). Table 3 shows that the D-clustering approach leads to sub-regions with low *H* values, which means that the obtained sub-regions are more homogeneous than those obtained using Ward method. The values of *GSIL* using both clustering approaches are close to 0.5 (Table 3). This means that the obtained sub-regions using both approaches are well-separated. Note that, the *GSIL= 0.47* using D-clustering is slightly higher than *GSIL= 0.43* obtained using Ward approach. This can be explained by the fact that both *GSIL* measure and Ward method are based on the Euclidean distance whereas the D-clustering approach is based on the spatial depth function. In addition, in terms of the number of sites, the sub-regions obtained using the D-clustering are more suitable for regional estimation than those obtained by Ward method (Table 3). In fact, in RFA, the

number of sites in a sub-region should be large enough to ensure the applicability of a regional estimation, but not too large to maintain cluster homogeneity. Hosking and Wallis [1997] suggested that homogeneous sub-regions for RFA must contain more than 9 sites. In this study, the two sub-regions 2 and 4 formed by Ward approach contain a limited number of sites (7 and 8 sites respectively), which is not suitable for RFA estimation [Hosking and Wallis, 1997]. However, all D-clustering sub-regions contain more than 9 sites.

Since the delineation step affects the regional estimation, the performance of the D-clustering and Ward method are compared in terms of flood estimation from the index-flood model. This comparison is based on the RB and the RRMSE, given in (14) and (15) respectively, of the flood quantiles for the nonexceedance probabilities t = 0.9, 0.99, 0.995 and 0.999. To this end, it is required to obtain the corresponding growth curves $q_t^{Sub_i}(.)$ (9) for each sub-region. Hence in this study, five different probability distributions commonly used in RFA are considered as candidates i.e., the Generalized Extreme Value (GEV), the Generalized Logistic (GLO), the Pearson type-III (PE3), the Generalized Pareto (GPA) and the Generalized Normal (GNO) distributions. To select the suitable distribution, the Z goodness-of-fit test (8) is computed. The obtained results for Ward and D-clustering approaches are shown in Table 4. The GEV, GNO and the PE3 distributions are passed the test. To select the appropriate distribution, the regional *LMR*s diagram (Figure 8) and the Akaike Information Criterion "AIC" (Table 5) are used. The proximity of the regional estimated L-moments to a particular candidate theoretical distribution in the (L_{skew} , L_{kur}) space indicates the appropriateness of that distribution to describe the regional data. According to Figure 8a, for the Ward sub-regions, the GEV is chosen as a distribution for sub-regions 1 and 2 and the PE3 is chosen for sub-regions 3 and 4. However, for the D-clustering sub-regions, the GEV is chosen for sub-region 1 and the PE3 for the sub-regions 2 and 3 (Figure

8b). These distributions have minimum AIC values for each sub-region (Table 5). The associate parameters are ξ, α and κ (respectively the location, scale and shape parameters). From (9), the regional parameter vector of growth curve is $\theta^R = \left(\kappa^R, \alpha^R, \xi^R\right)$. The obtained values of these parameters are presented in Table 6.

The RB and RRMSE given in (14) and (15) are computed using the growth curve and the regional parameters presented in Table 6. The results of the index-flood model performances related to the different delineation approaches are summarized in Table 7 (index-flood). As expected and shown in a number of studies, for a given region, the flood quantile estimation is generally more accurate for small nonexceedance probabilities *t*. Table 7 shows that the delineation of sub-regions using the D-clustering leads to more efficient estimation in terms of RB and RRMSE than those obtained using Ward one. As previously indicated, D-clustering sub-regions are larger and more homogeneous than those formed by Ward approach, which could be the reason of this improvement.

As previously indicated, the depth function is also included in the estimation step through the Depth-based index-flood model (10). The detailed steps to apply the latter, i.e. how compute the optimal weigh function φ and the regional growth curve parameters $\left(\hat{\theta}^R\right)_D$, are not presented in this study since they are already presented in Wazneh et al. [2013a]. However, the results of this application are presented in Table 7. These results indicate that, the inclusion of the depth function in the RFA stages improves the estimation results. More precisely, Table 7 shows that the inclusion of the depth function in the delineation step provides more significant performance improvement in terms of the RRMSE than in the estimation step for a given risk; which shows the importance of the delineation step in the RFA procedure. Indeed, for instance, given *t* =

0.999, the inclusion of the depth function in the estimation step allows to decrease the RRMSE from 17.33% to 14.43% corresponding respectively to Ward combined to index-flood and Ward combined to Depth-based index-flood. This improvement in terms of the RRMSE is due to the fact that the Depth based index-flood model takes into account the hydrological variability through the variability of at-site parameters across all sites of identified sub-regions which is not the case with the traditional index-flood model [Wazneh et al., 2013a]. However, considering depth in the delineation step decreases the RRMSE from 17.33% to 11.54 % corresponding to Ward combined to index-flood and D-clustering combined to index-flood. This result is due to the fact that sub-regions obtained using D-clustering are more homogeneous and more representative than those obtained using the traditional Ward method [Table 3]. Furthermore, Table 7 shows that the best combination of the two stages of RFA is the D-clustering approach with the Depth-based index-flood model (RRMSE = 10.17%). Note that the improvement performance using depth function increases with respect to the nonexceedance probabilities *t.*

To compare the relative errors of flood quantile estimates obtained by different delineation approaches and using the two estimation methods presented in Table 7. Figure 9 illustrates these errors with respect to the logarithm of the basin area. It is generally observed that using the D-clustering approach the relative errors are lower than those obtained with the Ward method. We also observe large errors using the Ward method for some sites, such as number 1, 9 and 37. These sites have high discordancy statistics D_i (Table 1; respectively 1.97, 2.60 and 2.90) and they are "far" in term of *LMRs* from the other sites of the data set (Table 1). These sites are misclassified using Ward method. However, using the D-clustering approach these sites have proportionally small relative errors. This can be explained by the robustness of the D-clustering approach. Consequently, the "far" sites are well classified using this approach.

5. Conclusions

In the present paper, a new robust and objective approach to identify the homogeneous sub-regions for RFA is proposed and applied. The proposed D-clustering approach is based on the statistical notion of depth function. This approach employs depth functions to distribute sites into the sub-regions. In fact, using the D-clustering approach, homogeneous sub-regions are formed by assigning each site in the data set to the sub-region that maximizes its depth value. Therefore, aside from leading to the best estimation results, the proposed approach allows the delineation of sub-regions to be more practical and does not require the user's subjective intervention. The user has only to select one criterion (such as the *H* heterogeneity used in this study) to obtain the sub-regions with respect to this criterion.

The obtained results from a set of sites in the North-West of Italy show that the D-clustering approach leads to more homogeneous and larger sub-regions than those obtained using Ward method. The sites with high discordancy value (far sites) are well clustered in their sub-regions using the D-clustering approach which is not the case using the traditional approach. Additionally, it leads to more efficient estimates in terms of RB and RRMSE than those obtained using the traditional approach. One of the findings is that the inclusion of the depth function in the delineation step provides more significant performance improvement than its inclusion in the estimation step. It is also concluded that the introduction of the depth function in the two stages of RFA leads to the best results.

Acknowledgment

Financial support for this study was graciously provided by the Natural Sciences and Engineering Research Council (NSERC) of Canada and the Canada Research Chair Program. The data sets used in this study are available in the R-package "nsRFA" at http://cran.r-

project.org/web/packages/nsRFA/index.html. The authors are grateful to the Editor, the Associate Editor, and the reviewers for their valuable comments and suggestions which helped improve the quality of the manuscript.

References

Acreman, M.C., Sinclair, C.D., 1986. Classification of drainage basins according to their physical characteristics; an application for flood frequency analysis in Scotland. Journal of Hydrology, 84(3-4): 365-380.

Baeriswyl, P.A., Rebetez, M., 1997. Regionalization of precipitation in Switzerland by means of principal component analysis. Theoretical and Applied Climatology, 58(1-2): 31-41.

Bárdossy, A., Singh, S.K., 2008. Robust estimation of hydrological model parameters. Hydrology and Earth System Sciences, 12(6): 1273-1283.

Bárdossy, A., Singh, S.K., 2011. Regionalization of hydrological model parameters using data depth. Hydrology Research, 42(5): 356-371.

Bargaoui, Z.K., Fortin, V., Bobee, B., Duckstein, L., 1998. Fuzzy approach to the delineation of region of influence for hydrometric stations. Revue des sciences de l'eau, 11(2): 255-282.

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

Bhaskar, N.R., O'Connor, C.A., 1989. Comparison of method of residuals and cluster analysis for flood regionalization. Journal of Water Resources Planning and Management, 115(6): 793-808.

Burn, D.H., Goel, N.K., 2000. The formation of groups for regional flood frequency analysis. Hydrological Sciences Journal, 45(1): 97-112.

Castellarin, A., Burn, D.H., Brath, A., 2001. Assessing the effectiveness of hydrological similarity measures for flood frequency analysis. Journal of Hydrology, 241(3-4): 270-285.

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

Chebana, F., Ouarda, T.B.M.J., 2009. Index flood-based multivariate regional frequency analysis. Water Resources Research, 45(10).

Chebana, F., Ouarda, T.B.M.J., 2011. Depth-based multivariate descriptive statistics with hydrological applications. Journal of Geophysical Research D: Atmospheres, 116(10).

Chen, Y., Dang, X., Peng, H., Bart Jr, H.L., 2009. Outlier detection with the kernelized spatial depth function. IEEE Transactions on Pattern Analysis and Machine Intelligence, 31(2): 288-305.

Chernick, M.R., 2012. The jackknife: A resampling method with connections to the bootstrap. Wiley Interdisciplinary Reviews: Computational Statistics, 4(2): 224-226.

Chiang, S.M., Tsay, T.K., Nix, S.J., 2002. Hydrologic regionalization of watersheds. I: Methodology development. Journal of Water Resources Planning and Management, 128(1): 3-11.

Ding, Y., Dang, X., Peng, H., Wilkins, D., 2007. Robust clustering in high dimensional data using statistical depths. BMC Bioinformatics, 8(SUPPL. 7).

Donoho, D.L., Gasko, M., Breakdown Properties of Location Estimates Based on Halfspace Depth and Projected Outlyingness. The Annals of Statistics, 20(4): 1803-1827.

Dutta, S., Ghosh, A.K., 2012. On robust classification using projection depth. Annals of the Institute of Statistical Mathematics, 64(3): 657-676.

Ghosh, A.K., Chaudhuri, P., 2005. On maximum depth and related classifiers. Scandinavian Journal of Statistics, 32(2): 327-350.

Haddad, K., Rahman, A., 2012. Regional flood frequency analysis in eastern Australia: Bayesian GLS regression-based methods within fixed region and ROI framework - Quantile Regression vs. Parameter Regression Technique. Journal of Hydrology, 430-431: 142-161.
Haddad, K., Rahman, A., Ling, F., 2014. Regional flood frequency analysis method for Tasmania, Australia: A case study on the comparison of fixed region and region-of-influence approaches. Hydrological Sciences Journal: null-null.
Hosking, J.R.M., Wallis, J.R., 1993. Some statistics useful in regional frequency analysis. Water Resour. Res., 29(2): 271-281.
Hosking, J.R.M., Wallis, J.R., 1997. Regional Frequency Analysis: An Approach Based on L-moments, Cambridge, U.K., 244 pp.
Hugg, J., Rafalin, E., Seyboth, K., Souvaine, D.L., 2006. An Experimental Study of Old and New Depth Measures, ALENEX, pp. 51-64.
Ilorme, F., Griffis, V.W., 2013. A novel procedure for delineation of hydrologically homogeneous regions and the classification of ungauged sites for design flood estimation. Journal of Hydrology, 492: 151-162.
Johnson, R.A., Wichern, D.W., 2002. Applied multivariate statistical analysis, 5. Prentice hall Upper Saddle River, NJ.
Jörnsten, R., 2004. Clustering and classification based on the L1 data depth. Journal of Multivariate Analysis, 90(1 SPEC. ISS.): 67-89.
Krauße, T., Cullmann, J., 2012. Towards a more representative parametrisation of hydrologic models via synthesizing the strengths of Particle Swarm Optimisation and Robust Parameter Estimation. Hydrology and Earth System Sciences, 16(2): 603-629.
Krauße, T., Cullmann, J., Saile, P., Schmitz, G.H., 2012. Robust multi-objective calibration strategies – possibilities for improving flood forecasting. Hydrology and Earth System Sciences, 16(10): 3579-3606.
Li, J., Cuesta-Albertos, J.A., Liu, R.Y., 2012. DD-classifier: Nonparametric classification procedure based on DD-plot. Journal of the American Statistical Association, 107(498): 737-753.
Liu, R.Y., Parelius, J.M., Singh, K., 1999. Multivariate analysis by data depth: descriptive statistics, graphics and inference,(with discussion and a rejoinder by Liu and Singh). The Annals of Statistics, 27(3): 783-858.
Lu, L.H., Stedinger, J.R., 1992. Sampling variance of normalized GEV/PWM quantile estimators and a regional homogeneity test. Journal of Hydrology, 138(1-2): 223-245.
Nathan, R.J., McMahon, T.A., 1990. Identification of homogeneous regions for the purposes of regionalisation. Journal of Hydrology, 121(1-4): 217-238.
Ouarda, T.B.M.J. et al., 2008. Intercomparison of regional flood frequency estimation methods at ungauged sites for a Mexican case study. Journal of Hydrology, 348(1–2): 40-58.
Ouarda, T.B.M.J., Girard, C., Cavadias, G.S., Bobée, B., 2001. Regional flood frequency estimation with canonical correlation analysis. Journal of Hydrology, 254(1-4): 157-173.
Ouarda, T.B.M.J., Hache, M., Bruneau, P., Bobee, B., 2000. Regional flood peak and volume estimation in northern Canadian basin. Journal of Cold Regions Engineering, 14(4): 176-191.
Oudin, L., Kay, A., Andréassian, V., Perrin, C., 2010. Are seemingly physically similar catchments truly hydrologically similar? Water Resources Research, 46(11): W11558.

Potter, K.W., Faulkner, E.B., 1987. CATCHMENT RESPONSE TIME AS A PREDICTOR OF FLOOD QUANTILES1. JAWRA Journal of the American Water Resources Association, 23(5): 857-861.
Rao, A.R., Srinivas, V.V., 2003. Some problems in regionalization of watersheds. IAHS-AISH Publication(280): 301-308.
Rao, A.R., Srinivas, V.V., 2006a. Regionalization of watersheds by fuzzy cluster analysis. Journal of Hydrology, 318(1-4): 57-79.
Rao, R.A., Srinivas, V.V., 2006b. Regionalization of watersheds by hybrid-cluster analysis. Journal of Hydrology, 318(1-4): 37-56.
Reed, D.W., Houghton-Carr, H.A., Richardson, B.D., 1999. Introducing the flood estimation handbook. Proceedings of the 1999 34th MAFF Conference of River and Coastal Engineers.
Rousseeuw, P.J., 1987. Silhouettes: A graphical aid to the interpretation and validation of cluster analysis. Journal of Computational and Applied Mathematics, 20(C): 53-65.
Shu, C., Ouarda, T.B.M.J., 2007. Flood frequency analysis at ungauged sites using artificial neural networks in canonical correlation analysis physiographic space. Water Resources Research, 43(7).
Tukey, J.W., 1975. Mathematics and the picturing of data. Proceedings of the International Congress of Mathematicians, 2: 523-531.
Viglione, A., 2010. Non-supervised Regional Frequency Analysis. http://cran.r-project.org/web/packages/nsRFA/index.html.
Viglione, A., Claps, P., Laio, F., 2007a. Mean annual runoff estimation in north-western Italy,. In: Loggia, L. (Ed.), Water Resources Assessment and Management Under Water Scarcity Scenarios. CDSU Publication, Milan.
Viglione, A., Laio, F., Claps, P., 2007b. A comparison of homogeneity tests for regional frequency analysis. Water Resources Research, 43(3).
Ward Jr, J.H., 1963. Hierarchical grouping to optimize an objective function. Journal of the American Statistical Association, 58(301): 236-244.
Wazneh, H., Chebana, F., Ouarda, T.B.M.J., 2013a. Depth-based regional index-flood model. Water Resources Research, 49(12): 7957–7972.
Wazneh, H., Chebana, F., Ouarda, T.B.M.J., 2013b. Optimal depth-based regional frequency analysis. Hydrology and Earth System Sciences, 17(6): 2281-2296.

Table 1. Catchment characteristics variables and summary statistics of annual maximum peak flood data used in this study. *RL* is the record length in years, Q_m is the average of annual maximum peak flood quantile in m^3/s, X_{bar} and Y_{bar} are the coordinates in the UTM system of the catchment centroids, H_m is the mean elevation of the drainage basin in m, the three L-moment ratios and the discordancy measure *D* for the whole region are also shown here.

Id site	*RL* (years)	Q_m (m^3/s)	X_{bar} (m)	Y_{bar} (m)	H_m (m)	L_{cv}	L_{sk}	L_{kur}	*D*
1	15	1569	8.397	46.375	2137	0.074	0.028	0.017	1.97
2	32	1380	8.225	46.149	1674	0.149	0.179	0.061	0.82
4	65	1409	8.652	46.169	1286	0.142	0.147	0.145	0.07
5	15	1728	8.456	46.035	1251	0.153	0.128	0.292	2.32
6	32	1590	8.206	45.888	1319	0.208	0.065	0.081	1.19
8	26	1271	7.936	45.838	2112	0.117	0.090	0.048	0.72
9	24	1427	8.091	45.833	1491	0.133	-0.09	-0.06	2.60
10	61	925	7.395	45.728	2090	0.116	0.256	0.160	1.52
11	44	1043	7.425	45.470	1924	0.149	0.123	0.100	0.08
12	49	1093	7.287	45.290	1773	0.157	0.115	0.140	0.08
14	34	694	7.084	44.963	1730	0.192	0.071	0.071	0.77
15	23	654	6.965	45.001	2144	0.179	0.203	0.182	0.21
16	30	662	6.851	44.932	2165	0.157	0.152	0.070	0.38
17	27	589	6.912	45.070	1867	0.157	0.149	0.079	0.28
18	39	1272	7.115	44.693	2261	0.154	0.094	0.049	0.43
19	44	506	7.398	44.736	924	0.207	0.130	0.133	0.61
20	41	829	7.240	44.403	1565	0.190	0.135	0.108	0.28
21	23	925	7.007	44.356	2074	0.154	0.250	0.183	0.61
22	20	1240	7.053	44.267	2138	0.153	0.182	0.123	0.17
23	16	1128	7.576	44.178	1677	0.170	0.118	0.298	2.83
24	23	1272	7.018	44.311	2194	0.119	0.007	0.005	1.29
25	18	1012	7.137	44.316	1814	0.178	0.187	0.172	0.15
28	58	515	8.064	44.548	651	0.200	0.222	0.130	0.54
29	34	1032	7.771	44.124	1576	0.190	0.235	0.088	1.02
30	31	900	7.901	44.179	1222	0.206	0.265	0.114	1.21
31	36	779	7.852	44.298	938	0.167	0.138	0.174	0.20
32	29	1065	7.828	44.226	1513	0.180	0.124	0.210	0.86
33	25	827	9.040	44.628	688	0.183	0.191	0.153	0.15
34	22	965	8.300	44.297	602	0.213	0.325	0.217	1.54
35	16	881	8.458	44.447	605	0.197	-0.01	0.013	2.20
37	18	783	9.112	44.668	867	0.285	0.268	0.214	2.90
41	17	897	7.177	45.718	2267	0.106	0.205	0.193	1.05
42	23	1351	7.830	45.855	2625	0.093	0.052	0.124	0.94
43	31	1648	6.970	45.672	2512	0.083	0.103	0.194	1.44
44	17	1023	7.151	45.828	2229	0.107	0.044	0.138	0.86
45	30	991	7.742	45.872	2631	0.104	0.169	0.277	2.00
46	24	1259	7.559	45.613	2352	0.150	0.232	0.142	0.58
47	15	1077	7.206	45.523	2723	0.137	0.246	0.206	0.81

Table 2. Sites that have negative depth deviance **"**DeD". These sites are moved from their initial sub-regions in the D-clustering approach.

Id site	Initial sub-region	*DeD*	Final sub-region
6	1	-0.10	3
11	1	-0.12	2
12	1	-0.21	2
14	2	-0.15	3
17	2	-0.13	3
20	2	-0.17	3
23	2	-0.21	3
29	2	-0.15	3
32	2	-0.13	3

Table 3. Heterogeneity measures *H*, the silhouette *SIL* and the global silhouette *GSIL* using traditional Ward and D-clustering approaches.

Delineation approach	Sub-region	Number of sites	*H*	*SIL*	*GSIL*
Ward	1	10	1.93	0.41	0.43
	2	7	1.96	0.39	
	3	13	-0.3	0.41	
	4	8	0.5	0.51	
D-clustering	1	14	1.03	0.44	0.47
	2	9	-0.6	0.49	
	3	15	0.014	0.48	

Table 4. Z measure values for the different candidate distributions and for each one of the proposed sub-regions using traditional Ward and D-clustering approaches

Delineation approach	Sub-regions	Distributions				
		GLO	GEV	GNO	PE3	GPA
Ward	1	1.68	**-0.24**	**-0.37**	**-0.86**	-4.1
	2	2.59	**0.74**	**0.89**	**0.68**	-2.96
	3	3.82	**1.26**	**1.21**	**0.65**	-4.11
	4	2.08	**0.57**	**0.26**	**-0.40**	-2.83
D-clustering	1	2.06	**-0.23**	**-0.30**	**-0.81**	-5.07
	2	3.56	**1.44**	**1.41**	**0.97**	-2.99
	3	3.07	**0.79**	**0.57**	**-0.12**	-4.15

Bold character indicates that the distribution may be accepted as a regional distribution for each sub-region.

Table 5. AIC values of the distributions which may be accepted as a regional distribution for each sub-region using both approaches.

Delineation approach	Sub-regions	Distributions		
		GEV	GNO	PE3
Ward	1	**-124.6**	-122.4	-123.5
	2	**38.6**	39.1	39.1
	3	129.1	129.6	**129.0**
	4	168.9	169.3	**168.6**
D-clustering	1	**-121.5**	-120.2	-119.7
	2	42.9	42.9	**42.6**
	3	279.8	278.9	**278.5**

Bold character indicates the best distribution for each sub-region i.e. distribution have the minimum AIC value (each line)

Table 6. The estimated parameters of growth curve using index-flood model for the two delineation approaches. Distribution represents the chosen growth curve distribution for each sub-region.

Delineation approach	Sub-regions	Distribution	Parameters		
			ξ^R (location)	α^R (scale)	κ^R (shape)
Ward	1	GEV	0.905	0.170	0.023
	2	GEV	0.882	0.246	0.108
	3	PE3	0.298	0.134	5.206
	4	PE3	0.359	0.220	2.905
D-clustering	1	GEV	0.902	0.181	0.041
	2	PE3	0.329	0.117	5.693
	3	PE3	0.283	0.178	4.015

Table 7. Quantile estimation results in % using the index-flood and Depth-based index-flood method for the studied region using D-clustering and Ward delineation approaches

Delineation approach	Estimation method	0.9		0.99		0.995		0.999	
		RB	RRMSE	RB	RRMSE	RB	RRMSE	RB	RRMSE
Ward	Index-flood	0.21	4.43	0.23	10.70	0.18	12.70	-0.05	17.33
	Depth-based index-flood	0.18	4.19	0.32	8.75	0.35	10.65	0.86	14.43
D-clustering	Index-flood	-0.78	3.25	0.18	9.05	0.32	9.40	-0.12	11.54
	Depth-based index-flood	-0.42	3.59	-0.21	8.00	0.44	8.77	0.07	10.17

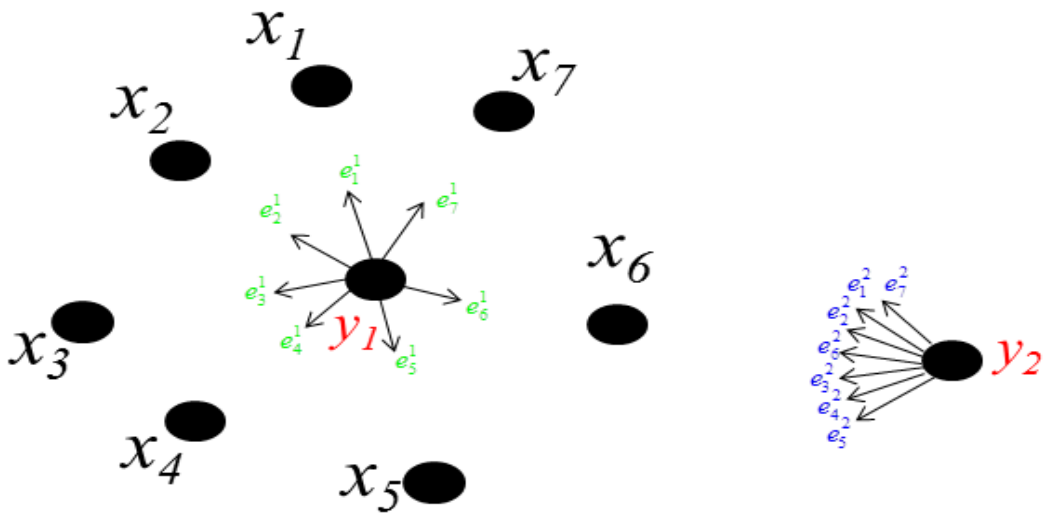


Figure 1. Illustration of the way the spatial depth function is computed in the bivariate case. $X = \{x_1,...,x_7\}$ represents the sample and e_i^s $(i = 1,..7; s = 1,2)$ represents the unit vector from y_s to x_i. $SPD(y_1, X)$ is close to 1 and $SPD(y_2, X)$ is close to 0.

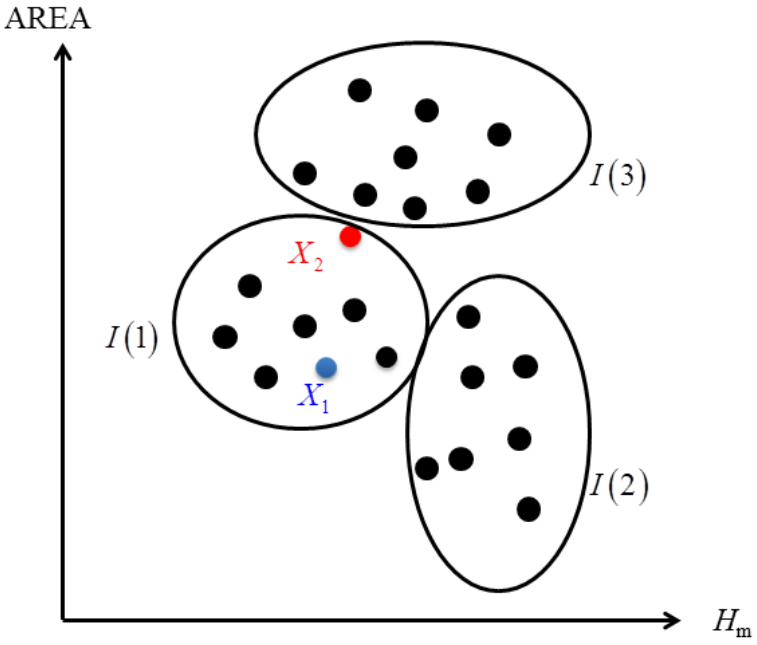


Figure 2. Illustrations of the way the within and between depths of sites are computed. Suppose that three initial sub-regions are formed using the traditional approach and the clustering variables are the basin areas (*AREA*) and the mean elevation (H_m). X_1 and X_2 are respectively the vector of attribute values for site 1 and 2. The within depth of X_1 and X_2 are respectively $D_1^w = SPD\left(X_1, I(1)\right)$ and $D_2^w = SPD\left(X_2, I(1)\right)$. The between depth of X_1 and X_2 are respectively $D_1^b = SPD\left(X_1, I(2)\right)$ and $D_2^b = SPD\left(X_1, I(3)\right)$.

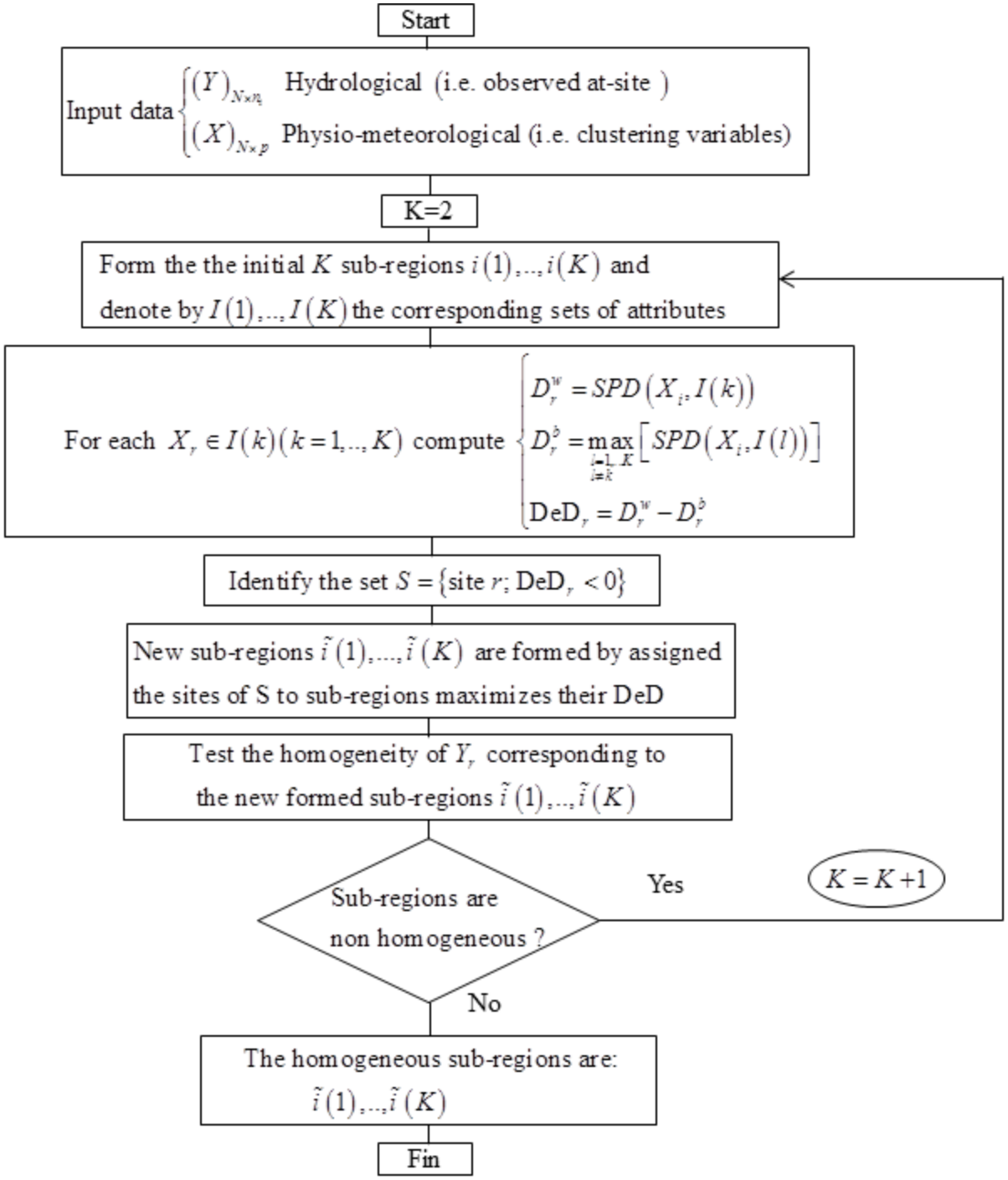


Figure 3. An overview diagram summarizing the delineation of hydrological homogeneous sub-regions using spatial depth function

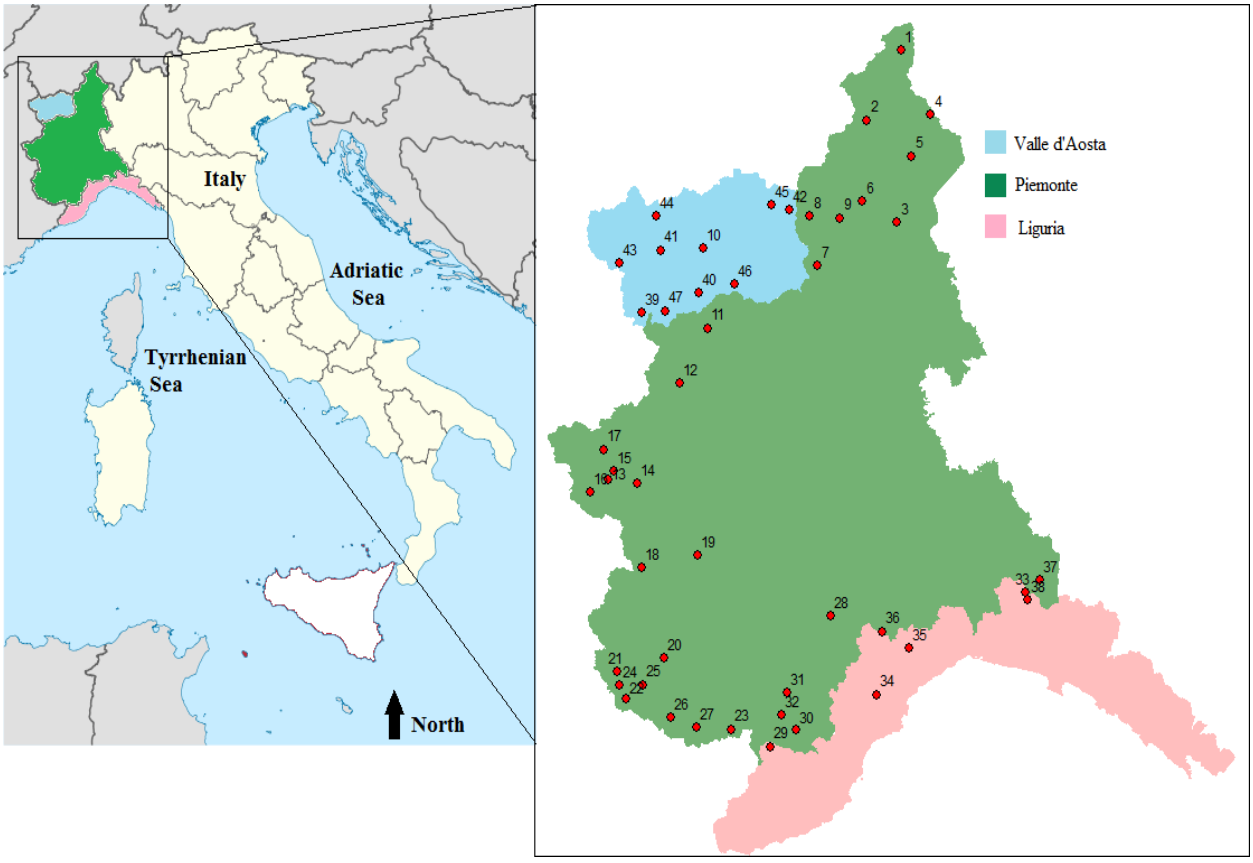


Figure 4. Geographical location of the studied sites in the three geographical regions.

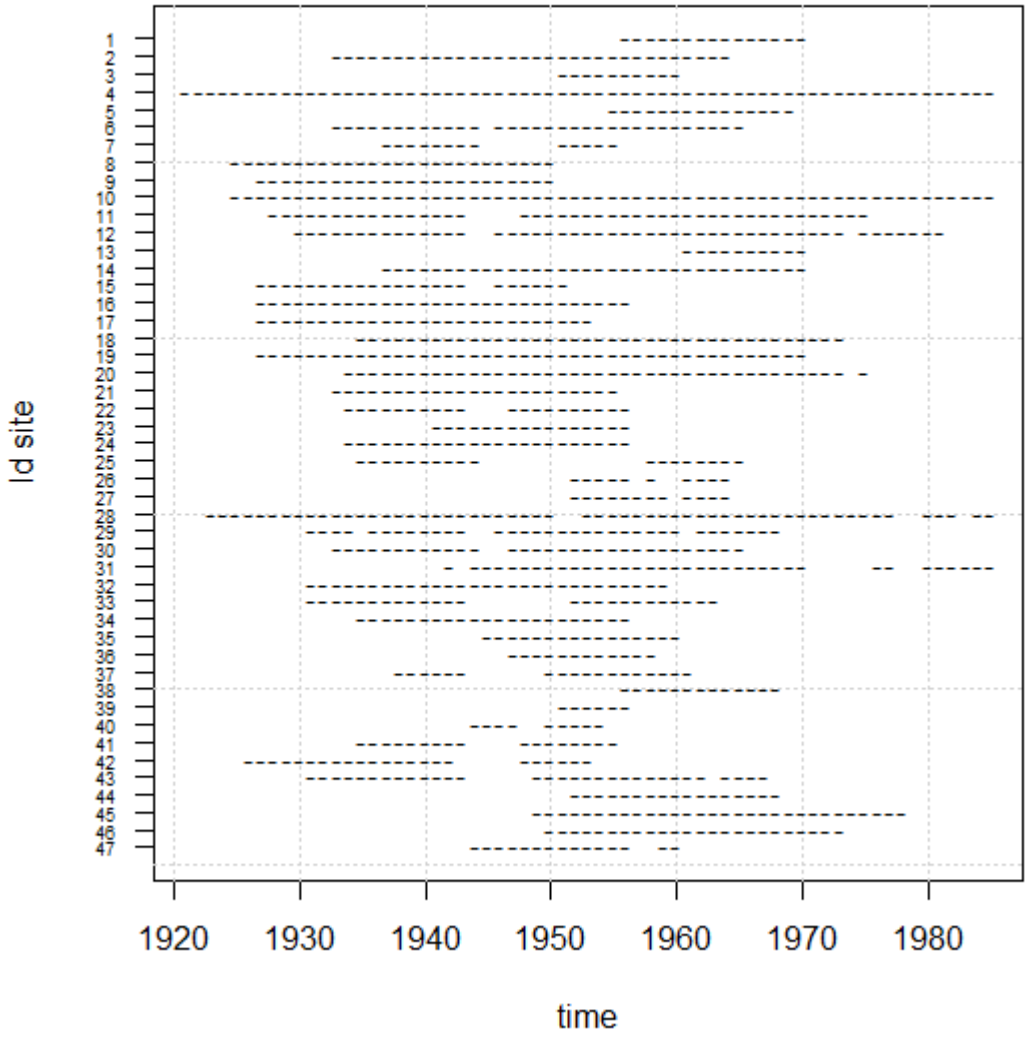


Figure 5. The observation period of gauged sites. Sites with Id numbers 3, 7, 13, 26, 27, 36, 38, 39 and 40 have a record length less than 15 years and they are removed from the data set.

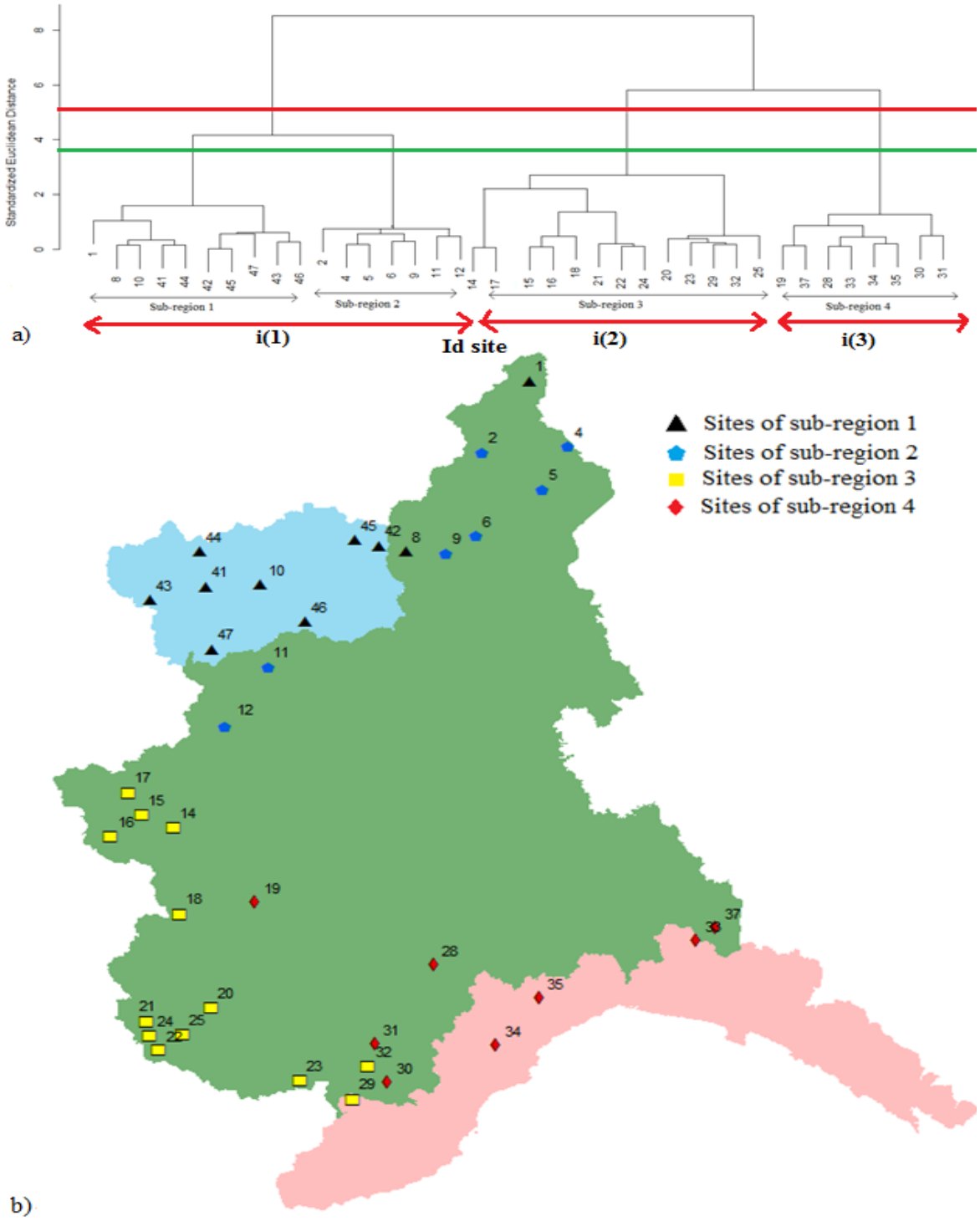


Figure 6. a) Hierarchical clustering of stations using Ward method. The green line represents the cut off to define the four homogenous sub-region used in the Ward approach. The reed line represent the cut off to define the three initial sub-region *i*(1), *i*(2), *i*(3) used for the D-clustering approach. b) Geographical location of sites of four homogeneous sub-regions formed using Ward method.

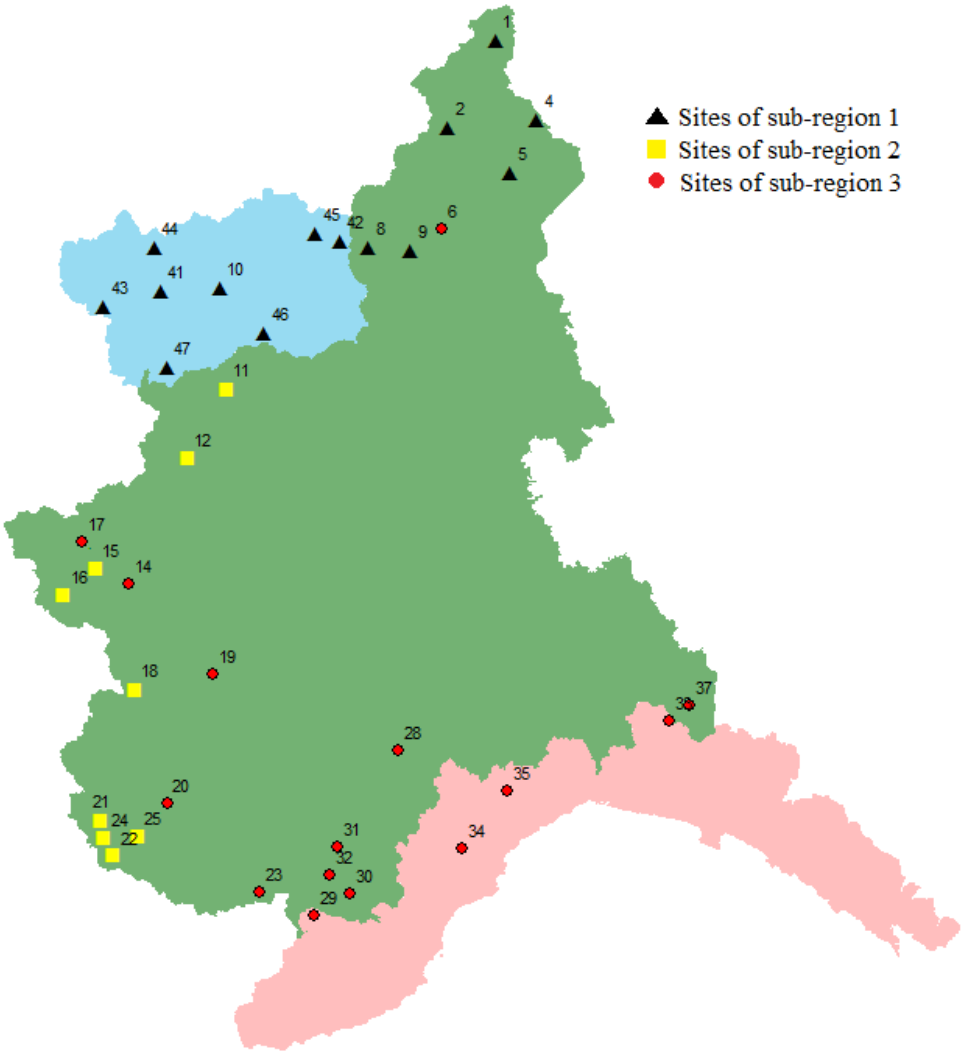


Figure 7. Geographical location of sites of final sub-regions formed using the D-clustering approach.

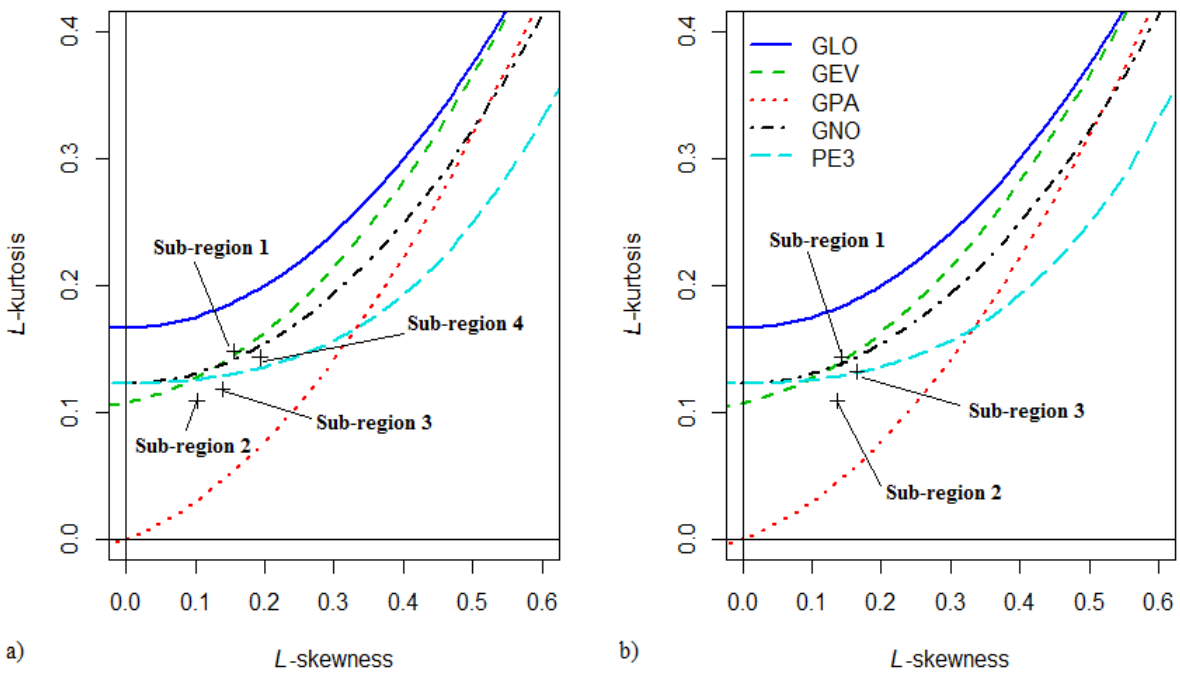


Figure 8. . Regional average L-moments ratio diagram: (a) for the four sub-regions formed using Ward method and (b) for the three sub-regions formed using D-clustering approach.

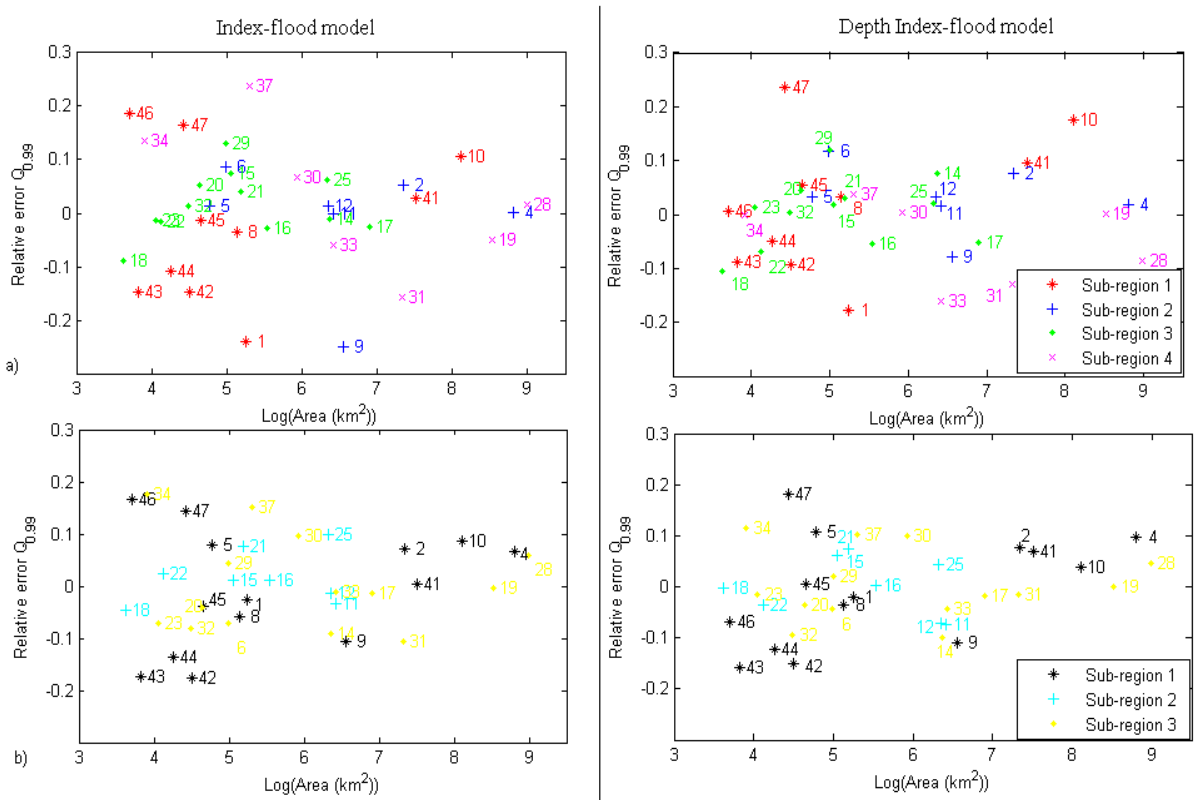


Figure 9. Relative quantile error of $Q_{0.99}$ using (a) Ward approach and (b) D-clustering approach. The first column illustrate the error using index-flood model and the second column illustrate the error using depth-based index-flood model.

CHAPITRE 4: MODÈLE D'INDICE DE CRUE RÉGIONAL BASÉ SUR LES FONCTIONS DE PROFONDEUR

Depth-based regional index-flood model

H. Wazneh*[1], F. Chebana[1] and T.B.M.J. Ouarda[2]

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

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

Accepted October 9th 2013

(Water Resources Research)

Abstract:

Regional flood frequency analysis aims to estimate flood risk at sites where little or no hydrological data are available. The index-flood model is one of the commonly employed models for this purpose. In this model, the predicted value depends on the growth curve and its regional parameters. The latter are estimated as weighted averages of the at-site parameters. Traditional approaches are mainly based on site record lengths or region size to define these weights. Hence, they are not representative of the hydrological similarity between sites within a region. In addition, they are not defined to reach optimality in terms of model performance. To overcome these limitations, the present paper aims to propose a new optimal iterative weighting scheme to the index-flood model. The proposed approach is based on a number of elements: a statistical depth function to introduce similarity between sites, a weight function to amplify and control the depth values, an iterative procedure to improve estimation accuracy, and an optimization algorithm to objectively automate the choice of the weight function. A data set from the Island of Sicily (Italy) is used to compare the proposed approach with traditional ones. On the basis of the L-moments and using cluster analysis techniques, the studied region is subdivided into three homogeneous sub-regions. The results indicate that the proposed approach performs significantly better than traditional ones both in terms of relative bias and relative root mean squares error. The proposed approach allows identification of cross-correlation in the region and provides a significant performance improvement.

Keywords: regional flood frequency analysis, index-flood, statistical depth function, regional growth curve, flood quantile.

1. Introduction and review

Hydrological data records are generally short and are not always available in the location of interest. Consequently, at-site hydrological frequency analysis is not always reliable or even possible. Regional frequency analysis (RFA) is commonly used to estimate extreme hydrological events at sites where little or no hydrological data are available. RFA is based on the transfer of the available information from gauged sites, within a homogeneous region, to the target site. The delineation of homogeneous hydrological regions and regional estimation are the two main steps of a RFA [GREHYS, 1996a].

Homogenous regions can be defined as geographically contiguous regions, geographically non-contiguous regions, or as hydrological neighborhoods [e.g., Ouarda et al., 1999]. The use of non-contiguous regions and neighborhoods was recommended in the literature for different parts of the world [Ouarda et al., 2008]. To define the non-contiguous regions, cluster analysis of site characteristics is one of the most practical method used in the literature [Hosking and Wallis, 1997]. However, two main methods were proposed for building hydrological neighborhoods, namely the region of influence [Burn, 1990a] and the canonical correlation analysis (CCA) approach [Ouarda et al., 2001].

In regional estimation, two main models were presented and widely employed in the literature, the index-flood approach [Hosking and Wallis, 1993] and multiple regression [Pandey and Nguyen, 1999]. The index-flood approach assumes that the frequency distributions at different sites of a homogeneous region are identical apart from a scale factor [Hosking and Wallis, 1997]. However, the multiple regression method is based on the concept that spatial variations in flood flow statistics are closely related with variations in regional catchment and climatic characteristics [Gupta and Dawdy, 1995].

The homogeneity of a region involves the similarity between its sites [Castellarin et al., 2001]. Generally, in the hydrological framework, there are two types of variables used to define similarity between sites in the region; the physiographic catchment characteristics (such as drainage area and catchment slope [e.g., Burn, 1990b]) and /or the hydrological variables (such as flood peak quantiles and statistical measures estimated from the available flood series [e.g., Ouarda et al., 1993]). Usually, this similarity is not explicitly incorporated in the regional estimation step, especially in the index-flood model.

The index-flood model was first introduced by Darlymple [1960]. In this model, flood peak series at different sites in the homogenous region are normalized by dividing the data by an at-site location parameter, called "index flood". The obtained normalized records are used to derive a regional distribution called growth curve. Flood quantiles at a target site are estimated by scaling the growth curve by an estimate of the at-site location parameter at the target site.

A number of methods have been employed to estimate the at-site parameters such as the maximum likelihood (ML) and L-moment methods. The estimated regional growth curve parameters are linear combinations, as weighted average, of their at-site counterparts. The contribution of each gauged site to the estimation of regional parameters is defined as a coefficient in these linear combinations. Consequently, the accuracy and the performance of estimated parameters depend on the weight coefficient values. More precisely, to compute the regional growth curve parameters, different types of weights were proposed in the literature. For instance, Hosking and Wallis [1997] attributed weights to sites proportionally to their record lengths. An alternative weighting, limit the influence of sites with longer records, was suggested by Stedinger et al. [1992]. However, an unweighted average is preferred for non-perfectly homogeneous regions [e.g., Jin and Stedinger, 1989].

The above types of weights proposed in the literature for the index-flood model are based solely on the record length without including an objective similarity measure. These weights do not take into consideration the real information content of the various series, the redundancy that may exist between sites, the similarity between the gauged sites of the region, and the similarity of the sites with the target one. In the common case where the region is not perfectly homogeneous proportional record weights could lead to a biased estimation [Hosking and Wallis, 1997]. On the other hand, in terms of model performance, the above types of weights do not necessarily lead to optimal results.

The aim of the present work is to propose a new optimal weighting scheme, which maximizes the performance of the index-flood model. The proposed approach is mainly based on the use of the statistical concept of depth functions and will be called depth weighted index-flood model. The proposed model is composed of a number of ingredients. A depth function is used to introduce the hydrological similarity in the model. A continuous weight function is applied to the depth for more generality. The estimated parameters are continuously improved through an iterative procedure. Finally, an optimization algorithm is proposed to obtain the optimal weight function according to a criterion representing the performance of the model, e.g., the relative root mean-squared error. Note that in this study, the hydrological similarity is considered in terms of at-site distribution parameters (such as location, scale and shape parameters for a Generalized Extreme Value (GEV) distribution), estimated from at-site peak flow series. This similarity is quantified through the depth values of these parameters and it is incorporated in the model through the weights of regional growth curve estimation. The at-site L-moment statistics can also be considered as an alternative instead of the distribution parameters.

Chebana and Ouarda [2008] introduced an analogous procedure in regional estimation through the regression model where the similarity is mainly based on at-site flood peak quantiles. They

showed that the effectiveness of their proposed procedure depends on the choice of the weight function. To optimize this choice in terms of model performance and automate it for practical use, an algorithm-based procedure is recently proposed by Wazneh et al.[2013].

The present paper is organized as follows. Section 2 summarizes the different elements of the background necessary to introduce the propose approach. Section 3 describes the proposed depth weighted index-flood model and its optimization in its general form. In Section 4, the proposed approach is applied and compared to the traditional approaches based on a data set from Italy. The last section is devoted to the conclusions of this work.

2. Background

In this section, the background material required to introduce and apply the depth weighting approach are briefly presented. This section contains a number of basic notions as well as a brief reminder of the classical index-flood model.

2.1. Regional frequency analysis with the index-flood model

The index-flood modeling involves the use of L-moments as a statistical tool, for instance, to define discordance and homogeneity measures as well as to select distributions and to estimate their parameters. A detailed or brief presentation of L-moments can be found, for instance, in Hosking and Wallis [1997] or Chebana and Ouarda [2007] respectively.

For a random variable X, for instance the annual flood peak, with a cumulative distribution function F, the $(r+1)^{th}$ L-moment is defined by:

$$\lambda_{r+1} = \sum_{v=0}^{r} p_{r,v}^{*} \beta_v \text{ where } p_{r,v}^{*} = (-1)^{r-v} \binom{r}{v} \binom{r+v}{v} \text{ and } \beta_v = E\left\{ X \left[F(X) \right]^v \right\} \quad (1)$$

In RFA, dimensionless ratios between L-moments, called L-moment ratios (*LMR*s), are particularly useful. In particular the L_{cv} is equal to $\tau = \lambda_2 / \lambda_1$ while the other *LMR*s (L_{skew} and L_{kur}) are given by:

$$\tau_r = \frac{\lambda_r}{\lambda_2} \; ; r = 3,4 \; (L_{skew} \text{ for } r = 3 \text{ and } L_{kur} \text{ for } r = 4) \tag{2}$$

A brief description of the main steps of the index-flood model is presented herein.

Step 1: initial screening of data

The discordancy measure *D* defined below provides an initial screening of the data and indicates sites where the data may need close examination. For a region with *N* sites, the discordancy measure D_i for site *i* is defined by:

$$D_i = \frac{N}{3}\left(u_i - \bar{u}\right)^{'} S^{-1}\left(u_i - \bar{u}\right) \tag{3}$$

where $u_i = \left[L_{cv}^i \; L_{skew}^i \; L_{kur}^i \right]'$, $\bar{u} = \frac{1}{N}\sum_{i=1}^{N} u_i$ and $S = \sum_{i=1}^{N}\left(u_i - \bar{u}\right)\left(u_i - \bar{u}\right)'$. Hosking and Wallis [1997] advised to examine the data for sites with the largest D_i values (or generally larger than 3).

Step 2: heterogeneity measure

The homogeneity of a region, after excluding the discordant sites, is measured by the following *H* statistic:

$$H = \frac{\left(V - \mu_V\right)}{\sigma_V} \text{ such that } V = \frac{\sum_{i=1}^{N} n_i \left(L_{cv}^i - \bar{L}_{cv}\right)^2}{\sum_{i=1}^{N} n_i} \tag{4}$$

where n_i is the sample size at site *i*, *N is* the number of sites within the region, L_{cv}^i and $\bar{L}_{cv}$ are L_{cv} at site *i* and the average regional one respectively, and μ_V and σ_V are the mean and the standard deviation of the simulated regions respectively. Regions are considered as "acceptably

homogenous" if $H<1$, "possibly homogenous" if $1\le H<2$ and "definitely heterogeneous" if $H\ge 2$.

Step 3: delineation of homogeneous regions (sub-regions)

Hosking and Wallis [1997] considered cluster analysis based on site characteristics to form sub-regions.

Step 4: regional growth curve

This step consists in selecting the appropriate distribution (growth curve) corresponding to each identified sub-region. To this end, for each sub-region, a set of candidate distributions is fitted to the pooled standardized data where at-site series are standardized by the associated location parameter, such as the mean. A candidate distribution is selected if the following goodness-of-fit statistic $\left|Z^{DIST}\right|$ is lower than the test threshold of 1.64 which corresponds to the 90% normal quantile:

$$Z^{DIST}=\frac{\left(\bar{L}_{kur}-L_{kur}^{DIST}\right)}{\sigma_{kur}} \tag{5}$$

where L_{kur}^{DIST} is the L_{kur} of the candidate distribution, $\bar{L}_{kur}$ is the regional average one and σ_{kur} is the standard variation obtained from appropriately simulated regions.

Step 5: estimate the flood quantile at the target-site

The index-flood model estimates the flood quantile corresponding to a nonexceedance probability *p* (i.e. return period $T=1/(1-p)$) through the expression:

$$Q_{i_0}(p)=\mu_{i_0}q_p\left(\theta^R\right)\quad ;\ 0<p<1 \text{ and } \theta^R=\left(\theta_1^R,..,\theta_L^R\right) \tag{6}$$

where i_0 is the identifier of the target site, μ_{i_0} is the scale factor called the index flood such as at-site location parameter, $q_p(.)$ is the growth curve function and θ^R is the vector with *L*

components of regional growth curve parameters. The number of components L is usually 2 or 3 since the majority of fitted distributions in RFA have 2 or 3 parameters (i.e. location and scale parameters such as for the Log-Normal (LN) distribution or location, scale and shape parameters such as for the GEV).

The estimation of flood quantiles at the target site i_0 requires estimates of μ_{i_0} and θ^R. The index flood μ_{i_0} can be estimated using, for instance, either regression or spatial interpolation [Chokmani and Ouarda, 2004; Grover et al., 2002; Skøien et al., 2006]. However, regional parameters θ^R are estimated using a linear combinations of their at-site counterparts (approaches are described below).

2.2. Regional growth curve estimation parameters in traditional index-flood

The l^{th} regional growth curve parameter is a linear combination of the l^{th} at-site gauged estimates:

$$\hat{\theta}_l^R = \sum_{h=1}^{N'} \omega_h \hat{\theta}_l^h, \quad l = 1,...,L \tag{7}$$

where $\hat{\theta}_l^h$ is the l^{th} parameter obtained from the at-site standardized distribution at the h^{th} gauged site and ω_h is the weight associated to site h for $h = 1,...,N'$ such that $\omega_h > 0$ and $\sum_{h=1}^{N'} \omega_h = 1$ where N' is the number of sites in the homogenous sub-region after removing the discordant sites. Therefore, the weights ω_h in (7) influence the estimated parameters and hence the predicted value $Q_{i_0}(p)$ of the target site. These weights quantify the contribution of the gauged sites to the estimation at the target one. This contribution should be related to the similarity between sites and based on the hydrological information.

A number of special cases of the weights ω_h are employed in the literature. A common choice of the proportional weights (PW) is:

$$\omega_h = \frac{n_h}{n}; \quad h = 1,.., N' \tag{8}$$

where n_h is the record length for site *h* and $n = n_1 + ... + n_{N'}$ is the total record length [Hosking and Wallis, 1997]. In the following, the model associated to (8) is denoted by PW-index-flood. If the region is not perfectly homogeneous, it is possible that (8) attributes undue influence to sites with long records [Hosking and Wallis, 1997]. Hence, the uniform weights (UW) are proposed as an alternative [e.g., Jin and Stedinger, 1989; Minghui and Stedinger, 1989]. They lead to the simple mean of the parameters where:

$$\omega_h = \frac{1}{N'}; \quad h = 1,..., N' \tag{9}$$

The corresponding model is denoted by UW-index-flood. Another choice which limits the weight assigned to sites with longer records was suggested by Stedinger et al. [1992]:

$$\omega_h = \frac{n_h K / (n_h + K)}{\sum_{h=1}^{N'} n_h K / (n_h + K)}; \quad h = 1,.., N' \tag{10}$$

where *K* is a positive constant generally taken to be $K \approx 25$. However, the optimal value of *K* depends upon both the heterogeneity of the region and the sample sizes [Stedinger et al., 1992]. It is interesting to use this weight in regions when some sites have much longer records. The corresponding model is denoted by KW-index-flood for K-weights.

2.3. Mahalanobis depth function

This section provides an overview of the Mahalanobis depth function. This particular case of depth functions is used in the present work. A detailed description of the way the Mahalanobis depth is used to introduce the similarity between sites is presented in section 3.

The absence of a natural order to classify multivariate data led to the introduction of depth functions [Tukey, 1975]. They are used in a number of research fields, and were introduced in

water sciences by Chebana and Ouarda [2008]. For a given cumulative distribution function F on $\Re^d$ $(d \geq 1)$, a depth function is any non-negative bounded function which possesses a number of suitable properties. These properties fit well the RFA requirements and constraints. They are detailed and discussed in Chebana and Ouarda [2008].

A number of depth functions are available in the statistical literature such as Tukey, Mahalanobis and Oja [Zuo and Serfling, 2000]. A simple illustration on the computation of Tukey depth is presented in Chebana and Ouarda [2011]. For high dimensions, most of the current depth functions are quite cumbersome to compute. In this study, the Mahalanobis depth function is used to quantify the similarity between sites of a sub-region. This function is used for its convenient properties, interpretable values (values between 0 and 1) and for the easiness of its calculation even in high dimensions. Moreover, it is connected to the Mahalanobis distance already used in RFA approaches to build hydrological neighborhoods [Ouarda et al., 2001].

For a given cumulative distribution function F on $\Re^d$, the Mahalnobis depth (*MHD*) of a point x is defined by:

$$MHD(x;F) = \left(1 + d_{\Sigma}^2(x,\eta)\right)^{-1} \qquad x \text{ in } \Re^d \tag{11}$$

where η is a location measure of F, Σ is a dispersion matrix of F, and $d_{\Sigma}^2(x,\eta) = (x-\eta)' \Sigma^{-1} (x-\eta)$ is the Mahalanobis distance. An empirical version of the *MHD* of x with respect to the sample $\mathrm{X} = \{x_1,..,x_M;\ x_i \text{ in } \Re^d\}$ is defined by replacing F by a suitable empirical distribution $\hat{F}_{\mathrm{X}}$. In the context of the present paper, the empirical notation $MHD(x;\hat{F}_{\mathrm{X}})$ is denoted by $MHD(x;\mathrm{X})$.

3. Approach development

This section describes the proposed approach in its general form. An algorithm is presented to facilitate the approach presentation.

3.1. Description

The depth weighted index-flood model will be denoted DW-index-flood in the remainder of the paper. The model aims first to generalize the classical index-flood model in terms of representativety as well as flexibility. In addition, the performance of the proposed model is optimized with respect to the weight function. Note that in the DW-index-flood, we use a new iterative procedure to estimate the regional growth curve parameters θ^R given in (7) but the scale factor μ_{i_0} is estimated with the traditional approaches (regression or spatial approach).

In the DW-index-flood, the weights are not related to site record lengths as in (8), (9) and (10). They are defined through a continuous increasing weight function and a depth function for a better representativeness of the hydrological similarity by including more information. Indeed, the depth function is introduced to quantify the similarity between sites. More precisely, Mahalanobis depth (11) of each vector of the at-site (gauged sites) parameters $\hat{\theta}^h = \left(\hat{\theta}_1^h, ..., \hat{\theta}_L^h\right)$ $h = 1, .., N'$, is computed with respect to the set $\Theta = \left\{\hat{\theta}^1, .., \hat{\theta}^h, .., \hat{\theta}^{N'}\right\}$ formed by N' vectors of parameters of all N' gauged sites of the considered sub-region. In the context of the present paper, the depth of each site *h* with respect to Θ is denoted by $MHD\left(\hat{\theta}^h; \Theta\right)$. These depth values are obtained by considering η in (11) as a location parameter of the set Θ. The regional growth curve parameter θ^R plays the role of η. To get an accurate estimate of θ^R, the proposed approach utilizes an iterative procedure where in each iteration, the weights are updated with respect to the estimated θ^R at the previous iteration (see algorithm section 3.3).

For more clarity, consider the example of the LN distribution. In this case, the vector of at-site parameters of site *h* is $\hat{\theta}^h = \left(\hat{\alpha}^h, \hat{\xi}^h\right)$ where $\hat{\alpha}^h$ and $\hat{\xi}^h$ are respectively the estimated scale and location parameters. The values of the Mahalanobis depth are calculated with respect to the set of vector parameters of all gauged sites within a region, which provides an outward ordering of these sites through (α, ξ) where the center is the regional parameter θ^R. A site with high depth value indicates that its parameter vector is "close" to θ^R. Therefore, it is expected that its hydrological response (e.g., T-year peak-flow) is close to the regional hydrological response. This indicates a high similarity between this site and the set of sites of the sub-region. Therefore, this site should have a high contribution to the estimation of θ^R where the corresponding weight should be high [Chebana and Ouarda 2008]. The opposite is valid for a site with a low depth value. Figure 1a illustrates the way the depth function is used to quantify the similarity.

For a better control and amplification of these depth values, a continuous, smooth and increasing weight function is applied to the depth values. This allows the intervention of the user, e.g., based on other available information or previous experience, to increase or reduce the weight allocated for each site with respect to the "sorting" of sites, since the weight function is increasing. Adding the weight function makes the associated index-flood model more flexible and general. In addition, the weight function is useful in that the depth function allows only to "sort" sites, but since the difference between the values of depth could be small, these values may not be unable to show notable differences between site contributions. Figure 1b illustrates the way by which the weights of sites are calculated in the DW-index-flood model. Furthermore, in the absence of external information or relevant previous experience of the hydrologist with the current case study, and in order to avoid the subjective selection of the weight function and to ensure the high performance of the model, an optimality algorithm is introduced.

3.2. Performance Criteria

One of the objectives of this study is to assess and compare the performance of various weighting index-flood approaches i.e. UW, PW, KW and DW. The bias and the root mean square error are commonly used criteria for the evaluation of the performance of a RFA approach based on the prediction error of quantile estimate $Q(p)$ given in (6) [e.g., Haddad et al., 2011].

Assume that $\hat{Q}_{i_0}^{D}(p)$ and $\hat{Q}_{i_0}^{Tr}(p)$ are the estimated flood quantiles corresponding to a nonexceedance probability p for a target site i_0 using DW and one of the traditional weights (PW, UW or KW) respectively. Using the index-flood model (6), these flood quantiles can be estimated by:

$$\hat{Q}_{i_0}^{D}(p) = \hat{\mu}_{i_0} q_p\left(\left(\hat{\theta}^R\right)_D\right) \quad \text{and} \quad \hat{Q}_{i_0}^{Tr}(p) = \hat{\mu}_{i_0} q_p\left(\left(\hat{\theta}^R\right)_{Tr}\right) \tag{12}$$

where $\left(\hat{\theta}^R\right)_D$ and $\left(\hat{\theta}^R\right)_{Tr}$ are the estimates of the regional growth curve parameters using depth (procedure described below) and traditional weights given in (8), (9) or (10) respectively. Note that the estimate value of $\hat{\mu}_{i_0}$ in (12) is independent of the growth curve parameters and, therefore, it is the same for all approaches (traditional and depth ones). The best weights are hence those based on minimizing of the error of the growth curve $q_p(.)$. To quantify this error, a jackknife resampling procedure is used [e.g., Chernick, 2012]. It consists in considering each site as an ungauged one by removing it temporarily from the sub-region and calculating the regional growth curve parameters excluding this site using the different proposed weights (DW, PW, UW and KW). Then, over all sites in the sub-region, for each approach, we consider the relative bias (RB) and the relative root mean square error (RRMSE) given respectively by:

$$RB_p = \frac{1}{N'}\sum_{i=1}^{N'}\left(\frac{q_p\left(\theta^i\right) - q_p\left(\hat{\theta}^{R_{<-i>}}\right)}{q_p\left(\theta^i\right)}\right) \tag{13}$$

$$RRMSE_p = \sqrt{\frac{1}{N'}\sum_{i=1}^{N'}\left(\frac{q_p\left(\theta^i\right) - q_p\left(\hat{\theta}^{R_{<-i>}}\right)}{q_p\left(\theta^i\right)}\right)^2} \tag{14}$$

where $q_p(.)$ is the growth curve function corresponding to a nonexceedance probability *p*, θ^i is the at-site parameter vector, $\hat{\theta}^{R_{<-i>}}$ is the regional estimated parameter to the sub-region excluding site *i* (using the weights of comparison approaches) and N' is the number of sites in the homogeneous sub-region (after removing discordant sites). The jackknife resampling procedure has been used in the RFA literature in the CCA-regression model context (e.g., Ouarda et al. [2000]). It is also employed for the index-flood model, for instance, by Grover et al [2002] and Merz and Blöschl [2005]. Note that the accuracy of estimated performance criteria (13) and (14) depends on the quality of at-site estimates. More precisely, the performances of these criteria are based on the general RFA assumption that the observations at gauged sites are enough to correctly estimate quantiles corresponding to high return periods. To validate this assumption, each site should have a long record of measured floods (at least 10 years) and its historical data must be homogeneous, stationary and independent [e.g., Chebana and Ouarda, 2008].

3.3. Regional growth curve estimation in DW-index-flood

After screening the data, determining homogenous sub-regions and choosing the growth curve, we propose to use the following iterative procedure to calculate the regional growth curve parameters that optimize the performance of the index-flood model. The main steps of the estimation using DW-index-flood model are illustrated in Figure 2.

Let $\left(\theta^R\right)_D = \left(\theta_1^R,..,\theta_L^R\right)_D$ be the regional growth curve parameter vector using DW-index-flood. The iterative procedure for estimating $\left(\theta^R\right)_D$ is composed of the following steps. Recommendations, examples and more details for each step are given at the end of the algorithm. The procedure algorithm is composed of the following 5 steps:

1. Required elements: select a weight function φ, a performance criteria ψ and set the number of depth iterations k_{iter};
2. Initialization $\left(k=1\right)$: Initialize the weights $\omega^1 = \left(\omega_1,..,\omega_{N'}\right)$ in (7) in order to calculate the preliminary regional growth curve parameters denoted by $\left(\hat{\theta}_1^R\right)_D = \left(\hat{\theta}_{1,1}^R,...,\hat{\theta}_{1,L}^R\right)_D$ using the parameters $\hat{\theta}^1,..,..,\hat{\theta}^{N'}$ of N' gauged sites in the homogeneous sub-region;
3. For each iteration *k*, $k = 2,3,..,k_{iter}$:

   3.1. Compute the Mahalanobis depth of each vector of at-site estimate parameters *h* $\left(h=1,...,N'\right)$ with respect to the set of N' at-site estimates:

$$MHD\left(\hat{\theta}^h;\Theta\right) \text{ where } \Theta = \left\{\hat{\theta}^i; i=1,..,N'\right\},\ \eta = \left(\hat{\theta}_{k-1}^R\right)_D \text{ and } \Sigma_{L\times L} = diag\left(Var\left(\Theta\right)\right) \tag{15}$$

   where $\Sigma_{L\times L}$ is the $L\times L$ diagonal matrix composed by the variances of Θ.

   3.2. Compute the weight vector $\omega^k = \left(\omega_1^k,...,\omega_{N'}^k\right)$ corresponding to iteration step *k* for φ and Mahalanobis depth by:

$$\omega_h^k = \varphi\left[MHD\left(\hat{\theta}^h;\Theta\right)\right];\ h=1,...,N' \tag{16}$$

   3.3. Update the vector of regional growth curve parameters using the new weight vector from step 3.2:

$$\left(\hat{\theta}^R_{k,l}\right)_D = \sum_{h=1}^{N'} \omega_h^k \theta_l^h, \quad l = 1,...,L \tag{17}$$

4. Use the regional growth curve parameters estimated at the last iteration $\left(\hat{\theta}^R_{k_{iter}}\right)_D = \left(\hat{\theta}^R_{k_{iter},1},...,\hat{\theta}^R_{k_{iter},L}\right)_D$ to calculate the pre-selected performance criterion ψ.

5. Optimize ψ with respect to the weight function φ. The outputs of this step are the optimal weight function φ, the optimal regional growth curve parameters $\left(\hat{\theta}^R\right)_D$ and the value of the selected criterion.

A possible option for the initial step is to allocate one of the classical weights defined in (8), (9) or (10). Note that the optimal value of $\left(\hat{\theta}^R\right)_D$ does not depend on the choice of the initial step. The latter affects only the iteration number for depth convergence k_{iter}, but not the performance of DW-index-flood model. In this paper, the PW in (8) is used for the initial step. The number of iterations k_{iter} is fixed to ensure the depth convergence, generally k_{iter} = 10 is appropriate. The convergence here means that the difference between consecutive values of the performance criterion according to the iterations becomes negligible (close to 0).

According to the above reasons (section 2.3), in this study the Mahalanobis depth function is used to order the gauged sites. In principal, other depth functions can be used for this purpose. However, in the literature, usually one depth function is adopted for a given study and there are no comparisons or selection rules [e.g., Bárdossy and Singh, 2008; Chebana and Ouarda, 2011].

By construction, the regional growth curve parameters $\left(\hat{\theta}^R_{k_{iter}}\right)_D$ estimated at the last iteration (step 4) depend on several factors and particularly on the choice of the weight function φ. Consequently, the performance of the model depends on φ. In order to avoid the subjective choice and naïve selection procedures of φ, an optimization step is required (step 5). This step

sets also practical reasons to automate the choice of φ and the optimal regional growth curve parameters according to the objective performance criterion ψ such as RB (13) or RRMSE (14). To construct the objective function (criterion to be optimized) we consider the above performance criteria as parameterized functions through the coefficients of the weigh function φ. The complex and non-explicit form of the criteria to optimize suggest the use of zero-order optimization algorithms such as Nelder-Mead [Nelder and Mead, 1965] or Pattern search [Hooke and Jeeves, 1961; Torczon, 2000]. These algorithms are appropriate when the objective function is not differentiable or the gradient is unavailable and should be calculated by numerical methods (e.g., finite differences). Note that, the optimization procedure used in this study and the description of optimization algorithms are presented in Wazneh et al. [2013] where they are applied to regressive model RFA.

Below are the definitions of the two classes of φ, Gompertz and logistic, to be considered in this paper. These families of functions are regular, flexible, and have other suitable properties. The Gompertz function, denoted φ_G, was originally formulated by Gompertz [1825] for modeling human mortality. The function φ_G is flexible and is given by:

$$\varphi_G(x) = c\exp\left\{-ae^{-bx}\right\} \quad a,b,c>0\ ;\ x\in\Re \tag{18}$$

where *c* is its upper limit, *a* and *b* are two coefficients which respectively allow to translate and change the shape of the curve.

The Logistic function has similar properties of those of φ_G. It is given by:

$$\varphi_{\text{logistic}}(x) = \frac{c}{1+ae^{-bx}} \qquad a,b,c>0; x\in\Re \tag{19}$$

where the coefficients *c*, *a* and *b* play the same role as in φ_G.

The above weight functions are increasing and have an S shape (i.e. three phases). The starting phase is for sites with very low similarity and hence low contribution in the estimation step, then the growth phase is for sites with intermediate contribution and finally the stationary phase is for highly contributing sites. The advantage of such kind of weight functions is to include sites gradually in the estimation of regional growth curve parameter.

4. Application

In this section, the DW-index-flood with optimal weight functions is applied on a real world data set and its performance is compared to the performance of the traditional approaches such as UW, PW and KW-index-flood. The considered data set is from the hydrometric station network of the Island of Sicily, Italy.

4.1. Case study

The case study on which the comparison was carried out concerns the catchments of Sicily, the largest island in the Mediterranean Sea, which extends over an area of 25,700 km^2 and is located between 38° 08' N, 013° 23' E in the southern part of Italy, North East of Tunisia. The mean annual rainfall over the island is about 715 mm (period 1921-2004).

Annual maximum peak flow data of 50 stream flow gauging sites are available with record lengths ranging from 10 to 65 years. The conditions to apply frequency analysis, i.e. homogeneity, stationary and independence, were tested on the historical data of these stations in several studies [e.g., Cannarozzo et al., 1995; Cannarozzo et al., 2009; Noto and La Loggia, 2009]. The area of these catchments is larger than 10 km^2 and less than 2000 km^2 and their annual peak flows vary from 0.78 to 2380 m^3/s.

The catchment characteristics used in this study come from SIRI (Sisteme informative Regionale Idrologico - Hydrological Regional Information System) and extracted from the study of Viola et al. [2011]. It includes: the coordinates in the UTM system of the catchment centroids (X_{bar} and

Y_{bar}) in m, the basin area (AREA) in km^2, the mean basin slope (MBS) in %, the mean annual precipitation of the watersheds (MAP) in mm and the average basin elevation (ABE) in m. The basic statistics of the variables AREA, MBS and MAP are summarized in Table 1.

4.2. Results

The delineation of sub-regions is first illustrated then the choice of the growth curve distribution for the formed sub-regions and the comparison results between traditional and DW approaches are presented.

By considering the entire set of 50 gauged sites in Sicily as a single region, the discordancy statistic D_i of four gauging sites are larger than 3 as the critical value (Table 2). For the sites with Id numbers 3, 31 and 45, the high D_i (Eq. 3) is due to the high values of *LMR*s (see Table 2). However, the discordancy of the site with Id number 46 is due to its small area (24 km^2 see Table 1 in Viola et al. [2011]) and its small *LMR*s values (see Table 2). These four sites have also been identified as discordant by Noto and La Loggia [2009] and then removed from their database because of their *LMR*s values. In addition, the heterogeneity measure *H* of the region decreases significantly (from 9.6 to 7.76) when the discordant sites are removed. For the above reasons, these sites are also removed from the analysis in the present study. Based on the data of the remaining 46 sites, the heterogeneity measure *H* is *H* =7.76 which indicates that the entire Sicily Island cannot be considered as a homogeneous region. Therefore, smaller sub-regions need to be identified.

To define homogeneous sub-regions, a hierarchical clustering with standardized Euclidean metric was applied on the basis of the six site characteristics described above (X_{bar}, Y_{bar}, AREA, MBS, MAP, ABE) as well as the at-site L_{cv}. Including L_{cv} in the clustering approach increases the homogeneity degree in the formed sub-region. In this case, to allocate an ungauged site to the

obtained sub-regions, its L_{cv} should be estimated. To this end, a regression model on the basis of a set of site descriptors (such as the mean elevation) can be used [Laio et al., 2011]. Note that this clustering approach was considered in a number of RFA studies [e.g., Burn and Goel, 2000; Ouarda et al., 2008]. In order to obtain normality, a required condition for clustering, nonlinear transformations were applied to some of the clustering variables: a logarithmic transformation to AREA, MBS and MAP and a square root transformation to ABE. These transformations lead to a more symmetrical distribution of the values of the site characteristics at the 46 sites. Each one of the clustering variables was then standardized by the associated standard deviation. The number of sites in a cluster should be large enough to ensure the applicability of a regional analysis, but not too large to maintain cluster homogeneity.

In this study, the cluster analysis leads to three different hydrometric homogeneous sub-regions as shown in Figure 3a. The geographical locations of sites of these sub-regions are shown in Figure 3b. We observe that these sub-regions are not geographically contiguous. In particular, most sites of sub-region 3 are located in the South and North West areas of the Island. However, sites of sub-regions 1 and 2 are located respectively in the Central and Northern areas. Note that sub-region 3 contains sites with low MAP and ABE values; whereas sub-regions 1 and 2 mostly contain sites with high MAP and ABE values (see also Table 1 in Viola et al. [2011]).

The discordancy statistic and the heterogeneity measure were applied to each one of the identified sub-regions. Within the latter, there are no discordant stations (results not presented due to space limitations). Table 3 gives the regional average *LMR*s values and the heterogeneity measures H of the three sub-regions. The values of H indicate that these sub-regions can be considered as homogeneous ($H < 1$). Therefore, they can be considered in the estimation step of the index-flood model in (6). Note that the negative value of H corresponding to sub-region 3

indicates that the data of this region have a dispersion less than the amount we expect for a homogeneous region, i.e. a positive correlation exists between the sites within the region [Hosking and Wallis, 1997; Rao and Hamed, 1999].

In this study, four different probability distributions commonly used in RFA are considered as candidates for the growth curve, that is; the Generalized Extreme Value (GEV), the Generalized Logistic (GLO), the Pearson type-III (PE3) and the Generalized Pareto (GPA) distributions. These distributions were considered in a number of studies that used the same database [e.g., Noto and La Loggia, 2009]. The *Z* goodness-of-fit test is computed by simulating 500 regions having no cross correlation between sites. The results for each candidate distribution and for the three sub-regions are shown in Table 4. The values of *Z* such that $|Z|<1.64$ indicate that the GEV and the PE3 distributions are suitable for sub-regions 1 and 2, while the GLO and the GEV are suitable for sub-region 3. To select the appropriate distribution, the regional *LMR*s diagram is used (Figure 4). In the literature, numerous authors have used this diagram to identify the regional distribution [e.g., Hosking and Wallis, 1995; Vogel et al., 1993]. The proximity of the regional estimated L-moments to a particular candidate theoretical distribution in (L_{skew} , L_{kur}) space indicates the appropriateness of that distribution to describe the regional data. According to Figure 4, the GEV is chosen as a distribution for the three identified sub-regions. Note that Noto and La Loggia [2009] found the same distribution for the study area by using different sub-regions. The expression of the GEV growth curve corresponding to the nonexceedance probability *p* is given by:

$$q_p(.)=\begin{cases}\xi+\dfrac{\alpha}{\kappa}\left[1-\left(-\log p\right)^{\kappa}\right] & \text{if } \kappa\neq 0\\ \xi-\alpha\log\left(-\log p\right) & \text{if } \kappa=0\end{cases} \qquad (20)$$

where ξ, α and κ are respectively the location, scale and shape parameters. These parameters are estimated using L-moments [Hosking and Wallis, 1997]. Following the notation in (7), the regional parameter vector of growth curve is $\theta^R = \left(\kappa^R, \alpha^R, \xi^R\right)$, where the number of components is L=3 and the sizes of sub-regions are $N' = 20, 18$ and 8 respectively.

The proposed DW-index-flood is applied to each one of the identified sub-regions then compared to the traditional ones (UW, PW and KW). This comparison is based on the RB and the RRMSE, given in (13) and (14) respectively, of the growth curve for the nonexceedance probabilities $p = 0.9,\ 0.99, 0.995$ and 0.999. The results related to the different weighting approaches are summarized in Table 5. The positive constant K for KW-index-flood in (10) is taken as 25 as suggested by Stedinger et al. [1992]. The optimal weight functions in DW-index-flood, from the Gompertz and logistic functions respectively given in (18) and (19), are obtained using the algorithm-based procedure proposed in section 3. The optimization is performed with respect to both criteria RRMSE and RB. As expected and shown in a number of studies, for a given region, the regional growth curve estimation is generally more accurate for small nonexceedance probabilities p (i.e. small return period) [e.g., Hosking and Wallis, 1997]. Table 5 shows that the DW-index-flood with optimal φ leads to more efficient estimates in terms of RB and RRMSE than those obtained using the other approaches. The improvement is more significant in sub-region 3 than in the two other regions and it increases for large nonexceedance probabilities p (i.e. large return periods). As previously indicated, sites of this sub-region present a significant cross-correlation, which could be the reason of this improvement. Therefore, the DW-index-flood performs better for regions with cross-correlation between sites, which is not the case with the traditional index flood model [Hosking and Wallis, 1997-P.8]. This can be explained by the fact that the Mahalanobis depth in (15) takes into account the variability of at-site parameters

(location, scale and shape) across all sites of a given sub-region. Furthermore, by comparing the results of the three traditional approaches UW, PW and KW, we observe that they are very close in terms of RB and RRMSE. This result is found to be the same for DW using the optimal weight functions φ_G or $\varphi_{\log istic}$.

The optimal weight functions corresponding to each sub-region for the estimation of $q_{0.995}(.)$ are shown in Figure 5. For both sub-regions 1 and 2, the two optimal weight functions using φ_G and $\varphi_{logistic}$, have an S shape whose lower boundary does not reach 0. This means that there are no stations to exclude within the corresponding sub-regions. Moreover, the upper extremity of the Gompertz curve (Figure 5a) in these two sub-regions does not reach 1. This behavior can be explained by the absence of sites which are perfectly similar to the target site. This is more realistic for some sub-regions. According to Figure 5, the optimal weight function (φ_G and $\varphi_{logistic}$) corresponding to sub-region 3 does not have the S shape. It has almost a straight line ranging from 0.1 and 0.2, which means that all the included sites have approximately the same weight in the estimation of the growth curve parameters. The previously indicated high positive correlation between sites could be the reason. Therefore, the DW-index flood recognizes the correlation between sites through the shape of the optimal weight function. The shape of the optimal function (16) is related to the Mahalanobis depth which takes into consideration more information such as the variability of the data at different sites through their parameters. This is not the case for the classical approaches which only employ the sites record length. Note that, in this study, the form of the weight function is independent of the value of the nonexceedance probability, i.e. nearly the same form for p = 0.9, 0.99, 0.995 and 0.999. This is true because of the small variation of the depth values according the variation of p.

In order to illustrate the various elements of the approach, a figure analogous to Figure 1 is developed. Figure 6 illustrates the Mahalanobis depth and the weight of sites of sub-region 1 using the Gompertz function. In particular, sites 14, 20 and 28 are located in the center of the GEV-parameters space. These sites have high depth values and hence have a high contribution (weight) in the estimation of the regional growth curve. These high depth values can be due to the low discordancy D_i values (see Table 2) of these sites. However, the opposite is true for sites 8 and 29 which have low depth values, are located in the border and have low weights.

Figure 7 presents the evolution of the performance criteria for nonexceedance probability p =0.995 for each sub-region as a function of the iteration number k. For the traditional PW-index-flood (the most commonly used in the literature), the criteria are represented by straight lines, since they are independent of the iteration number. This Figure illustrates the superiority of the DW over the PW-index-flood. Moreover, for sub-regions 1 and 2, we observe a rapid convergence to the criterion values in Table 5 after only 3 iterations (Figures 7a and 7b), whereas, for sub-region 3 (Figure 7c) it is reached after more than 7 iterations. These results could be due to the quality and size of sub-region 3. In other words, the absence of large correlation between data at sites of sub-regions 1 and 2 (value of H close to 0) have caused the rapid convergence of the algorithm in these two sub-regions, in contrast to sub-region 3.

To study the impact of depth iterations on the performance of the DW-index-flood, this model is applied to the three sub-regions but without iterations on the depth, i.e. $k_{iter} = 2$ in step 3 of the regional parameters estimation algorithm. The outputs of this application, with $\varphi = \varphi_G$ and p = 0.995, are shown in Figure 8 and are compared to the traditional PW model. These results indicate that, in both cases with and without iterations, the DW performs better than PW in terms of both RB and RRMSE. Furthermore, the iterated DW-index-flood produces leads to more

efficient parameter estimates than the direct method, which affects the model performance. Note that similar results are found for the logistic weight functions and for different nonexceedance probabilities *p*.

Using the DW-index-flood model, the optimal weight function φ is obtained by optimizing a criterion ψ which could vary with the nonexceedance probability *p* (step 5 of the estimation algorithm). Therefore, the weight vector in (16) could vary with *p*, resulting in the variation of regional parameters $\left(\hat{\theta}^R\right)_D$ in (17) with *p* as well. Figure 9 illustrates this variation for sub-region 1 with $\varphi = \varphi_G$. To draw this Figure, each site *i* ($i = 1,..,N'$) is temporally excluded from the sub region and then the corresponding regional parameters, denoted by $\theta^{R_{<-i>}} = \left(\kappa^{R_{<-i>}}, \alpha^{R_{<-i>}}, \xi^{R_{<-i>}}\right)$, are estimated using DW. This Figure shows a very small variation of the regional parameters as a function of *p* which could be due to numerical issues. More precisely, for different *p* values, we have almost the same shape κ and a very small variation for the scale α and the location ξ (variation ranges between 0 and 0.05). Note that the same result is found for the other two sub-regions and when using the logistic weight function.

Figure 10 illustrates the regional estimates in sub region 1 using different models. The parameters estimated by DW correspond to $p = 0.995$ and φ_G weight function. It is generally observed that these parameters are closer to the at-site values than those estimated by the traditional models. This is because of the flexibility in the weight calculation when using DW. Theoretically, this last result is validated by both criteria RB and RRMSE calculated in Table 6. Figure 10 shows also a negligible variation of regional parameters estimated by traditional models according to the excluded site. This means that, by using traditional models, the estimated regional parameters are

almost equal for any target site in the sub-region which then naturally affects the performance of these models.

5. Conclusions and future research directions

The present paper aims to optimize the estimation of flood quantiles when using the regional index-flood model. To this end, the statistical notion of depth function is introduced in the model. More precisely, the proposed approach employs depth functions to measure the similarity between sites and then weight the at-site estimate parameters in order to calculate the regional growth curve parameters. The proposed approach takes into consideration regional similarity within the region since it includes more hydrological information than the record length and does not adopt uniform weights. Furthermore, the iterative procedure allows improving the performance of the model.

The optimal weight function φ in the proposed DW-index-flood is obtained using the algorithm-based procedure with respect to both RB and RRMSE criteria. In addition to reaching the best model performance, the optimization of the weight function first allows the practical use of DW-index-flood and second avoids the subjective intervention of the user. The obtained results, from the Island of Sicily in Italy, show the superiority of the DW-index-flood model in terms of performance. The study of the three formed sub-regions shows an association between the inter-site correlation, the shape of the optimal weight function and the computation convergence speed. The proposed approach is able to identify cross-correlation in the region and provides a significant performance improvement.

Even though the proposed approach is superior to the traditional ones in terms of flexibility, representativeness, generality and optimality, it has some limitations. For instance, it is composed of several steps and includes several tools, and consequently, it could be time consuming. In the

present paper, most of the elements related to the proposed procedure are treated; however, others are worth developing in future efforts. For instance;

1- In this study, cluster analysis of site characteristics is used to define the homogeneous sub-regions. This method may be incompatible with the proposed estimation model which is based on the statistical depth function. Therefore, in future work, it would be reasonable to include the depth function in the clustering part as well (i.e. delineation step of RFA). Such a study would ensure compatibility between the two stages of RFA (estimation and delineation) and could provide improved results;

2- Another direction of future work could be a thorough sensitivity study of the impact of different factors which may affect the performances of the proposed model. Such factors include: the estimation method of the distribution parameters, the fitted regional distribution, the use of L-moments instead of the parameters to evaluate depths, as well as the choice of the depth function;

3- Further investigations are needed for the approach to study special situations. Such situations include regions with inter-dependence between sites, small number of sites or a low homogeneity level.

Acknowledgments

Financial support for this study was graciously provided by the Natural Sciences and Engineering Research Council (NSERC) of Canada and the Canada Research Chair Program. The authors are grateful to the Editor, the Associate Editor and the anonymous reviewers for their valuable comments and suggestions.

References

Bárdossy, A., Singh, S.K., 2008. Robust estimation of hydrological model parameters. Hydrology and Earth System Sciences, 12(6): 1273-1283.

Burn, 1990a. An appraisal of the "region of influence' approach to flood frequency analysis. Hydrological Sciences Journal/Journal des Sciences Hydrologiques, 35(2): 149-165.

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

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

Cannarozzo, M., D'Asaro, F., Ferro, V., 1995. Regional rainfall and flood frequency analysis for Sicily using the two component extreme value distribution. Hydrological Sciences Journal/Journal des Sciences Hydrologiques, 40(1): 19-42.

Cannarozzo, M., Noto, L.V., Viola, F., La Loggia, G., 2009. Annual runoff regional frequency analysis in Sicily. Physics and Chemistry of the Earth, 34(10-12): 679-687.

Castellarin, A., Burn, D.H., Brath, A., 2001. Assessing the effectiveness of hydrological similarity measures for flood frequency analysis. Journal of Hydrology, 241(3-4): 270-285.

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

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

Chebana, F., Ouarda, T.B.M.J., 2011. Depth-based multivariate descriptive statistics with hydrological applications. Journal of Geophysical Research: Atmospheres, 116(D10): D10120.

Chernick, M.R., 2012. The jackknife: A resampling method with connections to the bootstrap. Wiley Interdisciplinary Reviews: Computational Statistics, 4(2): 224-226.

Chokmani, K., Ouarda, T.B.M.J., 2004. Physiographical space-based kriging for regional flood frequency estimation at ungauged sites. Water Resources Research, 40(12): 1-13.

Dalrymple, T., 1960. Flood Frequency Analysis. Geological Survey water-supply paper (no. 1543-A) U.S. Geological Survey, Washington 77 pp.

Gompertz, B., 1825. On the nature of the function expressive of the law of human mortality, and on a new mode of determining the value of life contingencies. Philos. Trans. R. Soc. Lond., 115: 513-585.

GREHYS, 1996a. Presentation and review of somemethods for regional flood frequency analysis. Journal of Hydrology, 186: 63-84.

Grover, P.L., Burn, D.H., Cunderlik, J.M., 2002. A comparison of index flood estimation procedures for ungauged catchments. Canadian Journal of Civil Engineering, 29(5): 734-741.

Gupta, V.K., Dawdy, D.R., 1995. Physical interpretations of regional variations in the scaling exponents of flood quantiles. Hydrological Processes, 9(3-4): 347-361.

Haddad, K., Rahman, A., Weeks, W., Kuczera, G., Weinmann, P.E., 2011. Towards a new regional flood frequency analysis method for Western Australia, pp. 3788-3795.

Hooke, R., Jeeves, T.A., 1961. Direct search solution of numerical and statistical problems. Journal of the Association for Computing Machinery, 8(2): 212-229.

Hosking, J.R.M., Wallis, J.R., 1993. Some statistics useful in regional frequency analysis. Water Resources Research, 29(2): 271-281.

Hosking, J.R.M., Wallis, J.R., 1995. A comparison of unbiased and plotting-position estimators of L moments. Water Resources Research, 31(8): 2019-2025.

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

Jin, M., Stedinger, J.R., 1989. Flood frequency analysis with regional and historical information. Water Resources Research 25(5): 925-936.

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

Merz, R., Blöschl, G., 2005. Flood frequency regionalisation - Spatial proximity vs. catchment attributes. Journal of Hydrology, 302(1-4): 283-306.
Minghui, J., Stedinger, J.R., 1989. Flood frequency analysis with regional and historical information. Water Resources Research, 25(5): 925-936.
Nelder, J.A., Mead, R., 1965. A simplex method for function minimization. Comput. J., 7: 308-313.
Noto, L.V., La Loggia, G., 2009. Use of L-moments approach for regional flood frequency analysis in Sicily, Italy. Water Resources Management, 23(11): 2207-2229.
Ouarda, T.B.M.J., Ashkar, F., El-Jabi, N., 1993. Peaks over threshold model for seasonal flood variations. In: Hydrology, E. (Ed.). ASCE Publications, New York, USA, pp. 341-346.
Ouarda, T.B.M.J. et al., 2008. Intercomparison of regional flood frequency estimation methods at ungauged sites for a Mexican case study. Journal of Hydrology, 348(1-2): 40-58.
Ouarda, T.B.M.J., Girard, C., Cavadias, G.S., Bobée, B., 2001. Regional flood frequency estimation with canonical correlation analysis. Journal of Hydrology, 254(1-4): 157-173.
Ouarda, T.B.M.J., Hache, M., Bruneau, P., Bobee, B., 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., Lang, M., Bobée, B., Bernier, J., Bois, P., 1999. Analysis of regional flood models utilized in France and Quebec. Revue des Sciences de l'eau, 12(1): 155-182.
Pandey, G.R., Nguyen, V.T.V., 1999. A comparative study of regression based methods in regional flood frequency analysis. Journal of Hydrology, 225(1-2): 92–101.
Rao, R., Hamed, K., 1999. Flood Frequency Analysis. CRC Press, 330 pp.
Skøien, J.O., Merz, R., Blöschl, G., 2006. Top-kriging - Geostatistics on stream networks. Hydrology and Earth System Sciences, 10(2): 277-287.
Stedinger, J.R., Vogel, R.M., Foufoula-Georgiou, E., 1992. Frequency analysis of extreme events Handbook of hydrology. McGraw-Hill, New York, 18.1-18.66 pp.
Torczon, V., 2000. On the Convergence of Pattern Search Algorithms. SIAM Journal on Optimization, 7(1): 1-25.
Tukey, J.W., 1975. Mathematics and the picturing of data. Proceedings of the International Congress of Mathematicians, 2: 523-531.
Viola, F., Noto, L.V., Cannarozzo, M., La Loggia, G., 2011. Regional flow duration curves for ungauged sites in Sicily. Hydrology and Earth System Sciences, 15(1): 323-331; http://www.hydrol-earth-syst-sci.net/15/323/2011/hess-15-323-2011.pdf.
Vogel, R.M., McMahon, T.A., Chiew, F.H.S., 1993. Floodflow frequency model selection in Australia. Journal of Hydrology, 146(C): 421-449.
Wazneh, H., Chebana, F., Ouarda, T.B.M.J., 2013. Optimal depth-based regional frequency analysis. Hydrol. Earth Syst. Sci., 17(6): 2281-2296.
Zuo, Y., Serfling, R., 2000. General notions of statistical depth function. Annals of Statistics, 28(2): 461-482.

Table 1. Descriptive statistics of selected variables.

Variable	Minimum	Mean	Maximum	Standard deviation
AREA	9.40	218.23	1792	335.84
MBS	2.23	15.51	65.55	14.35
MAP	429.50	716.71	1187.40	172.76

Table 2. Summary statistics of annual maximum peak flood data used in this study (RL is the record length in years, Q_m is the average of annual maximum peak flood quantile in m^3/s, the three L-moment ratios and the discordancy measure *D* for the whole region).

Id site	*RL* (years)	Q_m (m^3/s)	L_{cv}	L_{skew}	L_{kur}	*D*
1	16	40.14	0.329	0.046	0.087	2.17
2	10	26.94	0.374	0.252	0.036	0.97
3	**10**	**43.00**	**0.753**	**0.825**	**0.818**	**3.23**
4	10	20.25	0.446	0.194	0.122	0.69
5	16	50.79	0.407	0.172	0.107	0.57
6	17	90.18	0.502	0.296	0.052	0.96
7	16	158.25	0.624	0.492	0.222	1.23
8	12	31.56	0.438	0.090	-0.105	1.56
9	22	142.63	0.385	0.398	0.169	1.79
10	54	462.3	0.358	0.249	0.157	0.17
11	17	62.26	0.470	0.226	0.101	0.46
12	45	11.08	0.263	0.153	0.077	0.63
13	32	17.48	0.326	0.101	0.124	1.40
14	24	30.96	0.433	0.321	0.177	0.09
15	64	99.29	0.359	0.326	0.204	0.45
16	21	36.22	0.372	0.311	0.153	0.50
17	33	48.25	0.446	0.245	0.084	0.29
18	22	40.06	0.763	0.728	0.625	1.78
19	21	16.85	0.735	0.751	0.624	1.45
20	19	12.79	0.463	0.334	0.286	0.30
21	16	97.08	0.371	0.118	0.025	0.59
22	16	14.38	0.655	0.672	0.533	0.84
23	20	79.70	0.415	0.409	0.253	0.57
24	31	215.65	0.392	0.244	0.215	0.43
25	13	60.49	0.364	0.301	0.212	0.23
26	25	137.73	0.352	0.353	0.259	0.58
27	31	63.66	0.385	0.390	0.358	0.61
28	31	557.6	0.467	0.304	0.192	0.08
29	22	55.29	0.559	0.597	0.410	0.86
30	21	34.01	0.438	0.359	0.252	0.03
31	**11**	**6.14**	**0.871**	**0.854**	**0.783**	**3.64**
32	14	15.11	0.388	0.143	-0.008	0.56
33	13	17.46	0.526	0.342	0.125	0.59
34	10	27.91	0.447	0.387	0.215	0.33
35	35	308.04	0.356	0.096	0.093	1.32
36	31	583.79	0.459	0.489	0.421	0.52
37	15	52.16	0.637	0.601	0.553	1.12
38	14	25.76	0.706	0.598	0.408	0.98
39	10	18.24	0.625	0.389	0.110	1.66
40	23	45.36	0.426	0.155	0.067	0.73
41	11	39.50	0.540	0.361	0.218	0.25
42	11	62.19	0.706	0.541	0.237	2.04
43	22	107.51	0.372	0.185	0.135	0.41
44	18	244.38	0.329	0.296	0.217	0.48
45	**17**	**92.79**	**0.470**	**0.726**	**0.628**	**3.28**
46	**12**	**21.32**	**0.202**	**0.345**	**0.371**	**3.15**
47	34	182.36	0.350	0.212	0.149	0.25
48	29	507.04	0.512	0.392	0.144	0.89
49	12	9.10	0.288	0.142	0.014	0.62
50	13	17.69	0.216	0.140	0.190	1.67

Bold character indicates the site has *D* that exceeds the critical value of 3 and could be considered discordant.

Table 3. Regional average *LMR*s values and heterogeneity measures *H* under the proposed hypothesis of three sub-regions.

	Number of sites	$\overline{L}_{cv}$	$\overline{L}_{skew}$	$\overline{L}_{kur}$	*H*
Sub-region 1	20	0.4397	0.3077	0.1917	0.01
Sub-region 2	18	0.3488	0.2321	0.1753	0.20
Sub-region 3	8	0.6891	0.5911	0.4532	-1.79

Table 4. *L*-kurtosis and *Z* measure values for the four different candidate distributions and for each one of the proposed sub-regions.

Distribution	Sub-region 1		Sub-region 2		Sub-region 3	
	$\overline{L}_{kur}$	*Z*	$\overline{L}_{kur}$	*Z*	$\overline{L}_{kur}$	*Z*
GLO	0.2455	2.22	0.2116	2.41	0.4682	**-1.27**
GEV	0.1996	**1.21**	0.1779	**0.75**	0.4660	**-0.74**
PE3	0.1887	**-1.26**	0.1592	**-0.79**	0.3526	-3.05
GPA	0.1471	-1.69	0.0958	-2.66	0.4536	-1.69

Bold character indicates the distribution may be accepted as a regional distribution for each sub-region.

Table 5. Growth curve estimation result in % with the various weighting approaches. RB and RRMSE are calculated versus locally estimated growth curve.

		nonexceedance probability *p*																			
		0.9	0.99	0.995	0.999	0.9	0.99	0.995	0.999	0.9	0.99	0.995	0.999	0.9	0.99	0.995	0.999	0.9	0.99	0.995	0.999
		UW $\omega_h = \frac{1}{N'}$				PW $\omega_h = \frac{n_h}{n}$				KW $\omega_h = \frac{\frac{n_h K}{n_h + K}}{\sum_{h=1}^{N'} \frac{n_h K}{n_h + K}}$				DW $\varphi_{\log istic}$				DW φ_G			
Sub-region 1	RB	-0.8	-1.8	-3.2	-9.0	0.2	-0.9	-2.4	-8.4	-0.3	-1.6	-3.1	-9.2	*0.1*	*0.4*	*0.5*	*0.8*	**0.1**	**0.4**	**0.4**	**0.8**
	RRMSE	6.5	25.6	33.3	55.0	6.2	24.8	32.6	53.3	6.2	25.1	33.0	54.1	*6.0*	*24.0*	**30.5**	**45.6**	**5.8**	**23.8**	*30.6*	*45.7*
Sub-region 2	RB	1.9	1.0	2.2	-0.9	0.2	-1.4	-2.7	-0.9	1.9	-1.3	-2.3	-6.0	*0.1*	*0.4*	*0.5*	*0.8*	**0.1**	**0.4**	**0.5**	**0.7**
	RRMSE	8.7	28.0	33.5	47.5	7.3	25.0	31.2	46.7	7.4	25.0	31.1	46.2	*6.9*	**22.8**	**28.2**	**40.5**	**6.9**	*23.0*	*28.5*	*40.9*
Sub-region 3	RB	-2.4	-5.5	-9.0	-21.1	-2.0	-6.3	-9.6	-22.8	-3.2	-5.5	-8.2	-19.4	*1.2*	**0.1**	**-0.1**	**0.3**	**-0.3**	*0.1*	*0.1*	*0.5*
	RRMSE	16.6	17.0	28.9	65.7	16.5	16.1	27.5	63.4	16.8	15.8	26.9	61.2	*13.6*	*11.5*	*17.9*	*34.9*	**13.4**	**11.3**	**17.8**	**34.6**

The best result for each sub-regions in character bold and the second result in italic.

Table 6. Parameters estimation result using different approaches

Parameters	Criterion	UW	PW	KW	DW
κ	RB	0.327	0.299	0.307	**-0.018**
	RRMSE	1.196	1.215	1.219	**0.427**
α	RB	-0.061	-0.035	-0.047	**0.015**
	RRMSE	0.057	0.052	0.054	**0.032**
ξ	RB	-0.033	-0.045	-0.038	**-0.007**
	RRMSE	0.024	0.025	0.024	**0.006**

Bold character indicate best criteria results

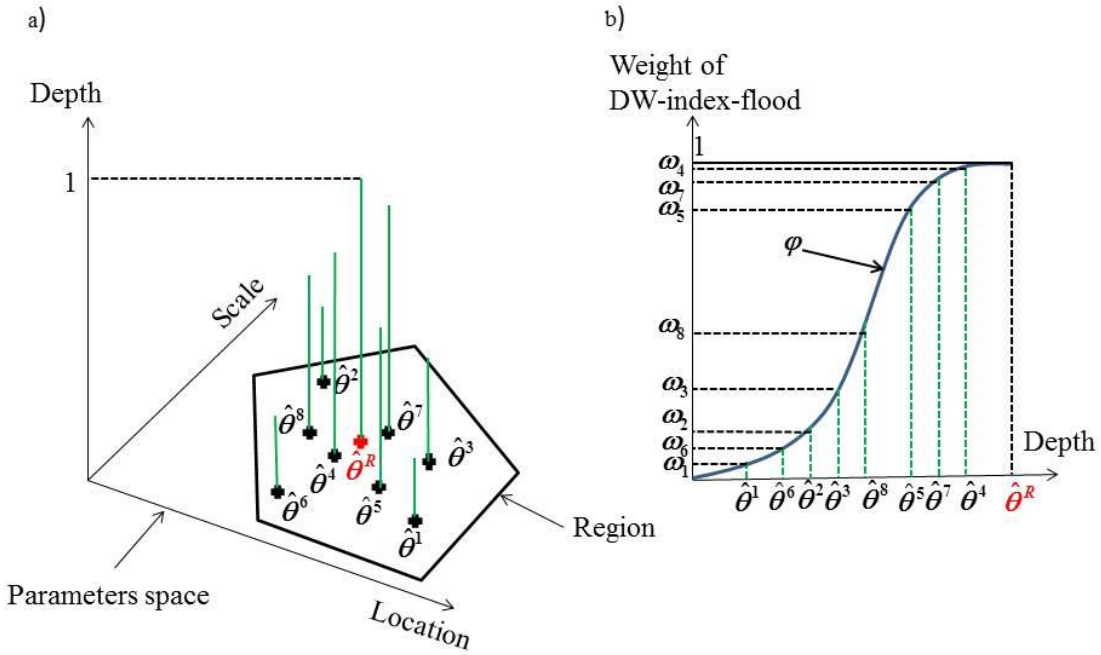


Figure 1. Illustration of the way by which a) the depth function is used to compute the similarity between sites of a sub-region and b) the weights are calculated in the DW-index-flood model. $\hat{\theta}^i\,(i=1,..,8)$ represent the at-site parameters and $\hat{\theta}^R$ represents the regional growth curve parameter.

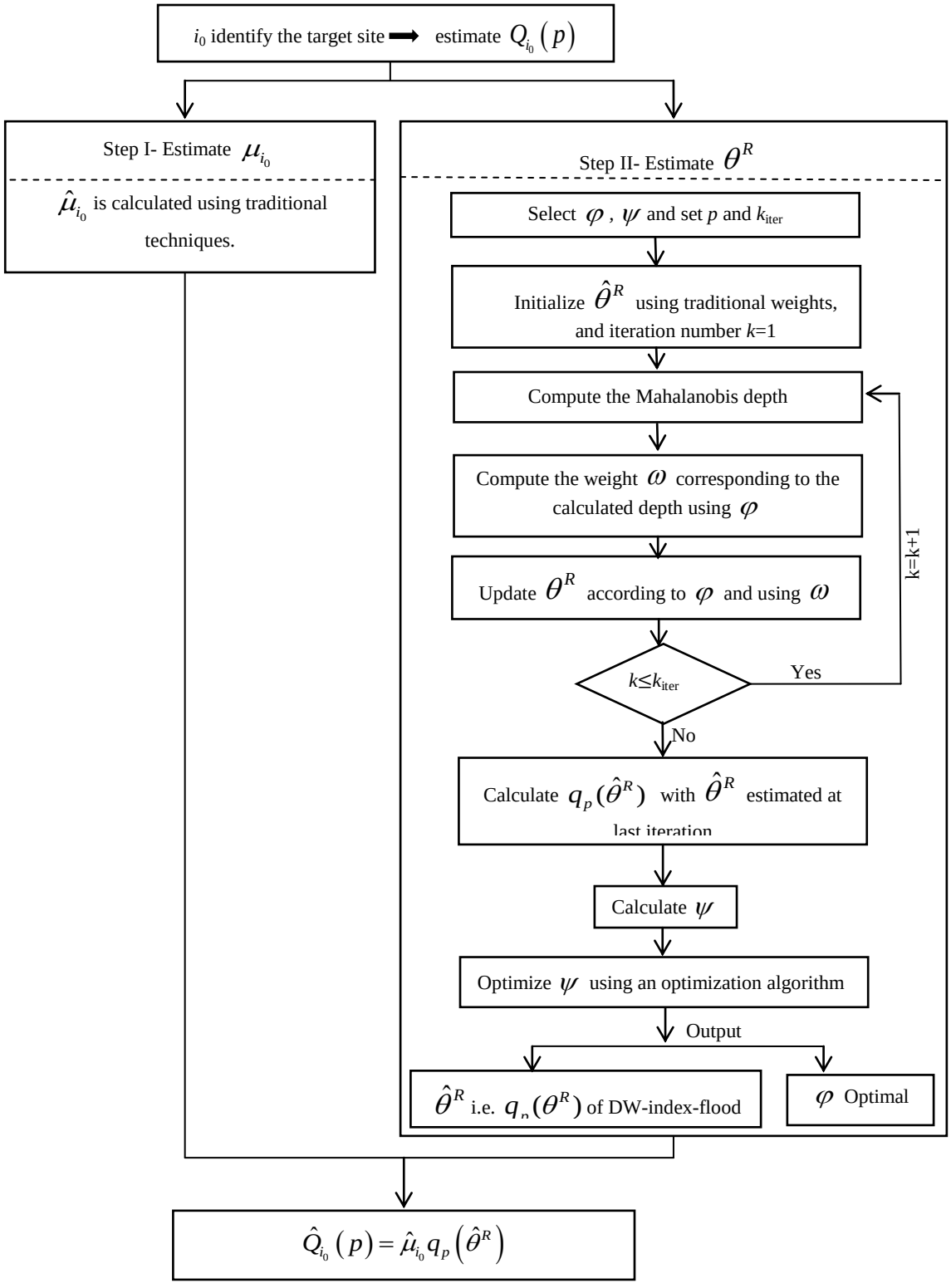
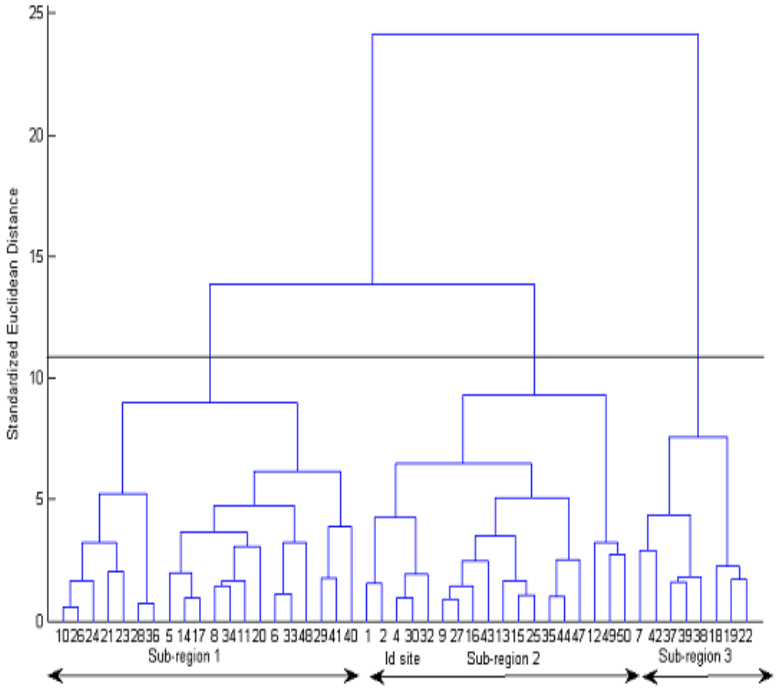


Figure 2. Estimation algorithm using DW-index-flood model.

a)



b)

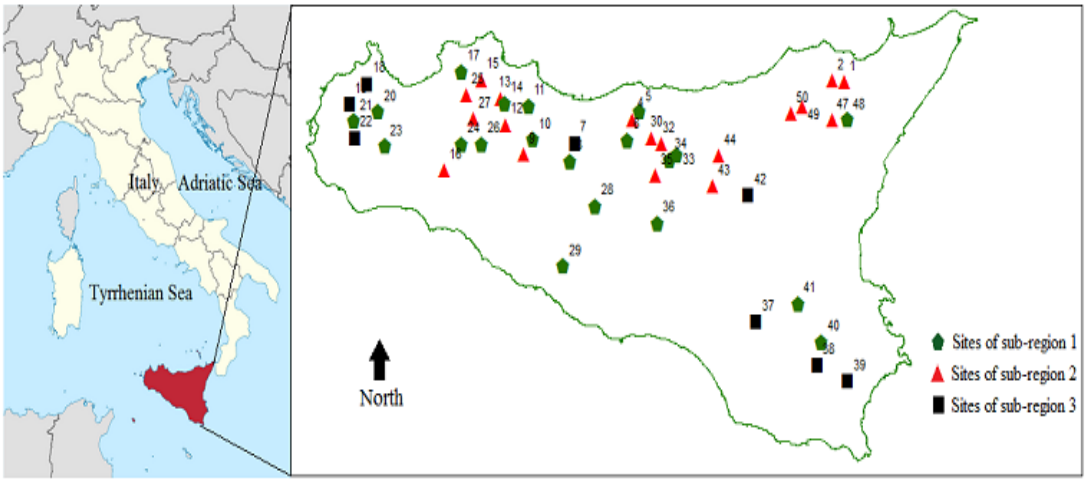


Figure 3. a) Hierarchical classification of stations. Sites with Id numbers 3, 31, 45 and 46 are not included in the dendrogram. b) Geographical location of the studied sites in the Sicily Island, Italy.

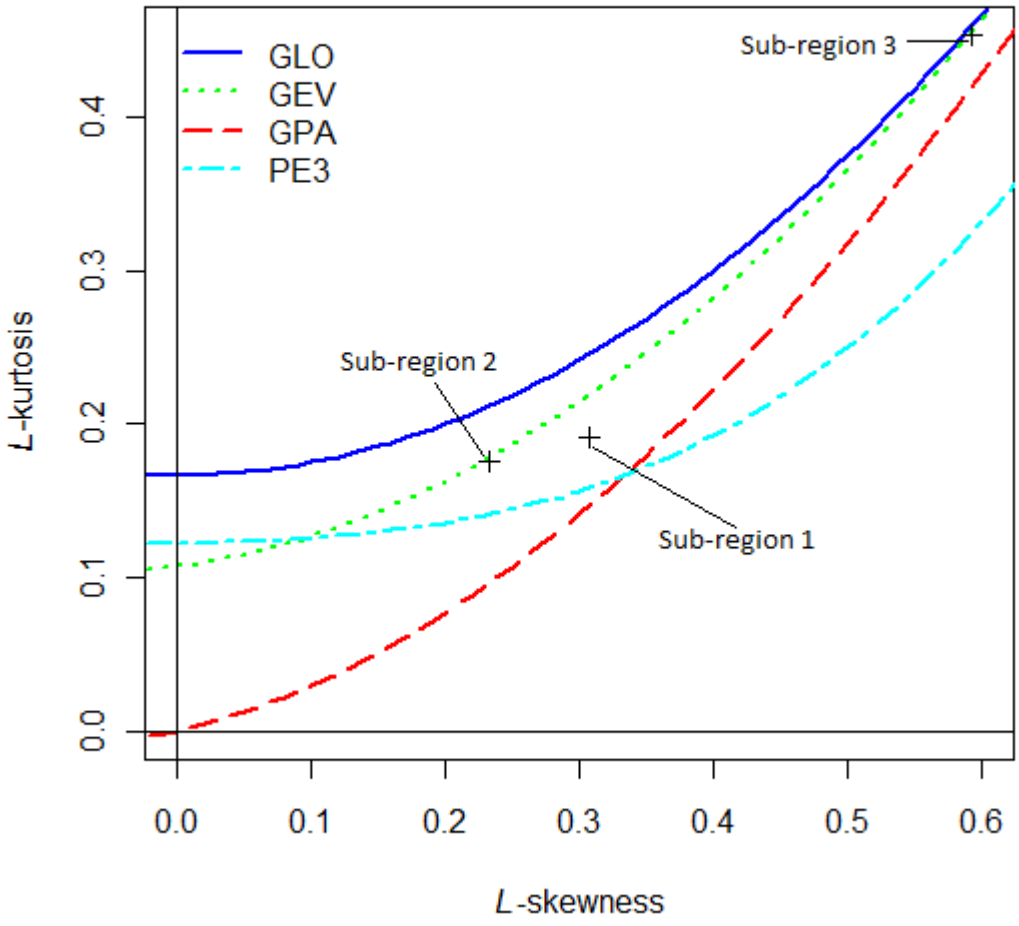
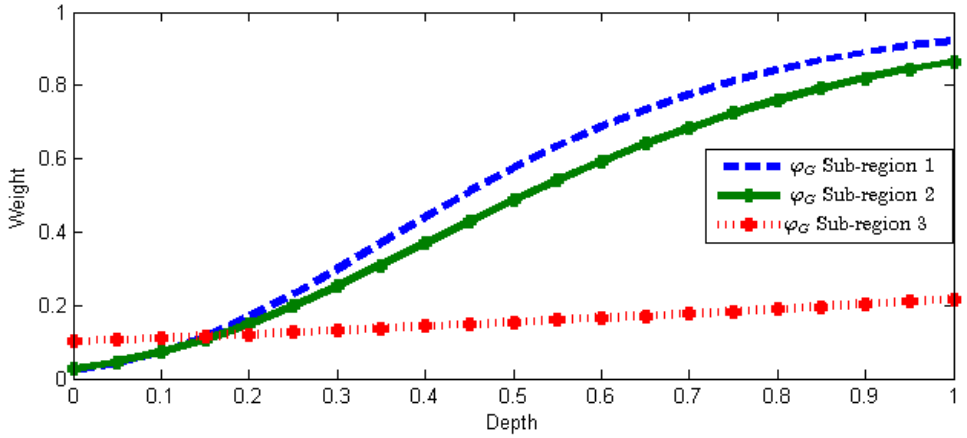


Figure 4. Regional average L-moments ratio diagram for the three formed sub-regions.

a)



b)

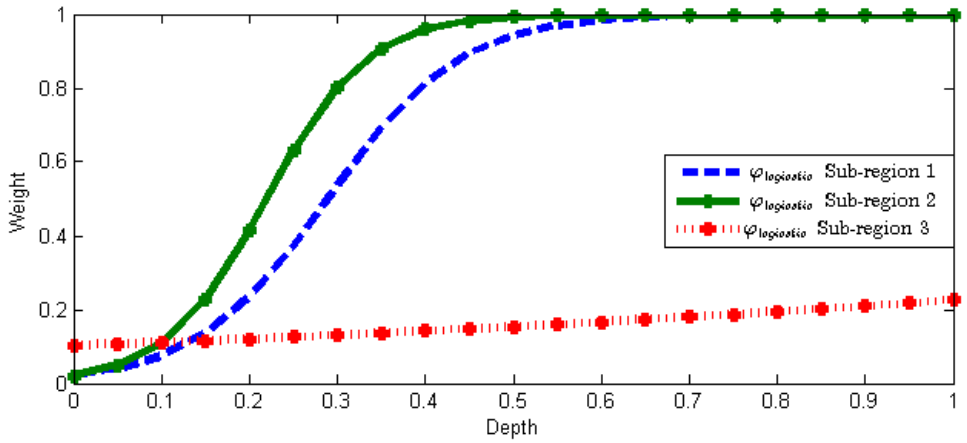


Figure 5. Optimal weight function in DW-index-flood of $q_{0.995}(.)$ for all regions using (a) φ_G and (b) $\varphi_{\log istic}$.

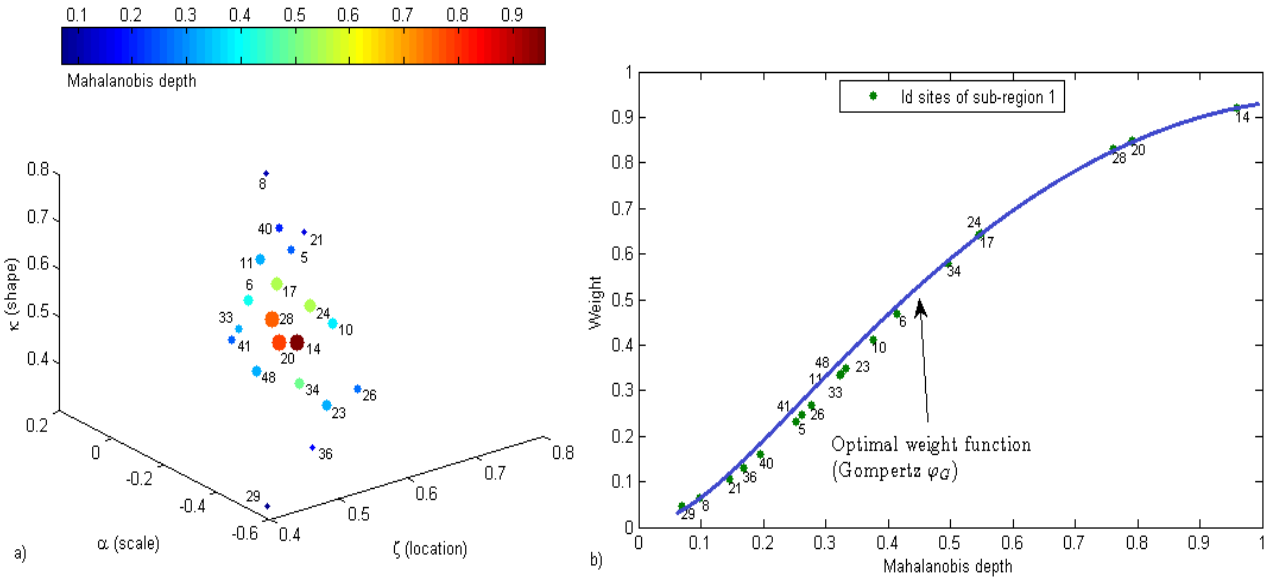


Figure 6. a) The Mahalanobis Depth of gauged sites of sub-region 1. The size of circle that represents the site is proportional to its depth. b) The associated Gompertz weights allocated for each site of sub-region 1.

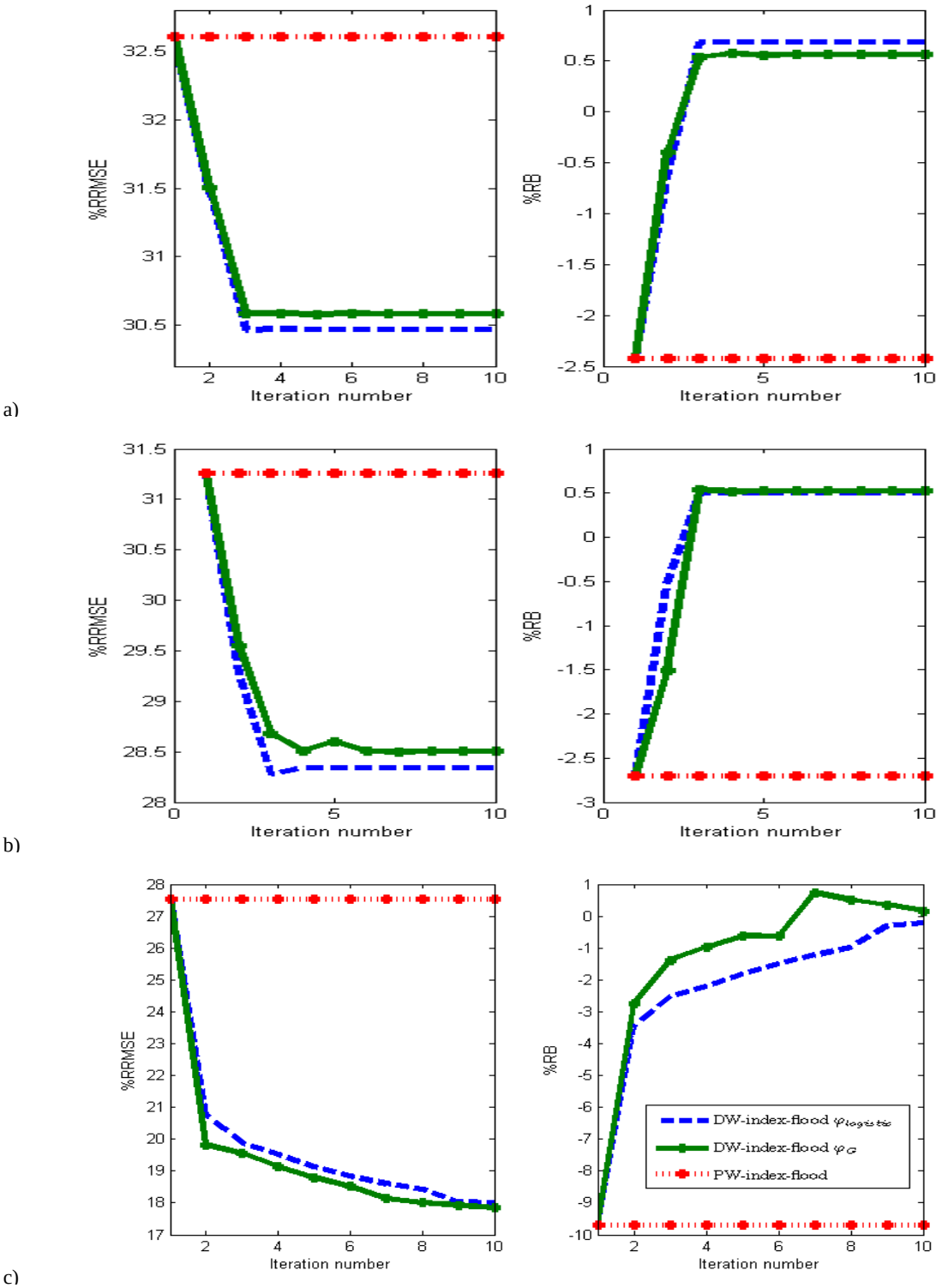


Figure 7. RRMSE and RB of $q_{0.995}(.)$ as a function of the depth iteration number and using various weighting ways in: (a) sub-region 1, (b) sub-region 2 and (c) sub-region 3.

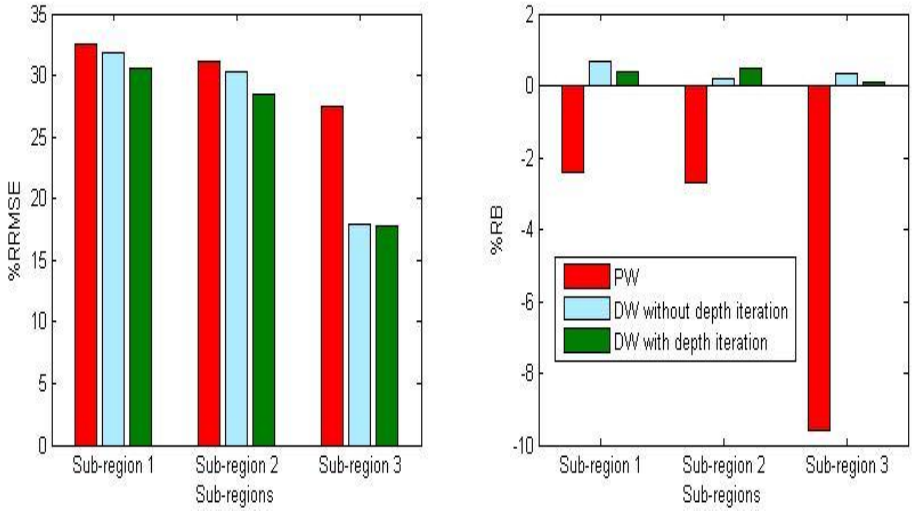


Figure 8. Result of growth curve estimation using PW and DW model with and without depth iterations. For the DW model p=0.995 and $\varphi = \varphi_G$.

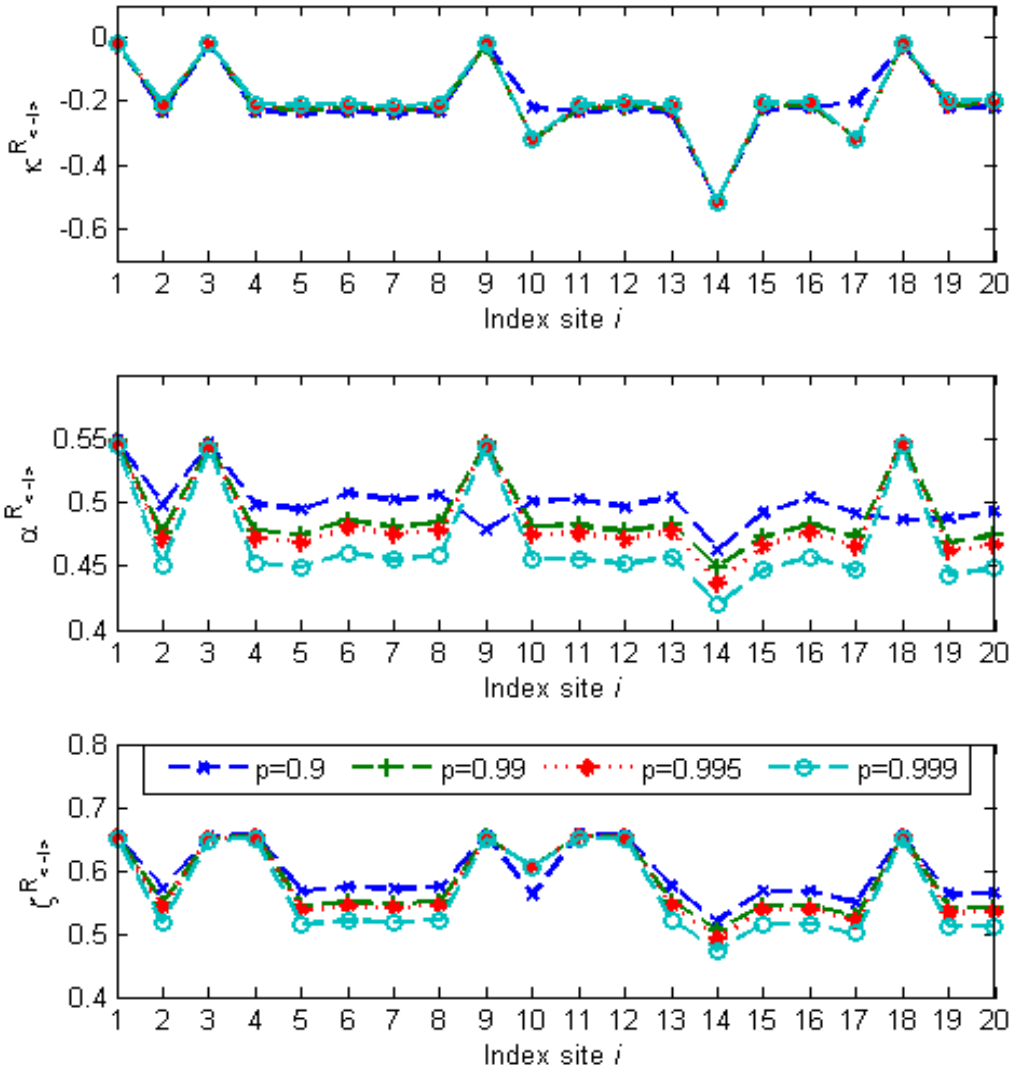


Figure 9. Regional growth curve parameters estimated by DW using $\varphi = \varphi_G$ in the sub-region 1 for different values of p.

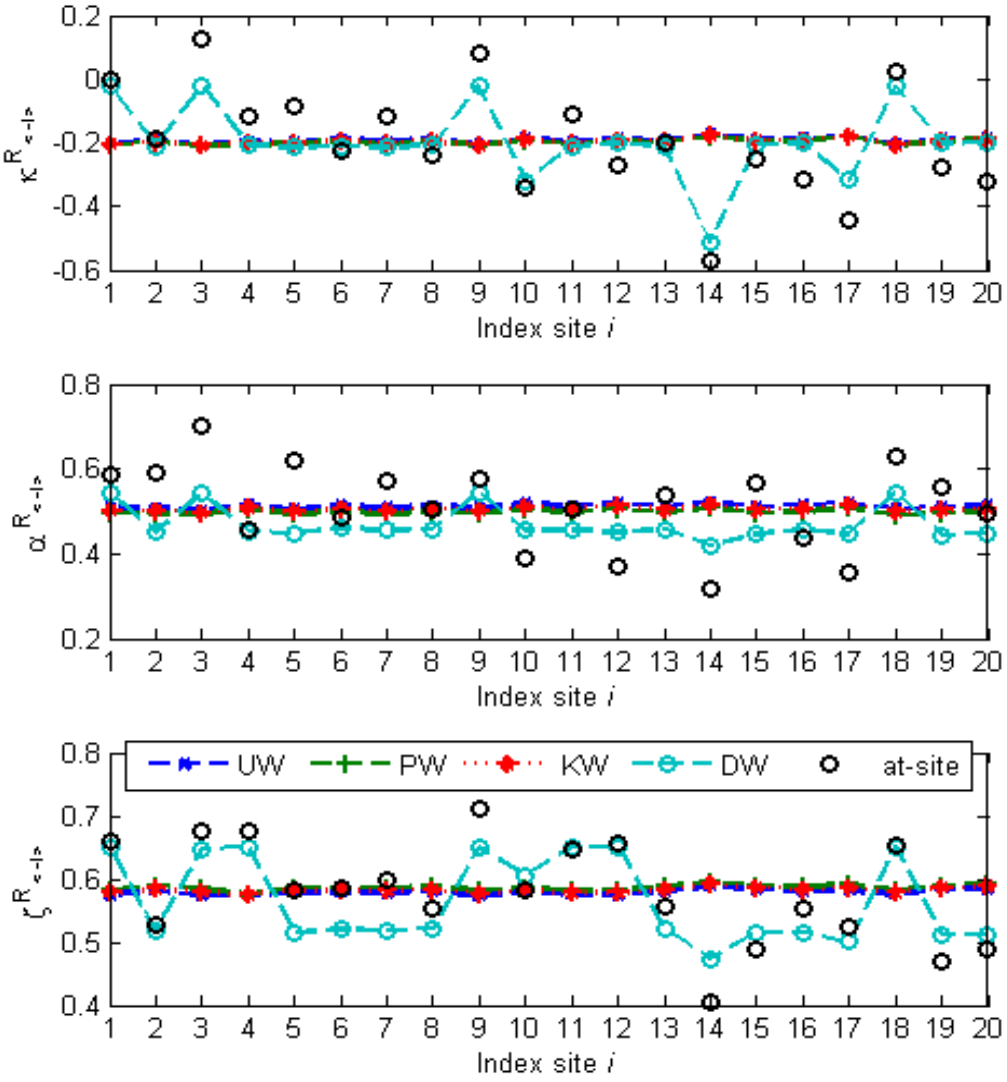


Figure 10. Regional growth curve parameters estimated by different approaches for the sub-region 1.

CHAPITRE 5: OPTIMISATION DU MODÈLE DE RÉGRESSION BASÉ SUR LES FONCTIONS DE PROFONDEUR

Optimal depth-based regional frequency analysis

H. Wazneh*[1], F. Chebana[1] and T.B.M.J. Ouarda[2,1]

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

[*]**Corresponding author:** Tel: +1 (418) 654 2530#4461
Email: hussein.wazneh@ete.inrs.ca

Accepted May 20th 2013

(Hydrology and Earth System Sciences).

Abstract:

Classical methods of regional frequency analysis (RFA) of hydrological variables face two drawbacks: 1) the restriction to a particular region, which can lead to a loss of some information and 2) the definition of a region that generates a border effect. To reduce the impact of these drawbacks on regional modeling performance, an iterative method was proposed recently, based on the statistical notion of the depth function and a weight function φ. This depth-based RFA (DBRFA) approach was shown to be superior to traditional approaches in terms of flexibility, generality and performance. The main difficulty of the DBRFA approach is the optimal choice of the weight function φ (e.g., φ minimizing estimation errors). In order to avoid subjective choice and naïve selection procedures of φ, the aim of the present paper is to propose an algorithm-based procedure to optimize the DBRFA and automate the choice of φ according to objective performance criteria. This procedure is applied to estimate flood quantiles in three different regions in North America. One of the findings from the application is that the optimal weight function depends on the considered region and can also quantify the region homogeneity. By comparing the DBRFA to the canonical correlation analysis (CCA) method, results show that the DBRFA approach leads to better performances both in terms of relative bias and mean square error.

Keywords: regional frequency analysis; statistical depth function; floods estimation; optimization; canonical correlation analysis; hydrology.

1. Introduction

Due to the large territorial extents and the high costs associated to installation and maintenance of monitoring stations, it is not possible to monitor hydrologic variables at all sites of interest. Consequently, hydrologists have often to provide estimates of design events quantiles *QT*, corresponding to a large return period *T* at ungauged sites. In this situation, regionalization approaches are commonly used to transfer information from gauged sites to the target site (ungauged or partially gauged) [e.g., Burn, 1990b; Dalrymple, 1960; Ouarda et al., 2000]. A number of estimation techniques in regional frequency analysis (RFA) have been proposed and applied in several countries [De Michele and Rosso, 2002; Haddad and Rahman, 2012; Madsen and Rosbjerg, 1997; Nguyen and Pandey, 1996; Ouarda et al., 2001].

In general, RFA consists of two main steps: (1) grouping stations with similar hydrological behavior (delineation of hydrological homogeneous regions) [e.g., Burn, 1990a] and (2) regional estimation within each homogenous region at the site of interest [e.g., GREHYS, 1996a; Ouarda et al., 2001; Ouarda et al., 2000]. The two main disadvantages of this type of regionalization methods are: i) a loss of information due to the exclusion of a number of sites in the step of delineation of hydrological homogeneous region, and ii) a border effect problem generated by the definition of a region.

To reduce or eliminate the negative impact of these disadvantages on the estimation quality, a number of regional methods have been proposed that combine the two stages (delineation and estimation) and use all stations [e.g., Ouarda et al., 2008; Shu and Ouarda, 2007; Shu and Ouarda, 2008]. One of these regional methods was developed recently by Chebana and Ouarda [2008]. This RFA method is based on statistical depth functions (denoted by DBRFA for depth-based RFA). The DBRFA approach focuses directly on quantile estimation using the weighted

least squares (WLS) method to estimate parameters and avoids the delineation step. It employs the multiple regression (MR) model that describes the relation between hydrological and physio-meteorological variables of sites [Girard et al., 2004].

After Chebana and Ouarda [2008], statistical depth functions are used in a number of hydrological and environmental studies. For instance, Chebana and Ouarda [2011a] used these functions in an exploratory study of a multivariate sample including location, scale, skewness and kurtosis as well as outlier detection. In another study, Chebana and Ouarda [2011b] combined depth functions with the orientation of observations to identify the extremes in a multivariate sample. Bardossy and Singh [2008] used the statistical notion of depth to detect unusual events in order to calibrate hydrological models. Recently, some studies present further developments of the approach that calibrate hydrological models by a depth function [e.g., Kraußе and Cullmann, 2012; Kraußе et al., 2012].

The DBRFA method consists generally of ordering sites by using the statistical notion of depth functions [Zuo and Serfling, 2000]. This order is based on the similarity between each gauged site and the target one. Accordingly, a weight is attributed to each gauged site using a weight function denoted φ. This function, with a suitable shape, eliminates the border effect and includes all the available sites proportionally to their hydrological similarity to the target site. Note that classical RFA approaches correspond to a special weight function with value 1 inside the region and 0 outside. The definition of a region in the classical RFA approaches becomes rather a question of choice of weight function φ according to a given criterion (e.g., relative root mean square error RRMSE).

By construction, the estimation performance in the MR model using the DBRFA approach depends on the choice of the weight functions φ. Chebana and Ouarda [2008] applied several

families of functions φ, where the corresponding coefficients were chosen arbitrary and after several trials. In addition, even though the obtained results are improvement of the traditional approaches, they are not necessarily the best ones.

The aim of the present paper is to propose a procedure to optimize the DBRFA approach over φ. This aim has theoretical as well as practical considerations. This procedure allows an optimal choice of the weight function φ and makes the DBRFA approach automatic and objective. It should be noted that Ouarda et al. [2001] determined the optimal homogenous neighborhood of a target site in the Canonical Correlation Analysis (CCA) based approach. In Ouarda et al [2001] the optimization corresponds to the selection of the neighborhood coefficient, denoted by α, according to the bias or the squared error. The optimal choice of weight functions has been the topic of numerous studies in the field of statistics [e.g., Chebana, 2004].

To optimize the choice of φ, suitable families of functions as well as algorithms are required. In the present context, four families of φ are considered: Gompertz (φ_G) [Gompertz, 1825], logistic ($\varphi_{\text{logistic}}$) [Verhulst, 1838], linear (φ_{Linear}) and indicator (φ_I). The three families φ_G, $\varphi_{\text{log}istic}$ and φ_{Linear} are regular, flexible, S-shaped and have other suitable properties.

Several appropriate algorithms can be considered [Wright, 1996]. They are appropriate when the objective function ξ (criterion to be optimized) is not differentiable or the gradient is unavailable and must be calculated by a numerical method (e.g., finite differences). Among these algorithms, the most commonly used are: the simplex method [Nelder and Mead, 1965], the pattern search method of Hooke and Jeeves [Hooke and Jeeves, 1961; Torczon, 2000] and the Rosenbrock methods [Rao, 1996; Rosenbrock, 1960]. These methods are used successfully in several domains, and are particularly popular in chemistry, engineering and medicine. Specifically, in this paper the simplex and the pattern search algorithms are used because of their advantages.

Indeed, they are very robust [e.g., Dolan et al., 2003; Hereford, 2001; Torczon, 2000], simple in terms of programming, valid for nonlinear optimization problems with real coefficients [McKinnon, 1999] and helpful in solving optimization problems with and without constraints [e.g., Lewis and Torczon, 1999; Lewis and Torczon, 2002].

In this study, the proposed optimization procedure is applied to the flood data from three different regions of the United States and Canada (Texas, Arkansas and southern Quebec). For each region, the obtained results are compared with those of the CCA approach.

The present paper is organized as follows. Section 2 describes the used technical tools including depth functions, the WLS method and the definitions of the considered weight functions. Section 3 describes the proposed procedure. Then, section 4 presents the application to the three case studies as well as the obtained results. The last section is devoted to the conclusions of this work.

2. Background

In this section, the background elements required to introduce and apply the optimization procedure of the DBRFA approach are briefly presented. This section contains a number of basic notions.

2.1. Mahalanobis depth function

The absence of a natural order to classify multivariate data led to the introduction of the depth functions [Tukey, 1975]. They are used in many research fields, and were introduced in water science by Chebana and Ouarda [2008]. Several depth functions were introduced in the literature [Zuo and Serfling, 2000]. Depth functions have a number of features that fit well with the constraint of RFA [Chebana and Ouarda, 2008].

In this study, the Mahalanobis depth function is used to sort sites where the deeper the site is the more it is hydrologically similar to the target site. This function is used for its simplicity, value

interpretability, and for the relationship with the CCA approach used in RFA. The Mahalanobis depth function is defined on the basis of the Mahalanobis distance given by $d_A^2(x,y)=(x-y)' A^{-1}(x-y)$ between two points $x, y \in R^d$ $(d \geq 1)$ where A is a positive definite matrix [Mahalanobis, 1936]. This distance is used by Ouarda et al. [2001] in the development of the CCA approach. The Mahalanobis depth of x with respect to μ is given by:

$$MHD(x;F)=\frac{1}{1+d_A^2(x,\mu)} \qquad x \text{ in } R^d \tag{1}$$

for a cumulative distribution function F characterized by a location parameter μ and a covariance matrix A. Note that the Mahalanobis depth function has values in the interval $[0,1]$.

An empirical version of the Mahalanobis depth of x with respect μ is defined by replacing F by a suitable empirical function $\hat{F}_N$ for a sample of size N [Liu and Singh, 1993]. In the context of the present paper, the notation in (1) is replaced by:

$$MHD_{\hat{A}}(x;\hat{\mu})=\frac{1}{1+d_{\hat{A}}^2(x,\hat{\mu})} \tag{2}$$

where $\hat{\mu}$ and $\hat{A}$ are respectively the location and covariance matrix estimated from the observed sample.

2.2. Weight functions

Below are the definitions of the four families of weight functions φ_G, $\varphi_{\text{logistic}}$, φ_{Linear} and φ_I considered in this paper along with special cases of functions φ for comparison purposes.

2.2.1. Gompertz function

The Gompertz function is usually employed as a distribution in survival analysis. This function was originaly formulated by Gompertz [1825] for modeling human mortality. A number of authors contributed to the studies of the characterization of this distribution [e.g., Chen, 1997;

Wu and Lee, 1999]. In the field of water resources, the Gompertz function was adopted by Ouarda et al. [1995] to estimate the flood damage in the residential sector. The function φ_G is increasing, flexible and continuous [Zimmerman and Núñez-Antón, 2001]. The Gompertz distribution has different formulations one of which is given by:

$$\varphi_G(x) = c\exp\left\{-ae^{-bx}\right\} \quad a,b,c>0\ ;\ x\in R \tag{3}$$

where *c* is its upper limit, *a* and *b* are two coefficients which respectively allow to translate and change the spread of the curve. Figure 1 shows the effects of these coefficients on the form of φ_G. Note that this function starts at zero (starting phase), then increases exponentially (growth phase) and finally stabilizes by approaching the upper limit *c* (stationary phase) with $0 \le \varphi_G(x) \le c$. The inflection point of this function is $\left(\frac{\ln a}{b}, \frac{c}{e}\right)$.

2.2.2. Logistic function

Verhulst [1838] proposed this function to study population growth. It is given by:

$$\varphi_{\text{logistic}}(x) = \frac{c}{1+ae^{-bx}} \qquad a,b,c>0; x\in R \tag{4}$$

where the coefficients *c*, *a* and *b* play the same role as in φ_G.

This function has similar properties to those of φ_G (increasing, flexible, continuous and with three phases). However, $\varphi_{\text{logistic}}$ is symmetric around its inflection point $\left(\frac{\ln a}{b}, \frac{c}{2}\right)$ which is not the case for φ_G.

2.2.3. Linear function

It is a simple function, linear over three pieces corresponding to the three previous phases. Explicitly it is given by:

$$\varphi_{Linear}(x)=\begin{cases}0 & \text{if } x \le d_1 \\ \dfrac{x-d_1}{d_2-d_1} & \text{if } d_1 \le x \le d_2, \qquad d_2 > d_1 > 0 \\ 1 & \text{if } x \ge d_2\end{cases} \tag{5}$$

This function is considered as a weight function in the study of Chebana and Ouarda [2008].

2.2.4. Indicator function

This function is given by:

$$\varphi_I(x)=\begin{cases}1 & \text{if } x \in A \\ 0 & \text{if } x \notin A\end{cases} \tag{6}$$

where A is a subset in R (set of real numbers), such as an interval. The subset A represents the neighborhood or the region in the classical RFA approaches. The weight is equal to 1 if the site is included in the region, otherwise, it is 0.

In the case where the set A is the interval $\left[C_{\alpha,p},1\right]$ with $C_{\alpha,p}=\dfrac{1}{1+\chi^2_{\alpha,p}}$ and $\chi^2_{\alpha,p}$ is the $(1-\alpha)$ quantile associated to the chi-squared distribution with p degrees of freedom, the DBRFA reduces to the traditional CCA approach [e.g., Bates et al., 1998]. The corresponding weight function is denoted by φ_{CCA}.

If $A=[0,1]$ i.e. $\alpha=0$, then the DBRFA represents the uniform approach which includes all available sites with similar importance. The corresponding weight function is denoted by φ_U.

2.3. Weighted Least Squares Estimation

In the RFA framework, the MR model is generally used to describe the relationship between the hydrological variables and the physiographical and climatic variables of the sites of a given region. This model has the advantage to be simple, fast, and not requiring the same distribution for hydrological data at each site within the region [Ouarda et al., 2001].

Let *QT* be the quantile corresponding to the return period *T*. It is often assumed that the relationship between QT, as the hydrological variable, and the physio-meteorological variables and basin characteristics $A_1, A_2, ...A_r$ takes the form of a power function [Girard et al., 2004]:

$$QT = \beta_0 A_1^{\beta_1} A_2^{\beta_2} ... A_r^{\beta_r} e \tag{7}$$

where *e* is the model error.

Let *s* be the number of quantiles *QT* corresponding to *s* return periods and *N* be the total number of sites in the region. A matrix of hydrological variables $Y = \left(QT_1, QT_2, ..., QT_s\right)$ of dimension $N \times s$ is then constructed. With a log-transformation in (7) we obtain the multivariate log-linear model in the following form:

$$\log Y = \left(\log X\right)\beta + \varepsilon \tag{8}$$

where $\log X = (1, \log A_1, \log A_2, ..., \log A_r)$ is the $N \times \left(r+1\right)$ matrix formed by (*r*) physio-meteorological variables series, β is the $(r+1) \times s$ matrix of parameters and $\varepsilon = \left(\varepsilon^1, ..., \varepsilon^s\right)$ is the $N \times s$ matrix that represents the model error (residual) with null mean vectors and variance-covariance matrix Γ:

$$E(\varepsilon) = \left(0, .., 0\right) \quad \text{and} \quad Var(\varepsilon) = \Gamma = \begin{pmatrix} Var\left(\varepsilon^1\right) & \dots & Cov\left(\varepsilon^1, \varepsilon^s\right) \\ \vdots & \ddots & \vdots \\ Cov\left(\varepsilon^s, \varepsilon^1\right) & \cdots & Var\left(\varepsilon^s\right) \end{pmatrix} \tag{9}$$

The parameter matrix β can be estimated, using the WLS estimation, by:

$$\begin{aligned} \hat{\beta}_w &= \arg\min_{\beta} \left(logY - logX\,\beta\right)' \Omega \left(logY - logX\,\beta\right) \\ &= \left((\log X)' \Omega \log X\right)^{-1} (\log X)' \Omega \log Y \end{aligned} \tag{10}$$

where $\Omega = \text{diag}(w_1, ..., w_N)$ is the diagonal matrix with diagonal elements w_i where w_i is the weight for the site *i*. The matrix Γ is estimated by:

$$\hat{\Gamma}_w = \frac{\left(\log Y - \log X \hat{\beta}_w\right)' \left(\log Y - \log X \hat{\beta}_w\right)}{N - r - 1} \quad (11)$$

Note that the log-transformation induces generally a bias in the estimation of *QT* [Girard et al., 2004].

3. Methodology

This section describes a general procedure for optimizing the DBRFA approach and treats special cases where this procedure is applied using the weight functions defined in section 2.2.

3.1. General procedure

In order to find the optimal weight function $\varphi_{Optimal}$ in the DBRFA approach, the procedure is composed of three main steps. They are summarized as follows:

i. For a given class of weight functions φ and a set of gauged sites (region), use a jackknife procedure to assess the regional flood quantile estimators (Eq. 8) for the sites of the region using the DBRFA approach. These estimators depend on the weight function φ through its coefficients;

ii. For a pre-selected criterion, calculate its value to quantify the performance of the estimates obtained from step i;

iii. Using an optimization algorithm, optimize the criterion (objective function) calculated in step ii. The parameters of the optimization problem are the coefficients of the weight function. The outputs of this step are $\varphi_{Optimal}$ and the value of the selected criterion.

3.2 Description of the procedure

In the first step of the procedure, we use a jackknife resampling procedure to assess the regional flood quantile estimators for the sites of the region. This jackknife procedure consists in considering each site l ($l = 1, ..., N$) in the region as an ungauged one by removing it temporarily from the region (i.e. we assume that the hydrological variable Y_l of site l is unknown and the physio-meteorological variable X_l is known since it can be easily estimated from existing physiographic maps and climatic data). Then we calculate the regional estimator $\left(\hat{Y}_l\right)_\varphi$ of site l by the iterative WLS regression, using the $N-1$ remaining sites, which is related to the given weight function φ. The parameters of the starting estimator (initial point) of DBRFA, denoted by $\hat{\beta}_{1,l}$ and $\hat{\Gamma}_{1,l}$, are calculated by assuming that $X = X^{<-l>}$, $Y = Y^{<-l>}$ and $\Omega = I_{N-1}$ in (10) and (11), where $X^{<-l>}$ represents the matrix of physio-meteorological variables excluding site l, $Y^{<-l>}$ is the matrix of hydrological variables excluding site l and I_{N-1} is the identity matrix of dimension $(N-1)\times(N-1)$. The starting estimator $\left(\hat{Y}_{1,l}\right)_\varphi$ is obtained by replacing β with $\hat{\beta}_{1,l}$ in (8). Then for each depth iteration k, $k = 2, 3, .., k_{iter}$, we calculate the Mahalanobis depth (2) of the gauged site i, $i = 1, ..., N-1$, with respect to the ungauged site l denoted by $\left(D_{k,(i,l)}\right)_\varphi = MHD_{\left(\hat{\Gamma}_{k-1,l}\right)_\varphi}\left(\log Y_i; \left(\log \hat{Y}_{k-1,l}\right)_\varphi\right)$. The number of iterations k_{iter} is fixed to ensure the convergence of the depth function (generally k_{iter} = 25 is appropriate). The weight matrix at iteration k is defined by applying the function φ to the depth calculated at this iteration. The parameters of the MR model at the k^{th} iteration are estimated by:

$$\left(\hat{\beta}_{k,l}\right)_\varphi = \left(\left(\log X^{<-l>}\right)' \left(\Omega_{k,l}\right)_\varphi \left(\log X^{<-l>}\right)\right)^{-1} \left(\log X^{<-l>}\right)' \left(\Omega_{k,l}\right)_\varphi \log Y^{<-l>} \quad (12)$$

$$\left(\hat{\Gamma}_{k,l}\right)_{\varphi}=\frac{\left(\log Y^{<-l>}-\left(\log X^{<-l>}\right)\left(\hat{\beta}_{k,l}\right)_{\varphi}\right)'\left(\log Y^{<-l>}-\left(\log X^{<-l>}\right)\left(\hat{\beta}_{k,l}\right)_{\varphi}\right)}{(N-1)-r-1} \quad (13)$$

where $\left(\Omega_{k,l}\right)_{\varphi}$ is a *N*-1 diagonal matrix with elements:

$$\varphi\left[\left(D_{k,(1,l)}\right)_{\varphi}\right],.....,\varphi\left[\left(D_{k,(N-1,l)}\right)_{\varphi}\right] \quad (14)$$

Note that all these parameters depend on φ. Then, the regional quantile estimator for the site *l* in this iteration is:

$$\left(\hat{Y}_{k,l}\right)_{\varphi}=\exp\left[\left(\log X_l\right)\left(\hat{\beta}_{k,l}\right)_{\varphi}\right] \quad (15)$$

In the second step of the procedure, we use the regional estimators at the last iteration since their associated estimation errors are the minimum possible by construction. Consequently, in order to simplify the notations in the rest of this paper, we denote $\left(\hat{Y}_1\right)_{\varphi}=\left(\hat{Y}_{k_{iter},1}\right)_{\varphi},...,\left(\hat{Y}_l\right)_{\varphi}=\left(\hat{Y}_{k_{iter},l}\right)_{\varphi},..,\left(\hat{Y}_N\right)_{\varphi}=\left(\hat{Y}_{k_{iter},N}\right)_{\varphi}$.

After calculating $\left(\hat{Y}_l\right)_{\varphi}$, $l=1,..,N$ in step i, we consider and evaluate one or several performance criteria in step ii. The considered criteria are employed as objective functions in the optimization step iii.

The relative bias (RB) and the relative root mean square error (RRMSE) are widely used in hydrology, particularly in RFA, as criteria to evaluate model performances. These two criteria are defined using an element-by-element division by:

$$RB_{\varphi} = 100 \times \frac{1}{N} \sum_{l=1}^{N} \left(\frac{Y_l - \left(\hat{Y}_l \right)_{\varphi}}{Y_l} \right) \quad (16)$$

$$RRMSE_{\varphi} = 100 \times \sqrt{\frac{1}{N-1} \sum_{l=1}^{N} \left(\frac{Y_l - \left(\hat{Y}_l \right)_{\varphi}}{Y_l} \right)^2} \quad (17)$$

where Y_l is the local quantile estimation for the l^{th} site, $\left(\hat{Y}_l \right)_{\varphi}$ is the regional estimation by DBRFA approach according to φ and excluding site *l*, and *N* is the number of sites in the region. The RB_{φ} measures the tendency of quantile estimates to be uniformly too high or too low across the whole region and the RRMSE_{φ} measures the overall deviation of estimated quantiles from true quantiles [Hosking and Wallis, 1997]. Note that other criteria can also be considered such as the Nash criterion (NASH) and the coefficient of determination (R^2). In the hydrological framework, the previously defined criteria are used as key performance indicators (KPI) to compare different RFA approaches [e.g., Gaál et al., 2008].

Finally in step iii, we apply an optimization algorithm on the selected and evaluated criterion in step ii. The algorithms to be considered are indicated in the introduction section. The formulation of the criteria to be optimized, generally complex and non-explicit, suggests the use of zero-order algorithms. The application of these algorithms allows to find the optimal function $\varphi_{Optimal}$ with respect to selected criteria. An overview diagram summarizing the optimization procedure of the DBRFA approach is illustrated in Figure 2.

The procedure described above aims to calculate $\varphi_{Optimal}$ according to the desired criterion. In order to estimate the quantile Y_u of an ungauged site *u* using the optimal DBRFA approach, the

user simply repeats step i of the procedure without excluding any site and while fixing the weight function, i.e. step i with $\varphi = \varphi_{Optimal}$.

Based on the optimization procedure of the DBRFA approach described previously, the parameters of the optimization problem are the coefficients of the weight function. Consequently, reducing the number of coefficients in φ can make the algorithm more efficient and less expensive in terms of memory and computing time. If the weight function is one of the two functions Gompertz (3) or logistic (4), the coefficient c represents the upper limit of these functions. As in the DBRFA approach, the upper limit of φ is 1, namely the gauged site is completely similar to the target site, hence the value $c = 1$ is fixed. In this case, the problem is reduced to find the couple $\left(\hat{a}_N, \hat{b}_N\right)$ that optimizes one of the pre-selected criteria, such as (16) and (17).

Moreover, in the classes $\varphi = \varphi_G$ or $\varphi = \varphi_{\text{logistic}}$, the optimization problem is applied in semi-bounded domain (i.e. $a > 0$ and $b > 0$) and without other constraints (linear or nonlinear). In this case, the Nelder-Mead algorithm can also be applied as well as the Pattern search one [Luersen and Le Riche, 2004].

On the other hand, in the case where $\varphi = \varphi_{Lineair}$ (5), the inequality constraint $d_2 > d_1 > 0$ is imposed. Therefore, the Nelder-Mead algorithm cannot be considered.

Theoretically and generally, the two optimization algorithms used in this paper (i.e. the Nelder-Mead and the pattern search algorithms) converge to a local minimum (or maximum) according to the initial point. To overcome this problem and make the algorithm more efficient, two solutions are proposed in the literature: a) for each objective function, use several starting points and calculate the optimum for each of these points; the optimum of the function will be the best value of these local optima [Bortolot and Wynne, 2005]; or b) use a single starting point and each

time the algorithm converges, the optimization algorithm restarts again using the local optimum as a new starting point. This procedure is repeated until no further improvement in the value of the objective function is obtained [Press et al., 2002].

4. Data sets for case studies

In this section we present the data sets on which the DBRFA approach will be applied the following section. These data come from three geographical regions located in the states of Arkansas and Texas (USA) and in the southern part of the province of Quebec (Canada). The first region is located between 45 ° N and 55 ° N in the southern part of Quebec, Canada. The data set of this region is composed of 151 stations, each with station has a flood record of more than 15 years. The conditions of application of frequency analysis (i.e. homogeneity, stationary and independence) are tested on the historical data of these stations in several studies [Chokmani and Ouarda, 2004; Ouarda and Shu, 2009; Shu and Ouarda, 2008]. Three types of variables are considered: physiographical, meteorological and hydrological. The selected variables for the regional modeling are also used in Chokmani and Ouarda [2004]. The selected physiographical variable are: the basin area (AREA) in km^2, the mean basin slope (MBS) in % and the fraction of the basin area covered with lakes (FAL) in %. The meteorological variables are the annual mean total precipitation (AMP) in mm and the annual mean degree days over 0°C (AMD) in degree-day. The selected hydrological variables are represented by at-site specific flood quantiles (*QST*) in m^3/km^2s, corresponding to return periods T = 10 and 100 years.

The two other considered regions correspond to a database of the United States Geological Survey (USGS). This database, called Hydro-Climatic Data Network (HCDN), consists of observations of daily discharges from 1659 sites across the United States and its Territories

[Slack et al., 1993]. The sites included in this database contain at least 20 years of observations. As part of the HCDN project, the United States are divided into 21 large hydrological regions.

In this study, the data of the states of Arkansas and Texas (USA) are used for comparison purposes. The applicability conditions of frequency analysis as well as the variables to consider are justified in the study of Jennings et al., [1994]. The physiographical and climatological characteristics are the area of drainage basin (AREA) in km^2, the slope of main channel (SC) in m/km, the annual mean precipitation (AMP) in cm, the mean elevation of drainage basin (MED) in m and the length of main channel (LC) in km. The selected hydrological variables in these two regions are the at-site flood quantiles (QT), in m^3/s, corresponding to the return periods T = 10 and 50 years.

The data set of the states of Arkansas is composed of 204 sites. These data and the at-site frequency analysis are published in the study of Hodge and Tasker [1995]. Tasker et al. [1996] used these data to estimate the flood quantiles corresponding to the 50 year return period by the region of influence method [Burn, 1990b].

The Texas data base is composed of 90 sites but due to the lack of some explanatory variables at several sites, modeling was performed with only 69 stations. The data-set used in this region is the same used by Tasker and Slade [1994].

5. Results

The results obtained from the CCA-based approach are first presented and then compared to those obtained by the optimized DBRFA approach.

The variations of the two performance criteria RB and RRMSE, obtained by the CCA approach, as a function of the coefficient α (neighborhood coefficient) for the three regions are presented in Figure 3. The complete range of α is the interval [0, 1]. However, in this application, the range is

[0, 0.30] for Quebec and Arkansas regions and [0, 0.17] for the Texas region. These upper bounds of α are fixed to ensure that all neighborhoods of the sites contain sufficient stations to allow the estimation by the MR model. Note that it is appropriate to have at least three times more stations than the number of parameters in the MR model [Haché et al., 2002]. Figure 3 indicates that, for a given region, the same value of α optimizes the two criteria for the various return periods, even though this is not a general result [Ouarda et al., 2001]. The optimal α values are 0.25, 0.01 and 0.05 respectively for Quebec, Arkansas and Texas.

The coefficients λ_1 and λ_2 correspond respectively to the correlations of the first and the second couples of the canonical variables. Their values for Arkansas ($\lambda_1 = 0.973$, $\lambda_2 = 0.470$) and Texas ($\lambda_1 = 0.923$, $\lambda_2 = 0.402$) are larger than those of Quebec ($\lambda_1 = 0.853$, $\lambda_2 = 0.281$). This corresponds to a large optimal value of α for the latter region. Indeed, the higher the canonical correlation, the smaller the size of the ellipse defining the homogeneous neighborhood [Ouarda et al., 2001]. The value of α should be small enough so that the neighborhood contains an appropriate number of stations to perform the estimation in the MR model, and large enough to ensure an adequate degree of homogeneity within the neighborhood.

Figure 4 shows the projection sites of the three regions in the two canonical spaces (V1, W1) and (V2, W2) corresponding respectively to λ_1 and λ_2. This figure shows that for these three regions, the relationship between V1 and W1 is approximately linear, in contrast to V2 and W2. The presentation of a site in the space (V1, W1) is useful for an a priori information on the estimation error of this site. For example, in the Quebec region, the two sites 66 and 122 are poorly estimated. By fitting a linear model between V1 and W1 for each region, it is seen that the linearity assumption is more respected in Arkansas and Texas than in Quebec (R^2_{Arkansas} = 0.94, R^2_{Texas} = 0.85 et R^2_{Quebec} = 0.73).

The previous results show that the values of λ_1, λ_2, α and R^2 can be used as indicators of the quality of the homogeneity in a given region. In this application, the lower values of λ_1, λ_2 and R^2 as well as the higher value of α for Quebec compared to the values of the other two regions indicate that the Quebec region is less homogeneous than the two others. This conclusion needs to be verified by other criteria or statistical tests.

The DBRFA approach is applied by using the Mahalanobis depth function (2). The optimal weight functions, from each one of the three considered families, are obtained on the basis of the indicated optimization algorithms (i.e. φ_G and $\varphi_{\text{logistic}}$ using Nelder-Mead and φ_{Linear} using pattern search). They are presented in Figure 5. The corresponding results are summarized in Table 1. The optimization is made with respect to the RB and RRMSE criteria. Note that, for a given region, the regional flood quantile estimation is more accurate for small return periods. This result is valid for local as well as regional frequency analysis approaches [Hosking and Wallis, 1997]. In addition, Table 1 shows that the worst estimates are obtained using the uniform approach (weight function φ_U). This justifies the usefulness of considering the regional approaches. Note that for all regions, DBRFA with $\varphi_{Optimal}$ leads to more accurate estimates in terms of RB and RRMSE than those obtained using the CCA approach with optimal α. These results show also that the optimal coefficients of a given weight function depend on the chosen criterion (objective function). Finally, for the southern Quebec region, the results of Chebana and Ouarda (2008) are very close to those in the present paper (Table 1). The reason for this closeness is that the above authors forced the DBRFA approach to provide good results by trying several different combinations of values of φ coefficients (i.e. iteration loop of coefficients). Consequently, their trials took a long time and did not ensure the optimality of the approach which is not the case for the present study.

According to Figure 5, the form of optimal weight function depends on the considered region. For instance, the steep S-curve (with long upper extremity) of the two regions Arkansas and Texas depicts a large number of gauged sites similar to the target one; however, the high S-curve (with short upper extremity) of Quebec shows a small number of gauged sites similar to the target one. This result supports the previously mentioned conclusion about the homogeneity level for these regions.

In order to visualize the influence of gauged sites on the regional estimation of a target site in the DBRFA and CCA approaches, assume that Texas site number 25 is a target site and has to be estimated using the remaining 68 gauged sites. Figure 6 illustrates the weights allocated to each gauged site in the canonical hydrological space (W1, W2) instead of the geographical space. The estimate is made with the optimal α for the CCA approach and the optimal φ_G for the DBRFA approach. We observe that the influence of a gauged site on the estimation of the target site in the DBRFA approach is proportional to the hydrological similarity between these two sites. Hence, the weight function takes a bell shape in a 3D presentation (Figure 6b). However, with the CCA approach, the weight function (6) takes only two values, 1 within the neighborhood of the target-site or 0 otherwise (Figure 6a).

To study the impact of depth iterations on the performance of the DBFRA method, this approach is applied to the three regions but without iterations on the Mahalanobis depth (i.e. $k_{\text{iter}} = 2$ in step i in the DBRFA optimization procedure). The outputs of this application, with $\varphi = \varphi_G$ and $\zeta(.) = RRMSE$, are shown in Table 2. These results indicate that the optimal weight function changes depending on the case (with or without iterations) but keeps the S shape (for space limitation, the associated figure is not presented). In addition, using the iterations, we observe an improvement in the performance of the DBRFA method. This improvement varies from one

region to another where it is more significant in Quebec than in Texas and Arkansas (Table 2). This is another result indicating a difference between Quebec and the two other regions. Note that similar results are found for other families of weight functions and for different optimization criteria. In conclusion, the depth iterative step in the DBRFA before weight optimization is important.

In order to examine the convergence speed in terms of the performance criteria, we present the variations of these criteria as a function of depth iteration for different weight functions (Figure 7). The employed coefficient values of the weight functions are those minimizing the RRMSE (Table 1). We observe a rapid convergence (5 iterations) to the RRMSE values in Table 1 for Arkansas and Texas (Figure 7b and 7c), whereas, for Quebec (Figure 7a) it requires more than 20 iterations to converge to the results in Table 1. These results could be again due to the level of homogeneity in the region.

To compare the relative errors of flood quantile estimates obtained by different approaches for the three regions, Figure 8 illustrates these errors with respect to the logarithm of basin area. The weight functions used are those optimizing the RRMSE. It is generally observed that the DBRFA relative errors are lower than those obtained with the CCA approach. We also observe large negative errors for some sites, such as number 64 and 66 in the southern Quebec, 180 and 175 in Arkansas and 62 and 69 in Texas.

In this paper, the optimal DBRFA approach is mainly compared with the basic formulation of one of the most popular RFA approaches, that is the CCA approach. However, different variants of the latter are developed and are available in the literature, such as the Ensemble Artificial Neural Networks-CCA approach (EANN-CCA) [Shu and Ouarda, 2007] and the Kriging-CCA approach [Chokmani and Ouarda, 2004]. In order to insure the optimality of the optimal DBRFA, it is of interest to expend the above comparison to those approaches. A comprehensive

comparison requires presentation of these approaches as well a number of data sets for the considered regions. Some of the data sets are not available for the regions of Texas and Arkansas, e.g. at-site peak flows to estimate at-site quantiles as hydrological variables. However, all these approaches are already applied to the region of Quebec in different studies. Table 3 summarizes the obtained results for all those methods along with those of the DBRFA approach. The results indicate that the optimal DBRFA performs better than the available approaches both in terms of RB and RRMSE, except a very slight difference of 1% in the RRMSE of QS10 with EANN-CCA. This could be related to the numerical approximations in the computational algorithms.

6. Conclusions

In the present paper, a procedure is proposed to optimize the selection of a weight function in the DBRFA approach. This procedure automates the optimal choice of the weight function φ with respect to a given criterion. Therefore, aside from leading to optimal estimation results, it allows the DBRFA approach to be more practical and usable without the user's subjective intervention. The user has only to select one or several objective performance criteria to obtain the model, the estimated performance and the weight functions for a specific region. One of the findings is that the optimal weight function can be seen as characterization of the associated region.

General and flexible families of weight function are considered, as well as two optimization algorithms to find $\varphi_{Optimal}$. The used algorithms can handle cases with or without constraints on the definition domain of the function φ.

The obtained results, from three regions in North America, show the utility to consider the DBRFA method in terms of performance as well as the efficiency and flexibility of the proposed optimization procedure.

The study of the three regions shows an association between the level of the homogeneity of the region, the form of the optimal weight function and the computation convergence speed. This result deserves to be developed in future work.

Acknowledgments

Financial support for this study was graciously provided by the Natural Sciences and Engineering Research Council (NSERC) of Canada and the Canada Research Chair Program. The authors are grateful to the Editor and the anonymous reviewers for their valuable comments and suggestions.

References

Bárdossy, A., Singh, S.K., 2008. Robust estimation of hydrological model parameters. Hydrology and Earth System Sciences, 12(6): 1273-1283.

Bates, B.C., Rahman, A., Mein, R.G., Weinmann, P.E., 1998. Climatic and physical factors that influence the homogeneity of regional floods in southeastern Australia. Water Resources Research, 34(12): 3369-3381.

Bortolot, Z.J., Wynne, R.H., 2005. Estimating forest biomass using small footprint LiDAR data: An individual tree-based approach that incorporates training data. ISPRS Journal of Photogrammetry and Remote Sensing, 59(6): 342-360.

Burn, 1990a. An appraisal of the "region of influence' approach to flood frequency analysis. Hydrological Sciences Journal/Journal des Sciences Hydrologiques, 35(2): 149-165.

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

Chebana, F., 2004. On the optimization of the weighted Bickel-Rosenblatt test. Statistics and Probability Letters, 68(4): 333-345.

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

Chebana, F., Ouarda, T.B.M.J., 2011a. Depth-based multivariate descriptive statistics with hydrological applications. Journal of Geophysical Research D: Atmospheres, 116(10).

Chebana, F., Ouarda, T.B.M.J., 2011b. Multivariate extreme value identification using depth functions. Environmetrics, 22(3): 441-455.

Chen, Z., 1997. Parameter estimation of the Gompertz population. Biometrical Journal, 39(1): 117-124.

Chokmani, K., Ouarda, T.B.M.J., 2004. Physiographical space-based kriging for regional flood frequency estimation at ungauged sites. Water Resources Research, 40(12): 1-13.

Dalrymple, T., 1960. Flood frequency methods. Water Supply Paper No. 1543 A.

De Michele, C., Rosso, R., 2002. A multi-level approach to flood frequency regionalisation. Hydrology and Earth System Sciences, 6(2): 185-194.

Dolan, E.D., Michael Lewis, R., Torczon, V., 2003. On The Local Convergence Of Pattern Search. SIAM J. OPTIM, 14(2): 567-583.

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

Girard, C., Ouarda, T.B.M.J., Bobée, B., 2004. Study of bias in the log-linear model for regional estimation. Étude du biais dans le modèle log-linéaire d'estimation régionale, 31(2): 361-368.

Gompertz, B., 1825. On the nature of the function expressive of the law of human mortality, and on a new mode of determining the value of life contingencies. Philos. Trans. R. Soc. Lond., 115: 513-585.

GREHYS, 1996a. Presentation and review of somemethods for regional flood frequency analysis. Journal of Hydrology, 186: 63-84.

Haché, M., Ouarda, T.B.M.J., Bruneau, P., Bobée, B., 2002. Regional estimation by canonical correlation analysis: Hydrological variable analysis. Estimation régionale par la méthode de l'analyse canonique des corrélations : Comparaison des types de variables hydrologiques, 29(6): 899-910.

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

Hereford, J., 2001. Comparison of four parameter selection techniques. Proceedings of SoutheastCon 2001, Clemson, SC, 30 March-1 April 2001: 11-16.

Hodge, S.A., Tasker, G.D., 1995. Magnitude and Frequency of Floods in Arkansas. U.S. Geological Survey Water-Resources Investigations Report.

Hooke, R., Jeeves, T.A., 1961. Direct search solution of numerical and statistical problems. Journal of the Association for Computing Machinery, 8(2): 212-229.

Hosking, J.R.M., Wallis, J.R., 1997. Regional frequency analysis: an approach based on L-moments. Cambridge University Press, Cambridge.
Jennings, M.E., Thomas W.O, Jr., Riggs, H.C., 1994. Nationwide summary of U.S. geological survey regional regression equations for estimating magnitude and frequency of floods for ungaged sites, 1993. USGS Water-Resources Investigations Rep. 94-4002.
Krauße, T., Cullmann, J., 2012. Towards a more representative parametrisation of hydrologic models via synthesizing the strengths of Particle Swarm Optimisation and Robust Parameter Estimation. Hydrology and Earth System Sciences, 16(2): 603-629.
Krauße, T., Cullmann, J., Saile, P., Schmitz, G.H., 2012. Robust multi-objective calibration strategies – possibilities for improving flood forecasting. Hydrology and Earth System Sciences, 16(10): 3579-3606.
Lewis, R.M., Torczon, V., 1999. Pattern search algorithms for bound constrained minimization. SIAM Journal on Optimization, 9(4): 1082-1099.
Lewis, R.M., Torczon, V., 2002. A globally convergent augmented Lagrangian pattern search algorithm for optimization with general constraints and simple bounds. SIAM Journal on Optimization, 12(4): 1075-1089.
Liu, R.Y., Singh, K., 1993. A quality index based on data depth and multivariate rank tests. J. Amer. Statist. Assoc., 88(421): 252-260.
Luersen, M.A., Le Riche, R., 2004. Globalized nelder-mead method for engineering optimization. Computers and Structures, 82(23-26): 2251-2260.
Madsen, H., Rosbjerg, D., 1997. Generalized least squares and empirical Bayes estimation in regional partial duration series index-flood modeling. Water Resources Research, 33(4): 771-781.
Mahalanobis, P.C., 1936. On the generalized distance in statistics. Calcutta Statist. Assoc. Bull., 14: 9.
McKinnon, K.I.M., 1999. Convergence of the Nelder-Mead simplex method to a nonstationary point. SIAM Journal on Optimization, 9(1): 148-158.
Nelder, J.A., Mead, R., 1965. A simplex method for function minimization. Comput. J., 7: 308-313.
Nguyen, V.T.V., Pandey, G., 1996. A new approach to regional estimation of floods in Quebec. Proceedings of the 49th Annual Conference of the CWRA: 587-596.
Ouarda et al., 2008. Intercomparison of regional flood frequency estimation methods at ungauged sites for a Mexican case study. Journal of Hydrology, 348(1-2): 40-58.
Ouarda, El-Jabi, N., Ashkar, F., 1995. Flood damage estimation in the residential sector. Water Resources and Environmental Hazards: Emphasis on Hydrologic and Cultural insight in the Pacific Rim, AWRA Technical Publication series (1995): 73-82.
Ouarda, Shu, C., 2009. Regional low-flow frequency analysis using single and ensemble artificial neural networks. Water Resources Research, 45(11).
Ouarda, T.B.M.J., Girard, C., Cavadias, G.S., Bobée, B., 2001. Regional flood frequency estimation with canonical correlation analysis. Journal of Hydrology, 254(1-4): 157-173.
Ouarda, T.B.M.J., Hache, M., Bruneau, P., Bobee, B., 2000. Regional flood peak and volume estimation in northern Canadian basin. Journal of Cold Regions Engineering, 14(4): 176-191.
Press, W.H., Flannery, B.P., Teukolsky, S.A., Vetterling, W.T., 2002. Numerical recipes in C: the art of scientific computing. 2nd ed.
Rao, S.S., 1996. Engineering Optimization-Theory and Practice, 3rd Ed., 9: 621-622.
Rosenbrock, H.H., 1960. An automatic method for finding the greatest or least value of a function. Comput. J., 3(3): 175-184.
Shu, C., Ouarda, T.B.M.J., 2007. Flood frequency analysis at ungauged sites using artificial neural networks in canonical correlation analysis physiographic space. Water Resources Research, 43(7).
Shu, C., Ouarda, T.B.M.J., 2008. Regional flood frequency analysis at ungauged sites using the adaptive neuro-fuzzy inference system. Journal of Hydrology, 349(1-2): 31-43.
Slack, J.R., Lumb, A.M., Landwehr, J.M., 1993. Hydro-Climatic Data Network (HCDN): Streamflow data set, 1874-1988. Hydro-climatic Data Network (HCDN): A U.S. Geological Survey Streamflow Data Set for the United States for the Study of Climate Variations, 1874-1988.

Tasker, G.D., Hodge, S.A., Barks, C.S., 1996. Region of influence regression for estimating the 50-year flood at ungaged sites. Journal of the American Water Resources Association, 32(1): 163-170.
Tasker, G.D., Slade, R.M., 1994. An interactive regional regression approach to estimating flood quantiles. Water Policy and Management: Solving the Problems, ASCE Proceedings of the 21st Annual Conference of the Water Resources Planning and Management Division: 782-785.
Torczon, V., 2000. On the Convergence of Pattern Search Algorithms. SIAM Journal on Optimization, 7(1): 1-25.
Tukey, J.W., 1975. Mathematics and the picturing of data. Proceedings of the International Congress of Mathematicians, 2: 523-531.
Verhulst, P.F., 1838. Notice sur la loi que la population pursuit dans son accroissement.
Wright, M.H., 1996. Direct search methods: Once scorned, now respectable. Dundee Biennial Conf. Numer.
Wu, J.W., Lee, W.C., 1999. Characterization of the mixtures of Gompertz distributions by conditional expectation of order statistics. Biometrical Journal, 41(3): 371-381.
Zimmerman, D.L., Núñez-Antón, V., 2001. Parametric modelling of growth curve data: An overview. Test, 10(1): 1-73.
Zuo, Y., Serfling, R., 2000. General notions of statistical depth function. Annals of Statistics, 28(2): 461-482.

1 **Table 1.** Quantile estimation result with the various approaches

2

Objective function ζ	Weight function φ	Region: Southern Quebec (Canada) — Optimal coefficients	QS10 RB (%)	QS10 RR MSE (%)	QS100 RB (%)	QS100 RR MSE (%)	Arkansas (United States) — Optimal coefficients	Q10 RB (%)	Q10 RR MSE (%)	Q50 RB (%)	Q50 RR MSE (%)	Texas (United States) — Optimal coefficients	Q10 RB (%)	Q10 RR MSE (%)	Q50 RB (%)	Q50 RR MSE (%)
-	φ_U	-	-8.60	55.00	-11.0	64.00	-	-13.2	65.48	-15.1	73.34	-	-9.70	46.50	-13.8	61.00
RRMSE or RB	φ_{CCA}	$\alpha = 0.25$	-7.54	44.62	-8.14	51.84	$\alpha = 0.01$	-7.80	48.16	-9.31	59.50	$\alpha = 0.05$	-1.20	42.30	-7.40	57.40
RRMSE	φ_G	a = 30.5 b = 7	-3.55	**38.70**	-2.20	**44.50**	a = 97 b = 25	-6.00	41.50	-6.33	47.70	a = 129.7 b = 35.4	-1.01	36.86	-6.00	50.79
	$\varphi_{\text{logistic}}$	a = 2537.5 b = 14.8	-3.85	39.20	-2.80	44.90	a = 11863 b = 54.149	-6.18	41.53	-6.52	47.65	a = 3618 b = 50.1	-0.90	**36.84**	-5.00	**49.50**
	φ_{Linear}	$C1$ = 0.30 $C2$ = 0.80	-3.60	38.94	-2.25	44.65	$C1$ = *0.157* $C2$ = *0.162*	-5.90	**40.90**	-6.37	**47.11**	$C1$ = *0.116* $C2$ = *0.152*	-2.81	38.20	-6.37	49.51
RB	φ_G	a = 55 b = 9	-3.50	39.10	-2.30	44.90	a = 23.950 b = 13.661	**-5.80**	41.52	**-6.29**	47.70	a = 2134 b = 43	-0.80	37.90	-6.20	52.17
	$\varphi_{\text{logistic}}$	a = 2791 b = 15	-3.70	39.30	-2.70	45.00	a = 19593.7 b = 58.417	-6.10	41.67	-6.49	47.70	a = 3618.2 b = 50.3	**-0.80**	37.70	**-4.90**	50.90
	φ_{Linear}	$C1$ = 0.296 $C2$ = 0.768	**-3.20**	38.90	**-1.90**	44.70	$C1$ = *0.093* $C2$ = *0.267*	-5.87	41.67	-6.35	47.74	$C1$ = *0.100* $C2$ = *0.112*	-0.90	39.20	-5.50	50.95

Best results for each region are in bold character.

Table 2. Results of the DBRFA Approach With and Without Depth Iterations using $\zeta(.) = RRMSE$ and $\varphi = \varphi_G$

	Region													
	Southern Quebec (Canada)					Arkansas (United States)					Texas (United States)				
		QS10		QS100			Q10		Q50			Q10		Q50	
	Optimal coefficients	RB (%)	RR MSE (%)	RB (%)	RR MSE (%)	Optimal coefficients	RB (%)	RR MSE (%)	RB (%)	RR MSE (%)	Optimal coefficients	RB (%)	RR MSE (%)	RB (%)	RR MSE (%)
With iteration	a = 30.5 b = 7	-3.55	38.70	-2.20	44.50	a = 97 b = 25	-6.00	41.50	-6.33	47.70	a = 129.7 b = 35.4	-1.01	36.86	-6.00	50.79
Without iteration	a = 66.50 b = 14.25	-6.60	47.05	-7.52	55.07	a = 721 b = 81	-7.24	42.87	-8.64	50.34	a = 186.7 b = 42.65	-1.60	38.29	-6.29	51.00

Table 3. Quantile estimation result for Quebec with available approaches and their references

Approach	Reference	QS10		QS100	
		RB (%)	RRMSE (%)	RB (%)	RRMSE (%)
Linear regression (LR)	Table 1 above	-9	55	-11	64
Nonlinear regression (NLR)	Shu and Ouarda [2008]	-9	61	-12	70
NLR with regionalisation approach	Shu and Ouarda [2008]	-19	67	-24	79
CCA	Table 1 above	-7	44	-8	52
Kriging-CCA space	Chokmani and Ouarda [2004]	-20	66	-27	86
Kriging-Principal Component Analysis space	Chokmani and Ouarda [2004]	-16	51	-23	70
Adaptive Neuro-Fuzzy Inference Systems (ANFIS)	Shu and Ouarda [2008]	-8	57	-14	64
Artificial Neural Networks (ANN)	Shu and Ouarda [2008]	-8	53	-10	60
Single ANN-CCA (SANN-CCA)	Shu and Ouarda [2007]	-5	38	-4	46
Ensemble ANN (EANN)	Shu and Ouarda [2007]	-7	44	-10	60
Ensemble ANN-CCA (EANN-CCA)	Shu and Ouarda [2007]	-5	**37**	-6	45
Optimal DBRFA	Table 1 above	**-3**	38	**-2**	**44**

Best results are in bold character

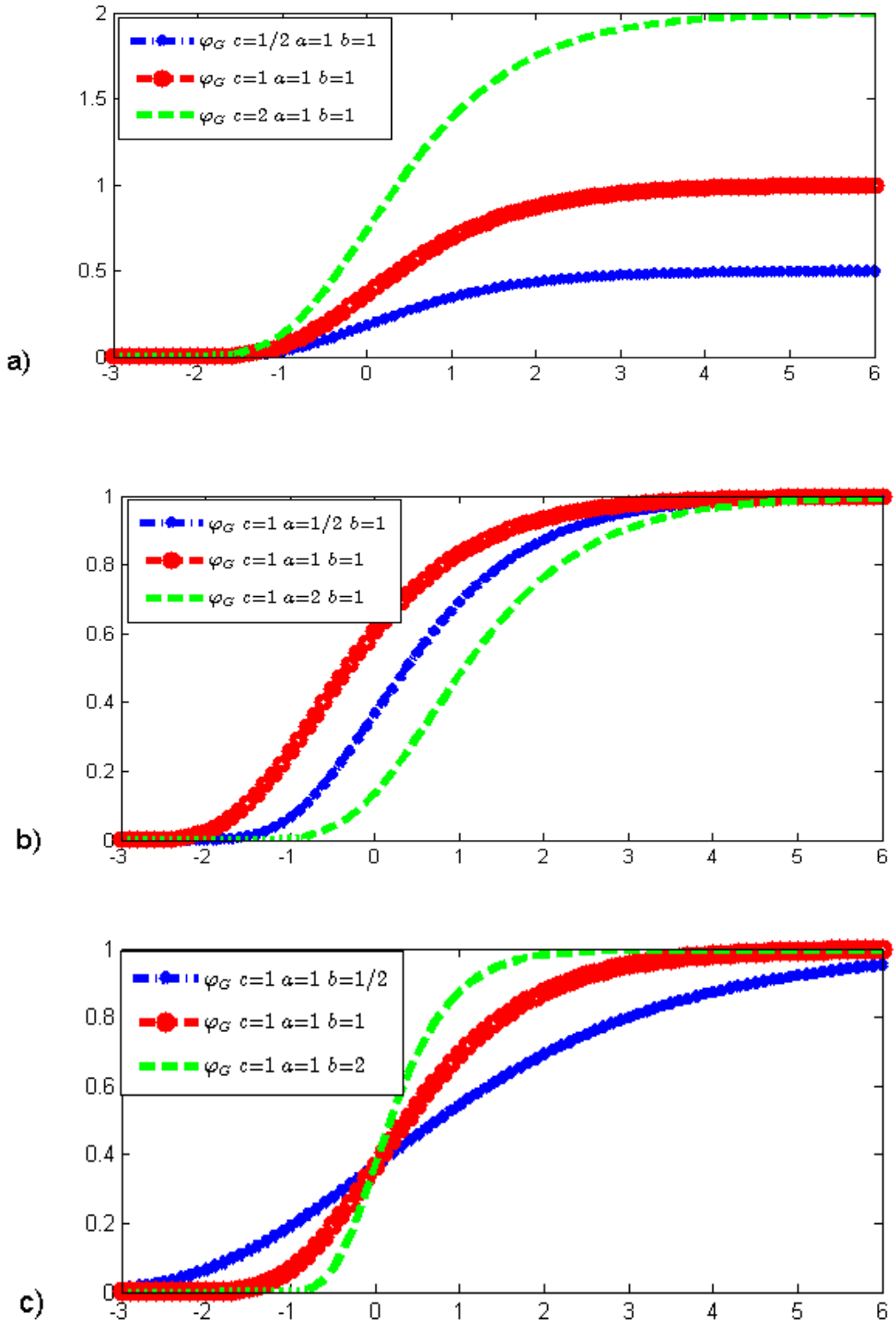


Figure 1. Illustration of Gompertz function: (a) *c* varies with fixed *a* and *b*, (b) *a* varies with fixed *b* and *c* and (c) *b* varies with fixed *a* and c.

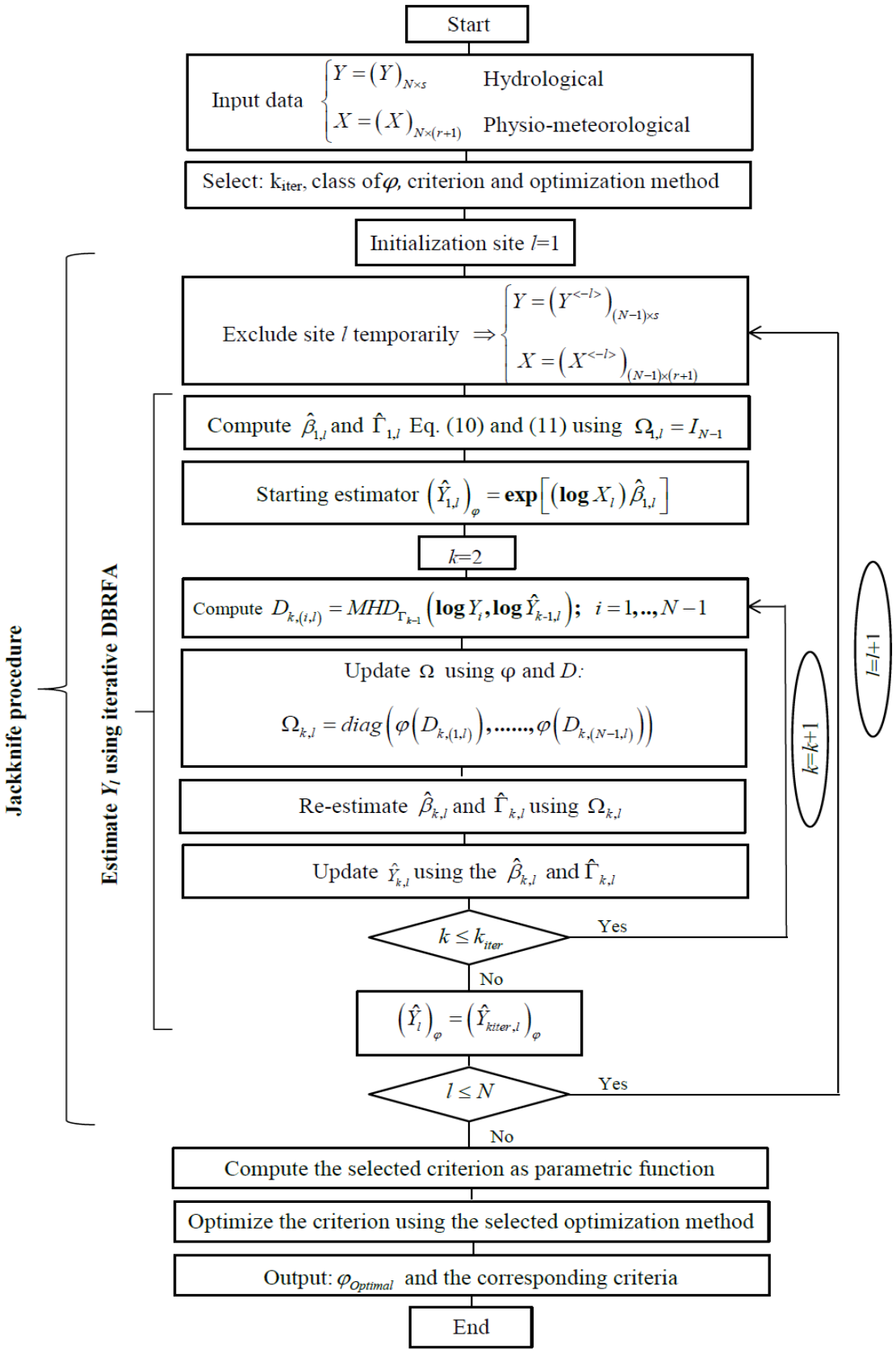


Figure 2. An overview diagram summarizing the optimization procedure of the DBRFA approach.

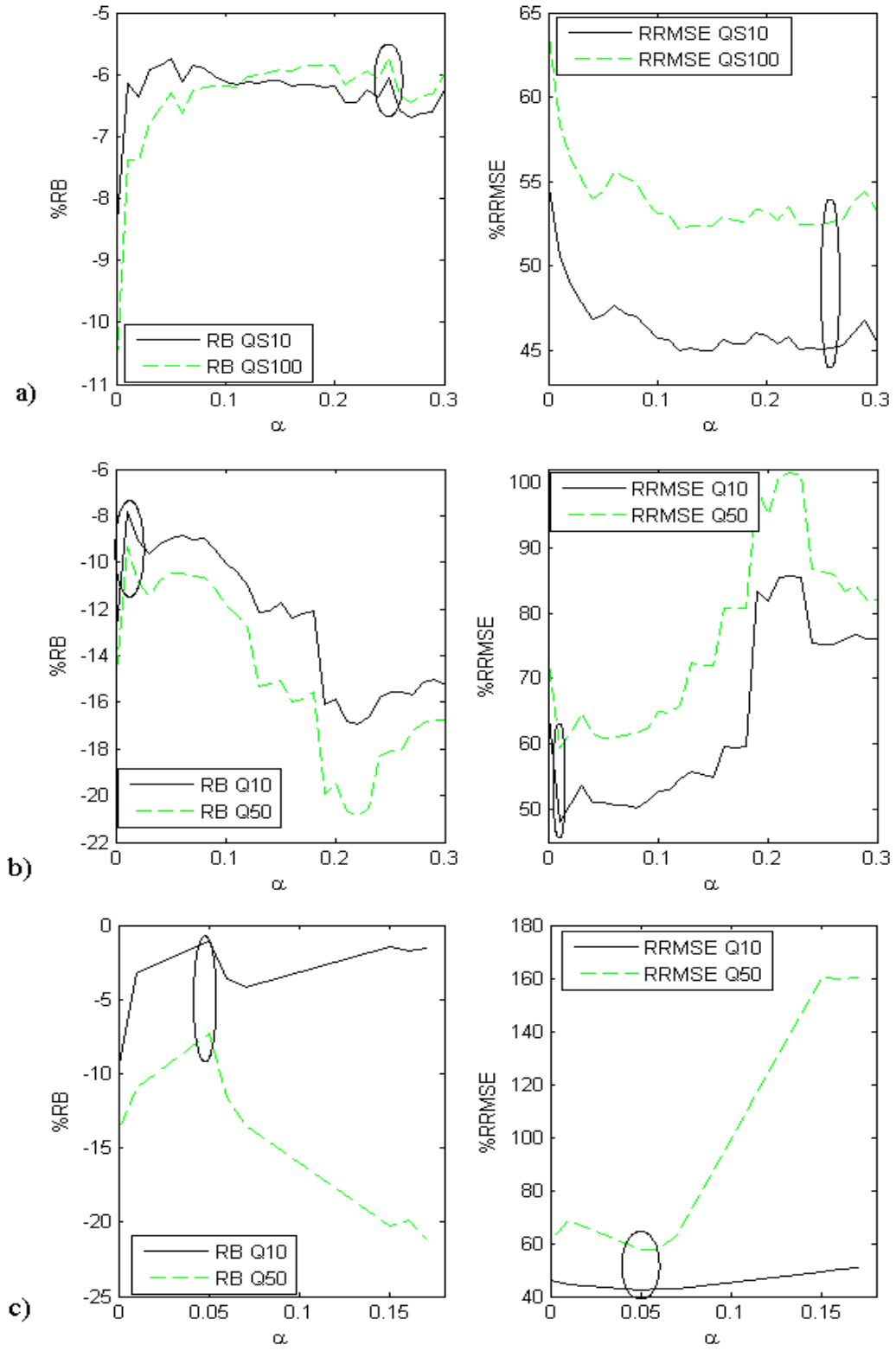


Figure 3. Optimal value of the neighborhood coefficient α for the CCA approach for: (a) Southern Quebec, (b) Arkansas and (c) Texas. The first column illustrates the RB and the second column illustrates the RRMSE.

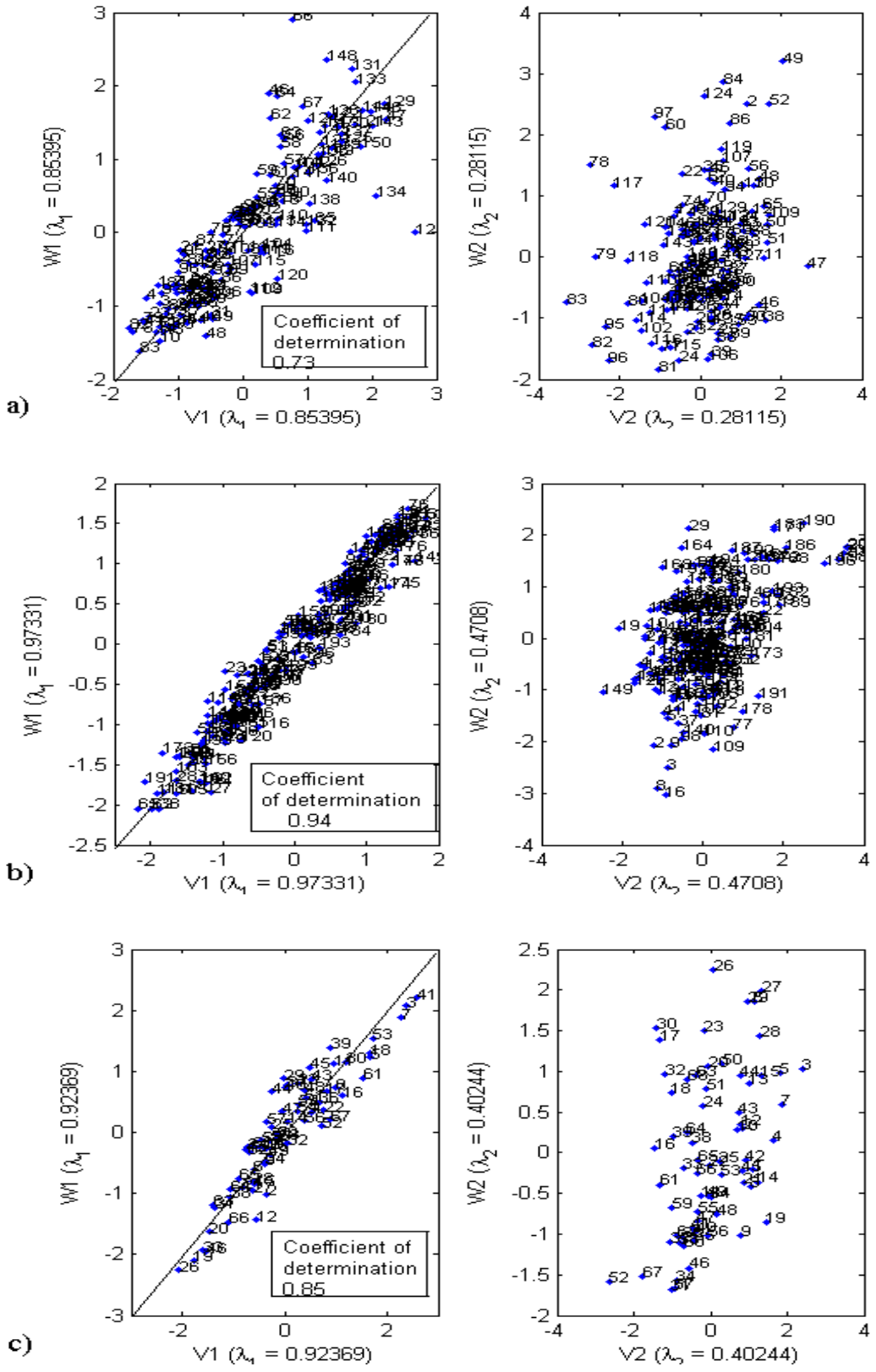


Figure 4. Scatterplot of sites in the canonical spaces (V1, W1) and (V2, W2) for: (a) Southern Quebec, (b) Arkansas and (c) Texas. The first column illustrates the canonical (V1, W1) space and the second column illustrates the (V2, W2) space.

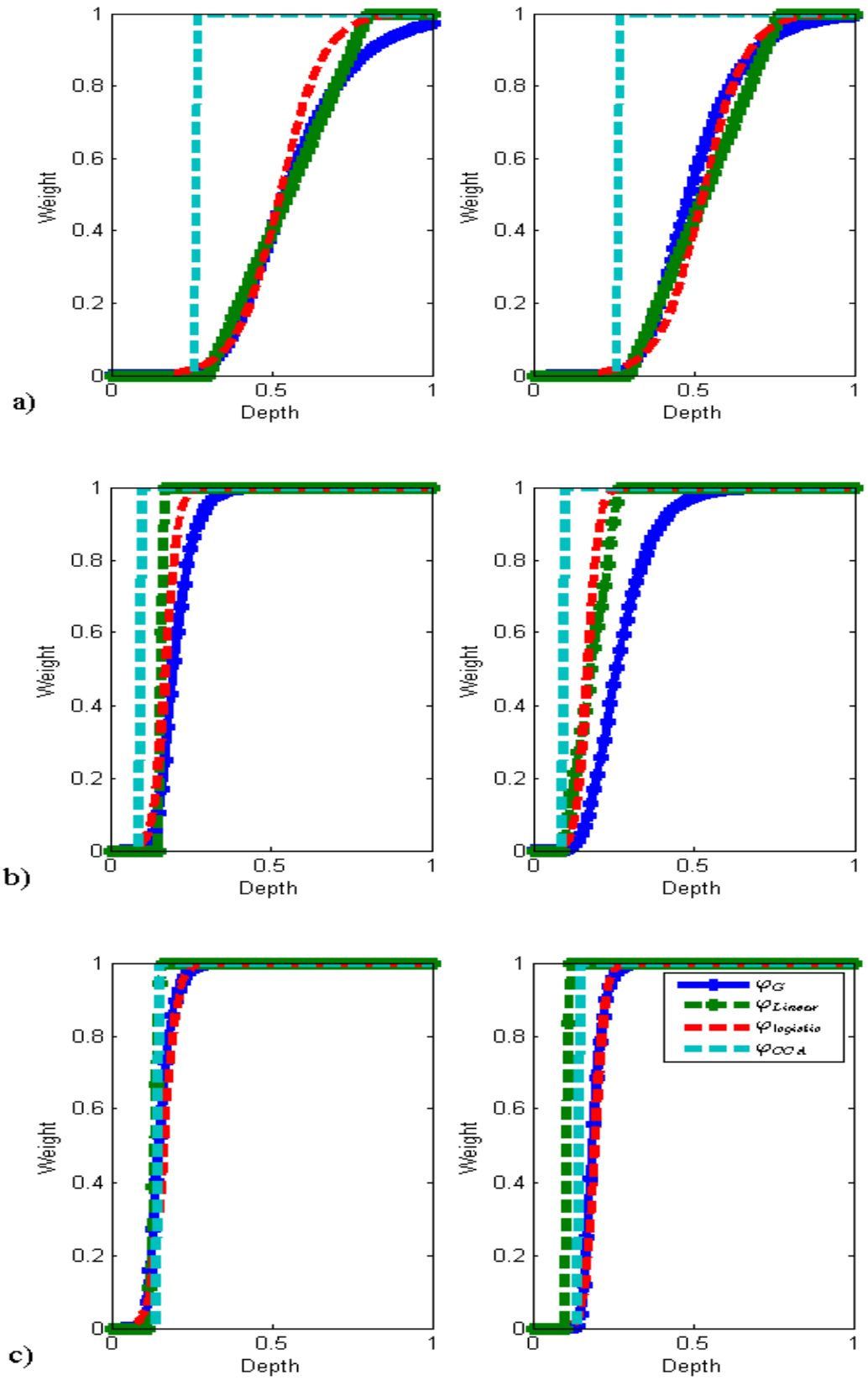
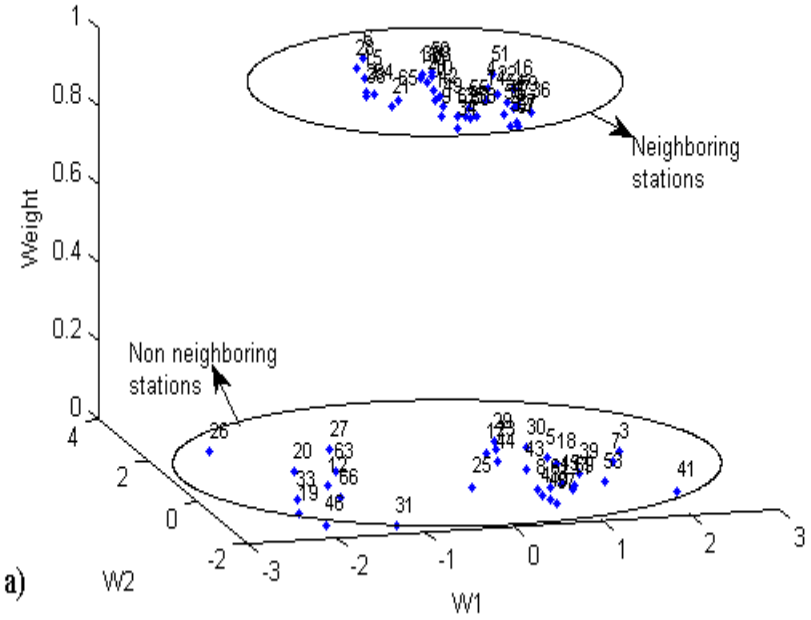


Figure 5. Optimal weight functions for: (a) Southern Quebec, (b) Arkansas and (c) Texas. The first column illustrates the weight functions optimal with respect to RRMSE and the second column illustrates the weight functions optimal with respect to RB.



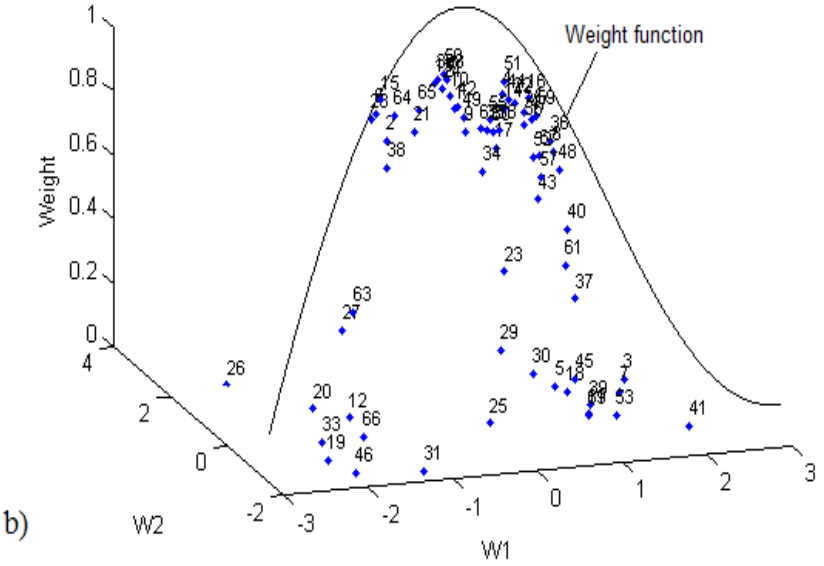


Figure 6. Weight allocated to each gauged-site to estimate the target-site number 25 in the Texas region in the Canonical hydrological space (W1, W2) using: (a) CCA with optimal α and (b) the DBRFA approach with optimal φ_G.

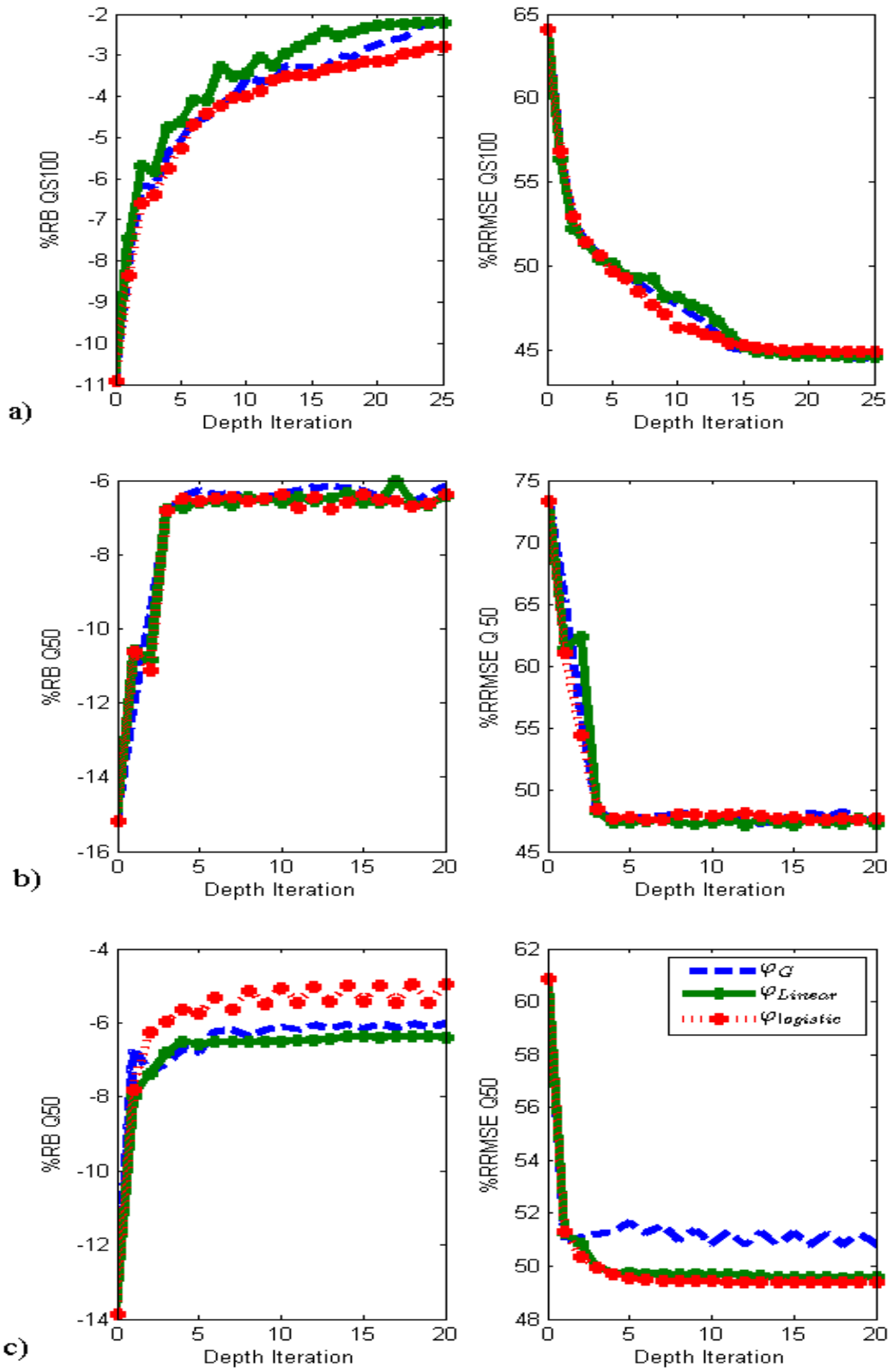


Figure 7. Variation of criteria (RB and RRMSE) as a function of the depth iteration number for the estimation of (a) QS100-Southern Quebec, (b) Q50-Arkansas and (c) Q50-Texas.

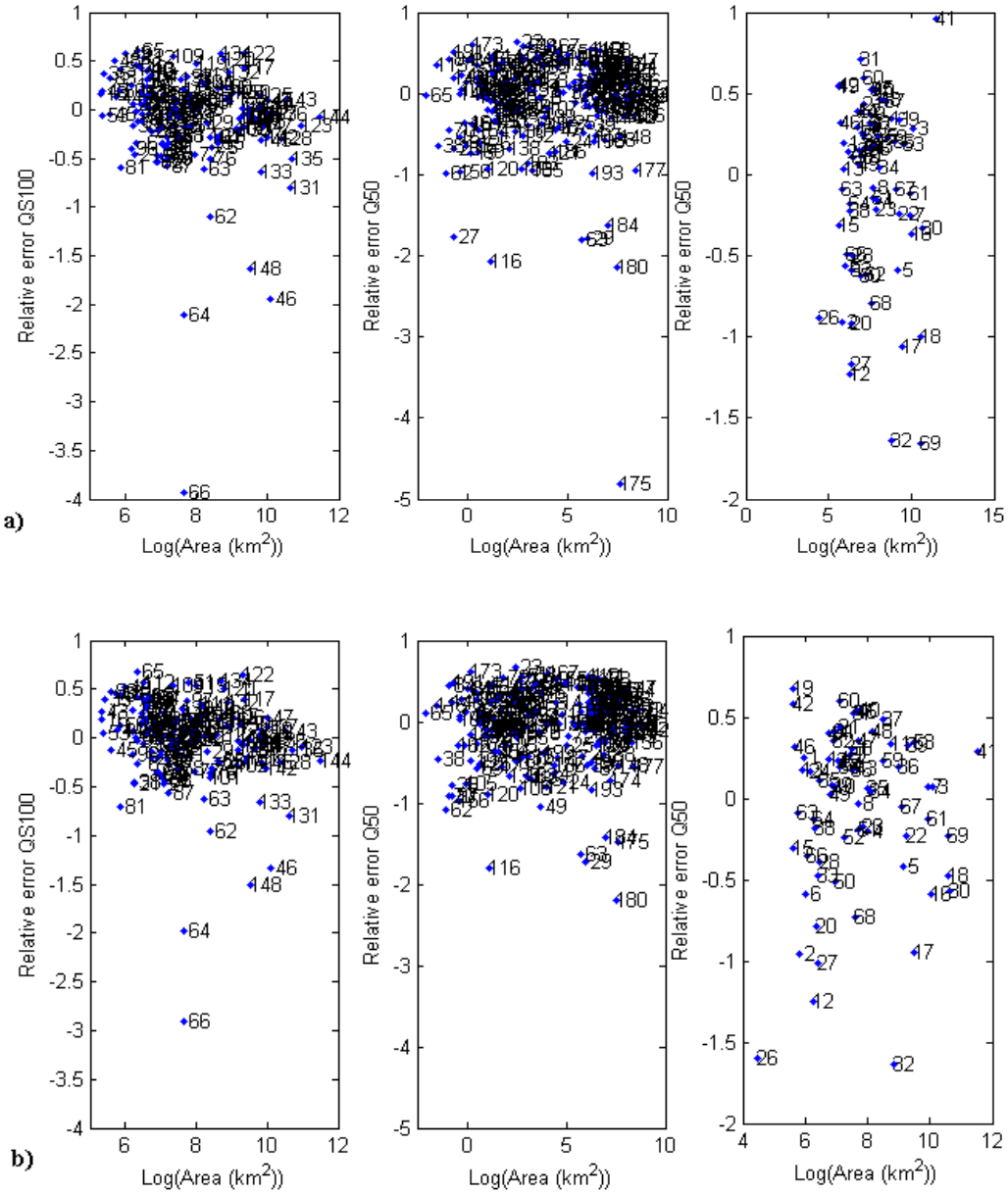


Figure 8. Relative quantile errors using: (a) φ_{CCA} and (b) φ_G. The first column illustrates the error of QS100 in southern Quebec, the second column illustrates the errors of Q50 in Arkansas and the third column illustrates the errors of Q50 in Texas.