TABLE DES MATIÈRES

LISTE DES TABLEAUX

Page

LISTE DES FIGURES

LISTE DES ABRÉVIATIONS, SIGLES ET ACRONYMES

AICA	Akaike Information Criterion Averaging
ASA	Adaptive Simulated Annealing
BICA	Bayes Information Criterion Averaging
BGA	Bates Granger Averaging
BMA	Bayesian Model Averaging
CD	Catchment Descriptor
CMAES	Covariance Matrix Adaptation Evolution Strategy
CRCM	Canadian Regional Climate Model
CS	Cuckoo Search
DDS	Dynamically Dimensioned Search
DE	Differential Evolution
GA	Genetic Algorithm
GRA	Granger – Ramanathan Averaging A
GRB	Granger – Ramanathan Averaging B
GRC	Granger – Ramanathan Averaging C
HS	Harmony Search
IDW	Inverse-Distance Weighting
MOPEX	MOdel Parameter EXperiment
NSE	Nash-Sutcliffe Efficiency
NR	Nash Ratio
PET	Potential EvapoTranspiration
PS	Pattern Search

PSO	Particle Swarm Optimization
RCM	Regional Climate Model
SAM	Simple Arithmetic Mean
SCA	Shuffled Complex Averaging
SCEUA	Shuffled Complex Evolution – University of Arizona
SR	Success Rate

CHAPITRE 1

INTRODUCTION GÉNÉRALE

1.1 Problématique de la thèse

L’hydrologie est la science qui étudie le cycle de l’eau et les interactions de l'eau avec l’environnement. Un des buts premiers de cette science, notamment pour les applications en ingénierie, est de prévoir les débits en rivière en tenant compte des événements météorologiques (pluie, fonte de neige, etc.) afin de gérer les systèmes hydriques (production hydroélectrique, contrôle des inondations) par la prévision des événements potentiellement dangereux (inondations, sécheresses, etc.) (Wurbs 1998). Depuis l’avènement de l’ordinateur, l’outil de prédilection de l’hydrologue est le modèle hydrologique (Pechlivanidis et al. 2011). Ces modèles permettent de simuler les processus hydrologiques sur un bassin versant et ainsi de fournir une estimation des débits en rivière (Singh et Woolhiser 2002). Traditionnellement, l’hydrologue collecte des données météorologiques observées dans le passé ainsi que l’hydrogramme observé sur la même période afin de calibrer le modèle hydrologique par l'ajustement de ses paramètres afin de produire l'hydrogramme simulé le plus similaire à l’hydrogramme observé (Duan et al. 1992). La méthodologie est schématisée à la figure 1.1.

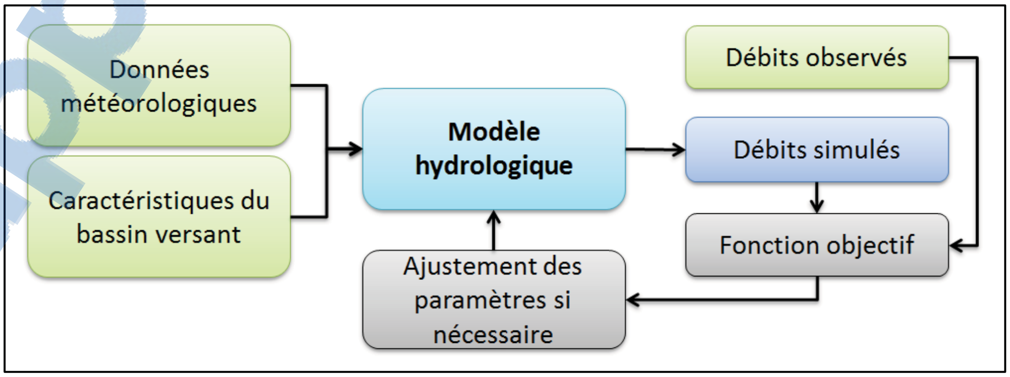

Figure 1.1 Modélisation pluie-débit traditionnelle, telle qu’à un site jaugé

Cette procédure est l'étape du calage du modèle. Lorsque les paramètres optimaux sont fixés, il est possible de valider le modèle sur une période distincte n'ayant pas servi au calage. Si la perforance du modèle est toujours adéquate sur cette nouvelle période, le calage peut être utilisé pour simuler des débits sur des périodes futures (à l'aide de prévisions météorologiques) ou passés (avec les archives météorologiques) en fonction des besoins particuliers du projet à l'étude.

Cependant, il arrive qu'il faille simuler les apports sur un bassin qui ne soit pas équipé d'une station limnimétrique permettant de mesurer les apports en rivière (Sivapalan et al. 2003). Il n'existe donc pas d'historique hydrologique sur ce bassin, que l'on qualifie de « non-jaugé ». L'étape du calage du modèle devient alors impossible puisqu'il n'existe pas de cible à atteindre lors de l'ajustement des paramètres. Par conséquent, la simulation des apports passés et/ou futurs ne peut se faire par la méthode traditionnelle, tel qu'illustré à la figure 1.2.

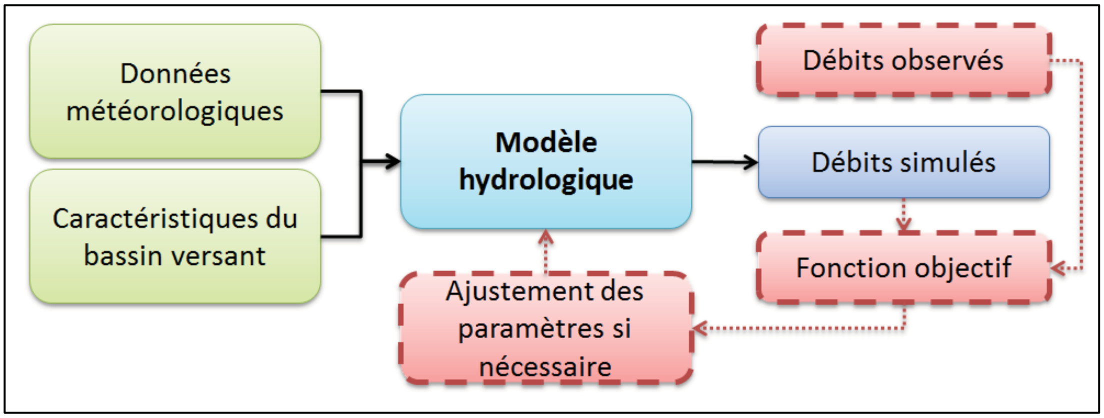

Figure 1.2 Schématisation du processus de modélisation à un site non-jaugé. En raison de l'absence de débits observés, il est impossible d'évaluer la fonction objectif et par conséquent de déterminer les paramètres du modèle hydrologique

Au fil des ans, des techniques particulières ont été développées pour estimer les apports aux sites non-jaugés (He et al. 2011, Parajka et al. 2013). Toutefois, ces méthodes sont encore très imparfaites et soumises à des contraintes inévitables, telles que la qualité des données observées et l'incertitude quant au paramétrage des processus physiques régissant le comportement hydrologique d'un bassin versant (Samuel et al. 2011).

Cette thèse a pour but d'analyser la problématique de la prévision des apports aux bassins non-jaugés, de comprendre les limitations des méthodes de régionalisation existantes et de proposer des améliorations aux méthodes existantes. De plus, diverses techniques permettant d'optimiser la modélisation hydrologique (multi-modèle, calage automatique des paramètres, données climatiques alternatives) seront analysées afin d'améliorer la prévision aux sites non-jaugés.

1.2 Organisation de la thèse

La thèse est séparée en onze chapitres. Suivant cette introduction, une brève revue de la littérature générale est étalée. Puis, chacun des sept chapitres suivants présente un article soit publié ou soumis dans un journal scientifique et ayant servi à tirer les conclusions de cette thèse. Finalement, une discussion, une conclusion générale ainsi que des recommandations sont présentées.

Le chapitre 3 présente l'article intitulé « A comparison of stochastic optimization algorithms in hydrological model calibration». Cette publication, parue dans le *Journal of Hydrologic Engineering*, compare 10 algorithmes d'optimisation en calage de modèles hydrologiques. Ces travaux ont permis de déterminer quels algorithmes permettaient d'obtenir les meilleurs paramètres pour les divers modèles utilisés pour la suite du projet et de réduire l'incertitude des jeux de paramètres attribuée au calage.

Le chapitre 4 présente l'article « A comparative analysis of 9 multi-model averaging approaches in hydrological continuous streamflow simulation ». Ce projet visait à analyser et sélectionner des algorithmes de pondération multi-modèle afin de combiner les hydrogrammes provenant de plusieurs sources et ainsi améliorer la qualité des simulations. Les conclusions de cet article, accepté pour publication dans le *Journal of Hydrology*, ont servi pour deux autres publications, dont une en prévision aux sites non-jaugés.

Le chapitre 5 présente la publication intitulée « Improving hydrological model simulations using multiple gridded climate datasets in multi-model and multi-input averaging frameworks». Cet article soumis au *Journal of Hydrology* propose une méthode pour moyenner les hydrogrammes provenant d'un modèle auquel on a fourni plusieurs sources de données sur grille. Les résultats mettent la table pour de nouvelles utilisations de données alternatives, telles que les réanalyses et les données issues de modèles régionaux de climat.

Dans le chapitre 6 se trouve l'article « Continuous streamflow prediction in ungauged basins : the effects of equifinality and parameter set selection on uncertainty in regionalization approaches ». Ce travail, publié dans la revue *Water Resources Research*, est le premier de cette thèse qui traite exclusivement de régionalisation et de prévision aux sites non-jaugés. L'effet de l'incertitude paramétrique sur la performance en régionalisation y est étudié, tout comme le sont deux nouvelles méthodes de régionalisation hybrides.

Le chapitre 7 expose la publication intitulée « Multi-model averaging for continuous streamflow prediction in ungauged basins ». Cet article, accepté pour publication dans le *Hydrological Sciences Journal*, utilise les méthodes de pondération multi-modèle afin d'améliorer la prévision aux sites non-jaugés. Il a été démontré que les méthodes multi-modèle ne performent pas aussi bien en régionalistion qu'en simulation classique. Par contre, il a été montré que la performance des ensembles multi-modèles dépend de la robustesse des modèles individuels en régionalisation.

Le chapitre 8 présente l'article « Analysis of continuous streamflow regionalization methods using a regional climate model environment framework ». Ces travaux ont été soumis pour publication au *Hydrological Sciences Journal*. Il s'agit d'une partie importante de cette thèse en raison du nombre et de la qualité des contributions qui en ont découlé. L'abstraction des incertitudes du monde réel et la richesse en données du monde virtuel ont permis d'analyser les méthodes de régionalisation classiques dans un environnement contrôlé. Des contributions sur le niveau de précision des descripteurs physiques ainsi que des limitations fondamentales des méthodes de régionalisation y ont été apportées.

Le chapitre 9 contient le septième et dernier article de cette thèse en tant qu'auteur principal. L'article, intitulé « Parameter dimensionality reduction of a conceptual model for streamflow prediction in ungauged basins», analyse la performance de deux modèles hydrologiques utilisés en régionalisation sous l'angle de l'équifinalité paramétrique. Les conclusions démontrent que l'équifinalité n'est pas nécessairement problématique et que dans l'incertitude entourant la définition des processus physiques, tenter de limiter le nombre de paramètres n'est pas bénéfique. Cette étude a été soumise pour publication à la revue *Advances in Water Resources*.

Deux autres articles connexes à cette thèse sont présentés en annexe. Il s'agit de travaux auxquels l'auteur a contribué mais pour lesquels je n'étais pas le chercheur principal. Le premier, intitulé « Potential of gridded data as inputs to hydrological modeling », est présenté à l'annexe I. Il détaille une analyse de quatre jeux de données sur grille en modélisation hydrologique. Ce sont les mêmes jeux de donnés que ceux utilisés dans l'article 3 présenté au chapitre 5. Il a été soumis au *Journal of Hydrometeorology*. Le second, trouvé à l'annexe II et intitulé « Reducing the parametric dimensionality for rainfall-runoff models : a benchmark for sensitivity analysis methods », a été soumis à *Advances in Water Resources* et analyse la méthode de sensibilité globale de Sobol' à l'aide d'une méthode stochastique plus robuste. La méthode de Sobol' est celle qui a été utilisée pour réduire le nombre de paramètres dans l'article 7 du chapitre 9 du présent document. Ces deux contributions sont donc directement liées aux travaux de cette thèse.

Finalement, plusieurs conférences, présentations orales et par affiches ont permis de disséminer les résultats et les conclusions des travaux effectués. La liste complète des contributions scientifiques liées aux travaux de cette thèse se trouve à l'annexe III.

CHAPITRE 2

REVUE DE LITTÉRATURE

Ce chapitre présente une revue de littérature générale de la prévision hydrologique aux sites non-jaugés, de ses diverses variantes, de sa problématique actuelle et des pistes de solution potentielles. La régionalisation paramétrique, la réduction de l'équifinalité ainsi que les approches multi-modèle seront traitées. Ces trois domaines se recoupent dans cette thèse et chacun d'entre eux est explicitement traité dans les articles respectifs. En conséquence, seule une vue d'ensemble est donnée ici.

2.1 Prévision des apports aux sites non-jaugés

La prévision des apports aux sites non-jaugés, aussi appelé « régionalisation paramétrique » ou simplement « régionalisation », est une approche de modélisation hydrologique qui permet d'estimer l'historique hydrologique d'un bassin versant n'étant pas équipé d'une station limnimétrique (He et al. 2011, Sivapalan et al. 2003). Plusieurs méthodes ont été proposées pour estimer les apports à de tels sites, avec un succès mitigé (Razavi et Coulibaly 2013, Parajka et al. 2013, He et al. 2011). Trois méthodes sont généralement reconnues comme étant les méthodes classiques de régionalisation : la régression linéaire multiple, la proximité spatiale et la similitude physique. Cette section présente brièvement les trois méthodes et donne un aperçu des approches alternatives.

2.1.1 Régression linéaire multiple

La première méthode à avoir été utilisée en régionalisation fut la régression linéaire multiple. Pour cette technique, un modèle de régression linéaire est construit à partir de descripteurs physiques (superficie, pente, élévation, couverture au sol, etc.) des bassins versants jaugés où le modèle hydrologique a été préalablement calé. Les descripteurs servent de prédicteurs pour le modèle de régression, alors que la valeur du paramètre est la valeur à prédire. Il est alors possible d'estimer la valeur du paramètre au site non-jaugé en se basant sur les caractéristiques de ce dernier, tel qu'illustré à la Figure 2.1. Wagener et Wheater (2006) ont exprimé la méthode selon la formulation suivante :

$$\hat{\theta}_L = H_R\left(\theta_R \middle| \phi\right) + v_R \tag{2.1}$$

où $\hat{\theta}_L$ représente la valeur estimée du paramètre au bassin non-jaugé, H_R est le modèle de régression liant les paramètres calés du modèle hydrologique sur les bassins jaugés θ_R aux descripteurs des bassins jaugés (ϕ) et v_R représente l'erreur résiduelle du modèle de régression.

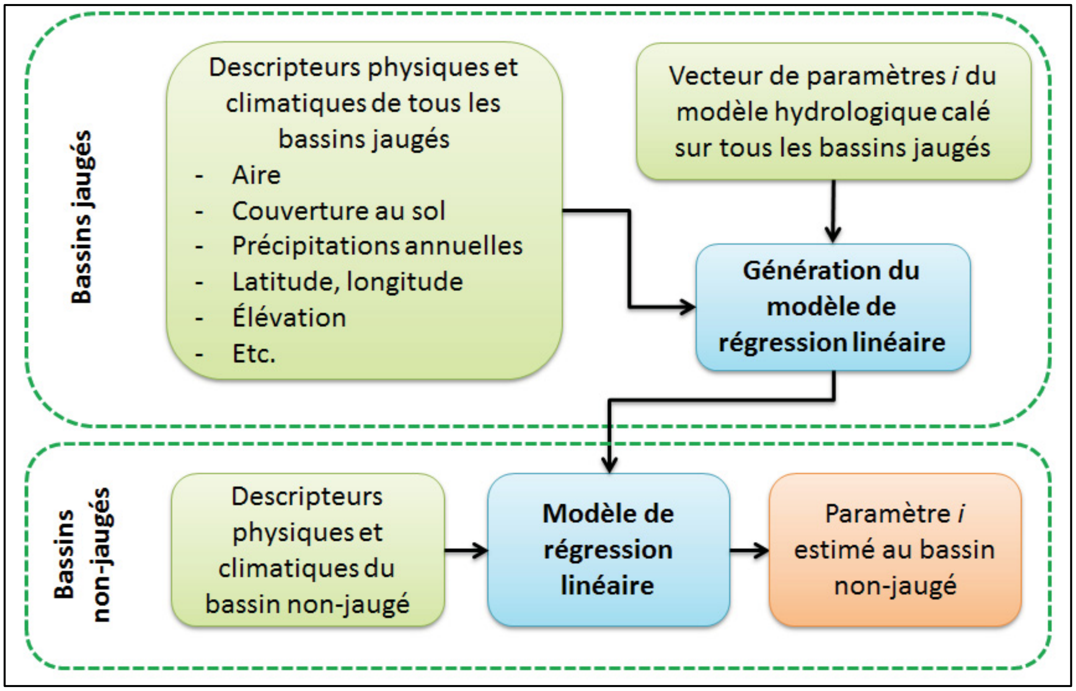
Bassins jaugés
Descripteurs physiques et climatiques de tous les bassins jaugés
- Aire
- Couverture au sol
- Précipitations annuelles
- Latitude, longitude
- Élévation
- Etc.
Vecteur de paramètres *i* du modèle hydrologique calé sur tous les bassins jaugés
Génération du modèle de régression linéaire
Bassins non-jaugés
Descripteurs physiques et climatiques du bassin non-jaugé
Modèle de régression linéaire
Paramètre *i* estimé au bassin non-jaugé

Figure 2.1 Schéma de la méthode de régression linéaire multiple

Le processus est répété pour chacun des paramètres, générant ainsi un jeu de paramètres complet à utiliser sur le site non-jaugé. Cette méthode a l'avantage d'être simple à mettre en œuvre, mais elle fait l'hypothèse que les paramètres ne sont pas corrélés entre eux et qu'ils sont des fonctions des descripteurs physiques. Effectivement, les paramètres sont estimés de manière indépendante. Cette méthode ne tient donc pas compte de l'interaction paramétrique qui est attendue dans un modèle hydrologique de grande dimensionnalité, soit qui contient plusieurs paramètres corrélés. Parajka et al. (2013) ont montré que la méthode de régression linéaire multiple était généralement moins performante que les autres approches, mais était en mesure d'égaler leur performance dans des régions arides.

2.1.2 Proximité spatiale

Les méthodes de proximité spatiale diffèrent des approches de régression puisqu'elles transfèrent des jeux de paramètres entiers, conservant la cohérence entre les paramètres. Son mode de fonctionnement est basé sur l'utilisation des paramètres du modèle hydrologique calé sur le bassin versant le plus près du site non-jaugé. L'hypothèse de cette méthode est que les bassins adjacents possèdent des caractéristiques physiques similaires (couverture végétale, type de sol, etc.) et, par conséquent, ont des régimes hydriques comparables. Elle ne nécessite donc pas de descripteurs physiques, outre la latitude et la longitude.

La première étape consiste à classer l'ensemble des bassins versants où le modèle a été préalablement calé en ordre croissant de distance entre ceux-ci et le bassin non-jaugé. Cette distance est habituellement calculée entre les centroïdes des bassins, selon la formulation suivante :

$$d = \sqrt{\left(X_G - X_U\right)^2 + (Y_G - Y_U)^2} \tag{2.2}$$

où d est la distance entre centroïdes, X_G et Y_G sont respectivement la longitude et la latitude du bassin jaugé et X_U et Y_U sont respectivement la longitude et la latitude du bassin non-jaugé. Le bassin minimisant la distance d est sélectionné en tant que bassin donneur. La figure 2.2 schématise le concept derrière la méthode de proximité spatiale.

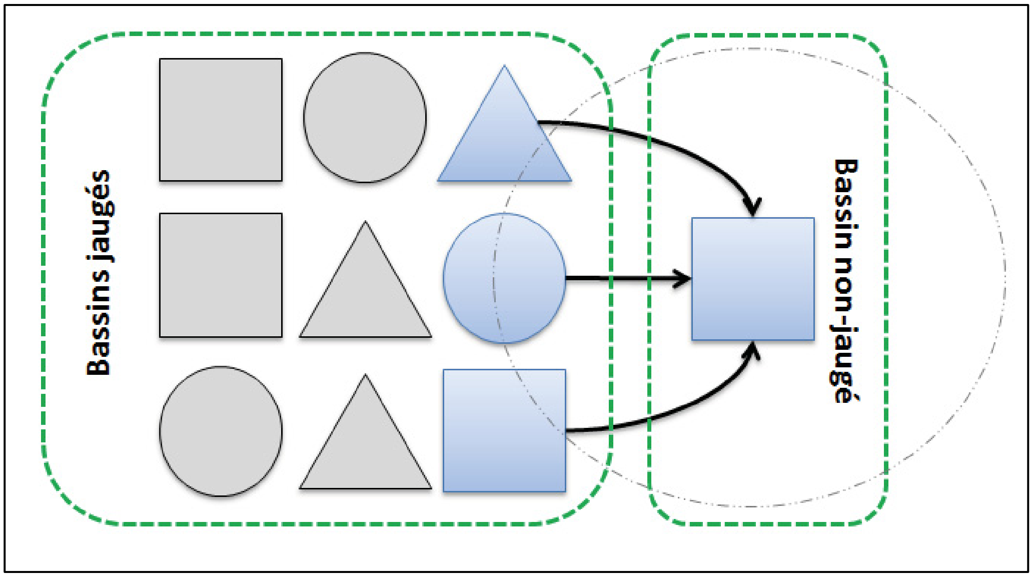

Figure 2.2 Schéma de la méthode de sélection des bassins versants donneurs pour la méthode de proximité spatiale

Les formes géométriques à la figure 2.2 représentent des bassins versants. Les bassins sélectionnés sont en bleu, les autres sont en gris. La méthode de proximité spatiale sélectionne les bassins versants les plus près pour le transfert de paramètres. Le cercle pointillé gris montre que les bassins sont sélectionnés en ordre de distance géographique.

Les paramètres du modèle hydrologique optimisé sur le bassin donneur sont ensuite transférés au bassin non-jaugé, sur lequel le modèle sera exécuté. Ceci permet de conserver une cohérence entre les paramètres. Cette méthode est utilisée lorsqu'il y a une forte densité de bassins versants (Parajka et al. 2005, Oudin et al. 2008) ou lorsqu'il y a peu de descripteurs physiques disponibles. La méthode de proximité spatiale est candidate aux techniques de moyenne des débits afin d'améliorer la qualité des simulations, tel que décrit à la section 2.2.

2.1.3 Similitude physique

La méthode de similitude physique est un croisement des méthodes de régression linéaire multiple et de proximité spatiale. Elle combine l'avantage de la cohérence paramétrique de la méthode de proximité spatiale avec l'hypothèse que les descripteurs physiques représentent adéquatement les processus hydrologiques simulés dans les modèles.

Comme la méthode de proximité spatiale, il faut d'abord classer les bassins versants par ordre croissant de distance avec le bassin non-jaugé. Cependant, la distance n'est pas calculée selon la position géographique, mais bien à l'aide d'un ensemble de descripteurs physiques. Burn et Boorman (1993) ont proposé la formulation de l'index de similitude afin de normaliser les valeurs des descripteurs et de classer les bassins par degré de similitude :

$$\Phi = \sum_{i=1}^{k} \frac{\left|X_i^G - X_i^U\right|}{\Delta X_i} \tag{2.3}$$

où ϕ est l'index de similitude, X_i^G et X_i^U sont les valeurs du descripteur physique i respectivement des bassins jaugé et non-jaugé, et ΔX_i est la plage de valeurs mesurée dans la base de données pour le descripteur i. Le bassin versant jaugé qui minimise l'index de similitude est sélectionné en tant que bassin versant donneur. La Figure 2.3 illustre ces propos de manière schématisée.

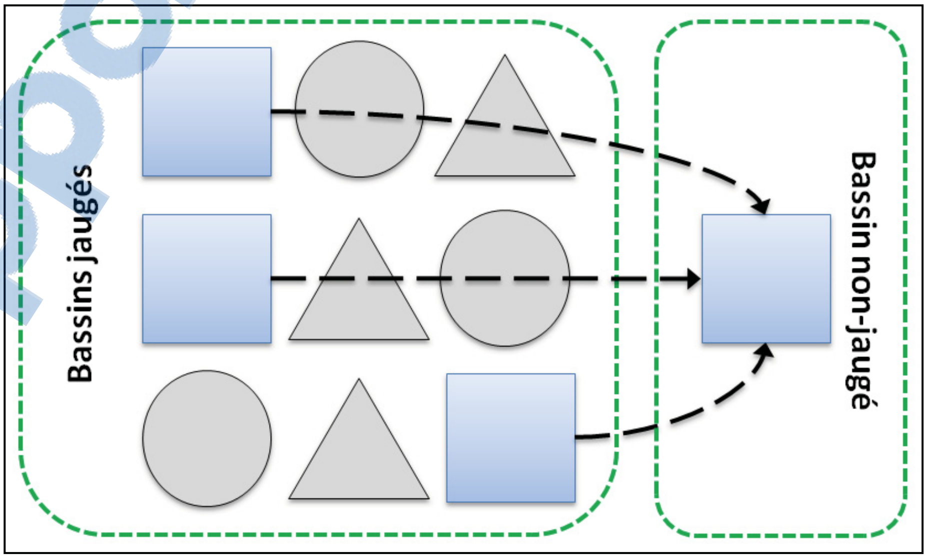

Figure 2.3 Schéma de la méthode de sélection des bassins versants donneurs pour l'approche de similitude physique

La méthode de similitude physique sélectionne donc les bassins les plus similaires comme donneurs. Les formes géométriques bleues à la figure 2.3 représentent les bassins versants sélectionnés tandis que les formes en gris ne l'ont pas été.

Les paramètres du modèle provenant de ces bassins sont transférés au bassin non-jaugé, où le modèle sera exécuté. Cette méthode requiert donc plus de données que la proximité spatiale. En contrepartie, il est attendu que les processus hydrologiques des bassins semblables soient plus similaires. Oudin et al. (2010) ont cependant montré qu'il ne s'agit pas nécessairement d'une hypothèse valide et que la relation entre les descripteurs physiques et les processus hydrologiques était plus complexe que prévu. La méthode de similitude physique peut également être utilisée dans l'optique de moyenner les apports provenant de plusieurs donneurs, tel que décrit à la section 2.2.

2.1.4 Autres méthodes

Outre ces trois approches classiques, certaines autres propositions ont été apportées dans la littérature. La plus commune est l'approche par krigeage, où les paramètres des modèles sont interpolés dans l'espace selon leur corrélation spatiale. D'excellents résultats ont été rapportés dans la littérature (Vandewiele et Elias 1995, Parajka et al. 2005). Cependant, les modèles utilisés dans ces études étaient généralement de dimensionnalité restreinte (11 paramètres et moins) et les bassins versants très rapprochés. Par exemple, Parajka et al. (2005) ont trouvé que le krigeage était la meilleure approche de régionalisation sur 320 bassins versants en Autriche. En contrepartie, certains auteurs ont montré que le krigeage n'était pas la meilleure approche. Par exemple, Samuel et al. (2011) ont utilisé la méthode du krigeage sur 94 bassins en Ontario, Canada. Ils ont trouvé que le krigeage performait de manière équivalente ou légèrement inférieure à l'approche de proximité spatiale pondérée par l'inverse de la distance, mais qu'il était beaucoup plus complexe à mettre en œuvre. Ces méthodes étaient d'ailleurs moins performantes qu'une variante de la proximité spatiale. De plus, ils ont filtré leurs bassins de manière à conserver uniquement ceux qui présentent une bonne performance en calage, ce qui pourrait avoir influencé les résultats.

Une autre technique employée est celle du calage local ou du calage global. Cette approche vise à réduire l'incertitude paramétrique en calant l'ensemble des bassins d'une même région d'un seul coup, trouvant ainsi le meilleur jeu de paramètres commun (Ricard et al. 2013). Un bassin non-jaugé se trouvant dans cette région reçoit le même jeu de paramètres que les autres bassins de sa région. Cette méthode comporte deux désavantages. Premièrement, elle ne permet pas de différencier certaines caractéristiques propres aux bassins individuels, donnant ainsi lieu à une qualité moindre en calage. En validation et en régionalisation, cette méthode repose sur l'hypothèse que les caractéristiques des bassins versants de la région sont similaires. Une variante de cette méthode est de caler les bassins indépendamment les uns des autres, puis de moyenner la valeur des paramètres au site non-jaugé se trouvant dans la dite région. Ceci néglige également l'interaction paramétrique. Les résultats sont souvent décevants selon cette méthode (Parajka et al. 2005, 2013).

Il existe, en parallèle, une catégorie de méthodes ne nécessitant pas de modèle hydrologique pour prévoir les apports aux sites non-jaugés. Ces méthodes requièrent des séries complètes à transférer et nécessitent généralement une forte densité de bassins afin de préserver le lien climatologique entre le bassin versant donneur et le bassin non-jaugé. Parmi ces approches, on peut noter les réseaux de neurones artificiels (Goswami et al. 2007), les ratios des aires contributrices et des courbes de durée-fréquence (Mohamoud 2008).

Plusieurs variantes des méthodes énumérées ici sont détaillées dans Razavi et al. (2013) et He et al. (2011) mais ne seront pas traitées dans cette thèse en raison de leurs similitudes intrinsèques.

2.2 Donneurs multiples

Les méthodes de régionalisation basées sur les bassins donneurs ont l'avantage de conserver la cohérence des paramètres lors du transfert au site non-jaugé. Cependant, suivant l'approche classique, seule l'information d'un seul donneur est prise en compte, contrairement à la méthode de régression linéaire. Pour pouvoir tirer profit d'une certaine diversité des bassins dans ces approches de régionalisation, il est possible d'utiliser plusieurs donneurs. Trois méthodes sont couramment employées pour y parvenir. Les figures 2.2 et 2.3 ont déjà montré la démarche pour sélectionner les bassins versants donneurs.

La première est la moyenne des paramètres des donneurs (Oudin et al. 2008). Elle repose sur l'idée que de simuler un seul hydrogramme basé sur la moyenne des paramètres permet de réduire l'erreur sur l'estimation du paramètre optimal. Cependant, cette méthode requiert idéalement une indépendance paramétrique du modèle afin que la moyenne des paramètres représente adéquatement les processus hydrologiques physiques. Ainsi, les valeurs des paramètres de *i* donneurs sont moyennées et fournies au modèle pour ne produire qu'un seul hydrogramme générée par le jeu de paramètres moyen.

La seconde approche, la moyenne arithmétique des apports simulés issus des donneurs, permet de conserver la cohérence paramétrique incluant les corrélations croisées entre paramètres (Oudin et al. 2008, Viney et al. 2009). Dans ce cas, le modèle hydrologique est exécuté au site non-jaugé autant de fois qu'il y a de donneurs en utilisant à chaque itération le jeu de paramètres optimal du bassin donneur. Il y a donc autant d'hydrogrammes simulés que de bassins donneurs. La dernière étape est de calculer la moyenne arithmétique des hydrogrammes simulés. Ceci permet de profiter de la diversité de l'ensemble et de réduire l'impact de simulations individuelles de piètre qualité. Par contre, l'ajout de trop de donneurs différents peut réduire la qualité de la simulation.

La dernière approche est la suite logique de la moyenne arithmétique des apports simulés, laquelle est remplacée par une moyenne pondérée par l'inverse de la distance des hydrogrammes simulés (Samuel et al. 2011, Zhang et Chiew 2009). Le bassin le plus près (ou le plus similaire) est donc pondéré plus fortement que le second, et ainsi de suite. La distance utilisée dans la pondération est soit la distance géographique pour la méthode de proximité spatiale ou la distance de similitude pour la méthode de similitude physique. Ceci assure que les bassins donneurs les plus éloignées (ou les moins similaires) ne soient pas considérés de façon aussi importante que les bassins les plus près (ou les plus similaires).

2.3 Analyse des approches de régionalisation

Les méthodes de régionalisation sont souvent évaluées selon leur capacité à reproduire les hydrogrammes d'un bassin versant (Parajka et al. 2005; Merz et Blöschl 2004). Étant donné qu'une telle comparaison requiert une série hydrométrique observée, il est impératif que ces vérifications soient effectuées sur des sites jaugés. L'évaluation se fait donc sur des sites « pseudo non-jaugés », où l'on prétend que le bassin est non-jaugé mais où on peut apprécier la qualité de la simulation. Le critère d'évaluation le plus commun est le critère de Nash-Sutcliffe (Nash et Sutcliffe, 1970). Le critère de Nash-Sutcliffe est généralement reconnu comme étant un bon compromis entre d'autres métriques telles que le coefficient de détermination (R^2), le biais ou l'erreur quadratique moyenne (RMSE), quoi qu'il pondère les crues plus fortement que les étiages en raison de sa nature quadratique. Il est également prisé en raison de son omniprésence dans la littérature, rendant ainsi son utilisation nécessaire pour des fins de comparaison avec d'autres projets similaires (Parajka et al. 2013). Le critère de Nash-Sutcliffe se calcule selon l'équation 2.4 :

$$NSE = 1 - \frac{\sum_{i=1}^{T}\left(Q_o^t - Q_m^t\right)^2}{\sum_{i=1}^{T}\left(Q_o^t - \overline{Q_o}\right)^2} \tag{2.4}$$

où *T* est le nombre de pas de temps, $Q^t{}_o$ est le débit observé au temps *t*, $Q^t{}_m$ est le débit simulé au temps *t* et $\overline{Q_o}$ est le débit observé moyen.

Puisque l'évaluation des méthodes de régionalisation doit être effectuée sur des bassins jaugés, la coutume est de tester les méthodes sur tous les bassins versants disponibles (où le modèle hydrologique a été préalablement mis sur pied). Les bassins sont donc considérés à tour de rôle comme étant pseudo non-jaugés et l'ensemble des autres bassins versants sont utilisés comme sources d'information soit pour la régression, soit en tant que bassins versants potentiellement donneurs. Il s'agit d'une approche de validation croisée dite de « leave-one-out » (Merz et Blöschl 2004). Les distributions des valeurs de Nash-Sutcliffe sur l'ensemble du jeu de bassins sont comparées entre les méthodes de régionalisation afin de tirer des conclusions sur la capacité de prévoir les apports aux sites non-jaugés.

2.4 Problèmes constatés en régionalisation

Malgré les percées récentes et la panoplie de méthodes de régionalisation disponibles, la prévision des apports aux sites non-jaugés demeure un problème compliqué auquel doivent faire face les hydrologues (Sivapalan et al. 2003). Plusieurs éléments expliquent cette difficulté.

Premièrement, la complexité des processus hydrologiques les rend difficilement modélisables, nécessitant des paramétrages pour pallier les lacunes de nos connaissances fondamentales (Wagener et Wheater 2006). Ces paramètres doivent donc être calibrés afin que le modèle hydrologique reproduise le plus fidèlement les observations hydrométriques. Mais en raison du nombre de paramètres et de processus modélisés, plusieurs paramètres deviennent corrélés entre eux, ajoutant un niveau de complexité quant à l'interprétation de leurs valeurs (Beven 2006a, 2006b). Il s'agit du phénomène d'équifinalité, où plusieurs jeux de paramètres peuvent retourner des performances similaires pour des raisons différentes. Il est impossible dès lors de savoir quel jeu de paramètres est le plus représentatif de la réalité. Wagener et Wheater (2006) notent que les interactions entre paramètres causent des difficultés d'interprétation des résultats en raison de l'incertitude générée. Pour la même raison, les choix faits lors de l'élaboration de la structure des modèles hydrologiques peuvent être une source d'incertitude supplémentaire (Lee et al. 2005).

Ensuite, la disponibilité et la qualité des données posent problème puisque les jeux de paramètres calés sur les bassins jaugés en sont directement dépendants (He et al, 2011). Il s'agit d'un problème persistant puisque la grande majorité des bassins versants ont des historiques hydrométriques et météorologiques relativement courts et sont certainement soumis à des biais et des erreurs d'observation. Ces problèmes ajoutent une nouvelle source d'incertitude dans les hydrogrammes simulés.

Finalement, la non-stationnarité climatique et hydrologique (météorologie, couverture au sol, occupation du territoire) implique des incertitudes supplémentaires lors de tentatives de prévision des apports aux sites non-jaugés. Peu d'études ont été entreprises sur le sujet qui reste évasif pour l'instant. La somme cumulée des sources d'incertitudes font en sorte qu'il est difficile de départager les forces et faiblesses des diverses méthodes de régionalisation et de prévoir leur niveau de performance sur un site réellement non-jaugé (Samuel et al. 2011).

2.5 Solutions potentielles

Afin de remédier aux problèmes courants en régionalisation paramétrique, plusieurs pistes de solution sont explorées. Ces solutions ont déjà été employées en simulation traditionnelle (à sites jaugés) et ont montré de bons résultats. Entre autres, la réduction de l'équifinalité (par l'amélioration du calage et en réduisant le nombre de paramètres) et les approches multi-modèle sont traitées.

2.5.1 Amélioration du calage

L'équifinalité liée aux jeux de paramètres optimaux des bassins versants donneurs apporte nécessairement une augmentation de l'incertitude. Pour restreindre le domaine des paramètres équivalents, des chercheurs ont proposé des algorithmes d'optimisation qui permettent de trouver des valeurs de paramètres de plus en plus près de l'optimum lors du processus de calage (Li et al. 2010, Moradkhani et Sorooshian 2009, Tolson et Shoemaker 2007).

Franchini et al. (1998) et Blasone et al. (2007) ont effectué des comparaisons entre algorithmes d'optimisation en calage automatique de modèles hydrologiques. Ils concluent que certains algorithmes sont mieux adaptés que d'autres en fonction du type de problème à résoudre. De plus, l'ajout de certaines informations sur le bassin dans le processus de calage permet de mieux cibler la zone optimale et de réduire l'équifinalité. Par exemple, Li et al. (2009) ont utilisé la variable LAI (« Leaf-Area Index ») afin d'estimer les propriétés liées à l'évapotranspiration pour forcer le calage dans une région restreinte de l'espace paramétrique. L'amélioration des stratégies de calage pourrait donc être un outil de réduction de l'équifinalité en régionalisation.

2.5.2 Réduction du nombre de paramètres

La notion de parcimonie est fortement encouragée dans la littérature en lien avec la régionalisation paramétrique (Razavi et Coulibaly 2013, Yadav et al. 2007). Un modèle ayant peu de paramètres devrait générer moins d'incertitude qu'un modèle surparamétré en raison du moins grand nombre d'interactions et corrélations possibles (Valéry et al. 2014).

Huang and Liang (2006) ont retravaillé le schéma de sous-surface de leur modèle hydrologique (VIC-3L) afin de réduire le nombre de paramètres à caler de 3 à 1. Ceci a été entrepris dans le but de réduire l'incertitude paramétrique pour des applications en régionalisation. La réduction du nombre de paramètres permet d'améliorer la robustesse du modèle et donc de réduire l'incertitude lors des simulations. Leurs résultats démontrent qu'il est possible de réduire l'incertitude (donc de produire des jeux de paramètres plus constants) en fixant les paramètres les plus influents, mais que le niveau de performance du modèle n'est pas amélioré.

Tang et al. (2007) ont analysé et comparé des méthodes de sensibilité globale en vue de permettre l'élimination de certains paramètres moins influents lors du processus de calage. Ils ont déterminé que la méthode de Sobol' était la plus robuste et la plus efficace (Sobol' 1993). Depuis, plusieurs études appliquent la méthode de Sobol' en contexte de modélisation pluie-débit classique afin de réduire l'incertitude paramétrique. Nossent et al. (2011) ont réduit le nombre de paramètres du modèle SWAT de 26 à 9 sans perte de performance avec cette méthode. Van Werkhoven et al. (2009) ont été en mesure d'éliminer de 30% à 40% des paramètres du modèle Sacramento sans aucun effet néfaste. Zhang et al. (2013) ont également appliqué la méthode de Sobol' avec succès sur le modèle SWAT. La réduction du nombre de paramètres semble donc être une avenue prometteuse pour des applications aux sites non-jaugés.

2.5.3 Modélisation multi-modèle

La modélisation multi-modèle est étudiée depuis une quinzaine d’années dans le domaine de la prévision météorologique (Mylne et al. 2002, Bowler et al. 2008, Davolio et al. 2008). Le concept est basé sur le fait que les modèles de prévision météorologique (et les modèles hydrologiques par le fait même) induisent des erreurs liées à la structure du modèle aux simulations. L’utilisation de plusieurs modèles permet de produire des simulations qui surestiment la réalité et d’autres qui la sous-estiment. En moyennant les sorties de modèles, il est possible de réduire l’erreur structurelle, tel qu’illustré à la figure 2.4.

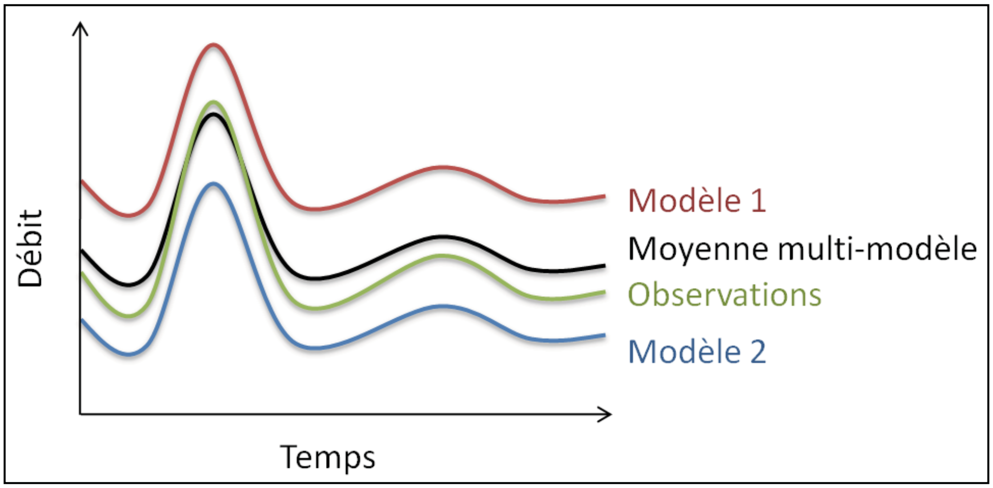

Figure 2.4 Hydrogrammes observé, simulés et moyennés pour les approches multi-modèles

Les méthodes de pondération multi-modèle permettent d’agréger plus efficacement les sorties de modèle en fonction de leur performance sur une période de calage.

En effet, les modèles sont lancés sur la période de calage, générant ainsi un hydrogramme par modèle utilisé. Puis, les méthodes de pondération ajustent les poids de chacun des modèles afin de minimiser l’écart entre la série moyennée et la série observée. La même pondération est ensuite appliquée en validation. Les résultats publiés dans la littérature font tous foi d’une amélioration en utilisant les approches multi-modèles. Une publication phare (Diks et Vrugt 2010) a comparé plusieurs algorithmes de pondération multi-modèle afin de déterminer laquelle était à privilégier en modélisation hydrologique. Ils ont trouvé que la méthode simple de Granger et Ramanathan (Granger et Ramanathan 1984), qui minimise l'erreur quadratique journalière, était plus performante et plus robuste que des méthodes plus complexes telles que les approches bayésiennes. Shamseldin et al. (1997) et Ajami et al. (2006) ont également montré que l'utilisation de modèles multiples améliore les simulations hydrologiques.

See et Openshaw (2000) ont appliqué un concept multi-modèle en prévision hydrologique sur des bassins jaugés. Leur approche permettait de sélectionner le meilleur modèle à chaque pas de temps simulé en fonction des apports passés et de la météo future. Il ne s'agit pas ici d'une approche de pondération, mais de changements de modèles successifs selon les conditions hydrologiques anticipées. Ce projet a toutde même permis de constater l'utilité d'avoir plusieurs modèles en fonction de leurs forces et faiblesses.

Dans un contexte d'analyse d'incertitude en régionalisation, McIntyre et al. (2005) ont usé de modèles multiples, Goswami et al. (2007) ont testé les approches multi-modèle en régionalisation et ont trouvé que les améliorations notées dans la période de calage ne se sont pas matérialisées lors des simulations en validation. Viney et al. (2009) ont utilisé 5 modèles en régionalisation et ont constaté que la qualité des simulations aux sites non-jaugés n'était pas significativement meilleure qu'avec les modèles individuels. Cependant, ils ont constaté que l'utilisation de donneurs multiples dans la sélection des paramètres du modèle jouait un rôle important. et améliorait les simulations.

La modélisation multi-modèle a donc des avantages indéniables et a fait ses preuves dans le domaine de la simulation mais les résultats en régionalisation sont peu concluants. Toutefois, une application en régionalisation à grande échelle, soit sur un grand nombre de bassins versants, n'a à présent jamais été entreprise.

CHAPITRE 3

ARTICLE 1 : A COMPARISON OF STOCHASTIC OPTIMIZATION ALGORITHMS IN HYDROLOGICAL MODEL CALIBRATION

Richard Arsenault[1], Annie Poulin[1], Pascal Côté[2] et François Brissette[1]

[1] Département de Génie de la Construction, École de technologie supérieure,
1100 Notre-Dame Ouest, Montréal, Québec, Canada H3C 1K3.

[2] Rio Tinto Alcan, 1954 Davis, Jonquière, Québec, Canada G7S 4R5.

Article publié dans la revue « Journal of Hydrologic Engineering » en 2014.

Abstract

Ten (10) stochastic optimization methods (Adaptive Simulated Annealing (ASA), Covariance Matrix Adaptation Evolution Strategy (CMAES), Cuckoo Search (CS), Dynamically Dimensioned Search (DDS), Differential Evolution (DE), Genetic Algorithm (GA), Harmony Search (HS), Pattern Search (PS), Particle Swarm Optimization (PSO) and Shuffled Complex Evolution (SCEUA)) were used to calibrate parameter sets for three hydrological models on ten different basins. Optimization algorithm performance was compared for each of the available basin-model combinations. For each model-basin pair, 40 calibrations were run with the 10 algorithms. Results were tested for statistical significance using a multi-comparison procedure based on Friedman and Kruskal-Wallis tests. A dispersion metric was used to evaluate the fitness landscape underlying structure on each test-case. The trials revealed that the dimensionality and general fitness landscape characteristics of the model calibration problem are important when considering the use of an automatic optimization method. It was shown that the ASA, CMAES and DDS algorithms were either as good as or better than the other methods for finding the lowest minimum, with ASA being consistently amongst the best. It was noted that SCEUA performs better when the model complexity is reduced, whereas the opposite is true for DDS. Convergence speed was also studied, and the same three methods (CMAES, DDS and ASA) were shown to converge faster than the other methods. SCEUA converged nearly as fast as the best methods when the model with the smallest parameter space was used, but was not as worthy in the higher-dimension parameter space of the other models. It was also noted that convergence speed has little impact on algorithm efficiency. The methods offering the worse performance were DE, CS, GA, HS and PSO, although they did manage to find good local minima in some trials. However, the other available methods have generally outperformed these algorithms.

Keywords: Hydrology, Model calibration, Stochastic optimization, Parameter search, Model complexity, Algorithm performance.

3.1 Introduction

Hydrologists have come to rely on hydrological models to foresee events that would otherwise be difficult to forecast: Estimating stream discharge after a given rainfall or snow melt, predicting the increase in discharge volume after a soil conservation treatment or even optimizing reservoir levels for hydropower production are all cases where hydrological models are very useful tools (Singh and Woolhiser 2002). However, for a model to predict events accurately, it must be adapted to the catchment being studied. In almost all cases, hydrological models are dependent on parameters that control certain aspects within the model and that cannot be estimated by measurements or prior information (Beven 2001). For example, some parameters dial down or increase the amount of evapotranspiration calculated by the internal equations. Models can be of varying complexity, ranging from the very simple lumped rainfall-runoff models with very few parameters (under 5), to the very complex physically based, distributed models with dozens or even hundreds of parameters (Moradkhani and Sorooshian 2009). For the model to operate as accurately as possible, these parameters must be fine-tuned through a calibration process. The best parameter set will result in the model replicating the historical discharges as closely as possible when fed the historical inputs required.

This task, even when few parameters are involved, can be a daunting one. Manual calibration is one option, but it is a timely and very laborious process. When complex models are used, it is practically impossible to perform a manual calibration that will find the best possible parameter set. It is more likely to return a local minimum, the quality of which is dependent on the experience of the operator. An alternative is to use automatic calibration algorithms (Moradkhani and Sorooshian 2009; Tolson and Shoemaker 2007). These heuristic or metaheuristic methods will search the parameter space (bounded by upper and lower parameter values for the model) to identify and, ideally, find the global minimum of the error-measure function (also called objective function). When the global minimum is attained, there can be no other parameter set that would allow the model to better represent the observed flows in calibration. However, there is no guarantee that the best parameters in calibration will yield the best results in validation, but this aspect is not covered in the scope of this study.

The main problem in finding the global optimum in hydrological model parameter optimization is that the problems are highly non-linear, multimodal and, most importantly, non-convex (Duan 1992). Since these problems have multiple local minima and it is impossible to prove that a minimum is global in non-convex problems, the algorithms will return parameters that are at least local minima or, in the worst case scenario, the lowest value measured even if it is not a local minimum (Fortnow 2009). The best methods will be the ones who converge to a better quality minimum, as rapidly as possible.

There are many automatic methods available and many new methods are proposed every year. However, they do not all offer the same performance level. Some were created primarily for hydrological model calibration, while others are derived from financial, economic and mechanical engineering backgrounds. Moreover, an algorithm's performance is always a relative measure. Many comparative studies have been put forward to prove or disprove the efficiency of a particular method (e.g. Franchini et al. 1998; Blasone et al. 2007). This study is a larger-scale and more thorough attempt to identify the best amongst 10 different optimization methods when used in a hydrological model automatic calibration context. The methods will be tested on ten different basins and with three hydrological models of different complexities to try and find trends and particularities. Ten basins were used to diversify the fitness landscape topography and thus challenge the optimization algorithms and to allow statistical analysis of the results while keeping computing costs reasonable. Forty (40) trials will be performed with each algorithm on every basin-model combination as independent calibration problems, allowing the application of statistical tests to compare the algorithms' performance. Finally, the algorithm performance is explained by comparing the fitness landscape to the algorithm characteristics. The aim of this study is to try and find the best optimization algorithm for a given hydrologic model calibration problem, based on the problem characteristics.

3.2 Optimization algorithms used in the study

Ten (10) stochastic optimization methods were used in the present study. These ten algorithms were selected based on their mainstream usage (SCEUA, GA, PSO), their adaptability to hydrological model calibration (DDS), their reputation in other fields (ASA, PS, CMAES, DE) and to compare with simpler algorithms such as HS and CS. Some other methods have been intentionally left out due to lack of availability of code. Others have been omitted because they are more basic versions of these "updated" algorithms (e.g. simulated annealing vs. adaptive simulated annealing). Any hybrid method which is a combination of "pure" algorithms was not investigated (e.g. Genetic Algorithm with simplex, Yen et al. 1998; Genetic Algorithm with Ant Colony Optimization, Chen and Lu 2005). All methods were either coded in Matlab or were used with a Matlab C++ wrapper to the optimization function. In this study, we have not considered multi-method approaches because our aim is at providing model users with simple, easy to implement algorithms for operational contexts. Readers who are interested in multimethod approaches are encouraged to read Vrugt and Robinson (2007).

3.2.1 Adaptive Simulated Annealing (ASA)

ASA (Ingber 1989, 1993, 1996) is a modified version of the widespread simulated annealing (SA) algorithm. While the basic SA was easily trapped in local optima unless many model iterations were performed with certain parameters, ASA provides methods to "tunnel" out of the local optima and yields a higher chance of success in finding the global optimum, or a better local minimum. As with basic SA, ASA simulates the process of annealing in metallurgy. Metals are heated then cooled repeatedly to promote the movement and organization of atoms in the metal. This causes them to become organized and improves the general quality of the metal. The heating temperatures and reannealing schedules are simulated in the algorithm and are applied to atoms (population) in the metal. The algorithm can perform parameter optimization in large, multi-dimensional search spaces. The reannealing process allows particles (solution sets) to move through the space before they are cooled down again, providing a method to escape local optima (Kirkpatrick et al. 1983). This method is a robust one when its parameters are correctly adjusted. However, there are many parameters and setting them requires experience. A self-optimizing option is available to help in this step. When properly tuned, it is considered a fast and efficient algorithm at finding global optima in verifiable problems (Ingber 1996).

3.2.2 Covariance Matrix Adaptation Evolution Strategy (CMAES)

CMAES (Hansen and Ostermeier 1997, 2001) is a continuous domain, non-convex, non-linear problem optimization algorithm. It is a second-order approximation method, but instead of using the fitness function directly (which is non-differentiable), it estimates the derivatives of the previously successful candidate solution distribution covariance matrix under a maximum-likelihood principle. By doing so, it tends to maximize the likelihood of the distribution. It can therefore be used in noisy, multimodal, non-smooth and non-continuous problems. It also optimizes its own parameters, therefore reducing the need to experiment. This method has been shown to find global optima more efficiently than other evolutionary strategies on functions where the global optimum is known (Hansen and Ostermeier 1997).

3.2.3 Cuckoo Search (CS)

CS (Yang and Deb 2009) is a relatively new algorithm for parameter optimization. It simulates the way the Cuckoo hens lay their eggs in other bird's nests as an obligate brood-parasite species. While laying their eggs, they remove the host's eggs from the nest (initial sampling). In turn, if the laid egg is of good quality, it will resemble the hosts' eggs (parameter set survives to the next round). The egg will be taken care of by the host so the Cuckoo does not have to expend energy taking care of its egg. However, if the egg is of poor quality, the host will find out and push the egg out of the nest (rejected parameter set). CS is based on the same natural occurrence in nature. The individuals (eggs) in the population (eggs in all nests) have a probability of being of poor quality (ejected eggs) or of surviving and breeding once again (best values). The next generation of cuckoos then lays its eggs elsewhere (using a random walk), and so on until the convergence criteria are met. Cuckoo search can also use Levy Flights instead of random walks, but the random walk version was used in this study.

3.2.4 Dynamically dimensioned search (DDS)

DDS (Tolson and Shoemaker 2007) is an algorithm designed to be an efficient tool for calibration of complex, large parameter space hydrological models. The creators claim that DDS outperforms the shuffled complex evolution – University of Arizona (SCE-UA) when many parameters must be optimized since it is designed for computationally expensive calibrations. The algorithm automatically scales the search space to reduce the number of model evaluations needed to attain the best quality local minimum region of the fitness function. It also has built-in systems to try and avoid local optimum traps.

3.2.5 Differential Evolution (DE)

DE (Storn and Price 1997; Pedersen 2010) is an iterative metaheuristic optimization algorithm. This gradient-free method uses vectors to evaluate candidate solutions, then permutes certain parameters in the vectors (known as agents) using given mathematical formulae. This process requires at least four agents because the algorithm uses one agent as a candidate and three others to compute its mathematical formulae. The process of agent selection is random. When the entire population of vectors is evaluated and the permutations have taken place, the process is cycled through again. The best candidates are continuously evolved while the worst are discarded. This method has been proven to be very effective in other fields when proper parameters are used. It is still the subject of much research.

3.2.6 Genetic Algorithm (GA)

The GA (Holland 1975; Goldberg 1989; Schmitt 2011) is a widely used optimization method that works by simulating the processes of natural selection. GA is commonly used for trying to determine the best local minimum of an objective function. The algorithm is, as all evolutionary strategies, population-based. An initial population is created in the search space (parameter sets), and each individual in this population is a model evaluation. The GA uses these results to create a second generation of individuals, who are independently evaluated once again. Mutations in genes (parameters) as well as intelligent crossovers are used to converge on found minima while other individuals are left searching for other possible minima.

3.2.7 Harmony Search (HS)

HS (Geem et al. 2001) is an evolutionary algorithm that simulates the process of musicians playing independently from one another. When the musicians play, they generate notes which, when combined, form a harmony. In parameter optimization problems, each parameter is a musician, and the local optimum is found when the best harmony is produced. Each model evaluation therefore is a combination of notes, and the parameters are fine-tuned to produce the harmony after the iteration. HS can be used to optimize discrete variables as well as continuous ones. In this study, only continuous variables are used. Moreover, the experience of previous attempts is considered when the parameters are adjusted.

3.2.8 Pattern Search (PS)

The PS algorithm is a direct search, which means minima are found by setting random points in the search space and then improved upon using local seraches. In this study, the initial parameter set was selected randomly within lower and upper boundaries for each model. PS uses a mesh around a given point to try and find a lower minimum in the surrounding neighbourhood. The process continues while improving the objective function with each new mesh. The pattern of the points in the mesh can be defined in many ways; however, for this study the mesh adaptive direct search (MADS) variation of PS was used. This variant does not require specific parameters to produce the mesh after each iteration since the MADS algorithm automatically generates the mesh parameters during optimization (Abramson et al. 2004; Audet and Dennis 2006).

3.2.9 Particle Swarm Optimization (PSO)

PSO (Kennedy and Eberhart 1995; Trelea 2003) is a very simple algorithm that was originally intended for simulating the social behaviour of living organisms, such as flocking birds. Further studies of the algorithm revealed it was actually performing optimization. The population (swarm) is a group of individuals (particles) moving through the search space. After they are evaluated with the objective function, bad particles are moved towards the best at different speeds depending on the distance and overall performance of the swarm. The worst particles are moved towards the best solutions at greater speed than those who are close. Each particle also has a momentum function which permits it to visit other areas in the search-space. If a better solution is found, the swarm will start to move towards this new solution. However, this method is known to be easily trapped in local minima (Fang et al. 2007).

3.2.10 Shuffled Complex Evolution – University of Arizona (SCE-UA)

SCEUA (Duan et al. 1992, 1993, 1994) has been of great use in hydrological model calibration and has arguably been the most popular algorithm to be used for this purpose. It uses groups of points (complexes) to evaluate a sample of the parameter space. A number of these complexes will work independently. Each complex is then updated by selecting a certain proportion of the points (sub-complex) and linking them in a geometric shape. For example, a complex may have 5 points and a sub-complex 3 points. When linked, three points will form a triangle. The point with the least value will be translated through the shape's geometric center. First, the point is mirrored. If the value is less than the original point, the point is translated half-way to the mirror line in the direction of the center of mass. If this point's evaluation is less than the original point, the point is simply mutated randomly in the search space. The process continues for each point in the complex. After every point has been updated, the points are shuffled according to performance and new complexes are formed. The process iterates until the search criteria are met.

3.3 Models, study area and data

This section first introduces the three hydrological models used in this paper. The study area is then briefly described, as is the data for each of the ten independent basins. Each of the models is known to have interdependent parameters, which is very frequent in hydrological modelling, but exerts additional stress on the optimization algorithms since the problem becomes non-separable.

3.3.1 Hydrologic models

Three models of varying complexity were used during this study. They are all coded in MATLAB so no external model had to be executed.

HSAMI

The HSAMI model (Fortin 2000; Minville 2008, 2009, 2010; Poulin et al. 2011; Chen et al. 2011, Arsenault et al. 2013) has been used by *Hydro-Quebec* for over two decades to forecast daily flows on many basins over the province of Quebec. It is a lumped conceptual model based on surface and underground reservoirs. It simulates the main processes of the hydrological cycle, such as evapotranspiration, vertical and horizontal runoffs, snowmelt and frost. Runoff is generated by surface, unsaturated and saturated zone reservoirs through two unit hydrographs: one for surface and another for intermediate (soilwater) reservoir unit hydrographs. The required inputs are spatially averaged maximum and minimum temperatures, liquid and solid precipitation and cloud cover fraction. The model has up to 23 calibration parameters, all of which were used for this study.

MOHYSE

MOHYSE is a simple lumped conceptual model that was first developed for teaching purposes (Fortin and Turcotte 2007). Since then, the model has been used in research applications (e.g. Velazquez et al. 2010) and operationally at the *Centre d'Expertise Hydrique du Québec*. MOHYSE simulates the main hydrometeorological processes that occur in Nordic watersheds, i.e. snow accumulation and melt, potential evapotranspiration (PET), runoff generation, vertical flow, and horizontal flow. The following modelling approaches are considered for each one of these processes, respectively: degree-days approach; PET estimation based on length of day (as a function of watershed mean latitude), and on absolute humidity of air at saturation point (as a function of mean temperature over a time step); runoff and infiltration separation using a simple calibrated threshold relationship; vertical water budget based on threshold and linear relationships between watershed surface and two underground reservoirs (unsaturated and saturated zones); unit hydrograph. MOHYSE can be run on a time scale that varies from sub-daily to multiple days. In the present study, a daily time step was used. The required input data are mean daily temperatures, total daily rain depth and total daily snow (expressed as water equivalent depth). All these values are aerial averages over the entire watershed since the model is lumped. Ten (10) parameters were to be calibrated in this study.

CEQUEAU

CEQUEAU (Charbonneau et al. 1977; Singh and Frevert 2001) is a distributed hydrological model based on physiographic data of the catchment. Soil use, altitude, slope, orientation and vegetation cover data are required to build the model. The precipitation and temperature data from weather stations is fed according to the measurement location. The catchment is first divided into "whole squares" (usually 10km x 10km) to form smaller hydrological units. In this step, hydrological processes such as precipitation, evapotranspiration and snowmelt are simulated through partially physically-based component models. Then, a second subdivision into "partial squares" occurs in order to efficiently calculate the routing and flow direction based on the physical parameters of the catchment. The routing in flow channels and production thresholds are reservoir-based. The CEQUEAU model has 25 free parameters which are calibrated in this paper.

3.3.2 Basins

Ten basins were used in this study, the HSAMI and MOHYSE models were used on all ten basins independently, but only two basins were used with the distributed model CEQUEAU: The Lac-St-Jean and Chute-à-la-Savane basins. These two basins are sub-basins of the larger Saguenay-Lac-St-Jean (SLSJ) basin. The CEQUEAU model was used on these basins only as it is a distributed model and the MOPEX dataset contains only aggregated meteorological data. Furthermore, its implementation is labour-intensive and the CS and LSJ basins were already modelled in CEQUEAU. The physiological data required to build a CEQUEAU model is not available in the MOPEX database, which would have rendered the task impractical. The SLSJ basin is located in the southern center of the province of Quebec in Canada. The Lac-St-Jean sub-basin (45432 sq. km) represents approximately 60% of the total area of the SLSJ basin. The annual mean flow is 850 m^3/sec. The Chute-à-la-Savane catchment is a small sub-basin (1300 sq. km) in the southernmost part of the SLSJ basin, and its mean flow is approximately 35 m^3/sec.

The hydrometeorological data was provided by Rio Tinto Alcan Company. The data for the Lac-St-Jean catchment spans the years 1988 to 1997 inclusively, while the data used for Chutes-à-la-Savane ranged from 2000 to 2009 inclusively.

The eight other basins used in this study were selected from the MOPEX database (Duan et al. 2006) because of their relatively different sizes and geographical locations. Furthermore, the basins were selected because they receive snowfall, which is a strong component in the three hydrological models.

Table 3.1 shows the basins characteristics such as their size, number and location. Datasets were available from 1948 to 2003 inclusively for all the MOPEX basins.

Table 3.1 Characteristics of the eight selected MOPEX and two Québec basins

Basin ID	**Catchment Name**	**State/Province**	**Area (km^2)**	**Mean Qobs (m^3/s)**
01060000	Royal River at Yarmouth	Maine (ME)	365	7.8
QC-CS	Chutes-à-la-Savane	Québec (CS)	1300	35.8
09132500	North Fork Gunnison river near Somerset	Colorado (CO)	1362	12.3
01076500	Pemigewasset river at Plymouth	New Hampshire (NH)	1610	38.8
12449500	Methow river at Twisp	Washington (WA)	3368	44.7
05520500	Kankakee river at Momence	Illinois (IL)	5939	60.7
01531000	Chemchung river at Chemchung	New York (NY)	6488	73.5
06191500	Yellowstone river at Corwin Springs	Montana (MT)	6791	90.8
03253500	Licking river at Catawba	Kentucky (KY)	8543	122.3
QC-LSJ	Lac-St-Jean	Québec (LSJ)	45432	868.6

Figure 3.1 shows the location of these selected basins. Since snowmelt processes are very important in the models used, the basins were selected where snow is not uncommon whilst maintaining an approximately random distribution.

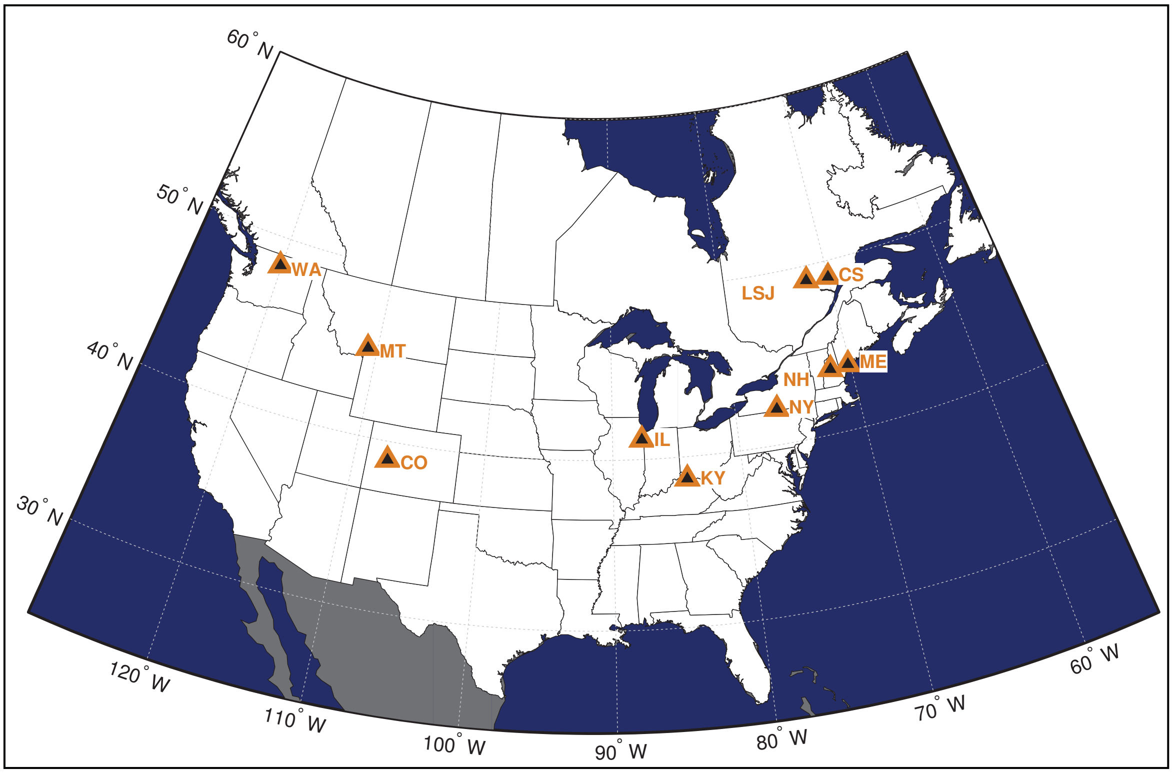

Figure 3.1 Selected catchment locations for the 8 MOPEX catchments (CO, IL, KY, ME, MT, NH, NY and WA) and 2 Québec catchments (CS and LSJ)

Basins NY, NH and ME were selected because of their spatial proximity and their different sizes. Hydrological conditions are expected to be somewhat consistent over this region, so differences in algorithm performance can be better linked to the size of the basin.

3.4 Benchmarking of the optimization methods

The actual testing of the methods was very straightforward. The Lac-St-Jean and Chutes-à-la-Savane catchments were set-up in all three models whereas the eight Mopex basins were set-up only in HSAMI and MOHYSE models. The 10 optimization methods were then programmed for each of the model-basin pairs using their default parameter values. Each optimization algorithm was used to complete 40 different model calibrations with each optimization run being limited to 25000 model evaluations.

The objective function value was saved for every model evaluation, thereby generating a trace for each calibration run. The objective function was [1 – Nash-Sutcliffe], defined as:

$$O.F. = 1 - NSE = 1 - \left[1 - \frac{\sum_{i=1}^{T} \left(Q_o^t - Q_m^t \right)^2}{\sum_{i=1}^{T} \left(Q_o^t - \overline{Q_o} \right)^2} \right] \qquad (3.1)$$

where T is the number of time steps, $Q^t{}_o$ is observed discharge at time t, $Q^t{}_m$ is simulated discharge at time t and $\overline{Q_o}$ is the mean observed discharge. The Nash-Sutcliffe Efficiency (NSE) value was chosen because it is arguably the most commonly used metric in hydrologic model calibration (Nash and Sutcliffe 1970). Many other objective functions exist and are used in hydrologic modelling, and each has its own properties. However, this study will concentrate solely on NSE to limit time and computational constraints. Since the basins and models were selected to simulate snow and snowmelt, a large part of the NSE value is based on the snowmelt peak discharge. The low-flows are less likely to negatively or positively impact the NSE value even if simulated poorly. Other objective functions would behave differently and should be analyzed independently.

Optimization methods will usually produce a good (possibly local or global minimum) value and then continue searching for better alternatives (sometimes finding better, but most of the time worse). This means the value of the objective function is almost always saw-toothed in time, as other candidate solutions are tested and some are inevitably worse than the previous trial. To overcome this, the vectors were modified to only follow a downward trend. This was done by verifying if the current objective function value is better than the previous one. If not, the current value was set to the previous evaluation value.

At times, an optimization model would not complete 25000 model evaluations due to other terminating criteria such as no gain in the last N iterations. For example, CMA-ES did not require 25000 evaluations to optimize the MOHYSE model parameters with its internal convergence stopping criteria.

3.5 Results

The general results for the test-cases are shown in this section. An overall comparison of the different algorithms performances under varying conditions is shown for convergence speed as well as for their ability to attain low objective function values.

3.5.1 Algorithm performance based on ranks

For each basin-model-algorithm combination, the best objective function value after 25000 simulations was selected. This operation was performed 40 times for all combinations to allow for a statistical significance test to be carried out. The averages of the 40 trial results are shown in Figure 3.2. The color-coded values are a direct indicator of the rank of each algorithm for each model-basin pair. This method allowed comparing the algorithms easily, however algorithms with similar performance can be ranked quite differently if their values are too similar.

Figure 3.2 shows the rank of each method for a model/basin case, where darker shades represent better rankings after averaging the 40 trial results. Clear patterns emerge from this figure, such as the strong (relative to other algorithms) performance of ASA, CMAES and DDS, which seem to be generally stronger than the other methods. For the CEQUEAU model, only two catchments were chosen to be modelled, hence the missing values for the MOPEX catchments. Here again the DDS and ASA algorithms perform very well, with HS and PS being tied for third place. It is also noteworthy that the SCEUA algorithm performed better than the others for the MOHYSE model, which has the smallest parameter space of the three, whereas CMAES, DDS and ASA were better when parameter space is larger. This is consistent with what is found in the literature.

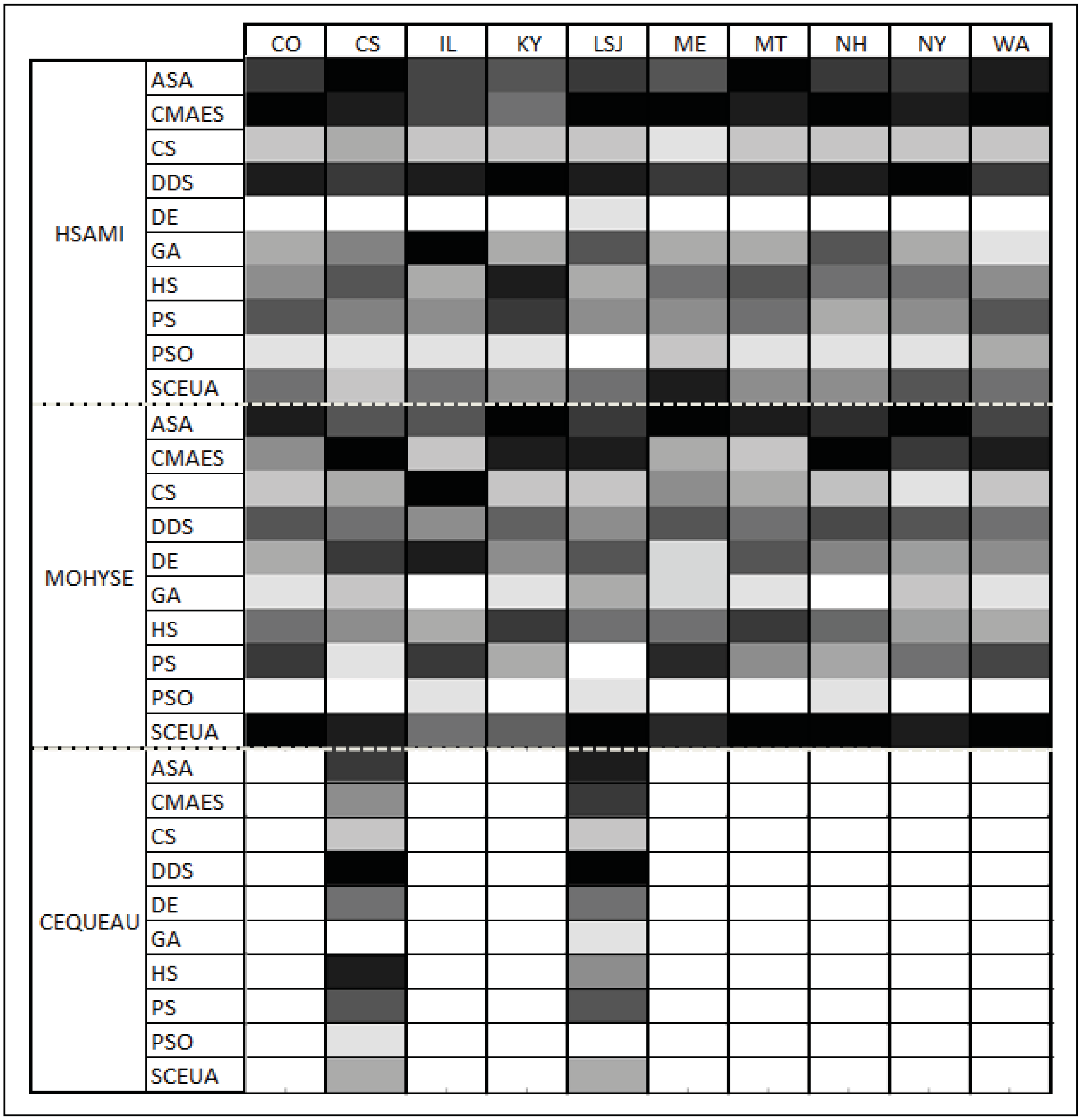

Figure 3.2 Color-coded rankings for algorithm performance for each test-case. Dark shades represent better rankings than pale shades

3.5.2 Algorithm performance based on convergence speed

The choice of an algorithm over another is usually attributed to the best overall objective function value. However, in an operational context such as continuous reservoir management for hydropower production, calibration speed is of the essence. In this respect, the convergence speed of each algorithm was tracked and compared to the overall best objective function value. This was done to prevent an algorithm which converges fast to a poor optimum to be ranked higher than it should.

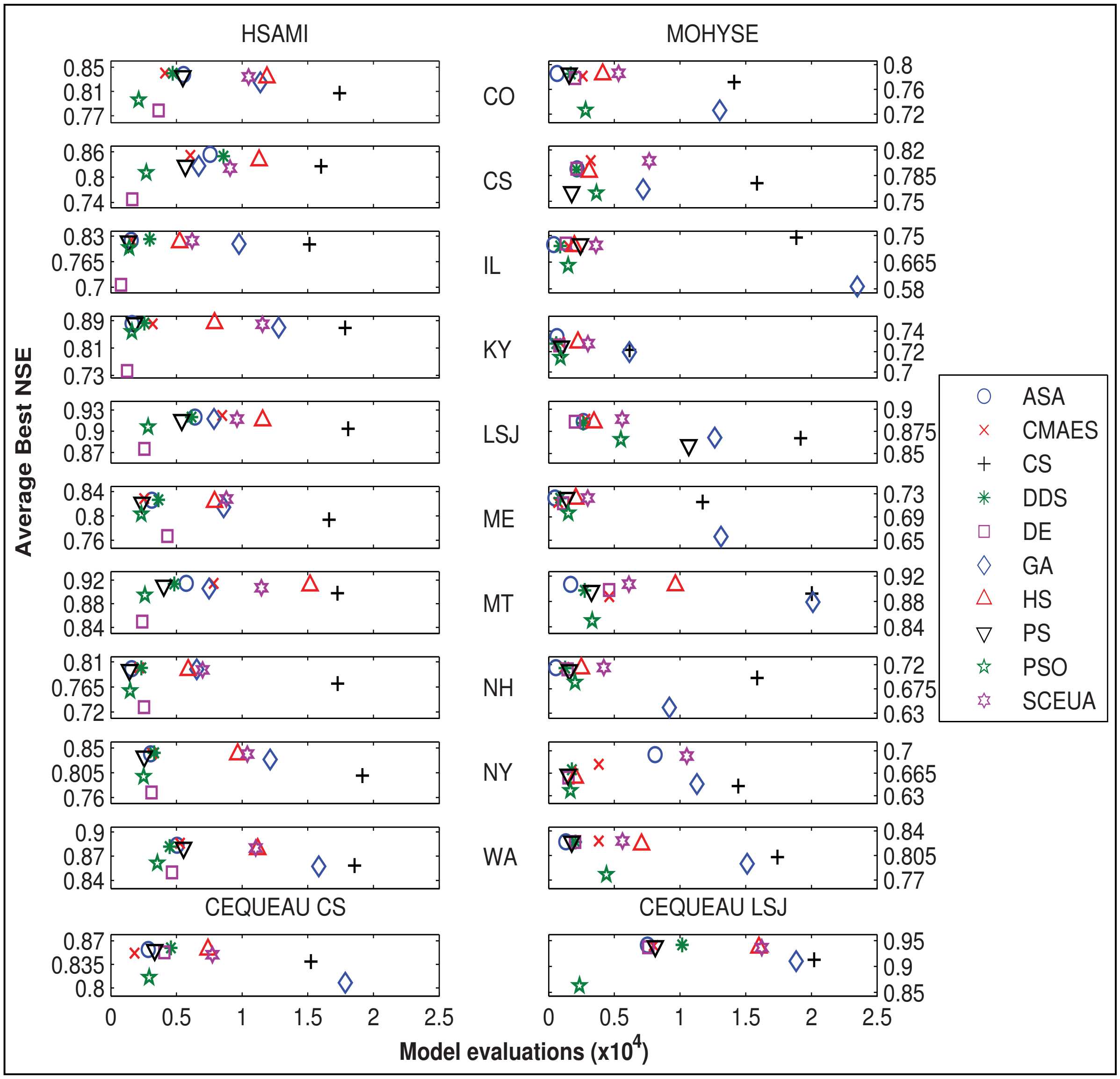

Figure 3.3 Average best NSE vs. model evaluations required to attain 95% of best NSE (measure of convergence speed). Leftmost points converge faster, and higher points attain better NSE values

Figure 3.3 illustrates the trade-off between the algorithms performances and their convergence speed. Each sub-plot represents a model/basin combination. The higher the algorithm is in the graph, the lower the average objective function value was (low objective value means high NSE). To quantify convergence speed, the number of model evaluations necessary to attain 95% of the best objective function value was noted, and the average number of runs was taken. The x-axis is represented by the average number of evaluations required (in 10^4 scale) to attain 95% of the best NSE value. Here, the fastest converging algorithms are leftmost. A good algorithm with fast convergence and a low minimum would be in the top-left corner. In Figure 3.3, it is clear that some algorithms are either slow to converge or simply not able to achieve minima as low as other methods. This is the case with CS and GA, which take a longer time to start converging. This in turn could be the reason why these methods are not amongst the better ranked ones; they simply did not converge after 25000 model evaluations. Perhaps these methods could attain good minima if given more time, but in a practical framework it seems unlikely that these algorithms can compete with the faster-converging ones.

It can also be seen in Figure 3.3 that there is a fierce competition between the different algorithms in the top-left area of the charts. The algorithms in this dense space are almost always comprised of ASA, CMAES, DDS and, to a lesser extent, PS and SCEUA, although the latter is dominant only when using the MOHYSE model.

3.5.3 Statistical significance tests

For HSAMI and MOHYSE models, the amount of data was too large to test individually, so a non-parametric, complete-block, two-way layout Friedman test was used to detect if there was a difference between the algorithms across the 10 selected catchments (nuisance factor) (Friedman 1937, 1940). In both cases, the tests showed that there was indeed a statistically significant difference in NSE medians between the groups of algorithms. A Bonferroni correction was used as a post hoc analysis to determine which algorithms are responsible for the difference between the groups (Hochberg 1988). Table 3.2 shows the pair-wise comparisons between the algorithms when the HSAMI model is used. A value of 1 means the two algorithms are significantly different from one another, whereas a value of 0 means they are statistically similar.

Table 3.2 Significant differences between algorithms for the HSAMI model

Optimization algorithms										
	ASA	**CMAES**	**CS**	**DDS**	**DE**	**GA**	**HS**	**PS**	**PSO**	**SCEUA**
ASA	---	---	---	---	---	---	---	---	---	---
CMAES	**0**	---	---	---	---	---	---	---	---	---
CS	1	1	---	---	---	---	---	---	---	---
DDS	**0**	**0**	1	---	---	---	---	---	---	---
DE	1	1	1	1	---	---	---	---	---	---
GA	1	1	1	1	1	---	---	---	---	---
HS	1	1	1	1	1	1	---	---	---	---
PS	1	1	1	1	1	1	1	---	---	---
PSO	1	1	**0**	1	1	1	1	1	---	---
SCEUA	1	1	1	1	1	1	**0**	**0**	1	---

The interpretation of these tables is as follows. First, all possible combinations between any two algorithms are assigned either a “1” or a “0”. The row/column order is unimportant. When a value of 1 is assigned, the two algorithms are from different groups (meaning they are statistically dissimilar). On the other hand, a value of 0 means the two algorithms are statistically the same. Furthermore, by looking at Figure 3.2, it can be seen that ASA, DDS and CMAES are ranked highly, thus confirming that they are in the top group and are statistically similar. In this case, and considering Figure 3.2, it is clear that the three methods ASA, CMAES and DDS are statistically similar but they are significantly better than the other tested algorithms in for the HSAMI model. The worst algorithm is DE, which can be seen in Figures 3.2 and 3.3. Table 3.3 shows the same results as table 3.2, but with the MOHYSE model instead.

When looking at table 3.3 with respect to Figure 3.2, it is clear that SCEUA is the best method for MOHYSE when taking into account the 10 basins the tests were carried out on. It is significantly better than all the other methods except ASA, which is a close second. This is somewhat expected since MOHYSEs’ small parameter space is directly in SCEUAs’ scope. CMAES and DDS are both close behind ASA. It is also clear, looking at Figure 3.2 and table 3.3, that the least effective methods are PSO and GA as they are statistically worse than the others.

Table 3.3 Significant differences between algorithms for MOHYSE model

Optimization algorithms										
	ASA	**CMAES**	**CS**	**DDS**	**DE**	**GA**	**HS**	**PS**	**PSO**	**SCEUA**
ASA	---	---	---	---	---	---	---	---	---	---
CMAES	**0**	---	---	---	---	---	---	---	---	---
CS	1	1	---	---	---	---	---	---	---	---
DDS	1	**0**	1	---	---	---	---	---	---	---
DE	1	**0**	1	**0**	---	---	---	---	---	---
GA	1	1	1	1	1	---	---	---	---	---
HS	1	**0**	1	**0**	**0**	1	---	---	---	---
PS	1	**0**	**0**	**0**	**0**	1	**0**	---	---	---
PSO	1	1	1	1	1	**0**	1	1	---	---
SCEUA	**0**	1	1	1	1	1	1	1	1	---

For the CEQUEAU model, only two basins were used. For this reason it was chosen to compare the algorithms for each basin separately. To compare the different methods, an analysis resting on the assumption of a normal distribution cannot be used. Therefore the Kruskal-Wallis test was used with α = 0.05 as Barrette et al. (2008) have demonstrated. The Kruskal-Wallis test is based on ranks and therefore does not suppose a normal distribution. It allows a comparison between samples from two or more groups (where each group is 40 trials for one algorithm) and to determine if they have statistically different medians (Kruskal and Wallis 1952). The confidence level is valid for the entire test as with the Friedman test, not for each pairwise comparison. A multiple comparison procedure is used to identify which methods are significantly different from one another.

The results displayed in Figure 3.4 show the confidence interval around each method's value. If two confidence intervals overlap, they are not statistically different. On the other hand, if they do overlap, they are different with confidence level 1- α. Figure 3.4 shows the multiple comparison tests for the Chutes-à-la-Savane (CS) basin for the CEQUEAU model.

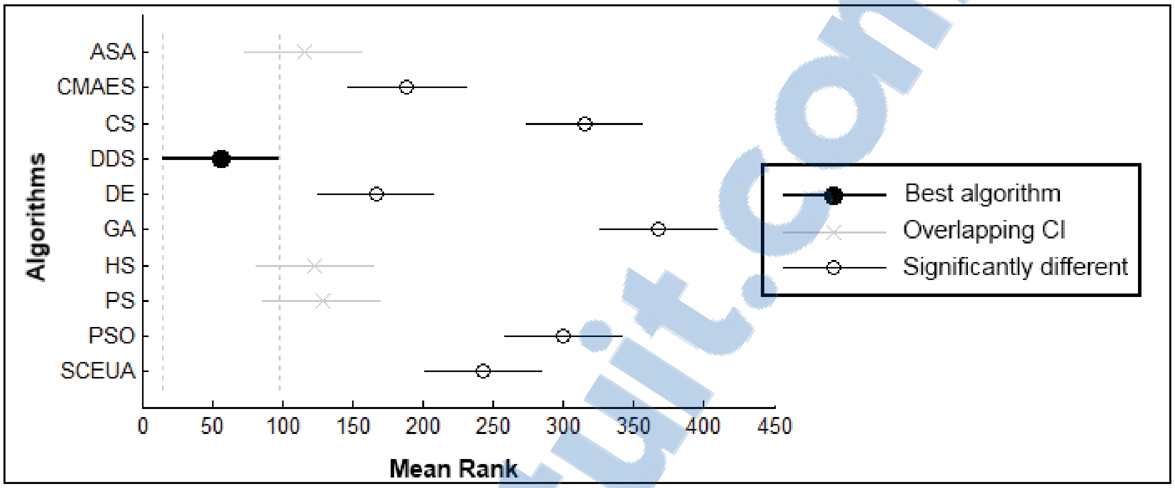

Figure 3.4 Multiple comparison test with confidence intervals on Chutes-à-la-Savane for the CEQUEAU model. Markers represent mean rank and lines represent 95% confidence intervals

For the CEQUEAU model simulating the Chutes-à-la-Savane basin, the best algorithm was DDS, which was statistically better than all the other algorithms except ASA, HS and PS. This was also expected as DDS claims to be most effective when the parameter space is large, as is the case with CEQUEAU. The worst algorithms were CS, GA and PSO. The results were the same with the Lac-St-Jean basin and the CEQUEAU model, except HS was substituted for CMAES in the multiple comparison test.

Overall, it is clear that ASA, CMAES and DDS are most efficient when the calibration problem parameter space is large, and SCEUA is best when the parameter space is smaller. It is important to note that these results are based on multiple catchments with three hydrological models with many trials each, which adds to their significance.

3.5.4 Dispersion Metric

Another analysis method was used to further investigate the causes of algorithm performance or lack thereof. As is widely mentioned in the literature, optimization algorithms cannot excel at all problems. Their inner structure is geared towards specific types of problems, and by understanding the fitness landscape produced by the hydrological models and the NSE objective function, it is possible to determine the algorithm that is most likely to perform well on a given calibration problem. The *dispersion metric* of Lunacek and Whitley (2006) was computed to do so. It uses iterative random sampling of the search space to measure the average pairwise Euclidian distance between *m*-best parameter sets from a population of *n* parameter sets, where the size of *m* is fixed and *n* is variable. A decrease in the average Euclidian distance when *n* is increased means the fitness landscape has a converging global structure. Further information on the dispersion metric is available in Lunacek and Whitley (2006).

The dispersion metric was measured for each of the model-basin pairs with *n* ranging from 100 to 100000 evaluations and *m*=100. The average Euclidian distances were normalized to ensure comparability. Figure 3.5 shows the results for the 22 test-cases. Note that the variable population size increases non-linearly.

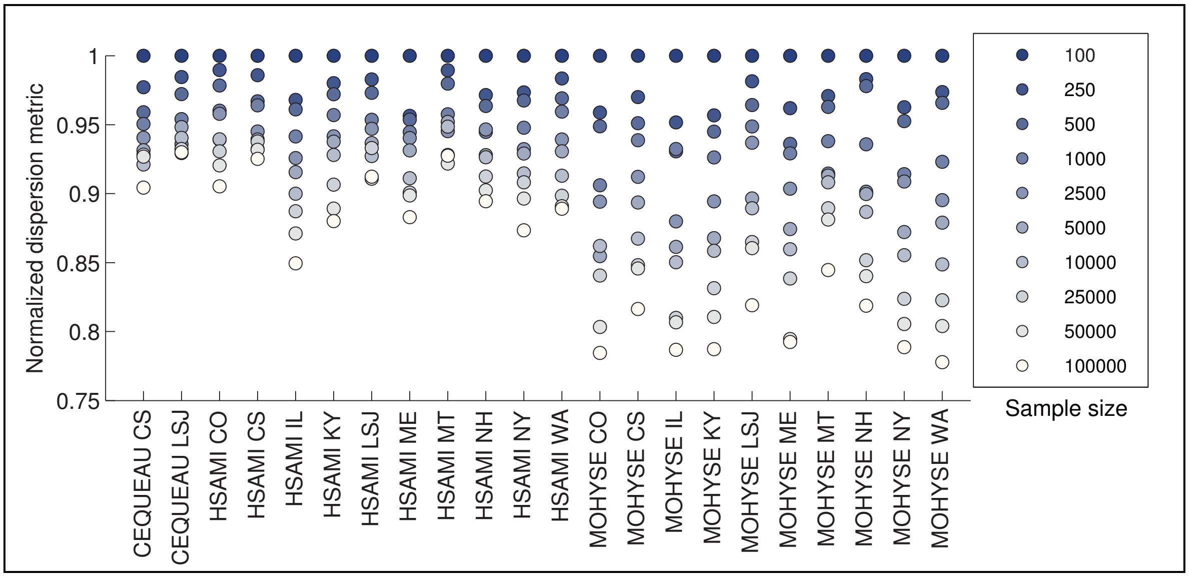

Figure 3.5 Normalized dispersion metric for 22 model-basin pairs and increasing *n* size for *m* = 100. Markers represent different *n*-sizes and are coded in increasing order from black to white

If the pale markers are lower than the dark markers, then the global structure is convergent. In this case, all models follow this pattern (a diverging structure would tend to increase with higher evaluations). All the test cases therefore follow a global structure which has a general optimum region, which could be populated by many local minima due to ruggedness and noise in the fitness landscape. It is also clear that the MOHYSE model converges faster as the markers are spaced further apart than they are for the other models, which is not surprising given the lower dimensionality of the model.

3.6 Discussion

The discussion is divided into five separate sections. A general performance assessment is made at first. Then, method performance with respect to model complexity, basin type, convergence speed and computing power are also addressed.

3.6.1 On overall performance

The first important aspect to be noted in this paper is that all the methods were able to find respectable NSE values at least once per model-basin pair. Some methods were significantly better than others. CMAES, ASA and DDS were almost always in the leading group of methods. SCEUA also showed its effectiveness when the parameter space is small, as is the case with the MOHYSE model (10 parameters).

On the other hand, SCEUA managed to converge rather quickly in all cases, but it seems to have difficulty in getting out of local minima when the parameter space is large. While it does an acceptable job at finding good parameter sets, it seems to be outdated compared to the best methods which are more recent.

3.6.2 On model complexity

Model complexity was found to play an important role in the performance of the optimization algorithms. First, the fact that the model parameters are interdependent means the optimization problem is non-separable and thus can be problematic for many optimization algorithms. It is widely known that canonical versions of PSO perform weakly on non-separable problems (Spears et al. 2010). The same can be said for DE (Ronkkonen et al. 2005), canonical GA (Salomon 1996) and HS (Ükler and Haydar 2012). Indeed, these methods did not fare well in the optimization problems in this study.

Secondly, PS (with Mesh Adaptive Direct Search) was shown to be very good on multimodal problems, but to have difficulty with problems that have globally clustered local optima, which seems to be the case regarding the dispersion metric (Whitley et al. 2006). In Figure 3.5, it is quite clear that CEQUEAU model has the least converging structure, and PS received its highest score on this problem, while it had difficulties with the faster converging fitness landcapes of the other two hydrologic models. One reason for this behaviour is because PS is biased towards exploration rather than exploitation.

The best algorithms (ASA, DDS and CMAES) are built for the test-cases in this study. They are adaptive, have measures to exploit the local optima and can handle high-dimension problems. CMAES in particular is known to perform well when the global structure has an optimum (Lunacek and Whitley 2006), while ASA improves on canonical SA with its adaptive parameters and tunnelling. DDS was designed to make use of the allowed number of model evaluations to efficiently balance the exploration/exploitation ratio.

Finally, SCEUA is shown to perform well on the smaller dimensionality and have difficulty when the search-space is large. CS was average in most test-cases, but did outperform other algorithms in one-test case (MOHYSE - Chutes-à-la-Savane). It was very close to the best group in other MOHYSE test-cases, but the gap increased significantly in the larger space problems.

Another difference in model complexity between HSAMI and CEQUEAU resides in the fact that HSAMI is lumped and CEQUEAU is distributed. However, the problem they pose to optimizing algorithms is their dimensionality, with 23 and 25 parameters respectively, which is evident for the SCEUA and CS algorithms for example.

3.6.3 On the effect of the basin on algorithm performance

In almost all cases, the selected basin does not alter the performance of the optimization algorithms. The results show that the same ranks are assigned to the methods for the majority of the basins.

The fact that Chutes-à-la-Savane is a sub-basin of Lac-St-Jean could be a factor in this seemingly strong correlation between the basins, especially for the CEQUEAU model which was only run on those two basins. The similarity of the physical aspects of the catchments could be at cause. However, it must be restated that two completely different time periods were used as input data. Therefore, the objective function landscape should be relatively different enough as to consider both calibration problems as independent.

As for the Mopex basins, they were selected to be as different as possible while ensuring snowfall/snowmelt was a significant component of their hydrologic cycle. This condition was essential since the models have strong snow models and having snow assured that they were operating in conditions they were designed to simulate. They were also selected because of their relatively heterogeneous sizes. In most cases, it is clear to see that the algorithms perform similarly on the different catchments, which means basin size and location is less crucial than model dimensionality when selecting an optimization algorithm. This is obviously true within the realm of chosen basins (temperate climate, presence of snow).

3.6.4 On convergence speed

The various algorithms in this study have been tested for their convergence speed as well as their capacity to find good objective function minima. The convergence speed is crucial for applications in any operational context. For example, if two algorithms perform similarly but one converges within 1000 simulations while the other requires ten times more simulations, it could be very convenient for the hydrologist to use the faster converging method. This is especially true for models with long running time. Complex distributed models may easily require more than 1 minute of computing time per year of simulation over large basins. In such cases, convergence speed is crucial and can be more important than the overall best optimum. As was shown in Figure 3.3, some algorithms attain the same best NSE value but attain 95% of that value at very different rates. An example of this is the HSAMI model on the NY basin. Clearly SCEUA and DDS have very similar best NSE values, but DDS uses approximately 3000 evaluations to converge on this value and SCEUA requires 10000 evaluations to arrive at the same result. For longer-running models, DDS would be preferred to SCEUA because of this net advantage in computing costs.

3.6.5 On computing power

In many optimization applications, the objective function is quickly evaluated. However, in hydrologic model calibration, the fitness function requires that the model be run to calculate the error between simulated and observed streamflow. Simulating the streamflow may take a lot of time, depending on the model. The models used in this paper were selected because of their evaluation speed. HSAMI and MOHYSE both took approximately 0.02 seconds to evaluate for any given basin due to them being lumped, whereas CEQUEAU required approximately 0.4 seconds for the small CS catchment and 3 seconds for the large LSJ basin on a 3.1GHz processor. These times are exceedingly fast in comparison with many other models, such as SWAT, WaSim and Hydrotel, (Neitsch et al. 2002; Schulla and Jasper 2000; Fortin et al. 2001) which may take 2 or 3 orders of magnitude more time to run. Even so, this project required over 2 months worth of calculations using as many as 42 computing cores (32x2.4GHz, 6x4.1GHz and 4x3.1 GHz). It would not have been possible to do so with more complicated models in a reasonable timeframe with current equipment for research purposes.

The methods that stand out in terms of convergence ability are ASA, CMAES, DDS and SCEUA, with PS close behind. If a model needed to be calibrated and resources were insufficient to perform a full-length calibration, CMAES, ASA and DDS have been shown to converge very close to the final optimum in less than 2000 evaluations. Even at 1000 evaluations, these methods would find relatively good values. For example, Figure 3.6 shows the convergence pattern for the ten algorithms for the Chutes-à-la-Savane basin using the

HSAMI model. The number of evaluations is limited to 2750 evaluations because no major change in the pattern appears for the rest of the trials. The curves represent the average (1-NSE) value for the 40 trials.

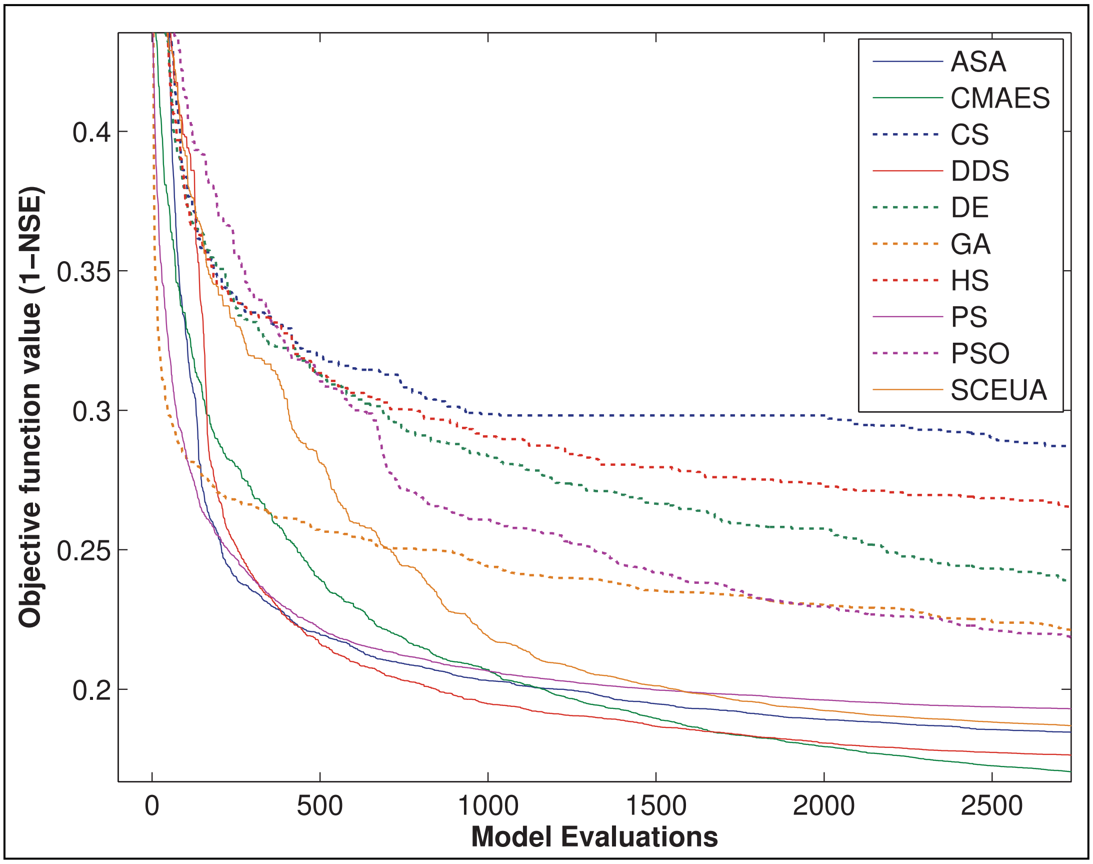

Figure 3.6 Convergence patterns using the HSAMI model on the Chutes-à-la-Savane basin

In comparison, for certain catchments, CS, DE, GA and HS had best values after 25000 evaluations which were worse than the best methods at 2000 evaluations. It is also important to note that PS and DDS are adaptive, meaning that they know how many evaluations they are allowed to run and adapt their search in response. For example, DDS could have converged even faster if it had been allowed only 2000 evaluations instead of 25000.

3.7 Conclusion

After reviewing the results of 22 test-cases, it seems clear that the three best methods to use for hydrlogic model calibration are ASA, CMAES and DDS for large parameter spaces and SCEUA for small parameter spaces, as they respectively find the better NSE values for the different calibration problems. This supposes that the hydrological model calibration fitness landscape has a definite, converging global structure. Depending on the convergence slope, different algorithms could outperform others. It must also be noted that for each model, some algorithms performed better than others depending on the basin. However, ASA requires a careful setup and has many parameters. This is both an advantage for adapting the method to particular needs, but also an inconvenience since the time required to find the best settings can be costly.

If many different types of calibration are regularly performed, the preferred choices of the authors would be CMAES and DDS, especially in an operational context. They require almost no parameter tuning and have proved to be amongst the best methods in every test. SCEUA would be the preferred choice for models with small dimensionality, assuming a similar underlying fitness landscape.

While the size and location of the basins played a very minor role on algorithm performance, the same cannot be said about the hydrologic model. The dimensionality of the models as well as the underlying structure play a key role in separating the best methods from the others. Other studies with independent-parameter models and non-converging global structures should also be undertaken to further investigate algorithm performance.

Of course, many more methods exist and new ones are published regularly. It is impossible to test all methods, but the selection made in this paper touches a wide range of the existing widespread algorithms. The next step will be to try and find other methods which promise to surpass the current ones and test them on a larger set of basins and hydrological models. For example, a variant of the Cuckoo Search (CS) called the Modified Cuckoo Search (MCS) promises to correct the faults in the regular CS (Tuba et al. 2011), and the same is valid for the new Modified Shuffled Complex Evolution (MSCE) (Mariani et al. 2011). Another possible avenue for future studies would be to test hybrid algorithms. A benchmark test of a combination of these methods would prove interesting.

3.8 Acknowledgements

The authors would like to thank Lester Ingber and contributors for the ASA source code, available at http://www.ingber.com/#ASA, as well as Shinichi Sakata for making public the Matlab routines to use ASA. The codes are available here: http://ssakata.sdf.org/software/. We would also like to thank Brian A. Tolson for supplying the DDS algorithm. A final thanks to the reviewers who have contributed to the quality of the present work.

3.9 References

Abramson, M. A., Audet, C. and Dennis Jr., J. E. (2004). “Generalized pattern searches with derivative information.” Mathematical Programming, Series B, 100, 3–25.

Arsenault, R., Malo, J., Brissette, F., Minville, M. and Leconte, R. (2013). “Structural and non-structural climate change adaptation strategies for the Péribonka water resource system.” Water Resour Manage. 13p. DOI: 10.1007/s11269-013-0275-6.

Audet, C. and Dennis Jr., J. E. (2006). “Mesh adaptive direct search algorithms for constrained optimization.” SIAM J. Optim. 17 (2), 188-217.

Barrette, M., Wong, T., de Kelper, B. and Côté, P. (2008). “Statistical multi-comparison of evolutionary algorithms” Bioinspired Optimization Methods and Applications, October 13-14, Ljubljana, Slovenia, pp. 71-80.

Beven, K. (2001). Rainfall-runoff modelling – The primer, John Wiley and Sons, Chichester, UK, 372p.

Blasone, R.S., Madsen, H. and Rosbjerg, D. (2007). “Parameter estimation in distributed hydrological modelling: comparison of global and local optimisation techniques.” Nordic hydrology, 38 (4-5), 451-476.

Charbonneau, R., Fortin, J.-P. and Morin, G. (1977). “The CEQUEAU model: description and examples of its use in problems related to water resource management / Le modèle CEQUEAU: description et exemples d'utilisation dans le cadre de problèmes reliés à l'aménagement.” Hydrological Sciences Bulletin, 22(1), 193-202.

Chen, M. and Lu, Q. (2005). “A hybrid model based on genetic algorithm and ant colony algorithm.” Journal of Information & Computational Science, 2, 647-653.

Chen, J., Brissette, F.P., Poulin, A. and Leconte, R. (2011). “Global uncertainty study of the hydrological impacts of climate change for a Canadian watershed.” Water Resour. Res., 47, W12509.

Duan, Q., Sorooshian, S. and Gupta, V. K. (1992). “Effective and efficient global optimization for conceptual rainfall runoff models.” Water Resour. Res., 24(7), 1163-1173.

Duan, Q., Sorooshian, S. and Gupta, V. K. (1993). “A shuffled complex evolution approach for effective and efficient optimization.” J. Optimiz. Theory Appl. 76(3), 501-521.

Duan, Q., Sorooshian, S. and Gupta, V. K. (1994). “Optimal use of the SCE-UA global optimization method for calibrating watershed models.” J. Hydrol. 158, 265-284.

Duan, Q., Schaake, J., Andreassian, V., Franks, S., Gupta, H.V., Gusev, Y.M., Habets, F., Hall, A., Hay, L., Hogue, T.S., Huang, M., Leavesley, G., Liang, X., Nasonova, O.N., Noilhan, J., Oudin, L., Sorooshian, S., Wagener, T. and Wood, E.F. (2006). “Model parameter estimation experiment (MOPEX): Overview and summary of the second and third workshop results.” J. Hydrol., 320, 3-17.

Fang, L., Chen, P., Shihua, L. (2007). “Particle swarm optimization with simulated annealing for TSP.” Proceedings of the 6th Conference on 6th WSEAS Int. Conf. on Artificial Intelligence, Knowledge Engineering and Data Bases - Volume 6. Corfu Island, Greece, World Scientific and Engineering Academy and Society (WSEAS), 206-210.

Fortin, J.-P., Turcotte, R., Massicotte, S., Moussa, R., Fitzback, J. and Villeneuve, J.-P. (2001). “Distributed watershed model compatible with remote sensing and GIS data. 1: Description of the model.” J. Hydraul. Eng., 6(2), 91-99.

Fortin, V. (2000). “Le modèle météo-apport HSAMI: historique, théorie et application.” Varennes: Institut de Recherche d’Hydro-Québec, 68p.

Fortin, V. and Turcotte, R. (2007). “Le modèle hydrologique MOHYSE (bases théoriques et manuel de l’usager).” Notes de cours pour SCA7420, Département des sciences de la terre et de l’atmosphère, Université du Québec à Montréal, Montréal, Canada, 17p.

Fortnow, L. (2009). “The status of the P versus NP problem.” Communications of the ACM, 52(9), 78–86. doi:10.1145/1562164.1562186

Franchini, M., Galeati, G. and Berra, S. (1998). “Global optimization techniques for the calibration of conceptual rainfall-runoff models.” Hydrological Sciences Journal, 43(3), 443-458.

Friedman, M. (1937). “The use of ranks to avoid the assumption of normality implicit in the analysis of variance.” Journal of the American Statistical Association, 32, 675–701.

Friedman, M. (1940). “A comparison of alternative tests of significance for the problem of m rankings.” Annals of Mathematical Statistics, 11, 86–92.

Geem, Z. W., Kim, J. H. and Loganathan, G. V. (2001). “A new heuristic optimization algorithm: Harmony search.” Simulation, 76(2), 60-68.

Goldberg, D. E. (1989). Genetic algorithms in search, optimization & machine learning, Addison-Wesley, Boston, 432p.

Hansen, N. and Ostermeier, A. (2001). “Completely derandomized self-adaptation in evolution strategies.” Evolutionary Computation, 9(2), 159-195.

Hansen, N. and Ostermeier, A. (1996). “Adapting arbitrary normal mutation distributions in evolution strategies: The covariance matrix adaptation.” In Proceedings of the 1996 IEEE International Conference on Evolutionary Computation, 312-317.

Hochberg, Y. (1988). “A sharper Bonferroni procedure for multiple tests of significance.” Biometrika, 75(4), 800-802.

Holland, J. (1975). Adaptation in natural and artificial systems. University of Michigan Press, Oxford, 183p.

Ingber, L. (1989). “Very fast simulated re-annealing.” Mathematical Computer Modelling, 12, 967-973.

Ingber, L. (1993). Adaptive simulated annealing (ASA), McLean, VA, Lester Ingber Research.

Ingber, L. (1996). “Adaptive simulated annealing (ASA): Lessons learned.” Control and Cybernetics, 25, 33-54.

Kennedy, J. and Eberhart, R. (1995). “Particle swarm optimization". Proceedings of IEEE International Conference on Neural Networks, 4. 1942–1948. doi:10.1109/ICNN.1995.488968

Kirkpatrick, S.; Gelatt, C. D.; Vecchi, M. P. (1983). "Optimization by simulated annealing". Science, 220(4598), 671–680. doi:10.1126/science.220.4598.671

Kruskal, W. H. and Wallis, W. A. (1952). “Use of ranks in one-criterion variance analysis.” J. Amer. Statist. Assn. 47, 583-621.

Lunacek, M. and Whitley, D. (2006). “The dispersion metric and the CMA evolution strategy.” Proceedings of the 8th annual conference on Genetic and evolutionary computation, July 08-12, 2006, Seattle, USA. doi:10.1145/1143997.1144085

Mariani,V.C., Luvizotto, L. G. J., Guerra, F. A. and Leandro D. S. C. (2011). “A hybrid shuffled complex evolution approach based on differential evolution for unconstrained optimization.” Applied Mathematics and Computation, 217(12), 5822-5829.

Minville, M., Brissette, F. and Leconte, R. (2008). “Uncertainty of the impact of climate change on the hydrology of a Nordic watershed.” J. Hydrol. 358(1-2), 70-83.

Minville, M., Brissette, F., Krau, S. and Leconte, R. (2009). “Adaptation to climate change in the management of a Canadian water-resources system.” Water Resour Manage. 23(14), 2965-2986.

Minville, M., Krau, S., Brissette, F. and Leconte, R. (2010). “Behaviour and performance of a water resource system in Québec (Canada) under adapted operating policies in a climate change context.” Water Resour Manage (2010) 24:1333–1352

Moradkhani, H. and Sorooshian, S. (2009). “General review of rainfall-runoff modeling: model calibration, data assimilation and uncertainty analysis.” In Hydrological Modelling and the Water Cycle: Coupling the Atmospheric and Hydrological Models, Sorooshian, S., Hsu, K.-l., Coppola, E., Tomassetti, B., Verdecchia, M., Visconti, G. (Eds.), Springer, 1-24.

Nash, J. E. and Sutcliffe, J. V. (1970). “River flow forecasting through conceptual models part I — A discussion of principles.” J. Hydrol., 10 (3), 282–290.

Neitsch, S.L., Arnold, J.G., Kiniry, J.R., Williams, J.R. and King, K.W. (2002). “Soil water assessment tool theoretical documentation.” Grassland, Soil and Water Research Laboratory, Temple, Texas. GSWRL Report 02-01.

Pedersen, M.E.H. (2010). Good parameters for differential evolution. Technical Report HL1002, Hvass Laboratories.

Poulin, A., Brissette, F., Leconte, R., Arsenault, R. and Malo, J.S. (2011). “Uncertainty of hydrological modelling in climate change impact studies in a Canadian, snow-dominated river basin.” J. Hydrol. 409(3-4), 626-636.

Ronkkonen, J., Kukkonen, S. and Price, K.V. (2005). “Real-parameter optimization with differential evolution.” Proc. IEEE Congr. Evolut. Comput., 506 -513.

Salomon, R. (1996). “Reevaluating genetic algorithm performance under coordinate rotation of benchmark functions.” BioSystems, 39, 263-278.

Schmitt, L. M. (2001). “Theory of genetic algorithms.” Theoretical Computer Science, 259, 1–61.

Schulla, J. And Jasper, K. (2000). “Model description WASIM-ETH (Water balance simulation model ETH)”, ETH-Zurich, Zurich, Switzerland.

Singh, V. P. and Frevert, D. K. (2001). “Mathematical models of large watershed hydrology.” Water Resources Publications. Chapter 13.

Singh, V.P. and Woolhiser, D.A. (2002). “Mathematical modeling of watershed hydrology.” J. Hydrol. Eng., 7, 270–292.

Spears, W.M., Green, D.T. and Spears, D.F. (2010). “Biases in particle swarm optimization.” International Journal of Swarm Intelligence Research, 1(2), 34-57.

Storn, R. and Price, K. (1997). “Differential evolution - a simple and efficient heuristic for global optimization over continuous spaces.” Journal of Global Optimization, 11, 341–359.

Tolson, B. A. and Shoemaker, C. A. (2007). “Dynamically dimensioned search algorithm for computationally efficient watershed model calibration.” Water Resour. Res., 43, W01413.

Trelea, I.C. (2003). "The particle swarm optimization algorithm: convergence analysis and parameter selection." Information Processing Letters, 85(6), 317–325.

Tuba, M., Subotic, M. and Stanarevic, N. (2011). “Modified cuckoo search algorithm for unconstrained optimization problems.” In Proceedings of the 5th European conference on European computing conference (ECC'11), Remi Leandre, Metin Demiralp, Milan Tuba, Luige Vladareanu, and Olga Martin (Eds.). World Scientific and Engineering Academy and Society (WSEAS), Stevens Point, Wisconsin, USA, 263-268.

Ülker, E.D. and Haydar, A. (2012). “Comparison of the performances of differential evolution, particle swarm optimization and harmony search algorithms on benchmark functions.” Academic Research International, 3(2), 85-92.

Velazquez, J.A., Anctil, F. and Perrin, C. (2010). “Performance and reliability of multimodel hydrological ensemble simulations based on seventeen lumped models and a thousand catchments.” Hydrology and Earth System Sciences, 14, 2303-2317.

Vrugt, J. A. and Robinson, B. A. (2007). “Improved evolutionary optimization from genetically adaptive multimethod search.” Proc. Natl. Acad. Sci. U. S. A., 104, 708–711.

Whitley, D., Lunacek, M., and Sokolov, A. (2006). “Comparing the niches of CMA-ES, CHC and pattern search using diverse benchmarks.” In Parallel problem solving from nature (PPSN IX), LNCS: 4193, 988–997.

Yang, X.-S. and Deb, S. (2009). “Cuckoo search via Levy flights.” In Proc. of World Congress on Nature & Biologically Inspired Computing, India. IEEE Publications, USA, 210-214.

Yen, J., Liao, J. C., Lee, B. and Randolph, D. (1998). “A Hybrid approach to modeling metabolic systems using genetic algorithms and simplex method.” IEEE Transactions on Systems, Man, and Cybernetics, 28, 173-191.

CHAPITRE 4

ARTICLE 2 : A COMPARATIVE ANALYSIS OF 9 MULTI-MODEL AVERAGING APPROACHES IN HYDROLOGICAL CONTINUOUS STREAMFLOW SIMULATION

Richard Arsenault[1], Philippe Gatien[1], Benoit Renaud[1],
François Brissette[1] et Jean-Luc Martel[1]
[1] Département de Génie de la Construction, École de technologie supérieure,
1100 Notre-Dame Ouest, Montréal, Québec, Canada H3C 1K3.

Article soumis à la revue « Journal of Hydrology » en février 2015.

Abstract

This study aims to test whether runoff modeling simulations are more accurate when determined by a weighted combination of several models, rather than individual ones. In addition, the project attempts to identify the most efficient model averaging method and the optimal number of models to include in the weighting scheme. In order to address the first objective, runoffs were simulated using five lumped hydrological models (HSAMI, HMETS, MOHYSE, GR4J-6 and GR4J-15), each of which were calibrated with three different objective functions on 429 watersheds. The resulting 15 hydrographs (5 models x 3 metrics) were weighted and combined with the help of 9 averaging methods which are the simple average (SAM), Akaike information criterion (AICA), Bates-Granger average (BGA), Bayes information criterion (BICA), Bayesian model averaging (BMA), Granger-Ramanathan average variant A, B and C (GRA, GRB and GRC) and the average by SCE-UA optimization (SCA). The same weights were then applied to the hydrographs in validation mode, and the Nash-Sutcliffe Efficiency metric was measured between the averaged and observed hydrographs. Statistical analyses were performed to compare the accuracy of weighted methods to that of individual models. A Kruskal-Wallis test and a multi-objective optimization algorithm were then used to identify the most efficient weighted method and the optimal number of models to integrate. Results suggest that the GRA, GRB, GRC and SCA weighted methods perform better than the individual members. Model averaging from these four methods were superior to the best of the individual members in 72% of the cases. Optimal combinations on all watersheds included at least one of each of the five hydrological models. None of the optimal combinations included all members of the ensemble of 15 hydrographs. The Granger-Ramanathan average variant C (GRC) is recommended as the best compromise between accuracy, speed of execution, and simplicity.

Keywords: model averaging; objective functions; averaging method comparison; model error reduction;

4.1 Introduction

Many aspects of daily operations in water resources management require an ability to predict future streamflows with the best possible accuracy. Over the years, numerous hydrological models have been proposed, each with its strengths and weaknesses. All adequate models have the capacity to predict streamflows, but none is able to consistently outperform others for all basin characteristics and heterogeneous climatology (i.e. the best all-around model). Recent literature has shown that on select catchments, weighted averages of multiple model simulations are more robust and more precise than their individual members. Cavadias and Morin (1985) introduced the concept of weighted multi-model averaging for streamflow determination using the Granger and Newbold method (Granger and Newbold, 1977). Shamseldin et al. (1997) then showed that multi-model averaging improved performance over individual model simulations using three averaging techniques: simple arithmetic mean, constrained ordinary least-squares weighting and a neural network averaging method. Shamseldin et al. (2007) compared three types of neural networks (Simple Neural Network, Radial Basis Function Neural Network and Multi-Layer Perceptron Neural Network) in a flow averaging study. They found that the neural networks outperform the models taken independently. However, neural networks are time-consuming to conduct and are prone to over-fitting. Other weighting schemes have been put forth which can combine streamflows in various manners to improve the averaged hydrograph. One such method, the Bayesian Model

Averaging method (BMA), computes weights based on the probability density function of the ensemble (Hoeting et al. 1999; Raftery et al. 1993, 2003, 2005). While weighted averaging was devised to incorporate the advantages of each individual member, it was shown that BMA is not appropriate if too many members are used (Neuman, 2003). BMA should therefore be limited to fewer and relatively similar member ensembles (Jefferys and Berger, 1992).

The seminal paper by Diks and Vrugt (2010) compared 7 model averaging methods: Equal Weights Averaging (EWA), Akaike/Bayes Information Criterion Averaging (AICA/BICA), Bates and Granger Averaging (BGA), Granger-Ramanathan-A Averaging (GRA), Bayesian Model Averaging (BMA) and Mallows Model Averaging (MMA). They conclude that the unconstrained methods (weights are not constrained to sum to unity) perform better than the constrained methods, and that the GRA method is the best overall since it is much faster and quicker to implement than MMA and BMA while offering the same performance.

Another study by Ajami et al. (2006) compared the EWA and constrained Ordinary-Least-Squares methods to the Multi-Model Super Ensemble (MMSE) and Modified MMSE (M3SE) methods using the Distributed Model Intercomparison Project Results (Smith et al. 2004). MMSE is used mostly in climate and weather forecasting but was applied to hydrological time series. M3SE is a frequency-based bias-corrected averaging method. These methods include bias correction and variance reduction to further improve simulation quality. The authors showed that the M3SE and MMSE methods are better than individual models, as previous studies have shown. They also showed that MMSE can sometimes produce unrealistic results (such as negative flows) because of the bias correction method implemented in the method.

Applications of multi-model flow prediction have been studied for over a decade. See and Openshaw (2000) proposed a probabilistic switching mechanism where the output from a single member was used at each time step, switching the donor member as hydrological conditions evolve. Hu et al. (2001) proposed a similar concept except model switching occurred based on discharge levels. Abrahart and See (2002) compared six flow amalgamation strategies (both switching and averaging) on two catchments. They determined that in flow forecasting, neural network methods improve predictive skill compared to the individual models if the flow regime is stable, whereas in volatile environments, a fuzzified probabilistic mechanism was the best tool. These applications are different from the simulation framework considered in this study as the averaging and prediction is balanced at each time step with the newly acquired information.

Other comparative studies have been published in the last few years on the subject of multi-model averaging (Bowler et al. 2008; Cavadias and Morin 1986; Mylne et al. 2002; Raftery and Zheng 2003; Raftery et al. 2005), especially in the hydrology and weather/climate prediction research fields. However most of these use either a limited set of basins, of models or of model averaging methods (or some combination thereof). In this paper, we compare 9 model averaging techniques on 429 catchments from the MOPEX database using 5 hydrological models calibrated with 3 objective functions. The 3 objective functions are used to produce different parameterizations of the models. This allows diversifying the models' ability to target different parts of the hydrograph. Oudin et al. (2006) noted that models calibrated with two different objective functions produced flows that improved the overall simulation performance when combined adequately. Consequently 15-member ensembles are available for the model averaging methods. This large sample size will allow a better understanding of which methods are to be used in future applications.

4.2 Data, models and multi-model averaging methods

4.2.1 Basins, hydrometric and climate data

The hydrometric and climate data were collected from the MOPEX (Model Parameter Estimation Experiment) database (Duan et al., 2006) for 429 catchments ranging in size from de 66 to 10324 km². The dataset covers years 1949-2003, but many of these years are incomplete or missing. All available data was used for each of the catchments. The MOPEX database was designed to have a minimal density of stations per catchment, ensuring a certain level of quality in the dataset. Even years were used for the calibration period and validation was carried out on the odd years in the available time series. The opposite (calibration on odd years and validation on even years) was also tested but the results were practically identical, and are thus not presented here. In all cases, the first year in calibration and in validation was sacrificed for model warm-up.

The geographical extents of the catchments as well as their mean annual precipitation (mm) are shown in figure 4.1.

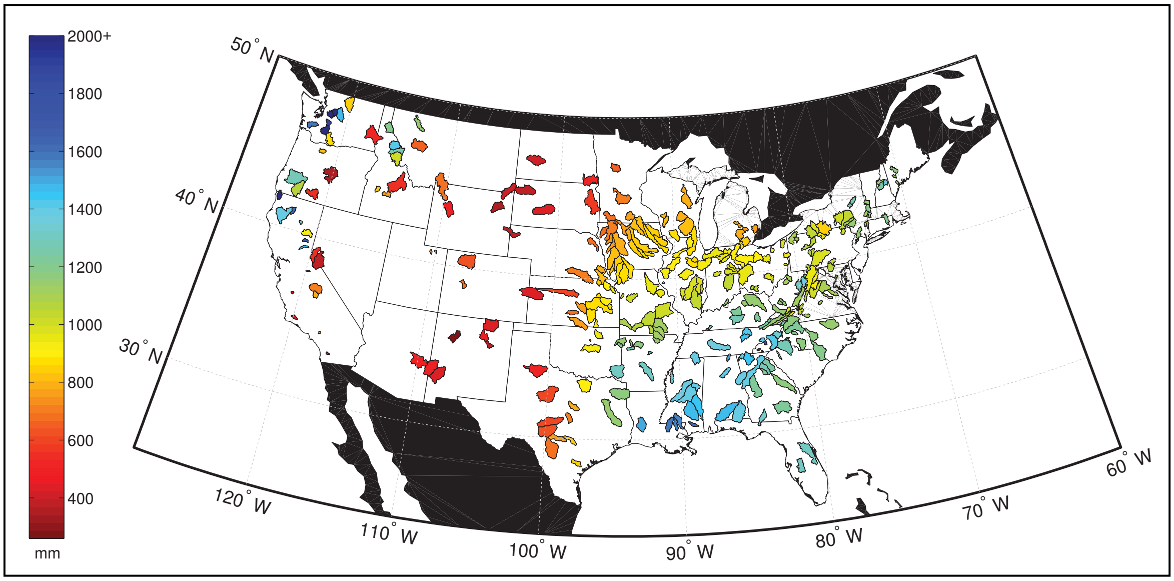

Figure 4.1 Spatial distribution of the 429 catchments from the MOPEX database used in this study and their total annual precipitation (mm)

It can be seen that annual precipitation varies greatly depending on the region, with clear gradients across the US. Some catchments in the west coast receive more than 2000 mm of precipitation, while arid regions in south-central US receive less than 400 mm. The east-west gradient is clear, with increasing precipitation values towards the east coast. Another lesser gradient is also observed in the north-south direction east of 95°W longitude. This information will be relevant for later analysis.

An overview of the hydrometeorological characteristics of the catchments in this study is presented in Table 4.1.

Table 4.1 Selected hydrometeorological descriptors for the basin set in this study. Minimum, maximum as well as 25th (Q1), 50th (Q2) and 75th (Q3) percentiles of the values are presented

	Min.	**Q1**	**Q2**	**Q3**	**Max.**
Number of years with available data	3	49	55	55	55
Area (km^2)	67	1048	2151	4304	10324
Annual precipitation (mm/yr) (P)	245	842	1001	1202	2748
Average Maximum daily temperature (°C)	7.5	14.6	16.9	19.9	28.9
Average Minimum daily temperature (°C)	-6.2	2.5	4.5	6.6	16.3
Average Mean daily temperature (°C)	0.8	8.6	10.7	13.2	22.6
Mean Potential Evapotranspiration (mm/yr) (PET)	598	775	917	1036	1757
Ratio of snow in total precipitation	0.00	0.04	0.08	0.13	0.73
Mean annual flow (mm/yr) (Q)	2	239	370	527	2642
Aridity index (PET/P)	0.19	0.85	1.15	1.43	4.02
Runoff Coefficient (Q/P)	0.005	0.261	0.367	0.456	0.981

The potential evapotranspiration (PET) is taken directly from the MOPEX database and is based on the based NOAA Freewater Evaporation Atlas. Different PET estimation methods would also impact the aridity index, which is the ratio of potential evapotranspiration to total precipitation.

4.2.2 Hydrological models

Since the project required calibrating a large number of hydrological model / objective function combinations on an even larger set of basins, distributed models were not considered for this study, and five lumped models were retained. The five models are presented here.

HSAMI

The HSAMI model (Fortin 2000; Minville et al. 2008, 2009, 2010; Poulin et al. 2011; Arsenault et al. 2013) has been used by *Hydro-Quebec,* Quebec's hydroelectric company, for over three decades to forecast daily flows on more than one hundred basins in the province. It simulates the entire hydrological cycle with a strong snow accumulation and melt model. Potential evapotranspiration is estimated using a proprietary formulation requiring only daily maximum and minimum temperatures. Runoff is simulated by surface, unsaturated and saturated zone reservoirs through two unit hydrographs: one for surface and another for intermediate (soil water) reservoirs. The model has 23 adjustable parameters, all of which were calibrated in this study.

MOHYSE

MOHYSE is a simple, 10 parameter model developed primarily for academic purposes (Fortin and Turcotte 2007). Since then, the model has been used in research applications (Arsenault et al. 2014; Velazquez et al. 2010) because of its ease of use as well as its execution speed. MOHYSE is geared towards cold climates and has custom snow accumulation and melt modules as well as a simple yet effective potential evapotranspiration formulation based on latitude, available effective daylight and temperature.

HMETS

HMETS is a daily model that uses two reservoirs for the saturated and vadose zones (Chen et al., 2011). It is Matlab-based and has 21 adjustable parameters which were all calibrated in this study. HMETS is similar to HSAMI as it shares some process functions, but they differ in the snowmelt, evapotranspiration and reservoir schemes. HMETS uses the Oudin evapotranspiration method, as was used for the GR4J model variants (Oudin et al. 2005). The snowmelt module is geared towards the northern climates which receive large quantities of snowfall.

GR4J-6 and GR4J-15

The GR4J model (Perrin et al. 2003) is an empirical and lumped, reservoir-based model. It was developed by the research group at CEMAGREF (now IRSTEA). It was conceived for water resources management and spring flood prediction for hydrologic applications. Initially, this model was parsimonious with only 4 parameters, with most secondary processes being represented by empirical constants. Since GR4J does not simulate snow accumulation or melt processes, a snow module (CEMANEIGE) was added to the basic model (Valéry, 2010, Valéry et al., 2014) to make it applicable in northern basins. The GR4J model with the snow model has 2 more calibrated parameters, for a total of 6. This is the GR4J-6 model.

Another version was also used, in which the empirical constants of the basic GR4J were replaced by 9 calibrated parameters, for a total of 15. This is the GR4J-15 model. They are therefore different versions of the same model. Evapotranspiration must be fed to the model, as it does not estimate it itself. The Oudin formulation (Oudin et al. 2005) was used to pre-process the evapotranspiration data for the GR4J model variants.

Table 4.2 summarizes the most important information regarding the models used in this study.

Table 4.2 Overview of the 5 hydrological models used in this study

Model	Main reference	Calibration parameters	Simulated processes (number of parameters)	Required input data
GR4J-6	Perrin et al. (2003) Valéry et al. (2010)	6	Flow Routing (1) Snow modeling (2) Vertical budget (3)	Tmax/Tmin/ P Potential evapotranspiration Median annual snowfall depth
GR4J-15	Perrin et al. (2003) Valéry et al. (2010)	15	Flow Routing (1) Snow modeling (11) Vertical budget (3)	Tmax/Tmin/ P Potential evapotranspiration Median annual snowfall depth
HMETS	Chen et al. (2011)	21	Evapotranspiration (1) Flow routing (4) Snow modeling (10) Vertical buget (6)	Tmax/Tmin Rain/Snow Daily radiation
HSAMI	Fortin (2000)	23	Evapotranspiration (2) Flow routing (5) Snow modeling (6) Surface runoff (3) Vertical budget (7)	Tmax/Tmin Rain/Snow
MOHYSE	Fortin and Turcotte (2007)	10	Evapotranspiration (2) Flow Routing (2) Snow modeling (2) Vertical budget (4)	Tmean Rain/Snow Latitude

4.2.3 Multi-model averaging methods

This section details the 9 multi-model averaging methods used in this study.

Simple arithmetic mean (SAM)

The SAM method is simply an unweighted average of each of the model members. While this method is unsophisticated and simplistic, it will serve as a reference for the other model averaging methods.

Akaike and Bayes information criteria averaging (AICA and BICA)

The AICA and BICA methods (Akaike 1974; Schwarz 1978; Buckland et al. 1997; Burnham and Anderson 2002; Hansen 2008) estimate the optimal probability of each model by using the mean of the logarithm of the member variances, to which a penalty term is added. The difference between the AICA and BICA methods lies in the penalty term calculation.

For AICA, the penalty is equal to double the number of calibrated parameters in the members. For BICA, the penalty is equal to the number of calibrated parameters times the logarithm of the number of time steps in the calibration period. In both cases, the weighting is a compromise between bias (which diminishes with more parameters) and model parsimony.

Bates-Granger averaging (BGA)

The BGA method, initially proposed by Bates and Granger (1969), aims to produce a combined ensemble by minimizing the Root Mean Square Error (RMSE). However, this method relies on the hypotheses that the ensemble members are not biased and their errors are not correlated. The weight of each member is calculated as the inverse of the member's variance.

Bayesian model averaging (BMA)

The BMA approach uses the members' probability distribution functions (PDFs) to determine the weights of each member. The combined distribution is corrected for bias and the difference between the distributions is minimized. The BMA method has been successfully applied in Gneiting et al. (2005), Neuman (2003), Raftery et al. (2005), Vrugt and Robinson (2007), Vrugt et al. (2007) and Ye et al. (2004). Readers are encouraged to consult one of these papers for more details about the mechanics of the BMA method.

Granger Ramanathan A, B and C (GRA, GRB and GRC)

The GRA approach (Granger and Ramanathan, 1984) sets weights based on the ordinary least squares (OLS) algorithm. It minimizes the RMSE but does not correct for bias. The GRB variant is similar to the A method, but the OLS algorithm is constrained such that the weights sum to unity. Finally, the GRC variant is unconstrained but the averaged streamflow values are bias corrected through the use of a constant term.

Shuffle complex averaging (SCA)

In this approach, a stochastic optimization algorithm was used to compute weights based solely on the maximization of the Nash-Sutcliffe Efficiency (NSE) metric. Weights were calibrated with the Shuffle Complex Evolution – University of Arizona (SCE-UA) algorithm (Duan et al., 1992). The NSE metric was computed between the calibration periods' averaged streamflow (using the weights as parameters) and the observed streamflow. The weights that maximize the NSE metric were used in the validation period afterwards. The weights were bounded from [-5;5] and were not constrained to sum to unity. Using wider boundaries [-10;10] did not yield any significant gain during calibration and sometimes ended up being detrimental to the validation period. In a few cases, one member of the ensemble was given a large negative weight potentially resulting in negative streamflows.

Table 4.3 summarizes the main characteristics of the 9 multi model averaging techniques used in this study.

Table 4.3 Summary of the model averaging algorithms used in this study

Acronym	Method description	Reference	Sums to unity	Negative weights possible	Bias correction	Iterative
SAM	Simple Arithmetic Mean	---	Yes	No	No	No
AICA	Akaike's Information Criterion	Akaike (1974)	Yes	No	No	No
BGA	Gates Granger Averaging	Bates and Granger (1969)	Yes	No	No	No
BICA	Bayes Information Criterion	Schwarz (1978)	Yes	No	No	No
BMA	Bayesian Model Averaging	Neuman (2003)	Yes	No	Yes	Yes
GRA	Granger-Ramanathan A	Granger and Ramanathan (1984)	No	Yes	No	No
GRB	Granger-Ramanathan B	Granger and Ramanathan (1984)	Yes	Yes	No	No
GRC	Granger-Ramanathan C	Granger and Ramanathan (1984)	No	Yes	Yes	No
SCA	Shuffled Complex Averaging	---	No	Yes	No	Yes

4.2.4 Model calibration

This project required calibrating 5 hydrological models using three objective functions over 429 catchments. The sheer volume of calibrations called for automatic optimization algorithms instead of manual calibrations. From the dozens of available algorithms, the CMA-ES (Hansen and Ostermeier 1996, 2001) was selected for its quick convergence speed and its ease of use (Arsenault et al. 2014). As mentioned earlier, the models were calibrated on the even years in the dataset and the odd years were used as the validation period.

The three objective functions to optimize during calibration were the Nash-Sutcliffe Efficiency (NSE) (Nash and Sutcliffe 1970), the NSE computed on the natural logarithm of the flow values (LN(NSE)) and a custom metric which is a combination of NSE, the coefficient of variation of the root-mean-square error (CV(RMSE)) and relative bias weighted equally, as shown in equation 4.1.

$$COMB. = \frac{1-NSE}{3} + \frac{CV(RMSE)}{3} + \frac{REL.BIAS}{3} \tag{4.1}$$

The rationale is that by using different objective functions, the calibrated parameter sets are geared to specifically target certain aspects of the hydrograph. The NSE metric targets high flows, the LN(NSE) targets lower flows and the combined metric aims to find an equilibrium between high and low flow performance. The advantage of using this method is that in a model averaging perspective, the simulated hydrographs will vary according to the parameter sets selected, which are dependent on the objective function during calibration. (Moriasi et al. 2007).

4.2.5 Multi-model averaging application

The first step in the model averaging approach was to generate the 15 streamflow members for the calibration and validation periods. Then, the 9 weighting schemes were fed the 15 members as well as the observed discharge for the calibration period. The weights were computed and then applied to the 15 members in the validation period. The resulting averaged streamflow was compared to the observed dataset in the validation period. The NSE metric was finally calculated to evaluate the model averaging methods' relative performances. The non-parametric Kruskal-Wallis statistical test (Kruskal and Wallis, 1952) was used to analyze the results and determine which methods should be investigated further. The Kruskal-Wallis test is used to determine if the data originate from the same distribution or if any groups do not come from the same distributions.

4.3 Results

4.3.1 Performance of the 15 ensemble members

The NSE scores in validation are important as poor validation NSE should limit the ability of the model averaging techniques to produce good results. Figure 4.2 shows the validation NSE values for the 15 model/objective function pairs.

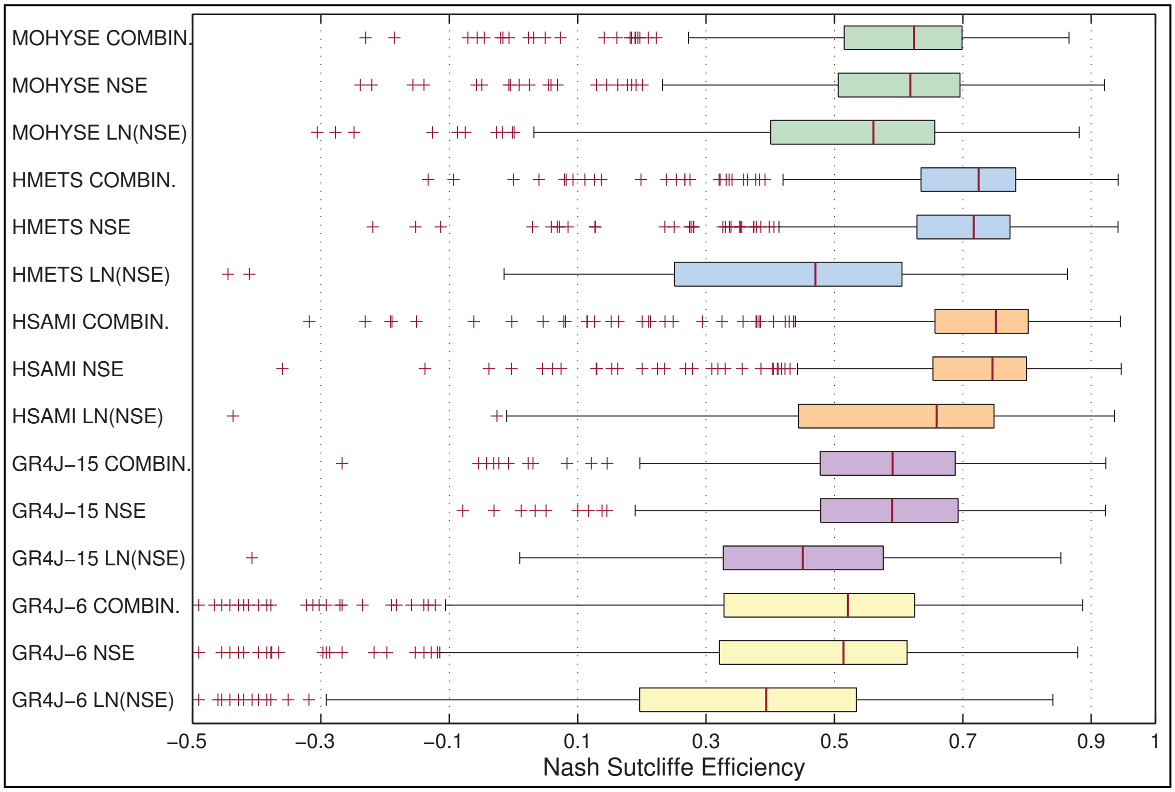

Figure 4.2 NSE values in validation for the 15 ensemble members (model/objective function pairs) computed on 429 catchments

The best individual members are the HSAMI and HMETS models calibrated with the NSE and combined metrics, while the GR4J-6 model scores lowest. The LN(NSE) metric is the worst for all models, as expected. The same reasoning was applied to the relative bias metric, which is often used in reservoir management situations. Figure 4.3 shows the relative bias for the 15 members.

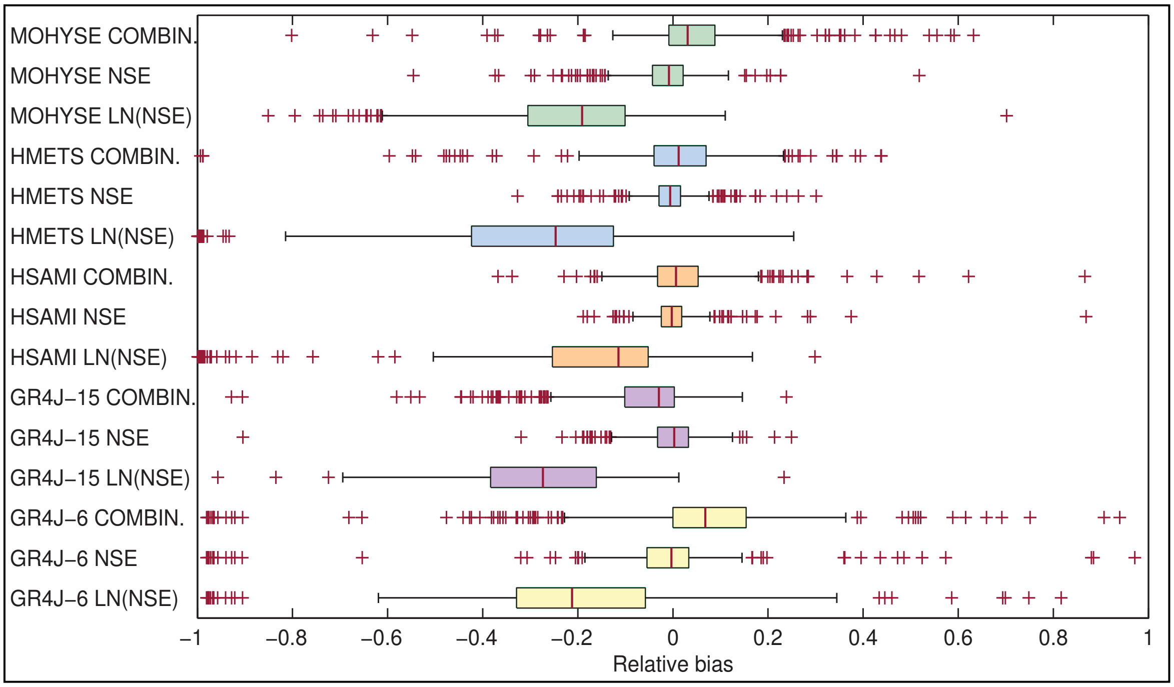

Figure 4.3 Relative bias values in validation for the 15 ensemble members (model/objective function pairs) computed on 429 catchments

The biases are similar for the HSAMI, HMETS and MOHYSE models. The LN(NSE) metric generates the most bias for all models. It is worth reminding that different objective functions were used during calibration, but all results presented in figures 4.2 and 4.3 are the NSE and relative bias values computed between the flows simulated with the calibrated parameter sets and the observations during the validation period. In all cases, and as expected, the models calibrated with the NSE and the combined metric (which includes an NSE component) show better NSE validation values. The HMETS and HSAMI models also perform better than the others in general. For the relative bias metric, HSAMI NSE and HMETS NSE are again the best members. The results in the calibration period are similar to those in validation and are not shown here.

4.3.2 Performance of the multi-model averaging methods

Using the streamflow series from the 15 ensemble members on the calibration period, as well as the observed streamflow values for the same period, the 9 multi-model averaging schemes were applied to calculate the optimal weights for each method. These weights were then applied in the validation period. The NSE and bias values were computed for the averaged flows based on the validation period observed streamflow. Figure 4.4 shows the validation NSE scores for each of the methods as well as the NSE scores for the best single member (HSAMI – NSE). The box-and-whisker plot is based on the 429 basins, however some methods could not generate weights for some basins. This is due to the heterogeneity of the members which negatively impacts some methods' performance in multi-model averaging, such as BMA.

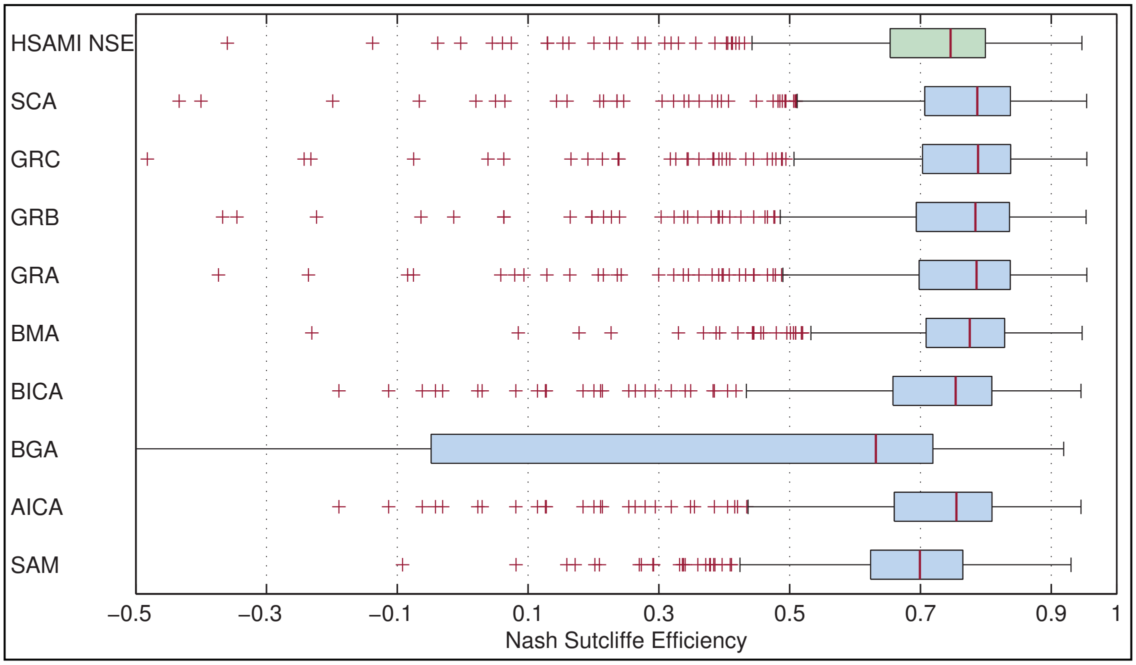

Figure 4.4 Multi-model averaging methods performance in validation for the NSE metric. Here the BMA method contains 349 catchments as it failed to converge on 80 catchments. The BGA method performs the worst

The same operation was conducted on the relative bias metric. The results are presented in figure 4.5, with the model averaging results compared to the HSAMI-NSE member.

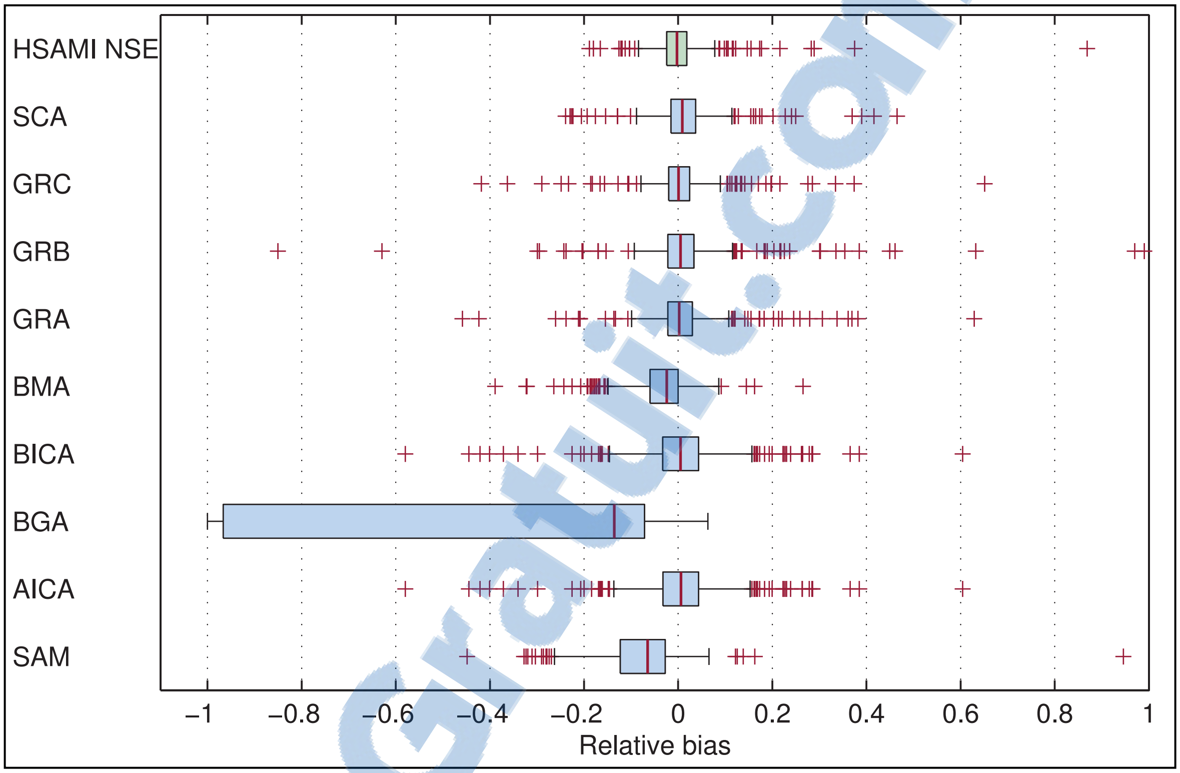

Figure 4.5 Multi-model averaging methods performance in validation for the Relative bias metric As in figure 4.4, the BMA method lacks 80 catchments on which it failed to converge and the BGA method shows the largest amount of bias

To get a better picture of the methods' relative performances, the number of times each method could not generate a set of weights was compiled. The BMA method generated 80 such errors, followed by the GRB (5), GRA and GRC (4), SCA (3), BGA (2) and AICA and BICA (1). The large number of members is known to impact the BMA method as will be discussed further.

These errors notwithstanding, it is still apparent that the GRA, GRB, GRC and SCA methods outperform the others. The BMA method follows closely, however the high failure rate discredits its performance somewhat. The BGA method generates poor weights in general with the 15 members and is considerably less capable than the other methods. It is important to note that other than the BGA method, the multi-model averaging schemes all perform better than the Simple Arithmetic Mean (SAM). Also, the four best methods perform better than the best single member of the ensemble for the NSE metric (HSAMI model calibrated on NSE), whereas for the relative bias, the HSAMI-NSE member is statistically similar to the best model averaging methods

A statistical test was performed to rank the methods along their significant differences. For the NSE metric, the Kruskal-Wallis tests in validation, after removing the problematic basins, shown in figure 4.6, confirm that the GRA, GRB, GRC and SCA methods all perform similarly.

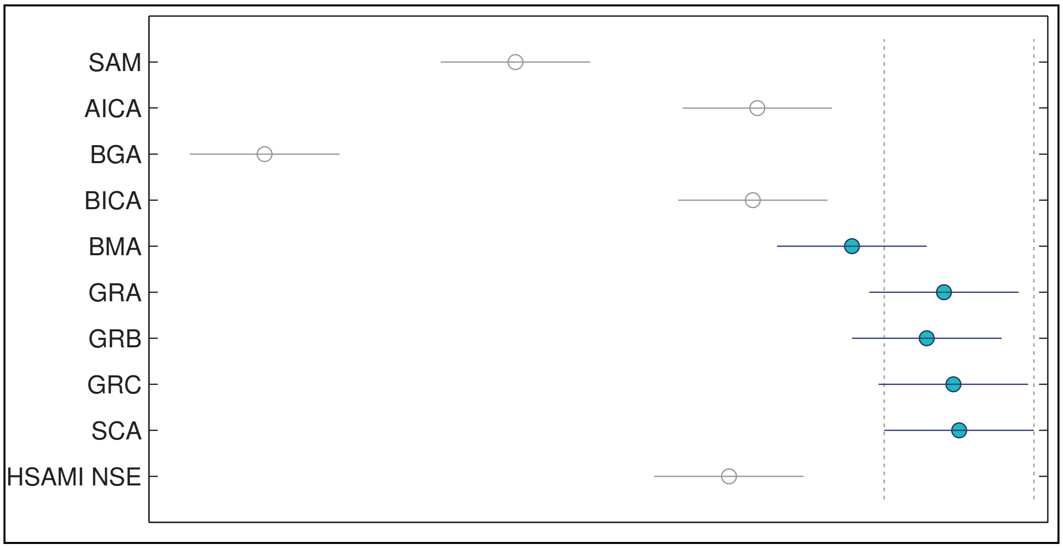

Figure 4.6 Multiple comparison test on ranks. The horizontal lines represent confidence intervals around the rank sum for each model averaging method. Overlapping confidence intervals signifies that the methods are not significantly different at the 5% level. The vertical lines are simply visual aids to see which methods are within the confidence interval of the best averaging scheme. The “x” axis represents rank values and has no use in this analysis

Also, the BMA method is amongst the best methods when it does not fail prematurely. The AICA and BICA methods are able to reproduce the performance of the HSAMI member. It can also be seen that the BGA method is unable to produce adequate results, and that SAM does not perform as well as the single model member. This says that while the information is available in the members, more sophisticated methods than the simple mean are needed to extract this information and contribute to increasing the overall performance in multi-model averaging. For the relative bias metric (results not shown), the HSAMI-NSE member performs statistically as well as the BMA, GRA, GRB, GRC and SCA methods. However, the biases are very low at this point and improving upon them is a challenge. For example, the HSAMI – NSE has an average relative bias value of 0.035 (3.5%) across the 429 basins. To satisfy the 95% statistical significance threshold, the model averaging methods would have to reduce the relative error by 20%, with an average error of less than 3%. For the remainder of this study, we will consider the NSE only as the bias is already minimal and the model averaging techniques are unable to improve upon the best-member performance. Furthermore, the best model averaging methods according to the relative bias are the same as for the NSE metric.The distribution of NSE values in validation excluding the problematic basins for BMA can be seen in figure 4.7.

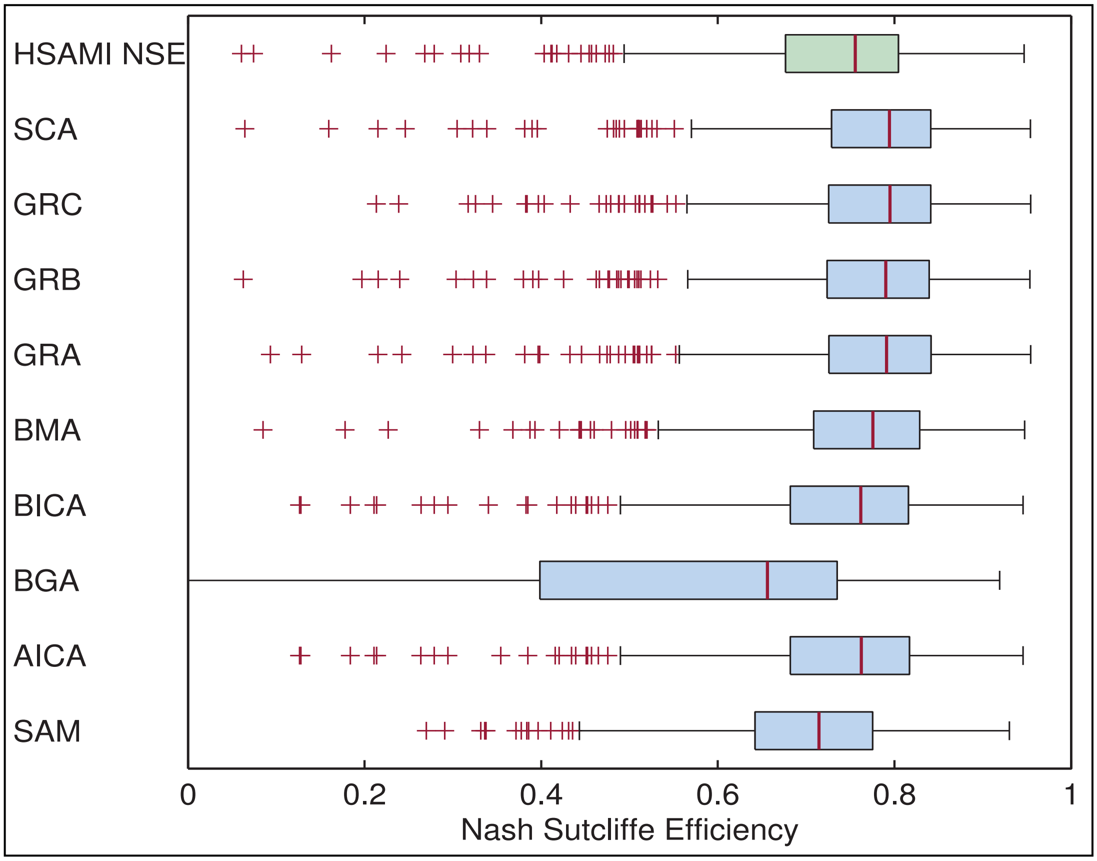

Figure 4.7 Multi-model averaging methods performance in validation for the NSE metric after removal of the catchments which caused the BMA method to fail. There remain 349 catchments for all methods in this figure. BGA is still the worst method but the difference with the other methods is less pronounced

Figure 4.7 shows that the AICA and BICA methods are approximately equal, BMA is slightly better and that the GRA, GRB, GRC and SCA methods are similar and offer the best performance. The BGA method is the only one worse than the SAM method. The same overall results are thus maintained, even without the problematic catchments.

4.3.3 Performance gain quantification

The next step was to compare the model averaging methods performances to the individual members in validation. Table 4.4 shows the frequency with which each model obtained the best score in validation. Table 4.4 includes the results of all basins as well as only the basins which do not cause the BMA method to fail.

Table 4.4 Frequency with which each model obtained the best score in validation, with all basins and with only non-problematic basins

Model Member	Frequency as best NSE (%) – All basins	Frequency as best NSE (%) – Only non-problematic basins
GR4J-6 LN(NSE)	0	0
GR4J-6 NSE	0	0
GR4J-6 COMBIN.	0	0
GR4J-15 LN(NSE)	0.9	0.9
GR4J-15 NSE	4.2	2.9
GR4J-15 COMBIN.	6.5	6.0
HSAMI LN(NSE)	3.7	3.7
HSAMI NSE	24.7	23.2
HSAMI COMBIN.	32.6	34.7
HMETS LN(NSE)	0.7	0.5
HMETS NSE	7.7	6.9
HMETS COMBIN.	17.0	18.9
MOHYSE LN(NSE)	0.2	0.3
MOHYSE NSE	0.7	1.2
MOHYSE COMBIN.	0.9	0.9

It can be seen that the HSAMI model calibrated on the NSE and combined metrics are the best members in validation in 57.3% of all cases. The other models share the remaining basins as best member, except for the GR4J-6 model which never outperforms the other members.

Another test was performed to determine the number of times the multi-model averaging techniques outperform the best available member in the ensemble. Table 4.5 shows the rate at which the weighting schemes surpass the best individual member in validation.

Table 4.5 Frequency with which the model averaging techniques surpass the best individual member

	SAM	**AICA**	**BGA**	**BICA**	**BMA**	**GRA**	**GRB**	**GRC**	**SCA**
Frequency (%) All basins	11.7	21.4	5.4	14.0	50.8	76.7	73.9	76.9	79
Frequency (%) Only good basins	10.3	18.4	5.1	12.1	50.8	66.2	63.4	66.2	67.6

Finally, the best individual member was selected for each catchment and the NSE value was compared to the NSE value obtained by the model averaging methods. Results of this analysis are presented in figure 4.8.

The best methods (GRA, GRB, GRC and SCA) increase performance by a small margin on most of the catchments. The catchments with low best-model NSE values are more volatile and seem more problematic for the model averaging methods. BMA seems to be the best suited to deal with low best-model NSE value catchments. The relatively low gain in performance must be taken in context. Here the “best member” value is selected from the 15 members independently for each catchment.

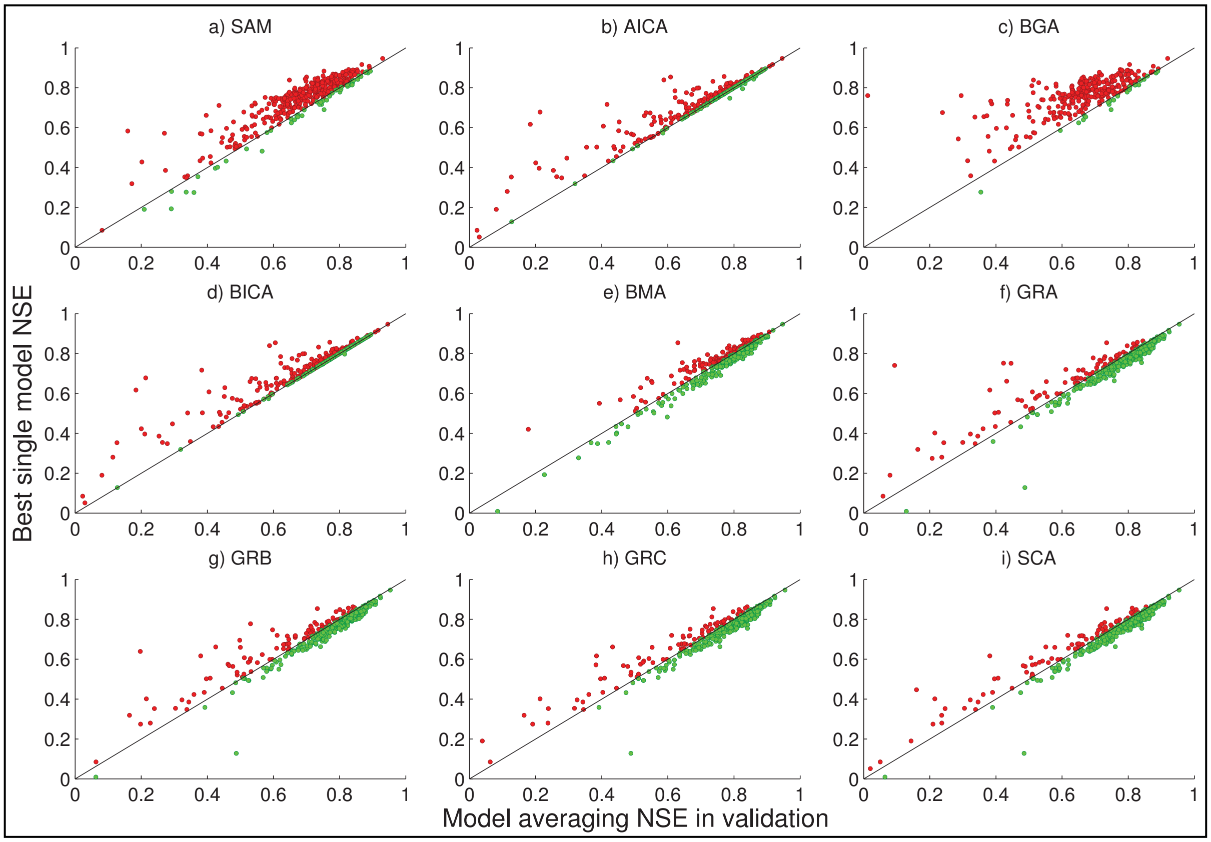

Figure 4.8 Comparison of the 9 multi model averaging method NSE values and the NSE values of the best single model-member for each of the catchments in this study. Model averaging that produces results better than the best member will generate markers under (or to the right of) the 45 degree line. Markers above the 45 degree line indicate that the model averaging performed worse than the best single model

It is clear from figure 4.8 that AICA and BICA prefer to heavily weigh the best member. To give an idea of the relative performance of the rapid degradation in member quality, the same analysis was performed, but this time with the second-best member for each of the catchments. Results are presented in figure 4.9.

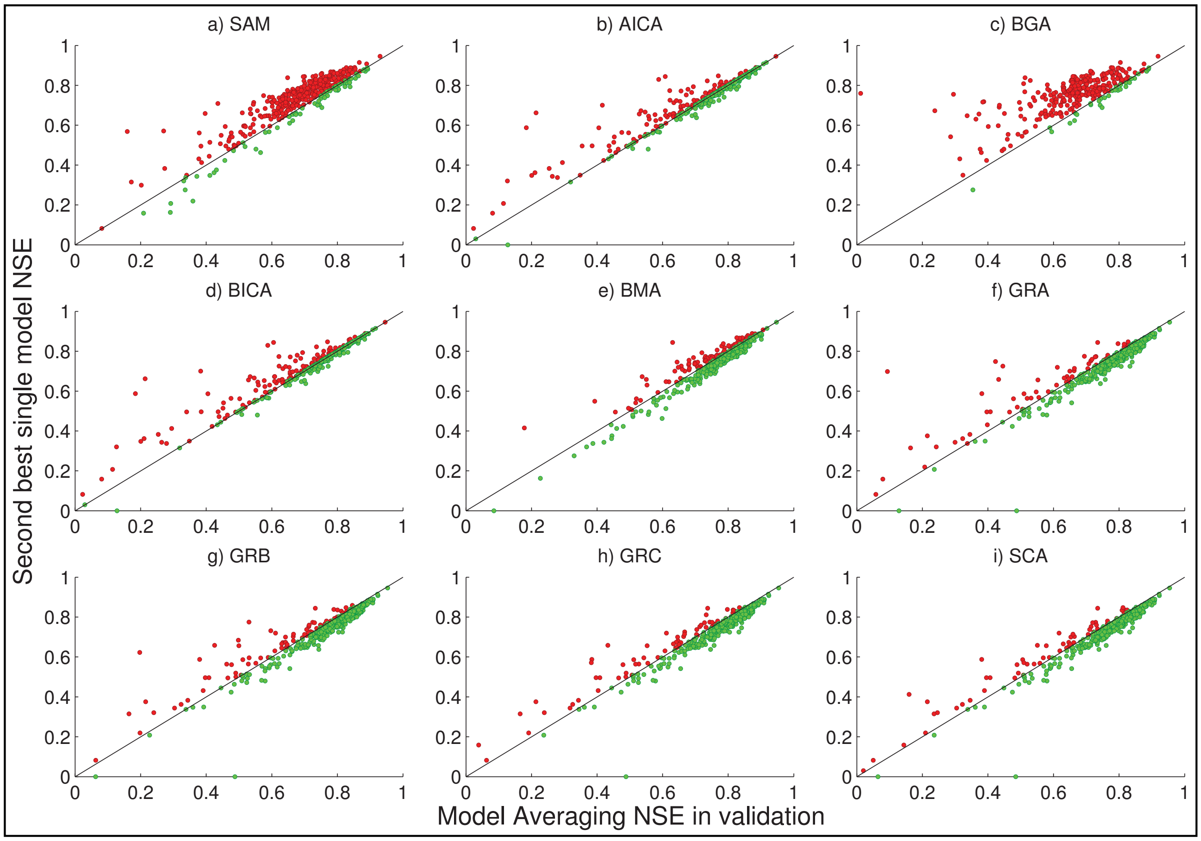

Figure 4.9 Comparison of the 9 multi model averaging method NSE values and the NSE values of the second-best single model-member for each of the catchments in this study. Model averaging that produces results better than the second-best member will generate markers under (or to the right of) the 45 degree line. Markers above the 45 degree line indicate that the model averaging performed worse than the second-best single model

It is clear from figures 4.8 and 4.9 that model averaging methods have the main advantage of being more consistent than the individual members, as well as outperforming even the best on many occasions.

4.3.4 Geographical analysis

The information obtained in this study was then analyzed from a geographical standpoint. The validation NSE values obtained with the GRC method were plotted on a map of the United-States to determine if there was a correlation between climatic zones and method performance, as shown in figure 4.10.

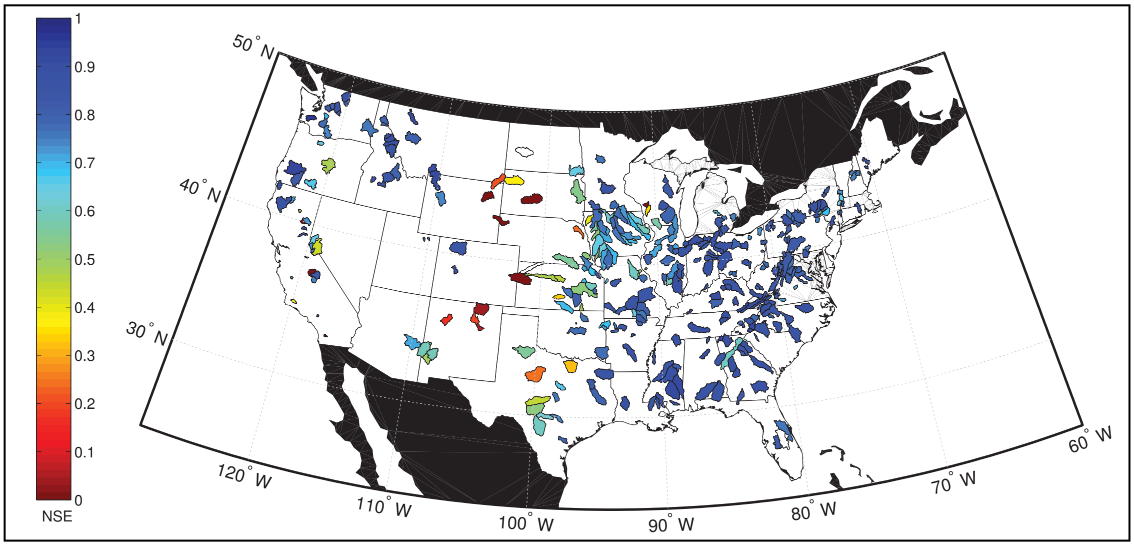

Figure 4.10 Geographic distribution of the basins and their NSE values in validation using the GRC model averaging method

It is clear from figure 4.10 that there are areas which allow for a better performance than others. These discrepancies will be discussed later on. Figure 4.11 differentiates the basins for which the GRC method was better than the best individual model (green) from the basins where the opposite is true (red).

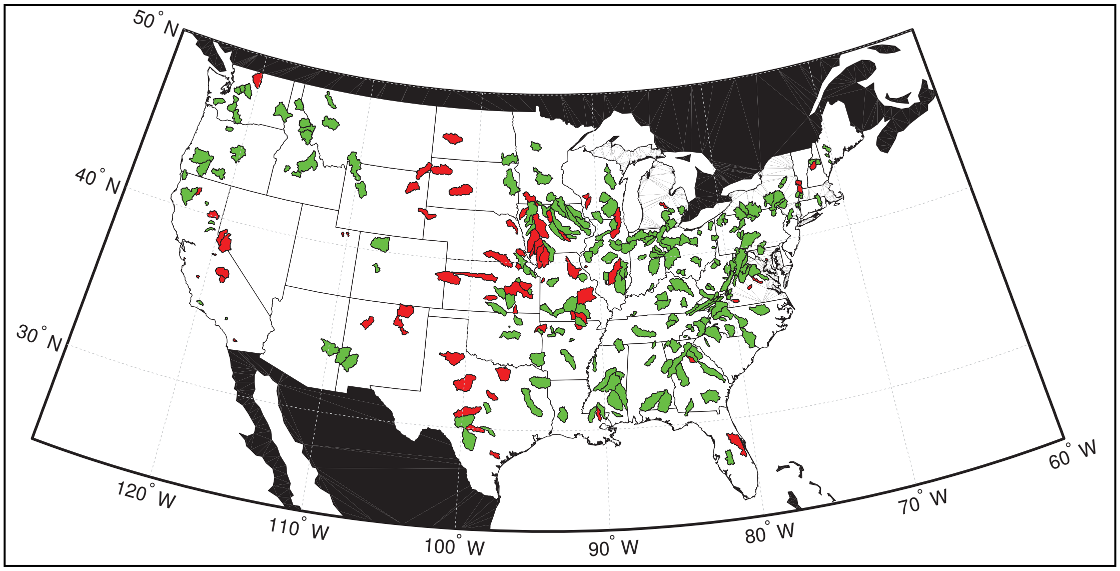

Figure 4.11 Geographic distribution of basins for which the GRC method performed better than the best individual model (green) and the basins for which the GRC was not as good as the best individual model (red)

There is a seemingly obvious area in the center of the US where the GRC method is not able to perform adequately. A few possible explanations will be given in the discussion.

4.4 Discussion

4.4.1 Individual model performance

During the calibration and validation of the 15 individual members, it was shown in figure 4.2 and table 4.4 that the HSAMI and HMETS models contribute respectively 58% and 25% of the best models for each catchment, while the GR4J-6 model did not contribute to the best member group at all. Furthermore, the LN(NSE) metric's contribution was minimal for all models. While it could be possible that the information included in these ensemble members was used in the multi-model averaging, the individual models did not perform as well as the others. This was predictable since the validation metric was the NSE, which favours the models calibrated directly on NSE or on the combined metric. However, this begs the question as to whether or not all 15 members are required to attain these performance levels in multi-model averaging. This question will be discussed in section 4.4.3.

4.4.2 Multi-model averaging method analysis

The reference method (SAM), even though it was meant only for comparative purposes, still did manage to improve upon the best individual members in 11% of cases (table 4.5). This shows that the errors between the different members and the observations are distributed on either side of the observations. More sophisticated methods (such as the ones used in this study) can therefore be expected to produce better results.

The BGA method was the least useful in this study, often performing much worse than the individual models. In only 5% of cases was it able to attain an NSE value equivalent or better than that of the best individual member. These results are not surprising, however, since BGA relies on the datasets being uncorrelated and unbiased, whereas the members are forcefully biased by the choice of objective functions used to produce the heterogeneous hydrographs. The BGA method does improve if only the 5 models calibrated on the NSE metric are used, but it is still outperformed by other methods. For this reason, BGA will not be considered for further analysis. In Diks and Vrugt [2010], BGA is the worst method except for simple averaging using 8 members on a single catchment. AICA, BICA, BMA and GRA we found to be better than BGA, and by a wide margin. The poor BGA performance was thus expected in our study.

The BMA method was found to be an excellent technique when the conditions are met, especially when the ensemble members offer low performance as seen in figure 4.8e). However, the high failure rate (>18%) caused by the incapacity of the method to properly converge on acceptable weights renders the process inefficient. This can happen when the probability distribution functions (PDFs) of the different members are too dissimilar; the expectation-maximization algorithm is then unable to converge towards a solution. In this study, the PDFs vary by large margins due to the different objective functions used in calibration. The positive aspect is that when the method fails, it does not produce weights at all, thus eliminating the risk of using very poor weights and creating a flawed average flow. Furthermore, the BMA method is the longest to execute because of its iterative nature. These conclusions are in line with Diks and Vrugt (2010), who state that the GRA method is as efficient as the BMA method, but that it is much simpler to implement and execute.

The AICA and BICA methods are more robust than the BMA method, but their performance is somewhat lower. They are only able to improve upon the best individual member's score in approximately 21% of cases for AICA and 14% for BICA. Statistically speaking, they are not as efficient as the GRA-GRB-GRC-SCA group as can be seen in figure 4.6. This could be due to the fact that they do not include bias-correction mechanisms, even if the calibration and validation is based on the odd-and-even year method, where one would expect the bias to be averaged out over time. More importantly, AICA and BICA have a tendency to heavily weigh the best individual member and neglecting the others. Diks and Vrugt (2010) drew the same conclusions for AICA and BICA. The resulting averaged flow is therefore similar to the best individual member, but not much better. This is reflected in figures 4.8b) and 4.8d), where the performance is approximately equal to the best member performance.

The remaining methods (GRA, GRB, GRC and SCA) perform quite similarly. The SCA and GRC methods offer the same level of performance, but they differ largely in their complexity. The GRC method is essentially a matrix multiplication operation whereas the SCA method is a stochastic optimization algorithm-based method which requires iterating and evaluating a fitness function. The difference in speed, setup difficulty and ease-of-use favours the GRC method. However, the SCA method is a novel approach which was tested here for the first time. Other algorithms could perhaps be used, or the weight boundaries could be modified to allow more freedom in the optimization routine or even be asymmetrical instead of limiting them to [-5:5]. More research in this area could shed light on this topic. It is also important to stress that the optimization problem is a very simple one that only uses hydrological model outputs. No model evaluation is required and convergence is very rapid. The computational cost of optimizing the weights is negligible compared to the task of having to initially calibrate all ensemble members.

Amongst the Granger-Ramanathan methods, the GRC method shows slight improvements over the others. As they are all extremely simple to implement, the GRC method was selected as being the best overall method in our case study. It has the advantage of producing an unbiased averaged streamflow, which is an important feature for water management. The rest of the analysis is therefore achieved with the GRC method.

4.4.3 Member contribution in multi-model averaging

Each multi-model averaging schemes works in specific ways, thus making it difficult to compare the member selection process. Seeing how the members produce very different hydrographs, it seems interesting to determine which ones contribute the most to the averaged hydrograph. However, since the weights can be small but non-zero in many cases, it can be difficult to determine a threshold defining the non-contribution of a member. Therefore, a multi-objective optimization approach was used. In essence, two conflicting objective functions were minimized simultaneously using the NSGA-II multiobjective optimization algorithm, thus creating a Pareto Front (Deb, 2001). The two conflicting functions were (1) the number of members used and (2) the calibration NSE in multi-model averaging using the GRC method. Using this approach, it was possible to see which models were selected when a certain number of members were required, as well as the associated NSE value. The validation NSE was then computed for the Pareto-optimal solutions using the same weights as defined in the multi-objective optimization process. The Pareto Front and the associated validation NSE values were generated for each of the catchments. The results for 4 randomly selected catchments in this study are presented in figure 4.12.

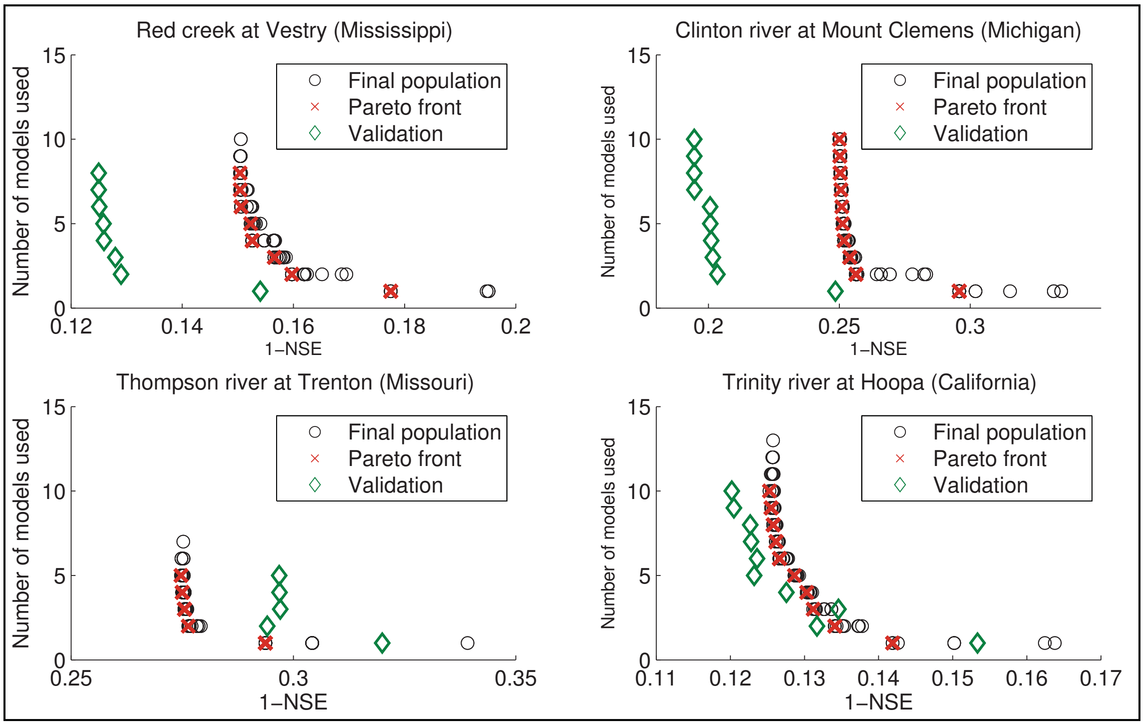

Figure 4.12 Pareto Fronts and associated validation skill for 4 basins with number of models used and (1-NSE) as conflicting objectives

Each solution comprises of a (1-NSE) value and a corresponding number of contributing models in the averaging scheme. In this manner, it was possibleThis allows counting the number of times the models were selected. As can be seen in figure 4.12, the Pareto Front was limited to less than 15 members, which means that adding more members degrades the performance rather than improving it. In fact, no basins required the 15 members to maximize the NSE metric. Furthermore, in many cases the gains made by adding members are negligible. In the leftmost portion of the Pareto Front, the NSE values that lie within a 0.005 (1-NSE)-wide window starting at the best value were removed and the solution with the lowest number of members within this window was kept. This allowed removing the members that do not contribute significantly to an increase in performance during multi-model averaging. Table 4.6 presents the number of times that each member was selected, after removal of the superfluous Pareto-optimal points.

Table 4.6 Number of times each member is selected in the Pareto front, after removal of the superfluous Pareto-optimal points

	GR4J-6			**GR4J-15**			**HSAMI**			**HMETS**			**MOHYSE**			**Total**		
Number of members	**LN(NSE)**	**COMBINED**	**NSE**	**LN(NSE)**	**COMBINED**	**NSE**	**LN(NSE)**	**COMBINED**	**NSE**	**LN(NSE)**	**COMBINED**	**NSE**	**LN(NSE)**	**COMBINED**	**NSE**	Number of occurrences	Number of occurrences / Number of members	Basins with max. number of members
1	0	0	1	3	15	30	14	91	151	7	27	78	5	4	3	429	429	8
2	6	3	10	23	24	52	72	105	177	62	84	167	20	14	23	842	421	79
3	30	39	24	40	40	69	103	94	136	111	90	144	43	32	31	1026	342	128
4	24	50	53	42	44	59	70	74	92	92	68	82	36	36	34	856	214	92
5	22	38	38	39	34	42	50	44	51	68	43	55	29	29	28	610	122	72
6	10	20	21	13	20	25	20	18	22	28	17	27	16	20	23	300	50	34
7	5	5	10	3	5	9	7	9	4	9	8	10	8	9	11	112	16	9
8	5	6	5	1	2	4	2	1	3	5	3	4	3	5	7	56	7	6
9	1	1	1	0	1	1	0	0	1	1	1	0	0	1	0	9	1	1
Total	103	162	163	164	185	291	338	436	637	383	341	567	160	150	160	---	---	429

Note that there are no cases requiring more than 9 members to maximize the NSE with the GRC method. Also note that, surprisingly, even if the GR4J-6 and GR4J-15 models were consistently amongst the worst in the individual member analysis, they regularly contribute to the optimal combination of members during multi-model averaging. Unfortunately, since all the models contribute at some point when more than 2 or 3 members are used, it is not recommended to remove these members from the available ensemble. If it were possible reduce the number of members, a lesser amount of model calibrations would be required in the initial setup of the project since each member is composed of 429 independently calibrated models. Real-world applications should therefore make use of all available models and let the GRC method sort the useful ones from the lot. Also, from figure 4.12, it is clear that the gains made in calibration are mostly retained in the validation mode. Although the figure depicts the results for only 4 catchments, the trend is maintained throughout the entire dataset. Therefore the multiple models serve their purpose and the GRC average seems robust in validation even though the models are all conceptual and similar in their process simulation. A more diverse set of models (such as physically based and distributed models) could be used to analyze the effects of the model structure on model averaging performance.

4.4 Use of multiple objective functions in calibration

The idea of using multiple objective functions during model calibration and considering their outputs as independent members was tentative at first, but the fact that the solution sets often incorporated these modified versions of the same model indicates that they do indeed deserve a role in this type of project. The modified outputs are based on the same model structure but with slightly different characteristics due to the way the parameter sets target diverse aspects of the hydrograph. The choice of objective functions was intended to introduce a variation in the parameter sets which could be helpful to the model averaging methods. However more work could be done to determine better suited objective functions to incorporate in future projects in order to diversify the ensemble members. For example, the LN(NSE) objective function was selected for its ability to better compensate for low flows and to lower the weights on the peak flows. The fact that it was selected even if the validation NSE for the individual models was poor means that the information can be – and is – used effectively by the GRC method. As the results in table 4.6 demonstrate, the best model (HSAMI) benefits more from the addition of multiple objective functions than the addition of hydrological models. This can be seen as when a second and third member are added, they are most often selected from the HSAMI LN(NSE) and HSAMI Combined metric members. Poorer models, on the other hand, would benefit from extra models rather than more objective functions since they are unable to rank well on their own. The best option seems to be to use as many of each as possible in the given time frame, while taking into consideration the relative model performances.

4.4.4 Geographic analysis

By comparing the mean annual precipitation rates for the basins in figure 4.1 to the NSE values obtained in multi-model averaging with GRC (figure 4.10), there seems to be a trend correlating low precipitation values to poor multi-model performance. In the central US, the drier climates are more evident. Figure 4.13 shows the GRC validation NSE versus the yearly average precipitation.

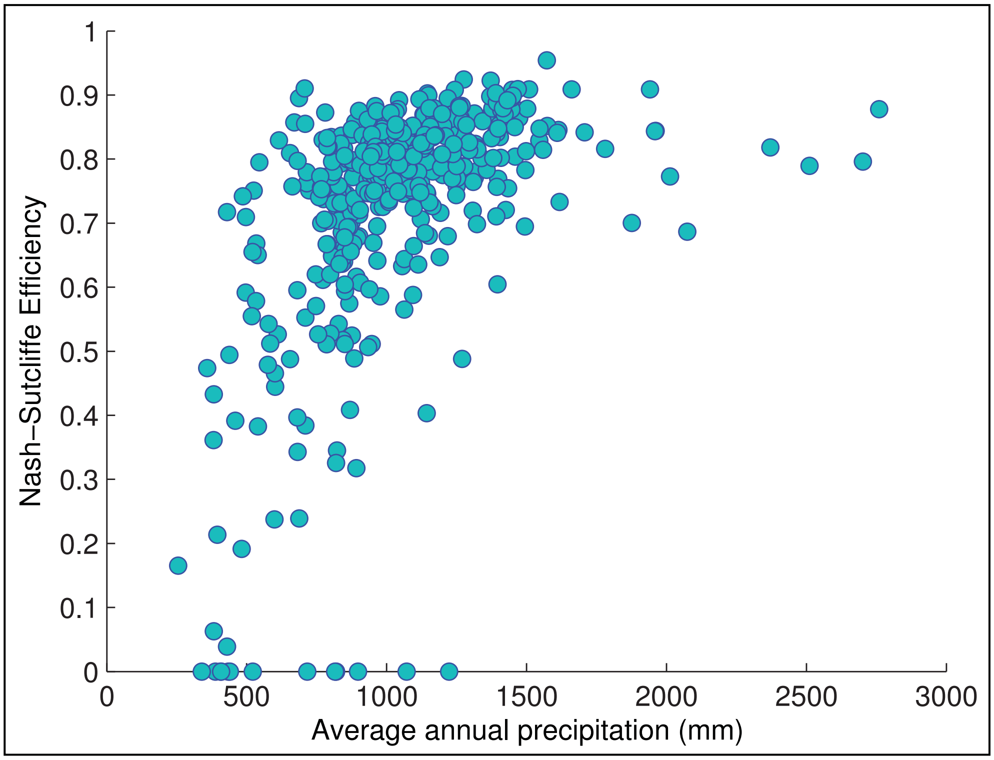

Figure 4.13 Correlation between the GRC validation NSE and the average annual precipitation (mm). The basins with an NSE value of 0 had negative values but were forced on the x-axis for display purposes

There seems to be a definite low-performance zone (below 500~600 mm/yr) in which the worst performances can be found. This can be linked to the hydrological models used in the study. The models are generally used for humid climates and have strong snow accumulation and snowmelt components. The arid catchments in southwestern and central US pose great difficulties for the models used in humid, except perhaps for GR4J variants as GR4J was designed for snow-less conditions. Therefore, the age-old adage “Garbage In, Garbage Out” applies to the multi-model averaging schemes. If the members are unable to correctly simulate the streamflows, it is unrealistic to expect them to combine into an acceptable hydrograph. This is also linked to the results in figure 4.11, where the model averaging methods are unable to improve performance upon the best individual member. It is also noteworthy that the best members on the arid catchments (precipitation lower than 600mm, 33 catchments) were models calibrated on the combined metric in 52% of cases as opposed to 38% in all the other basins. Furthermore, the GR4J model variants are the best members in almost all the catchments with low precipitation. Within the GR4J group, the combined metric performs the best on the arid catchments. The fact that the combined metric contains a bias term could explain its success in low volume streams.

4.5 Conclusion

The present study aimed at evaluating and confirming the pertinence of using multi-model averaging schemes in hydrologic prediction and comparing such methods to traditional mono-model approaches. Some of the methods tested herein (BGA, BICA, AICA) showed performance levels that were not up to expectations, often being unable to beat the best individual models. The BMA method failed to produce sensible results in 18% of cases, while it did perform quite well on the other basins. Adding to this its particularly long execution time due to its iterative nature eliminates it as being the optimal method. The SCA method was shown to be very effective and amongst the lead group, however it is also iteration-based and longer to execute. The fact that it was devised in this study as a proof of concept makes it an interesting candidate for future testing and analysis. The other methods in the lead group (GRA, GRB and GRC) are very similar in terms of complexity and speed, and the GRC method offers similar performance as its counterparts while providing unbiased averaged streamflows. For this reason, the GRC method is proposed as the best multi-model averaging scheme for hydrologic applications.

This study also shows that using multiple objective functions in model calibration produces hydrographs that differ significantly in validation and that these new members allow for an important increase in performance in multi-model applications. It is also shown that the large number of members does not contribute to lowering the predictive skill, but rather leads to an improvement.

Finally, the GRC method produced results that were better than any individual model in almost 80% of cases. In the problematic basins, the calibration skill of the members was usually poor to begin with. This leads to the conclusion that if the members are individually able to produce satisfactory results, the multi-model averaging should produce good results, thus eliminating some risk in the application of the GRC method in hydrological prediction.

4.6 Acknowledgements

The authors would like to thank Dr. Jasper Vrugt for his helpfulness in sharing the BMA MATLAB codes used in this study. We also thank the anonymous reviewers for their fruitful comments.

4.7 References

Abrahart, R. J., and See, L., 2002. Multi-model data fusion for river flow forecasting: an evaluation of six alternative methods based on two contrasting catchments. Hydrol. Earth Syst. Sci. 6(4), 655-670.

Ajami, N.K., Duan, Q., Gao, X., and Sorooshian, S., 2006. Multimodel combination techniques for analysis of hydrological simulations: application to Distributed Model Intercomparison Project results. J. Hydrometeorol. 7, 755–768.

Akaike, H., 1974. A new look at the statistical model identification. IEEE T. Automat. Contr. 19(6), 716-723.

Arsenault, R., Malo, J., Brissette, F., Minville, M. and Leconte, R., 2013. Structural and non-structural climate change adaptation strategies for the Péribonka water resource system. Water Resour. Manag. 27(7), 2075-2087. doi: 10.1007/s11269-013-0275-6.

Arsenault, R., Poulin, A., Côté, P. and Brissette, F., 2014. Comparison of stochastic optimization algorithms in hydrological model calibration. J. Hydrol. Eng. 19(7), 1374-1384. Doi: 10.1061/(ASCE)HE.1943-5584.0000938.

Bates, J.M. and Granger, C.W.J., 1969. The Combination of Forecasts. Oper. Res. Quart. 20(4), 451-468.

Bowler, N.E., Arribas, A., Mylne, K.R., 2008. The Benefits of Multianalysis and Poor Man's Ensembles. Mon. Wea. Rev., 136, 4113–4129. doi:10.1175/2008MWR2381.1

Buckland, S.T., Burnham, K.P. and Augustin, N.H., 1997. Model Selection: An Integral Part of Inference. Biometrics. 53(2), 603-618.

Burnham, K.P. and Anderson, D.R., 2002. Model Selection and Multi Model Inference: A Practical Information-Theoretic Approach, Second Edition. United-States: Springer-Verlag, New-York. 487p.

Cavadias, G. and Morin, G. 1986. The Combination of Simulated Discharges of Hydrological Models. Nord. Hydrol. 17(1), 21-32.

Chen, J., Brissette, F.P., Poulin, A. and Leconte, R., 2011. Uncertainty of downscaling method in quantifying the impact of climate change on hydrology. J. Hydrol. 401(3-4), 190-202.

Deb, K., 2001. Multi-Objective Optimization using Evolutionary Algorithms, 1st edition. Wiley-Interscience series in systems and optimization. Chichester, United-Kingdom: John Wiley & Sons, Ltd, 497p.

Diks, C.G.H. and Vrugt, J.A., 2010. Comparison of point forecast accuracy of model averaging methods in hydrologic applications. Stoch. Env. Res. Risk A. 24(6), 809-820.

Duan, Q., Sorooshian, S. and Gupta, V. K., 1992. Effective and efficient global optimization for conceptual rainfall runoff models. Water Resour. Res. 24(7), 1163-1173.

Duan, Q., Schaake, J., Andreassian, V., Franks, S., Gupta, H.V., Gusev, Y.M., Habets, F., Hall, A., Hay, L., Hogue, T.S., Huang, M., Leavesley, G., Liang, X., Nasonova, O.N., Noilhan, J., Oudin, L., Sorooshian, S., Wagener, T. and Wood, E.F., 2006. Model Parameter Estimation Experiment (MOPEX): An overview of science strategy and major results from the second and third workshops. J. Hydrol. 320, 3-17.

Fortin, V., 2000. Le modèle météo-apport HSAMI: historique, théorie et application. Varennes : Institut de Recherche d'Hydro-Québec, 68p.

Fortin, V. and Turcotte, R., 2007. Le modèle hydrologique MOHYSE (bases théoriques et manuel de l'usager). Note de cours pour SCA7420, Département des sciences de la terre et de l'atmosphère, Montréal : Université du Québec à Montréal, 1-17.

Gneiting, T., Raftery, A.E. and Westveld, A.H., 2005. Calibrated Probabilistic Forecasting Using Ensemble Model Output Statistic and Minimum CRPS Estimation. Mon. Wea. Rev. 133, 1098-1118.

Granger, C.W. and Newbold, P., 1977. Forecasting economic time series, First Edition. New-York, United-States : Academic Press, 333p.

Granger, C.W.J. and Ramanathan, R., 1984. Improved methods of combining forecasts. J. Forecasting. 3(2), 197-204.

Hansen B.E., 2008. Least-squares forecast averaging. J. Econometrics. 146(2), 342–350.

Hansen, N. and Ostermeier, A., 2001. Completely Derandomized Self-Adaptation in Evolution Strategies. Evol. Comput. 9(2), 159-195.

Hansen, N. and Ostermeier, A., 1996. Adapting arbitrary normal mutation distributions in evolution strategies: The covariance matrix adaptation. In Proceedings of the 1996 IEEE International Conference on Evolutionary Computation, 312-317.

Hoeting, J.A., Madigan, D. and Raftery, A.E., 1999. Bayesian Model Averaging: A Tutorial. Stat. Sci. 14(4), 382-401.

Hu, T.S., Lam, K.C. and Ng, S.T., 2001. River flow time series prediction with a range-dependent neural network. Hydrolog. Sci. J., 46, 729–745.

Jefferys, W.H. and Berger, J.O., 1992. Ockham's Razor and Bayesian Analysis. Am. Sci. 80(1), 64-72.

Kruskal, W. H. and Wallis, W. A., 1952. Use of ranks in one-criterion variance analysis. J. Amer. Statist. Assn. 47 (260), 583-621, doi:10.1080/01621459.1952.10483441.

Minville, M., Brissette, F. and Leconte, R., 2008. Uncertainty of the impact of climate change on the hydrology of a Nordic watershed. J. Hydrol. 358(1-2), 70-83.

Minville, M., Brissette, F., Krau, S. and Leconte, R., 2009. Adaptation to Climate Change in the Management of a Canadian Water-Resources System. Water Resour. Manag. 23(14), 2965-2986.

Minville, M., Krau, S., Brissette, F. and Leconte, R. 2010. Behaviour and Performance of a Water Resource System in Québec (Canada) Under Adapted Operating Policies in a Climate Change Context. Water Resour. Manag. (2010) 24, 1333–1352.

Moriasi D.N., Arnold, J.G., Van Liew, M.W., Bingner, R.L., Harmel, R.D. and Veith, T.L., 2007. Model evaluation guidelines for systematic quantification of accuracy in watershed simulations. Trans. Am. Soc. Agric. Eng. 50(3), 885-900.

Mylne, K.R., Evans, R.E. and Clark, R.T., 2002, Multi-model multi-analysis ensembles in quasi-operational medium-range forecasting. Q.J.R. Meteorol. Soc., 128, 361–384. doi: 10.1256/003590002604989 23

Nash, J. E. and Sutcliffe, W. H., 1970. River flow forecasting through conceptual models: Part 1. A discussion of principles. J. Hydrol. 10(3), 282-290.

Neuman, S.P., 2003. Maximum likelihood Bayesian averaging of uncertain model predictions. Stoch. Env. Res. Risk A. 17(5), 291-305.

Oudin, L., Andréassian, V., Mathevet, T., Perrin, C., and Michel, C., 2006. Dynamic averaging of rainfall-runoff model simulations from complementary model parameterizations. Water Resour. Res. 42(7), W07410, doi: 07410.01029/02005WR004636.

Oudin, L., Hervieu, F., Michel, C., Perrin, C., Andréassian, V., Anctil, F. and Loumagne, C., 2005. Which potential evapotranspiration input for a rainfall-runoff model? Part 2 – Towards a simple and efficient PE model for rainfall-runoff modelling. J. Hydrol. 303(1-4), 290-306, DOI: 10.1016/j.jhydrol.2004.08.026.

Perrin, C., Michel, C. and Andréassian, V., 2003. Improvement of a parsimounious model for streamflow simulation. J. Hydrol. 279(1-4), 275-289.

Poulin, A., Brissette, F., Leconte, R., Arsenault, R. and Malo, J.S., 2011. Uncertainty of hydrological modelling in climate change impact studies in a Canadian, snow-dominated river basin. J. Hydrol. 409(3-4), 626-636. doi:10.1016/j.jhydrol.2011.08.057.

Raftery, A.E., 1993. Change point and change curve modeling in stochastic processes and spatial statistics. Journal of Applied Statistical Science, vol.1, no°4, p. 403-424.

Raftery, A.E. and Zheng, Y., 2003. Discussion: Performance of Bayesian Model Averaging. J. Am. Statist. Assoc. 98(464), 931-938.

Raftery A.E., Gneiting, T. and Bakabdaoui, F., 2005. Using Bayesian Model Averaging to Calibrate Forecast Ensembles. Mon. Wea. Rev. 133(5), 1155-1174.

Schwarz, G.E., 1978. Estimating the dimension of a model. Annals of Statistics 6 (2): 461–464.

See, L. and Openshaw, S., 2000. A hybrid multi-model approach to river level forecasting, Hydrolog. Sci. J., 45, 523–536.

Shamseldin, A.Y., O'Connor, K.M., and Nasr, A.E., 2007. A comparative study of three neural network forecast combination methods for simulated river flows of different rainfall-runoff models. Hydrol. Sci. J.-J. Sci. Hydrol. 52(5), 896-916.

Shamseldin, A., O'Connor, K., and Liang, G., 1997. Methods for combining the output of different rainfall-runoff models. J. Hydrol., 197, 203-229.

Smith, M. B., Seo, D.-J., Koren, V. I., Reed, S., Zhang, Z., Duan, Q., Moreda, F., and Cong, S., 2004. The distributed model intercomparison project (DMIP): Motivation and experiment design. J. Hydrol., 298, 4–26.

Valéry, A., Andréassian, V., and Perrin, C. 2014. 'As simple as possible but not simpler': What is useful in a temperature-based snow-accounting routine? Part 2 - Sensitivity analysis of the Cemaneige snow accounting routine on 380 catchments, J. Hydrol., 517(0), 1176-1187.

Valery, A., 2010. Modélisation precipitations – debit sous influence nivale. Élaboration d'un module neige et évaluation sur 380 bassins versants. Agro Paris Tech., 417p.

Velazquez, J.A., Anctil, F. and Perrin, C., 2010. Performance and reliability of multi-model hydrological ensemble simulations based on seventeen lumped models and a thousand catchments. Hydrol. Earth Syst. Sci. 14, 2303-2317.

Vrugt, J.A. and Robinson, B.A., 2007. Treatment of uncertainty using ensemble methods: Comparison of sequential data assimilation and Bayesian model averaging. Water Resour. Res. 43, W01411, doi:10.1029/2005WR004838.

Vrugt, J.A., Gupta, H.V., Bouten, W. and Sorooshian, S., 2003. A Shuffled Complex Evolution Metropolis algorithm for optimization and uncertainty assessment of hydrologic model parameters. Water Resour. Res., 39, 1201, doi:10.1029/2002WR001642, 8.

Ye, M., Neuman, S.P. and Meyer, P.D., 2004. Maximum likelihood Bayesian averaging of spatial variability models in unsaturated fractured tuff. Water Resour. Res., 40, W05113, doi:10.1029/2003WR002557.

CHAPITRE 5

ARTICLE 3 : IMPROVING HYDROLOGICAL MODEL SIMULATIONS USING MULTIPLE GRIDDED CLIMATE DATASETS IN MULTI-MODEL AND MULTI-INPUT AVERAGING FRAMEWORKS

Richard Arsenault[1], Gilles R.C. Essou[1] et François Brissette[1]

[1] Département de Génie de la Construction, École de technologie supérieure,
1100 Notre-Dame Ouest, Montréal, Québec, Canada H3C 1K3.

Article soumis à la revue « Journal of Hydrology » en janvier 2015.

Abstract

In this study we examine the possibility of using gridded climate datasets as inputs to hydrological models as a means to generate distinct members for multi-model averaging. Three hydrological models and four climate datasets were combined to produce multi-model/multi-input, multi-model/mono-input and mono-model/multi-input averaged flows using a weighting scheme that minimizes the RMSE error between the averaged streamflow and the observed hydrograph. The results show that model averaging improves performance significantly and that multi-input averaging provides better results than classical multi-model averaging. A combination of all models run with all datasets (12 members in total) produced the best results with the averaged hydrograph being better than any single member on 70% of the catchments. The median Nash-Sutcliffe Efficiency metric under the multi-model/multi-input framework increased by 0.07 overall. The improvements were shown to stem from the reduction of structural error in the models and in the climate data sources. Tests to remove the climate data biases prior to the hydrological modelling by pre-averaging them proved to be inconclusive. Finally, a few possible improvements to the method and further research options are detailed.

Keywords: Multi-input; multi-model averaging; model structural error; gridded climate data

5.1 Introduction

Hydrological model output averaging has been used extensively in the past and it was shown that considerable increases in hydrological modeling performance can be made using model averaging schemes (Ajami et al., 2006; Cavadias and Morin, 1985; Diks and Vrugt, 2010; Shamseldin et al., 1997; Vrugt and Robinson, 2007). The usual approach is to set-up a given basin in a few hydrological models. The models are then calibrated and the optimal parameter sets are recorded for future use. Model averaging schemes are then typically used to find the optimal weights for each of the model outputs in order to minimize the error between the weighted combined flow and the observed streamflow time series. These weights can then be applied in a validation or prediction mode. The models then simulate flows on the validation or prediction period, and the weights are applied once again to produce a new weighted streamflow series. In most studies, it was shown that the model average generally performed better than any hydrological model taken individually. This technique has been used extensively outside the hydrology community; climate and weather forecast disciplines have been using model averaging for decades (Bowler et al., 2008; Bougeault et al., 2010; Raftery et al., 2005). The premise behind the use of model averaging is that each member has a residual error between it and the "real" observations. If these errors are equally distributed around the real value, a mean of many members will eliminate the overall error and thus give the best possible prediction. However, the errors are seldom equally distributed around the observations and model averaging techniques are used to minimize the overall error.

Over the course of the years, many model averaging schemes have been proposed to improve the prediction in model averaging. The first is the simple arithmetic mean (SAM), which is the "poor man's" option and the baseline against which others are evaluated. Then there are the various weighted options, such as the constrained and unconstrained Granger-Ramanathan averaging (GRA, GRB, GRC) methods, Multi-model Super-Ensemble (MMSE) and the Akaike and Bayes Information Criterion averaging (AICA, BICA). Finally, there are other more exotic approaches, such as the Shuffled-Complex averaging (SCA), Neural-Networks methods (NNM) and Bayesian Model averaging (BMA). Previous work has shown that the GRC method is amongst the most effective, yet it is easy to implement and quick to run. A detailed description of the inner workings of each of these methods is out of the scope of this paper and the reader is encouraged to read Diks and Vrugt (2010) and Arsenault et al. (2014b) for more information.

As the model averaging discipline has progressed, more attempts at finding the best averaging scheme and constraining the associated uncertainty have been put forth. Some studies have attempted to perform model averaging with a single model, but by first calibrating it with different objective functions, with promising results (Arsenault et al., 2014b). The aim is to find parameter sets that target different parts of the hydrograph, thus allowing the model averaging scheme to combine the best of the individual members into a single, better hydrograph. Results have shown that doing so can improve predictive skill, although there is no consensus on whether it is better to use a single model with multiple objective functions or multiple models with a single objective function.

Until now, climate data fed to the model has always been taken for granted and has been somewhat overlooked, although a wealth of studies in the literature attempts to estimate uncertainty due to errors in the climate data. In this paper, we use various sources of climate data within the same model to generate more members for model averaging techniques. Since climate data (of any source) is laden with errors compared to the real historic climate, the use of various types of climate data should make it possible to create a more precise weighted streamflow series by using the hydrological model as an integrator. This has been used in meteorological forecasting, where different forecast models and input data are averaged together in a “multi-system” approach (Mylne et al., 2002), but an equivalent has yet to be proposed in hydrological modelling. This is not to be confused with ensemble streamflow prediction methods which rely on perturbed weather forecast members to estimate the predicted envelope of future streamflow, as in Davolio et al. (2008) and Velasquez et al. (2011). Instead, this paper focuses on combining modeled flows using different sources of weather observations as inputs to obtain a single deterministic hydrograph.

5.2 Catchments and data

This study was performed on a set of 424 catchments selected from the MOPEX (Model Parameter Estimation Experiment) database (Duan et al., 2006). These catchments are distributed over the continental United-States and their mean annual precipitation calculated from the MOPEX database are shown in figure 5.1.

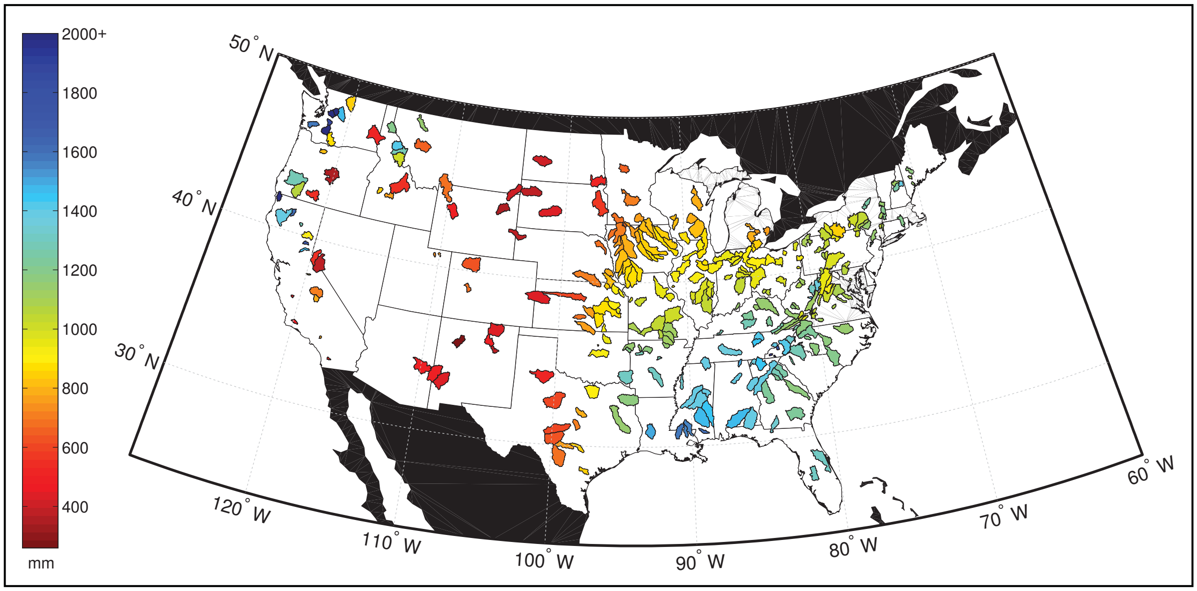

Figure 5.1 Locations of the 424 catchments and their mean annual precipitations

The MOPEX database includes daily climate data (precipitation, minimum and maximum temperature) as well as daily hydrometric time series. The database covers the years 1948-2003. Its conception stems from the National Climatic Data Center (NCDC) weather station observations. MOPEX climate data are averaged observation values for each of the different catchments in the database. An inverse distance weighting method was implemented to estimate the lumped MOPEX data from the weather station observations. A detailed description of this data source is available in Schaake et al. (2006). The MOPEX dataset is available online: ftp://hydrology.nws.noaa.gov/pub/gcip/mopex/US_Data.

Three gridded climate datasets were also used as inputs to the hydrological models for the multi-input aspect of this project. Each one has its own properties and unique interpolating algorithm which creates spatial heterogeneity over the study area. A summary of the datasets is presented here. A detailed comparison between the gridded datasets and the MOPEX lumped observations can be found in Essou et al. (2014).

Santa Clara gridded data

The University of Santa Clara gridded dataset contains daily precipitation and temperatures (minimum and maximum) for the years 1949-2003. They were interpolated on a 0,125° x 0,125° grid using weather measurement data. The interpolation algorithm is based on the Synergraphic Mapping System (SYMAP) by Shepard (1984) and implemented as proposed by Widmann and Bretherton (2000). The Santa Clara dataset is available online:
http://hydro.engr.scu.edu/files/gridded_obs/daily/ncfiles_2010.

Climate Prediction Center gridded data

The Climate Prediction Center (CPC) data contains precipitation data only for the years 1949-2003 with a spatial resolution of 0,25° x 0,25°. The interpolation uses three main sources of observation data such as cooperative network stations, daily NCDC observations and Hourly Precipitation Dataset values (Higgins et al. 2000). The interpolation uses the Cressman method (Cressman, 1959). The CPC dataset is available online:
http://www.esrl.noaa.gov/psd/data/gridded/data.unified.daily.conus.html.

Since CPC only produces precipitation values, it was coupled with the MOPEX temperatures to produce a complete climate dataset which was used in this study.

Daymet gridded data

The Daymet dataset includes daily maximum and minimum temperatures as well as precipitation data for the period 1980-2003. They are produced using the Daymet suite, an ensemble of algorithms and software designed to interpolate (and extrapolate) values at grid points with a 1km x 1km resolution (Thornton et al., 2012). Daymet uses a Gaussian weighting scheme to perform the interpolation on the observation network data. A detailed description of Daymet is available in Thornton et al. (1997). The Daymet dataset is available online: http://daymet.ornl.gov.

Because of Daymet's shorter data availability time period, only the common years for all datasets were used throughout the entire study (1980-2003).

5.3 Models and Methodology

This section describes the hydrological models, the parameter calibration process and the model averaging technique used in this study. The project methodology is also presented.

5.3.1 Hydrological models

Three lumped rainfall-runoff models were used for their ease of use and fast execution speed. They all require the same climate inputs: Maximum and minimum temperature (only the mean for MOHYSE) and separate rain and snow precipitation data at a daily time step.

HSAMI

The HSAMI model (Fortin 2000; Minville et al. 2008, 2009, 2010; Poulin et al. 2011, Arsenault et al. 2013) has been used by *Hydro-Quebec* for over two decades to forecast daily flows on many basins over the province of Quebec. Runoff is generated by surface, unsaturated and saturated zone reservoirs through two unit hydrographs: one for surface and another for intermediate (soil water) reservoir unit hydrographs. The model has 23 calibration parameters, all of which were used for this study.

MOHYSE

MOHYSE is a simple model that was first developed for academic purposes (Fortin and Turcotte 2007). Since then, the model has been used in research applications (e.g. Velazquez et al. 2010). MOHYSE is specifically built to handle Nordic watersheds and has a custom snow accumulation and melt as well as potential evapotranspiration (PET) modules. Ten adjustable parameters require calibration.

HMETS

HMETS is a model that uses two reservoirs for the vadose and phreatic zones (Chen et al., 2011). HMETS is a Matlab based model which has 21 parameters. HMETS' structure resembles that of HSAMI as it accounts for snow accumulation, snowmelt, soil freezing/thawing and evapotranspiration using the hydrometeorological data available to simulate the streamflow at the outlet. It was fitted with a more complex snowmelt model than HSAMI, which could improve simulations in the catchments with particular snow regimes.

5.3.2 Model parameter calibration process

The first step in this study was to calibrate all the catchments with the MOPEX climate and hydrometric data. The calibration period was the odd years from 1980-2003, whereas the validation was calculated based on the even years only. All calibrations were performed using the CMAES algorithm (Hansen and Ostermeier, 1996, 2001) as it was shown that it was able to consistently find good parameter sets for the models in this study (Arsenault et al., 2014a). The objective function used was the Nash-Sutcliffe Efficiency metric (Nash and Sutcliffe, 1970) as it is the most well-known continuous streamflow performance measure and is adequate in most cases over long time series. It does have drawbacks, such as heavily weighting the peak flows, but it is still the best suited metric for this project.

Furthermore, each model was calibrated on all of the basins using the four climate inputs. Therefore there were a total of (424 basins x 3 hydrological models x 4 climate datasets) = 5088 model calibrations, and each catchment is modelled with 12 model-climate pairs. It is worth noting that a previous attempt at using a single climate dataset to perform the calibrations was unsuccessful, as when the alternative datasets were used in validation mode, the results were much poorer then when the models were recalibrated independently.

5.3.3 Model averaging technique

Model averaging techniques rely on different strategies to optimize the prediction skill of the ensemble. Recent comparisons of such methods in the literature seem to point to a class of algorithms that is more robust and performs better than the others in hydrological modelling applications, which is the Granger-Ramanathan (GR) weighting schemes (Granger and Ramanathan, 1984; Diks and Vrugt, 2010). There are three variants of the GR algorithms (GRA, GRB, GRC), but the GRC method stands out as the best in the comparative studies as it performs as well if not better than all the other algorithms and it is more robust, easier to implement and much quicker to run, which is important in the current study. GRC was furthermore found to be the best method in a model averaging comparative study on the same catchments as the ones in this paper (Arsenault et al., 2014b). The GRC approach sets unconstrained weights based on the ordinary least squares (OLS) algorithm. It minimizes the RMSE and uses a constant term to bias-correct the averaged streamflow values.

5.3.4 Multi-model and multi-input averaging application

For each catchment, the hydrological models were run using their optimal parameter set and the corresponding climate data on the calibration period, for a total of 12 (3 hydrological models x 4 datasets) simulated hydrographs. The GRC weighting approach was then applied to generate weights based on these simulated and the observed hydrographs for the calibration period. Once the weights are computed, the hydrological models are run once again, this time on the validation period, which returns another 12 simulated hydrographs. The same weights are applied on the 12 members to generate a single weighted hydrograph. The Nash-Sutcliffe Efficiency metric is finally computed between the weighted and observed hydrographs.

Different scenarios were tested, ranging from all possible model/climate data combinations to mono-model/multi-climate data combinations and vice-versa. Also, statistics were computed on the improvements due to multi-input averaging and on the comparison between multi-input and multi-model approaches. Finally, one model (HSAMI) was run with a single dataset (MOPEX) multiple times with slightly different parameter sets found through multiple calibrations. This allowed finding 10 different parameter sets which can be seen as 10 different model versions, thus adding to the diversity of the ensemble in model averaging.

5.4 Results

The first step in this work was to produce the streamflows for the catchments using each of the model/climate data pairs. In doing so, it was found that for 25 catchments at least one model/climate data pair could not successfully generate acceptable hydrographs. Validation period hydrograph unavailability, poor parameter set and model inability to adapt to the inputs are the main causes for these failures. Therefore the total number of basins used in the remainder of this study is 399, which translates to a 94.1% success rate.

The cornerstone of this study is the hypothesis that model averaging techniques can improve upon single model streamflow simulation. To verify that the model averaging approach has its merits, the classic multi-model averaging was tested. As can be seen in figure 5.2, the multi-model averaging was conducted on the four climate datasets independently. The boxplots (on the left) show the distribution of the NSE values in validation for the four test cases. It can be seen that the GRC averaging performs better than the other individual models in all cases. To confirm these results, a non-parametric Kruskal-Wallis test was performed to measure the statistical significance (alpha=0.05) of the difference between groups (Kruskal and Wallis, 1952). The results of the statistical tests are presented on the right panels in figure 5.2, in-line with the boxplots. Overlapping confidence intervals signifies a non-significant difference between the overlapping groups, and the opposite is true for non-overlapping confidence intervals. It can be seen that the GRC average is significantly better than the individual models in all cases.

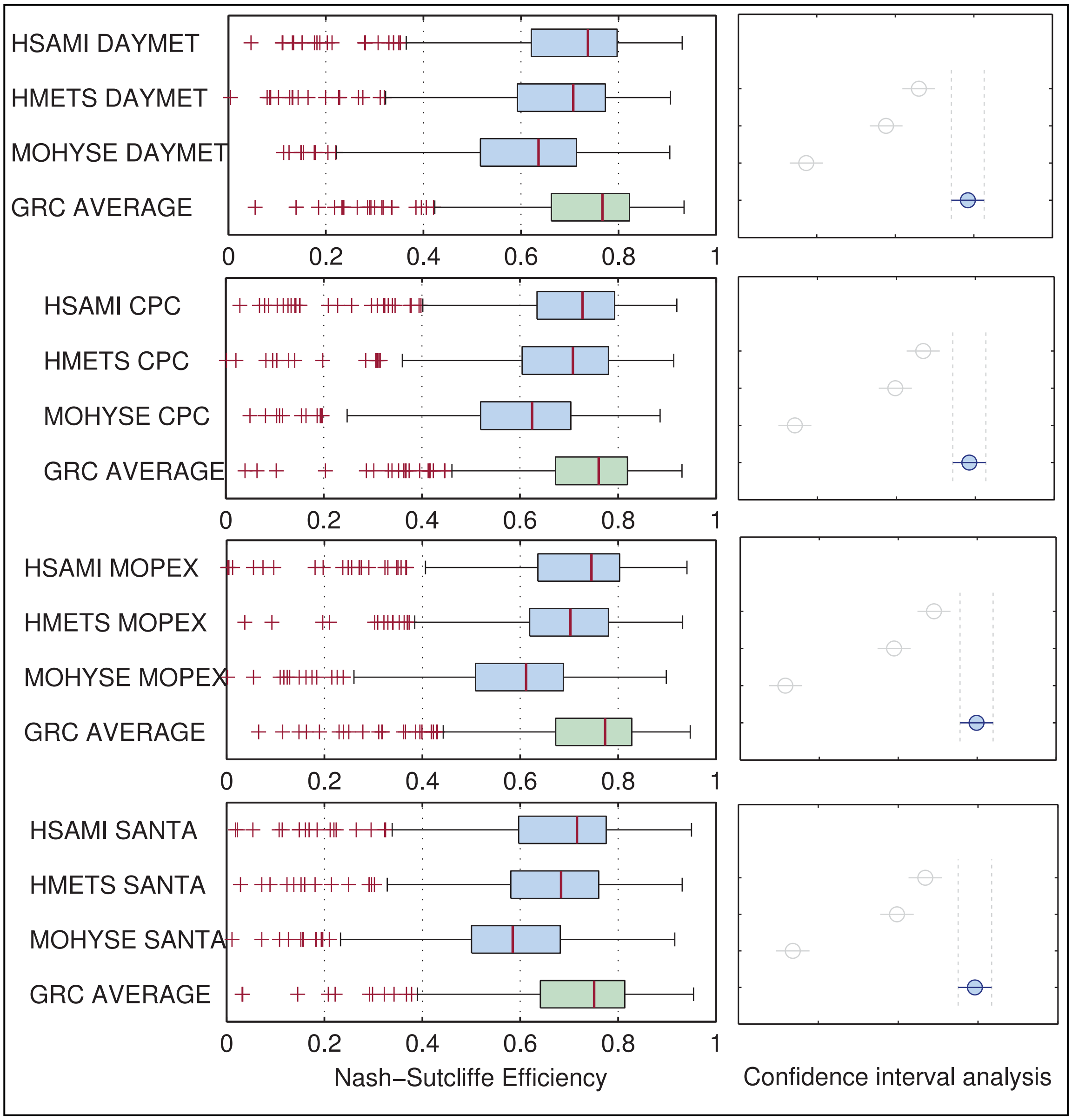

Figure 5.2 Multi-model, mono-input results with statistical significance confidence intervals. The box edges represent the 25th and 75th percentiles, the center marker is the median and the line endpoints are the value of the last element within 2.7 standard deviations (left panels). Outliers are represented by individual marks. Non-overlapping confidence intervals imply statistically significant differences between the groups (right panels)

It is clear that this type of multi-model averaging increases performance, as the body of literature would suggest. The next step was to measure the effects of mono-model/multi-input averaging. In this case, the three hydrological models were run with the four climate datasets, producing 4 members each. Figure 5.3 shows the NSE distributions of the results as well as the statistical significance tests for the mono-model/multi-input averaging.

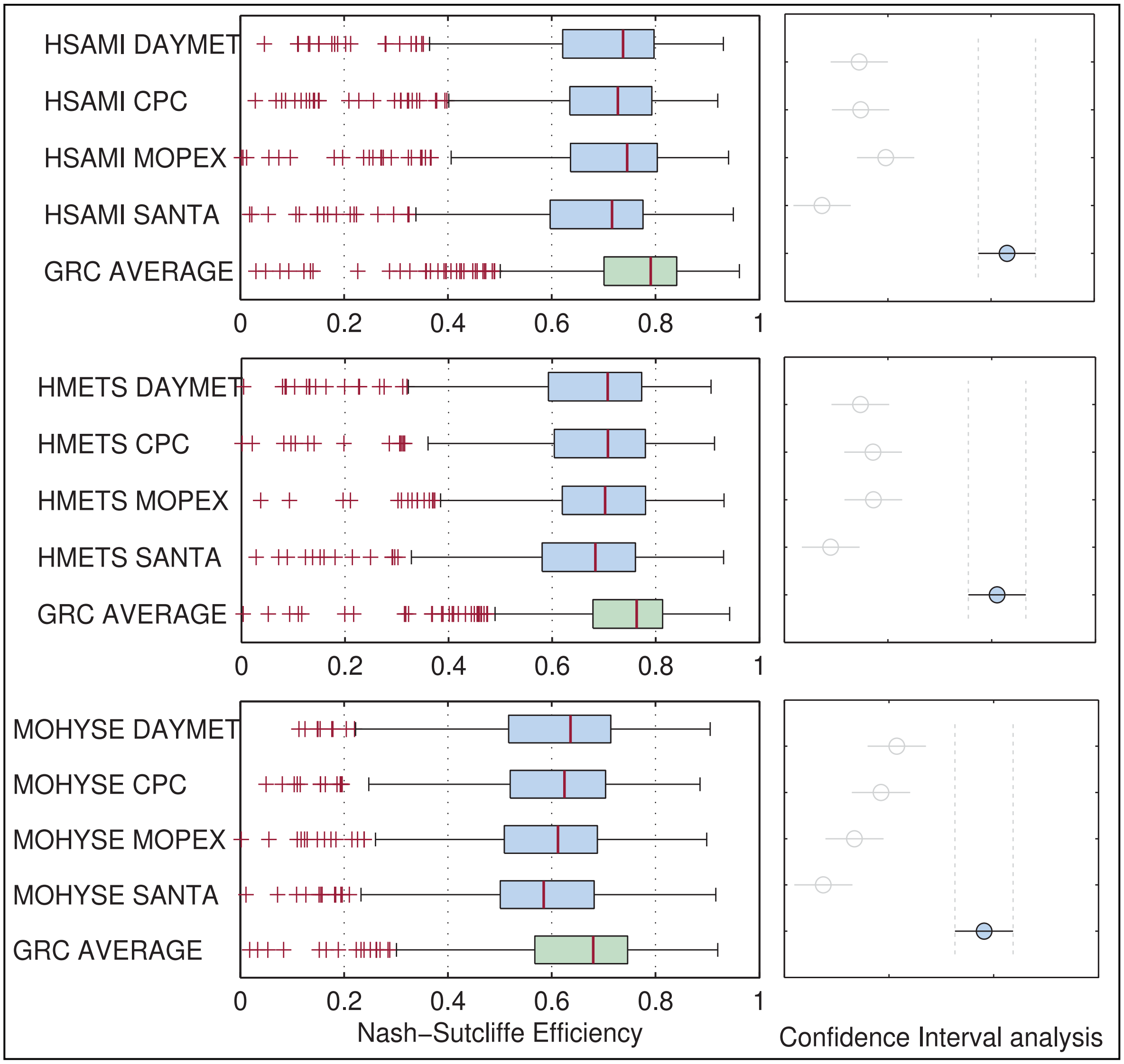

Figure 5.3 Mono-model, multi-input averaging results with statistical significance confidence intervals. The box edges represent the 25th and 75th percentiles, the center marker is the median and the line endpoints are the value of the last element within 2.7 standard deviations(left panels). Outliers are represented by individual marks. Non-overlapping confidence intervals imply statistically significant differences between the groups (right panels)

As was the case with the multi-model/mono-input trials, the GRC average again performs better than any individual member of the ensemble. The results are also statistically significant, which is predictable given the important increase in validation NSE for the three models. It can be seen that the median NSE for the GRC average is often better than the 75th percentile NSE of the individual models. The gains seem to be to a greater extent than the multi-model/mono-input averaging. It is important to note that four members were used here, rather than the 3 members for the multi-model approach. This could lead to a certain bias as literature suggests using more members to increase performance. Nonetheless, the multi-input averaging method seems at least equivalent to the classic multi-model averaging methods.

In an attempt to maximize the gains made with multi-input averaging, the 12 individual members were pooled in a single ensemble and a multi-model/multi-input averaging test was performed. Results are presented in figure 5.4.

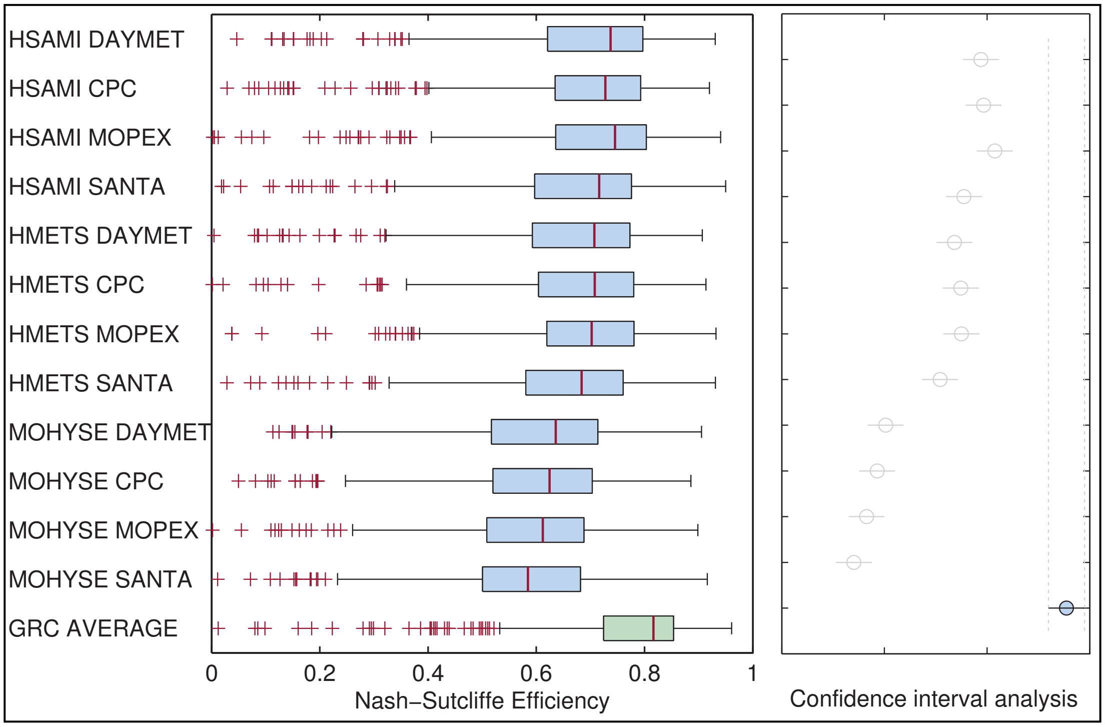

Figure 5.4 Multi-model, multi-input averaging results with statistical significance confidence intervals. The box-and-whisker box edges represent the 25th and 75th percentiles, the center marker is the median and the line endpoints are the value of the last element within 2.7 standard deviations (left panel). Outliers are represented by individual marks. Non-overlapping confidence intervals imply statistically significant differences between the groups (right panel)

It is quite clear from figure 5.4 that the multi-model/multi-input approach improves performance by a good margin. The box-and-whisker plot shows that the difference is larger than when fewer members were used. The median NSE value increased by 0.07 in validation over the best performing single member of the ensemble. The same conclusions can be held with the 25th and 75th percentiles.

The individual catchment gain (or loss) in performance for each model/climate data pair is shown in figure 5.5.

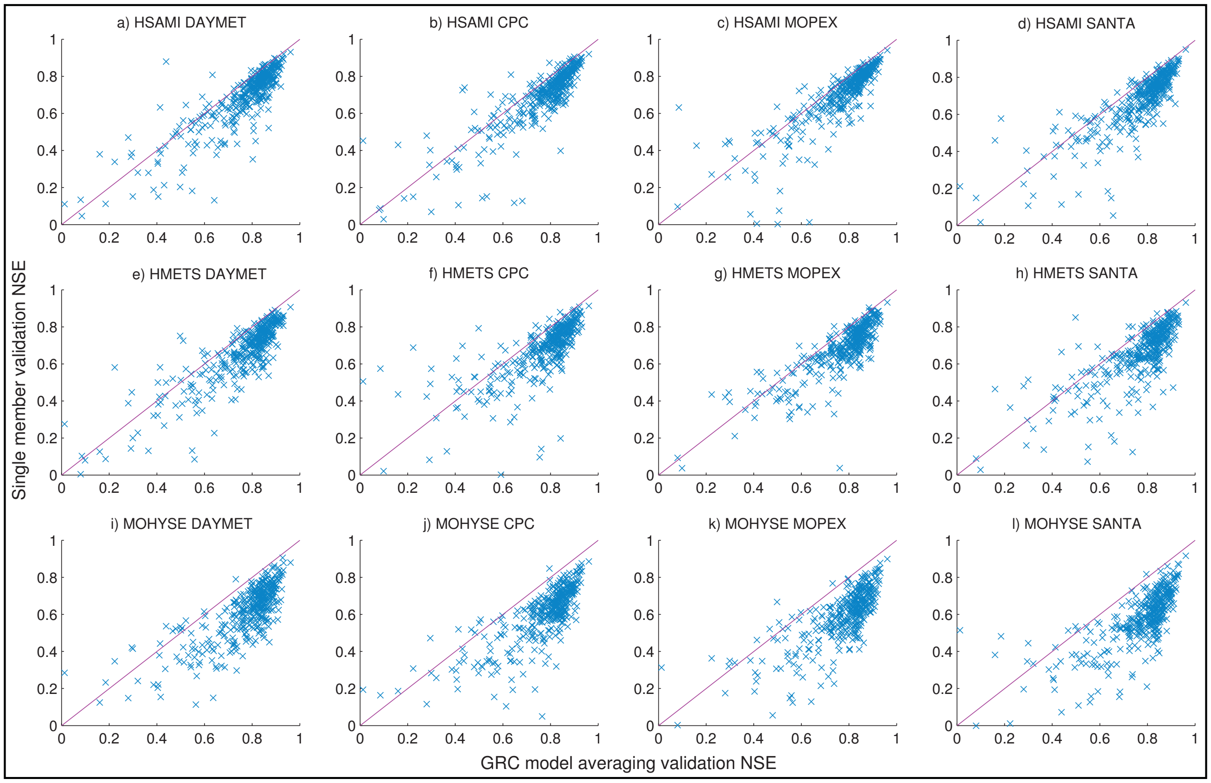

Figure 5.5 GRC model averaging validation NSE compared to the single member performance for each of the catchments. Data points under the 45 degree line represent catchments whose GRC validation NSE is superior to the single model/climate data pair

It can be seen in figure 5.5 that the GRC average outperforms the individual members in most cases. Only a few cases are problematic for GRC in which the average streamflow is worse than the individual member, and these cases are found to be in the central United-

States, which has a semi-arid climate (not shown here). Of these cases, even fewer have a relatively important impact (more than 0.01 in NSE). The HSAMI-CPC and HMETS-CPC members are the most affected by this phenomenon. Unsurprisingly, the MOHYSE model is rarely better than GRC. This was expected since the model in itself is poorer than its two counterparts.

One possible explanation that could discredit this method would be if multiple parameter sets of the same model/climate data pair were used to generate multiple simulated hydrographs. It is conceivable that the error due to the parameter set uncertainty could be eliminated (or reduced) by averaging these flows. A test was performed using 10 equifinal parameter sets for the HSAMI model using the MOPEX climate data as it reflects the observations more than the gridded climate datasets. It was found that the performance is marginally improved in validation and a Kruskal-Wallis test showed that no group was different than another at a 90% significance level. The p-value for this test was 0.28, thus negating the parameter set impact on the multi-input averaging results.

Finally, in an attempt to explain the seemingly large performance increases, single member and averaged hydrographs were compared to observed hydrographs. Figure 5.6 shows a sample of these analyses.

The NSE between the GRC average and the observations in this case is 0.84, whereas the best individual model has an NSE of 0.73. This particular basin was selected because its increase in NSE was the largest when using the GRC method, which makes it easier to see how the averaged hydrograph outperforms the individual members. The same results can be seen on the other basins, but to a lesser extent.

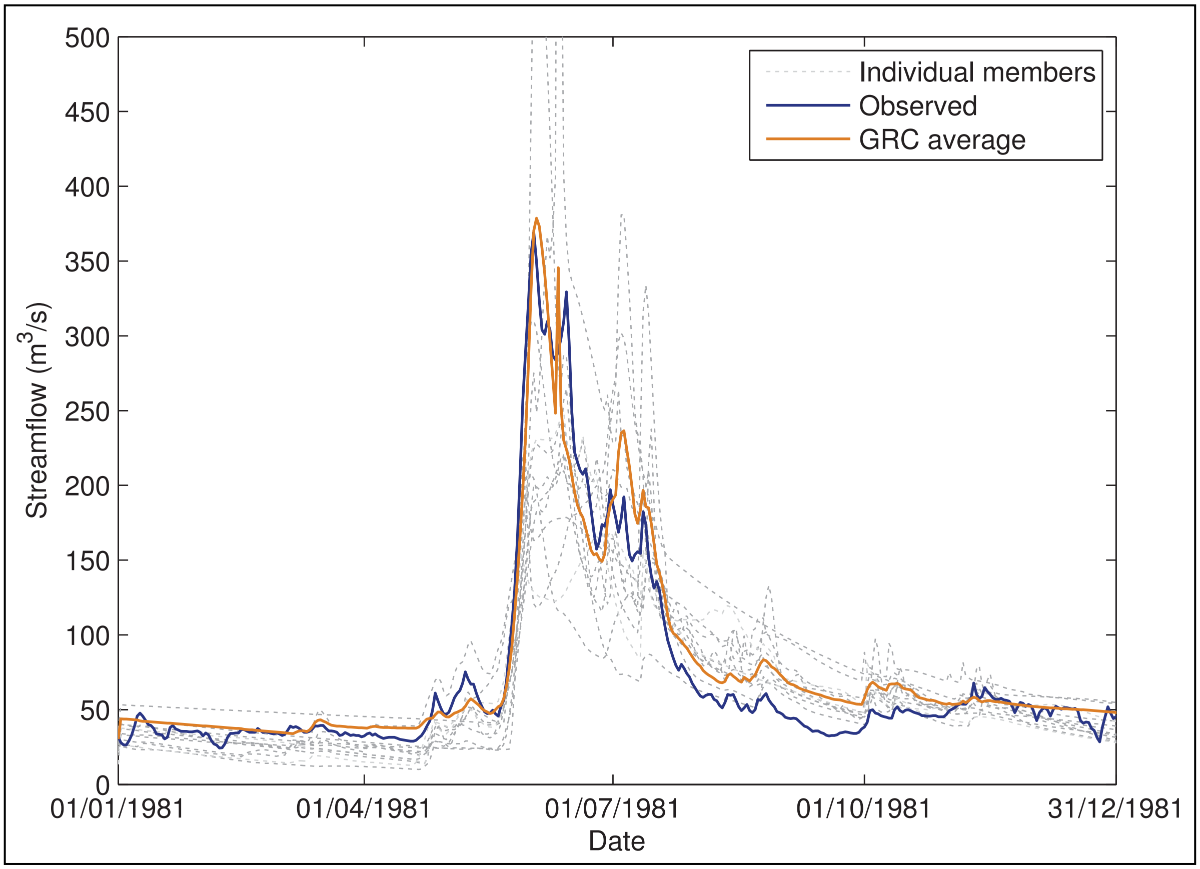

Figure 5.6 Hydrographs of 12 members and the GRC average compared to the observations for the catchment with the most improvement

The relative increase in performance was analyzed further to determine the number of times the multi-model and/or multi-input averaging was better than the best individual member in the ensemble. The first approach was to compare the averaged flow to the best member within each group; the second was to compare the 12-member ensemble average to the best member from each of the 7 other groups (3 multi-input groups + 4 multi-model groups). The results are presented in table 5.1.

Table 5.1 Number of catchments on which the GRC average performs better than the best member in the group (out of 399 basins)

	Group GRC average vs. best within group	**12 member GRC average vs. best within group**
12-member multi-model/multi-input	278	278
4-member HSAMI model	275	304
4-member HMETS model	286	335
4-member MOHYSE model	245	371
3-member MOPEX data	313	329
3-member CPC data	305	332
3-member SANTA data	295	337
3-member DAYMET data	301	338

From table 5.1, it is clear that the 12 member average outperforms the best member from each of the groups more often than the GRC average from within the same groups. This means that the information provided by the 12 members is actively used to improve performance. For example, the 4-member HSAMI GRC average was better than the best member on 275 catchments, whereas the 12-member GRC average was better than the best HSAMI member on 304 catchments. The 29 extra catchments are the result of the information brought by the other model/climate data pairs. It is also interesting to note that the improvements using the GRC average are proportional to the overall model performance. The MOHYSE model, which offers the least stellar performance, gains the most when the 12-member GRC average is compared to the best member of that group. By extrapolation, we can see that the same is applicable to the gridded datasets, where the Santa Clara members gain most from the 12-member GRC average. However, the gains are smaller in the multi-model categories than in the multi-input categories, which confirms that the multi-input averaging is able to significantly improve modelling performance.

5.5 Discussion

5.5.1 Results analysis

The results in this paper show that the model averaging approach increases performance in multi-model averaging, multi-input averaging and in a combination of both. The results are consistently better with GRC averaging than with the individual models, with only a few cases where the averaging fares worse than the individual members. From figures 5.2 and 5.3, it can be seen that the three hydrological models perform quite differently, with HSAMI having a median NSE approximately 0.1 higher than that of the MOHYSE model for a given climate input. However, the variability due to climate inputs is much lower for a given model, and the three hydrological models perform approximately equally well with the different climate inputs. This suggests that the multi-input approach can be a powerful tool given a few different sets of climate data, and that the addition of a model, even if it is generally weaker than others used by hydrologists, can still increase the performance by bringing more information to the averaging scheme.

A tentative test was performed in which the HSAMI model was calibrated with multiple biased MOPEX climate data series. The idea was to generate new precipitation values by reducing or increasing the MOPEX precipitation by 1, 2, 5 and 10%, for a total of 9 climate data series (4 reductions, 4 increases and the original MOPEX data). This would remove the need to work with 4 different datasets as the biases could be generated on-the-fly during the hydrological modelling process. The HSAMI model was calibrated for each of the 9 new climate series and the GRC averaging was applied to the 9 members. Surprisingly, there was an increase in performance in validation, although the improvement was not statistically significant. Perhaps more research in this area could produce interesting results, for example by modifying the precipitation deltas and also modifying the temperature value, but this is out of the scope of this study.

An interesting aspect of this work is that the calibrated hydrological model must be used with its corresponding climate data source for the multi-input method to be successful. An attempt was made to reduce computing time by calibrating the HSAMI model using the MOPEX database and then simply running the model with different climate data but with the same calibrated parameter set. This produced streamflows that when used in the GRC weighting scheme could not produce better results than the individual models. This leads us to believe that the calibrated model and climate data offer more information when they are used together than when used separately. Indeed, if a model is run with a given parameter set while being fed climate data from a different source, the hydrology model could be incapable of assimilating this information to produce meaningful flows. The error generated in the simulated streamflow series via the hydrological model are therefore not representative of the entire series or are not consistent within it. The GRC averaging scheme cannot use this information if the errors are inconsistent within the ensemble as the weights would not translate easily into the validation mode. Further validation of this theory is the fact that a hydrological model run with 10 equifinal parameter sets using a unique climate data source does not increase performance significantly when used in a model averaging scheme. This indicates that the climate data's intrinsic and structural errors are what allow the GRC method to transpose the data from one period to the next. Therefore, the models were calibrated independently using each of the climate data sources. It is also noteworthy that the calibration period was the odd years between 1980 and 2003 and the validation period was the even years in the same time span. The addition or removal of weather stations or general modification of data collection methods should therefore not cause any of the effects shown in this study.

The multi-input model averaging method described in this study shows good promise to improve rainfall-runoff modeling, especially if combined with multiple models. One aspect which was not evaluated in the current study was the added benefit of using distributed models, as their structure is quite different from the lumped models used here. It could be possible that the added diversity, especially with multiple climate data sources, would increase performance even further, but this would need to be verified in a further study.

Also, it must be stated that the computational burden of adding multiple climate datasets to a model averaging project is linearly dependent on the number of climate data series used. This is due to the fact that the hydrological models must be calibrated independently for each of the climate data sources. However, in our case, we consider the 0.07 increase in median NSE to be worth the extra calculations. In this study, the results are impressive even with very simple lumped models. Therefore we would recommend that any such application include multiple models and multiple gridded climate datasets as even the simplest models contribute to the increase in performance.

5.5.2 Pre-averaging of climate data

The basic premise of model averaging is that errors between multiple simulations and observations are distributed around the observations. Therefore, an attempt was made to pre-average the climate data series to eliminate errors before being fed into the hydrological model. A first trial was performed using equal weights, which returned rather poor results. In fact, it performed worse than any individual model. This is not surprising given the fact that the hydrological models were shown to perform poorly when they were used with different climate data series than with which they were calibrated. With a completely different climate source stripped of its coherency with the model parameters, this was to be expected.

However, another trial was done, this time by weighting the climate input series before feeding it to the hydrological model. Furthermore, the model was recalibrated at each step to preserve the climate data / parameter set coherency. A fixed random number generator seed was used in the recalibration loop to guarantee the reproducibility of the results when the weights were varied. Without this contraption, it would be impossible to identify if improvements were due to a better initial calibration seed (random starting point) or better climate data weights. By using this setup it was found that even the best possible weighting scheme over repeated tests with varying seed numbers could not achieve better results than other, unweighted individual members. We realize this method would be much longer than the streamflow averaging in practice, but even so the results are not as good as with multi-model or multi-input averaging (or a combination thereof). This signifies that the diversity in simulated streamflow series is important, more so than eliminating errors before the hydrological model processes come into play.

5.5.3 Possible further improvements

There are a few aspects that were left out of this paper but that could lead to further improvements. First, the various model averaging methods rely on weighting different members in such a way that the error distribution is minimized. For example, with multi-model/mono-input tests, the climate data is identical for all the members. Therefore the gains in validation must come from the reduction of model structural error. As was discussed previously, the parameter sets could have been culprits but were shown to have little effect on the model averaging performance. Second, the multiple inputs are known to contain biases as they are all different from one another, and they are probably all different from the real observations. The interpolating function and the underlying observed data sources contribute to producing structural error in the gridded climate datasets. Any improvements in mono-model/multi-input averaging validation skill must therefore be the result of the reduction of structural errors in the datasets around the observations. Reducing this error to zero, theoretically, would not necessarily mean that the equivalent of the real, error and bias free observations is used. Rather the averaging reduces the error caused by uncertainties in the inputs between the simulated and observed hydrograph to a minimum. Otherwise, we could apply the weights to the datasets and reverse-engineer the climate observations, which we showed to be untrue.

Applying these findings to other structural error sources could lead to better simulations and to an indirect way to measure relative uncertainty in the modelling process. As discussed in Liu and Gupta (2007), there are up to 7 model components in a hydrological model system. We have looked at three (model structure, inputs and parameters), but the initial states, the time-variant states (or a function thereof) and outputs themselves could be addressed individually to reduce the uncertainty in the complete modelling process. Adjusting initial states may have an impact on shorter simulations, such as in hydrologic prediction applications. Time-variant states could be looked upon as an extra source of information from the model and corrections to the states using the model averaging approaches in real-time could lead to reducing error in longer-term prediction. As for the outputs, one could imagine a scenario where the individual models are conditioned on different outputs, but this would require distributed models to take full advantage of. We could also argue that the “optimal” parameter set would depend on the objective function used to calibrate the model. Multiple parameter sets optimized for different objective functions could therefore be used to generate more distinct hydrographs which could, in theory, help to cancel out errors due to the parameters (Arsenault et al. 2014c).

Another possibility to improve performance would be to set thresholds, either by dividing the simulation in hydrologic seasons or by flow rates (See and Openshaw 2000; Hu et al. 2001). In either case, the weighting scheme could be applied to better select the models that contribute to that particular section of the hydrograph. For example, some models are more adapted to peak floods and others to low-flow events. By separating the hydrographs by regime type, the averaging scheme could be able to better adapt and make better simulations. However this has not been tried. It is expected that the weights determination would increase the averaged flow NSE in the calibration period because of the added degrees of freedom, but the performance gain remains to be verified in validation. This method has not been used in this study because of two possible drawbacks. The first is that there could be inconsistencies in the hydrographs when switching from one section to another. It could be possible to algorithmically correct these jumps, but the hydrological soundness would be debatable. Second, there is a risk of parameter overfitting if there are too many categories of flow regime. Defining the maximum number of sections to be used without losing validation performance to overfitting should be addressed by further research.

Finally, it is important to note that multi-input averaging works well in simulation, but that the uncertainty reduction in operational prediction has yet to be determined. Two reasons explain this. First, the datasets used in this work were produced over a long time period with interpolating methods. The biases they generate are supposed constant for the entire series. Therefore the model averaging can eliminate or reduce the bias in validation mode. For real-time prediction, meteorological forecast biases could be much less stable given the short prediction timeframe. Second, multiple sources of meteorological forecasts would be needed in real-time, which would add a level of operational complexity. However the increase in performance using multi-input and multi-model averaging should entice hydrologists to analyse this option further.

5.6 Conclusion

This study was based on the classical multi-model averaging framework, except hydrological models were fed climate data from 4 different climate sources to act as independent simulated streamflow sources. This allowed reducing the impacts of structural error from the models as well as from the climate data. The results have shown that the gains in validation are much larger than with traditional multi-model averaging, with a median NSE increase of 0.07. It is also shown that the multi-input averaging produces equivalent or better results than multi-model averaging, while requiring only one hydrological model. This makes the multi-input approach appealing for operational applications, however the predictive skill of the framework is unknown in operational prediction. This work also shows that it is important to recalibrate the model with each dataset for the models to truly act independently and that multi-parameter averaging produced negligible gains. However it does not seem possible to pre-average the climate data prior to using it in the hydrological models as a means to preserve computing resources.

A few ideas are left for future research, such as combining the multi-model, multi-input, multi-parameter and multi-objective function members into a large ensemble, however there is the risk of overfitting due to the higher number of degrees of freedom. Future projects should investigate the idea of model averaging for the other possible sources of uncertainty to improve overall simulation - and eventually prediction - skill.

5.7 Acknowledgements

The authors would like to thank the four institutions who made their data available for this study. The Daymet dataset is available at http://daymet.ornl.gov/, The MOPEX dataset is available at: ftp://hydrology.nws.noaa.gov/pub/gcip/mopex/US_Data, The Santa Clara dataset is available at: http://hydro.engr.scu.edu/files/gridded_obs/daily/ncfiles_2010, and the CPC dataset is available at: esrl.noaa.gov/psd/data/gridded/data.unified.daily.conus.html.

5.8 References

Ajami, N.K., Duan, Q., Gao, X., and Sorooshian, S., 2006. Multimodel combination techniques for analysis of hydrological simulations: application to Distributed Model Intercomparison Project results. J. Hydrometeorol. 7, 755–768.

Arsenault, R., Malo, J., Brissette, F., Minville, M. and Leconte, R., 2013. Structural and non-structural climate change adaptation strategies for the Péribonka water resource system. Water Resour. Manag. 27(7), 2075-2087. doi: 10.1007/s11269-013-0275-6.

Arsenault, R., Poulin, A., Côté, P. and Brissette, F., 2014a. Comparison of stochastic optimization algorithms in hydrological model calibration. J. Hydrol. Eng. 19(7), 1374-1384. Doi: 10.1061/(ASCE)HE.1943-5584.0000938.

Arsenault, R., Gatien, P., Renaud, B., Brissette, F. and Martel, J.L. 2014b. A comparative analysis of 9 multi-model averaging approaches in hydrological continuous streamflow simulation. J. Hydrol., (Under re-review).

Bougeault, P. and Coauthors, 2010. The THORPEX Interactive Grand Global Ensemble. Bull. Amer. Meteor. Soc. 91, 1059–1072. doi:10.1175/2010BAMS2853.1

Bowler, N.E., Arribas, A., Mylne, K.R., 2008. The Benefits of Multianalysis and Poor Man's Ensembles. Mon. Wea. Rev., 136, 4113–4129. doi:10.1175/2008MWR2381.1

Cavadias, G. and Morin, G. 1986. The Combination of Simulated Discharges of Hydrological Models. Nord. Hydrol. 17(1), 21-32.

Chen, J., Brissette, F.P., Poulin, A. and Leconte, R., 2011. Uncertainty of downscaling method in quantifying the impact of climate change on hydrology. J. Hydrol. 401(3-4), 190-202.

Cressman, G. P. (1959). An operational objective analysis system. Mon. Wea. Rev., 87(10), 367-374.

Davolio, S., Miglietta, M. M., Diomede, T., Marsigli, C., Morgillo, A., and Moscatello, A., 2008. A meteo-hydrological prediction system based on a multi-model approach for precipitation forecasting. Nat. Hazards Earth Syst. Sci., 8, 143–159, doi:10.5194/nhess-8-143-2008.

Diks, C.G.H. and Vrugt, J.A., 2010. Comparison of point forecast accuracy of model averaging methods in hydrologic applications. Stoch. Env. Res. Risk A. 24(6), 809-820.

Duan, Q., Schaake, J., Andreassian, V., Franks, S., Gupta, H.V., Gusev, Y.M., Habets, F., Hall, A., Hay, L., Hogue, T.S., Huang, M., Leavesley, G., Liang, X., Nasonova, O.N., Noilhan, J., Oudin, L., Sorooshian, S., Wagener, T. and Wood, E.F., 2006. Model Parameter Estimation Experiment (MOPEX): An overview of science strategy and major results from the second and third workshops. J. Hydrol. 320, 3-17.

Essou, G.R.C., Arsenault, R. and Brissette, F. 2014. Potential of gridded data as inputs to hydrological modeling. J. Hydrometeor. (Under Review)

Fortin, V., 2000. Le modèle météo-apport HSAMI: historique, théorie et application. Varennes : Institut de Recherche d'Hydro-Québec, 68p.

Fortin, V. and Turcotte, R., 2007. Le modèle hydrologique MOHYSE (bases théoriques et manuel de l'usager). Note de cours pour SCA7420, Département des sciences de la terre et de l'atmosphère, Montréal : Université du Québec à Montréal, 1-17.

Granger, C.W.J. and Ramanathan, R., 1984. Improved methods of combining forecasts. J. Forecasting. 3(2), 197-204.

Hansen, N. and Ostermeier, A., 2001. Completely Derandomized Self-Adaptation in Evolution Strategies. Evol. Comput. 9(2), 159-195.

Hansen, N. and Ostermeier, A., 1996. Adapting arbitrary normal mutation distributions in evolution strategies: The covariance matrix adaptation. In Proceedings of the 1996 IEEE International Conference on Evolutionary Computation, 312-317.

Higgins, R. W., Shi, W. Yarosh, E. and Joyce, R. 2000. Improved United States precipitation quality control system and analysis. NCEP/Climate Prediction Center ATLAS N°6.

Hu, T.S., Lam, K.C. and Ng, S.T., 2001. River flow time series prediction with a range-dependent neural network. Hydrolog. Sci. J., 46, 729–745.

Kruskal, W. H. and Wallis, W. A. 1952. Use of ranks in one-criterion variance analysis. J. Amer. Statist. Assn. 47 (260): 583-621, doi:10.1080/01621459.1952.10483441.

Liu, Y., and H. V. Gupta (2007), Uncertainty in hydrologic modeling: Toward an integrated data assimilation framework, Water Resour. Res., 43, W07401, doi:10.1029/2006WR005756.

Minville, M., Brissette, F. and Leconte, R. 2008. Uncertainty of the impact of climate change on the hydrology of a Nordic watershed. J. Hydrol. 358(1-2): 70-83

Minville, M., Brissette, F., Krau, S. and Leconte, R. 2009. Adaptation to Climate Change in the Management of a Canadian Water-Resources System. Water Resour. Manag. 23(14): 2965-2986.

Minville, M., Krau, S., Brissette, F. and Leconte, R. 2010. Behaviour and Performance of a Water Resource System in Québec (Canada) Under Adapted Operating Policies in a Climate Change Context. Water Resour. Manag. (2010) 24:1333–1352

Mylne, K. R., Evans, R. E. and Clark, R. T. (2002), Multi-model multi-analysis ensembles in quasi-operational medium-range forecasting. Q.J.R. Meteorol. Soc., 128: 361–384. doi: 10.1256/00359000260498923

Nash, J. E., and Sutcliffe, W. H. (1970) River flow forecasting through conceptual models: Part 1. A discussion of principles. J. Hydrol. 10(3), 282-290.

Poulin, A., Brissette, F., Leconte, R., Arsenault, R. and Malo, J.S., 2011. Uncertainty of hydrological modelling in climate change impact studies in a Canadian, snow-dominated river basin. J. Hydrol. 409(3-4), 626-636. doi:10.1016/j.jhydrol.2011.08.057.

Raftery A.E., Gneiting, T. and Bakabdaoui, F., 2005. Using Bayesian Model Averaging to Calibrate Forecast Ensembles. Mon. Wea. Rev. 133(5), 1155-1174.

Schaake, J., Cong, S., and Duan, Q. 2006. The U. S. MOPEX Data Set, IAHS Publication 307 (2006), 9-28.

See, L. and Openshaw, S., 2000. A hybrid multi-model approach to river level forecasting, Hydrolog. Sci. J., 45, 523–536.

Shamseldin, A., O'Connor, K., and Liang, G. (1997) Methods for combining the output of different rainfall-runoff models, J. Hydrol., 197, 203– 229.

Shepard, D.S. 1984. Computer mapping : The SYMAP interpolation algorithm, Spatial Statistics and Models, G, L, Gaile and C, J, Willmott, Eds,, D, Reidel, 133–145.

Thornton, P.E., Running, S.W., White, M.A. 1997. Generating surfaces of daily meteorological variables over large regions of complex terrain. J. Hydrol. 190, 214-251. doi:10.1016/S0022-1694(96)03128-9

Thornton, P.E., Thornton, M.M., Mayer, B.W., Wilhelmi, N., Wei, Y., Cook, R.B. 2012. Daymet: Daily surface weather on a 1 km grid for North America, 1980 - 2012. Acquired online (http://daymet.ornl.gov/)

Velazquez, J.A., Anctil, F. and Perrin, C., 2010. Performance and reliability of multi-model hydrological ensemble simulations based on seventeen lumped models and a thousand catchments. Hydrol. Earth Syst. Sci. 14, 2303-2317.

Velazquez, J.A., Anctil, F., Ramos, M. H., and Perrin, C., 2011. Can a multi-model approach improve hydrological ensemble forecasting? A study on 29 French catchments using 16 hydrological model structures. Adv. Geosci., 29, 33–42, doi:10.5194/adgeo-29-33-2011.

Vrugt, J.A. and Robinson, B.A., 2007. Treatment of uncertainty using ensemble methods: Comparison of sequential data assimilation and Bayesian model averaging. Water Resour. Res. 43, W01411, doi:10.1029/2005WR004838.

Widmann, M., Bretherton, C.S. 2000. Validation of mesoscale precipitation in the NCEP reanalysis using a new gridcell dataset for the northwestern United States, J. Climate, 13, 1936–1950.

CHAPITRE 6

ARTICLE 4 : CONTINUOUS STREAMFLOW PREDICTION IN UNGAUGED BASINS : THE EFFECTS OF EQUIFINALITY AND PARAMETER SET SELECTION ON UNCERTAINTY IN REGIONALIZATION APPROACHES

Richard Arsenault[1] et François Brissette[1]

[1] Département de Génie de la Construction, École de technologie supérieure,

1100 Notre-Dame Ouest, Montréal, Québec, Canada H3C 1K3.

Article publié dans la revue « Water Resources Research » en 2014.

Abstract

This paper focuses on evaluating the uncertainty of three common regionalization methods for predicting continuous streamflow in ungauged basins. A set of 268 basins covering 1.6 million square kilometres in the province of Québec was used to test the regionalization strategies. The multiple linear regression, spatial proximity and physical similarity approaches were evaluated on the catchments using a leave-one-out cross-validation scheme. The lumped conceptual HSAMI hydrological model was used throughout the study. A bootstrapping method was chosen to further estimate uncertainty due to parameter set selection for each of the parameter set/regionalization method pairs. Results show that parameter set selection can play an important role in regionalization method performance depending on the regionalization methods (and their variants) used, and that equifinality does not contribute significantly to the overall uncertainty witnessed throughout the regionalization methods applications. Regression methods fail to consistently assign behavioural parameter sets to the pseudo-ungauged basins (i.e. the ones left out). Spatial proximity and physical similarity score better, the latter being the best. It is also shown that combining either physical similarity or spatial proximity with the multiple linear regression method can lead to an even more successful prediction rate. However even the best methods were shown to be unreliable to an extent as successful prediction rates never surpass 75%. Finally, this paper shows that the selection of catchment descriptors is crucial to the regionalization strategies' performance and that for the HSAMI model, the optimal number of donor catchments for transferred parameter sets lies between 4 and 7.

Keywords: Regionalization, streamflow at ungauged sites, equifinality, regression-augmented similarity, hydrological modeling.

6.1 Introduction

One of the most fundamental, still unresolved problems facing the hydrological sciences community in the past decades has been predicting continuous streamflow in ungauged basins. *Sivapalan et al.* [2003] reinvigorated the quest to find acceptable solutions to this problem, as the IAHS issued the 2003-2012 decade on prediction in ungauged basins (PUB). This led to a multitude of studies aimed at finding methods that would yield satisfactory results. Despite this research impetus, there is still no accepted unique approach to predicting streamflow in such conditions [*Parajka et al.,* 2013; *Razavi and Coulibaly*, 2013].

One of the main tools used to predict flows in ungauged basins is regionalization. The term "regionalization" is somewhat vague and has been interpreted in many ways over the years [*He et al.*, 2011]. In this paper, regionalization refers to the art of finding behavioural parameter sets for hydrological models run on ungauged catchments. Most studies involve at least one of three common regionalization approaches. The three most utilized methods are the regression-based approach, the spatial proximity approach and the physical similarity approach.

Comparative studies have been performed on small to large datasets, on multiple hydrological models and with many variants which will be discussed further [*Merz and Blöschl*, 2004; *Parajka et al.*, 2005; *McIntyre et al.*, 2005; *Bardossy*, 2007; *Yadav et al.*, 2007; *Oudin et al.*, 2008; *Zhang and Chiew*, 2009]. *Razavi and Coulibaly* [2013] performed a thorough review of regionalization studies made in the past decade. The most notable studies have had some diverging results, which is part of the reason why more conclusive evidence is required [*Merz et al.*, 2006; *Oudin et al.*, 2008].

In the regression based regionalization schemes, gauged catchments are calibrated and catchment descriptors (CD) are used to predict individual parameter values to the ungauged catchment, based on its physical properties. *Wagener and Wheaton* [2006] describe the method as follows:

$$\hat{\theta}_L = H_R\left(\theta_R|\phi\right)+v_R \tag{6.1}$$

Here $\hat{\theta}_L$ is the estimated parameter value at the ungauged catchment, H_R is the regression model that links gauged basin parameter values (θ_R) to the catchment descriptors (ϕ) and v_R is the regression model error term. One regression model is built for each parameter, thus the complete parameter set is comprised of independently estimated parameters.

The physical similarity approach is similar to the regression based in that it uses catchment descriptors (CD) to identify gauged catchments similar to the ungauged one. The methods differ in that the physical similarity uses the single (or few) most similar donor catchment(s) to transfer entire parameter sets to the ungauged basin. This method has the advantage of keeping coherent parameter sets during the transfer. *Burn and Boorman* [1993] proposed a method to define similarity between catchments by using the similarity index:

$$\Phi = \sum_{i=1}^{k}\frac{\left|X_i^G - X_i^U\right|}{\Delta X_i} \tag{6.2}$$

Here i is the catchment descriptor identifier, X^G is the CD value for the gauged catchment, X^U is the CD value at the ungauged catchment and ΔX is the range of possible values taken by the respective X^G. The gauged catchment that minimizes the similarity index ϕ is used as the donor catchment.

The spatial proximity method is the simplest method as it makes assumptions about the catchment characteristics instead of requiring the collection of various data to compute CD values. In fact, the spatial proximity method supposes that the proximity alone should ensure a certain level of homogeneity throughout CD values and that any differences are random and fall under measurement uncertainty. As is the case with the physical similarity method, entire parameter sets are transferred from the donor catchment to the ungauged basin.

In both the spatial proximity and physical similarity methods, two options are possible if more than one donor basin is used. The first is parameter averaging, where the parameter sets are averaged before being fed into the model. The second is model output averaging, which averages the individual streamflows predicted by each independent parameter set.

6.1.1 Equifinality

One of the reasons why parameter regionalization is troublesome is because of its reliance on calibrated model parameter sets for initializing the different approaches. The calibration process is hindered by "equifinality", which is defined as having multiple, differing parameter sets that are equally acceptable during the model calibration and validation processes (*Beven*, 2006). For hydrological models with small parameter spaces and low parameter interdependence, equifinality is often not a problem. However, if the hydrological model has the opposite attributes, many parameter sets can be found as behavioral during calibration. Under the equifinality assumption, two very similar catchments could then have parameter sets that are uncorrelated, which could clearly be problematic for regionalization studies. In this case, model parameters are only loosely correlated to catchment attributes, which undermines the basic hypothesis of the regression-based and physical similarity based methods. Therefore, most studies utilize simple hydrological models with few parameters in order to preserve a high level of independence between parameters and a good correlation between parameters and catchment descriptors.

6.2 Scope and aims

With all the current studies in the literature, it is important to justify another publication on the subject. The main reason is that we are still far from understanding all the strengths and weaknesses of the different methods: Uncertainty lies in every aspect of PUB studies, parameter set selection is misunderstood and model structure uncertainty compounds the problem. By adding more test-cases to the body of literature, we hope to contribute to finding acceptable approaches for hydrologically different systems.

This paper will shine new light on a few areas not covered in the previous studies. Original contributions can be summarized in 4 points:

1- The 268 catchments cover the majority of the province of Québec, with a total area of 1.6 million km^2. In perspective, this is the equivalent of the combined land area of France, Germany, Spain and the United Kingdom, allowing for a large and diverse topographical, climate and land cover dataset. Furthermore, the catchments are very heterogeneous in size and attributes, with areas ranging from 30 to 69191 square kilometres;
2- Estimation of parameter set selection uncertainty on the regionalization strategies' performance;
3- Combination of regression-based and similarity/proximity methods;
4- Use of a hydrological model with high dimensionality and parameter interdependence.

Point 4 is not an innovation in itself, however using high dimensionality models is unusual in regionalization studies. Furthermore, the fact that many studies have shown differing results shows the need for more large-scale attempts at improving predictive skill in ungauged basins [*Oudin et al.*, 2008; *Bao et al.*, 2012].

6.3 Study area and data

This section describes the study area and the data used for the hydrological modelling on the 268 basins.

6.3.1 Study area

The study area is composed of 268 gauged catchments covering the province of Québec, Canada. Figure 6.1 shows the location of the catchments, their mean annual precipitation as well as their relative sizes. Some basins are sub-basins of larger basins which are included in the study. Therefore the largest basins do not appear whole on figure 6.1, because the sub-basins are in front.

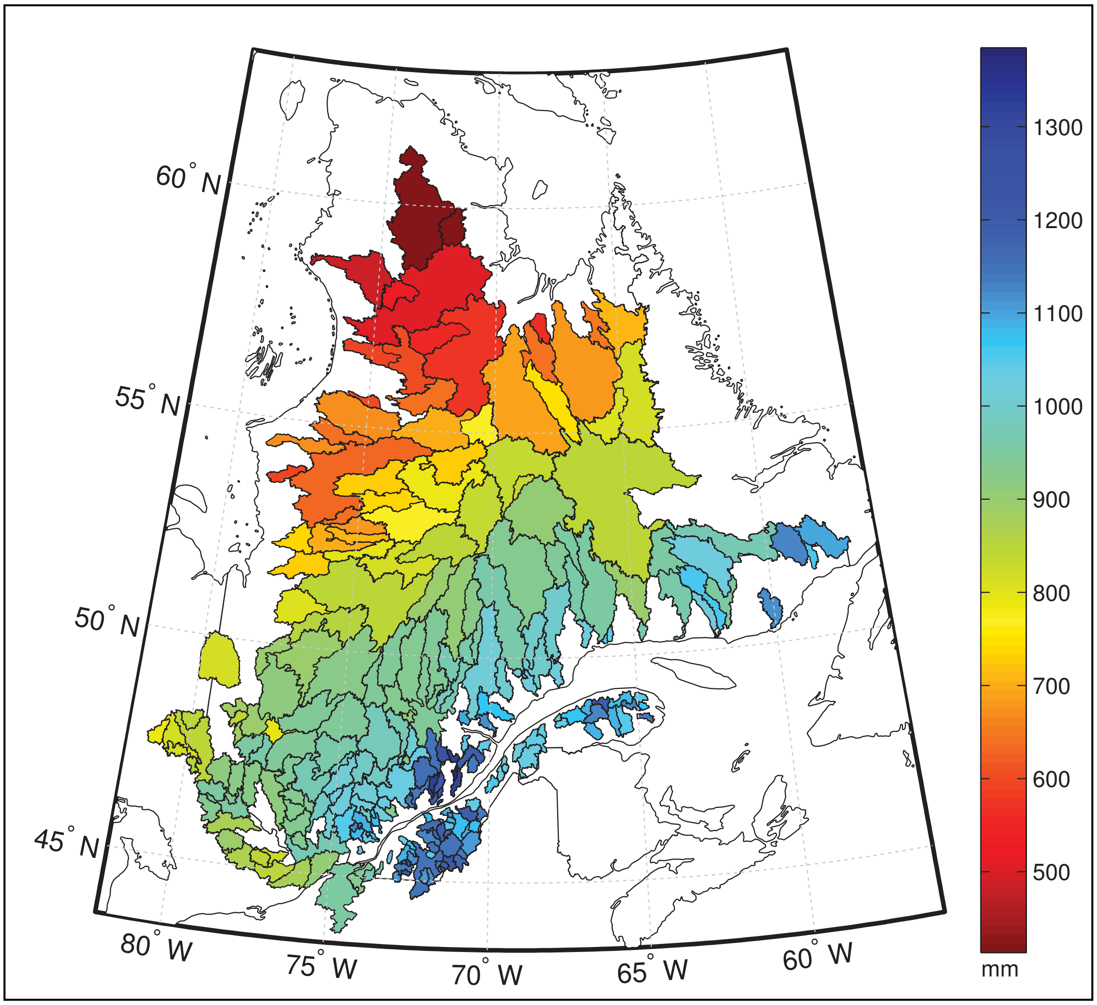

Figure 6.1 Catchment locations from the CQ2 database and mean annual precipitation

The basins range in size from 30 to 69191 km², and cover most of the province. They are therefore heterogeneous in hydroclimatic terms. The catchment descriptors' statistics are presented in table 6.1. The list is based upon the paper by *He et al.* [2011] which counted the number of times catchment descriptors were used in their review of regionalization methods. In order, the most oft-used descriptors are: Area (11 times), Slope (10), Percentage of area covered by various terrain types (10), Soil classification (6), Elevation (5) and Drainage density (5). Others are used less often, such as climate descriptors which are quoted less than 2 times in 15 studies. From this list, we use Area, Elevation, Slope and Land use percentages. Soil classification was impossible due to lack of data in the province of Quebec, where there is mostly only very rough estimates, so it was left out. The drainage density was another possibility, however it was decided to forego this particular descriptor as it required a very large investment in time to achieve (268 basins, some very large, in ArcGIS, with multiple manipulations). In fact, we started to extract the data but preliminary tests showed that physical similarity without drainage density was still better than the other methods, so it was not required to be the best method altogether.

Table 6.1 Statistics of Catchment Descriptors (CDs) used in this study

Catchment descriptors	**Maximum**	**Minimum**	**Average**
Area (km²)	69191	30	6832
Slope (%)	51.9	1.1	10.7
Elevation (m)	916	52	383
Land Cover - Crop (%)	83.1	0	8.7
Land Cover – Forest (%)	96	0	65.2
Land Cover - Grass (%)	65.5	0	13.6
Land Cover - Urban (%)	16.4	0	1.2
Land Cover - Water (%)	35.6	0	9.3
Land Cover - Wetlands (%)	17.1	0	1.2
Mean annual precipitation (mm)	1412	413	965
Longitude (degrees)	-57.9	-81	-72
Latitude (degrees)	59.9	44.5	49
Aridity index	0.99	0.31	0.61

The latitude and longitude descriptors act as surrogates for various characteristics that are either unknown or strongly correlated to the basin location. For example, hydrogeological properties are unknown over most of the territory, so it is supposed that there is somewhat of a continuity between two adjacent catchments regarding soil composition and conductivity. In the same manner, there are strong precipitation, temperature, evapotranspiration and snow cover gradients in the province, thus making the latitude and longitude rough surrogates of these climatic properties.

6.3.2 Meteorological and hydrological datasets

A newly created hydrometric database called CQ2 was used as the starting point for our study. CQ2 is a partnership between various province and industry partners who combined all their respective datasets into one large uniform set, and made available for certain research applications.

The observed climate data was replaced by the Canadian National Land and Water Information Service (NLWIS) 10km gridded dataset as many catchments had no observed data at all within their boundaries. In previous studies, NLWIS was shown to be an adequate substitute to observed data for hydrological modelling purposes [*Chen et al.*, 2013]. This dataset was created by interpolating station data using a thin plate-smoothing spline surface fitting method [*Hutchinson et al.*, 2009]. The observed datasets were generated by averaging all grid points within each watershed.

6.4 Methodology

This section describes the methods and regionalization strategies used in this study. Model calibration techniques, as well as implementation of the common approaches are described. The regression-augmented approach is also defined and described.

6.4.1 HSAMI hydrological model and calibration

The HSAMI model [*Fortin*, 2000; *Minville et al.*, 2008, 2009, 2010; *Poulin et al.*, 2011; *Chen et al.,* 2011; *Arsenault et al.*, 2013] has been used by *Hydro-Québec*, Québec's hydroelectric company, for over two decades to forecast daily flows on more than 100 basins over the province of Quebec. It is a lumped conceptual model based on surface and underground reservoirs. It simulates the main processes of the hydrological cycle, such as evapotranspiration, vertical and lateral runoffs, snowmelt and frost. Runoff is generated by surface, unsaturated and saturated zone reservoirs through two unit hydrographs: one for surface and another for intermediate (soilwater) reservoir unit hydrographs. The required inputs are spatially averaged maximum and minimum temperatures, liquid and solid precipitation and cloud cover fraction. The model has up to 23 calibration parameters, all of which were used for this study. The model is known to have interdependent parameters, which is very frequent in hydrological modelling, but adds additional uncertainty on the regionalization approaches.

The first step in this study was to calibrate the model on all the catchments to obtain parameter sets to be transferred. Ordinarily, a single calibration is made and the parameter set is used on another time period for validation. While calibration and validation were performed in this study, 10 calibration sets were generated instead of a single one. The 10 calibrated parameter sets for each catchment were only accepted if the NSE value was within 0.01 of the best NSE value for that basin to ensure equifinality was present. This series of calibrations will allow studying uncertainty due to parameter set selection using a bootstrapping method. More calibrations could be performed, however in our experience it would be unnecessary as the 10 sets should be different enough from one another to adequately sample the parameter set uncertainty under equifinality constraints, but this has yet to be proven definitively. Nonetheless, the 10 parameter set approximation is used in this study. For more insight on the HSAMI model uncertainty in calibrated parameter sets, see *Arsenault et al.* [2013]. Furthermore, calibrating the 268 basins required approximately 1 day on our 40 core cluster. Performing 10 calibrations thus required 10 computing days. For the bootstrapping (which will be discussed further), a full run would require approximately 4-6 hours per test case (for 1000 bootstraps). Therefore calibrating the model takes much more time than the bootstrapping of regionalization methods. With 1000 bootstraps, we found that the envelope of simulated hydrographs was stable (increasing from 500-1000 changed marginally (mostly the extremes)), but increasing from 1000 to 10000 changed practically nothing. For this reason, 1000 bootstraps and 10 calibrations were the optimal numbers as this was the most calibrations we could do with the allotted time on the cluster. Doing so on a personal computer would require 10x more time (4 core computer).

All calibrations for the HSAMI model were performed using CMAES [*Hansen and Ostermeier*, 1996, 2001] as it was shown that this particular algorithm was the best for the optimization problem at hand using the methodology proposed in *Arsenault et al.* [2013]. The objective function used was the Nash-Sutcliffe Efficiency metric [*Nash and Sutcliffe,* 1970] computed on daily discharge values, as is the case in most regionalization studies. NSE values range from 0.12 to 0.98, with a median of 0.84. Lower-scoring basins were not removed from the study in order to keep as much information as possible for the regionalization strategies. However, the approaches were tested with and without the basins whose NSE values were below 0.7. There are 31 basins (11.5% of the ensemble) whose calibration NSE value is lower than 0.7. It is important to note that the NSE was used since it is an adequate metric for continuous streamflow simulations, although it does put more weight on the peak floods. It was impossible to use other variables than streamflow since nothing other than streamflow is measured on the basins, but it could have been possible to use a transformation of the streamflow as a proxy. However, since the parameter sets perform equally, it is hypothesized that the difference in regionalization performance is due to parameter set selection and not calibration performance.

6.4.2 Uncertainty analysis

In each of the trials detailed above, a bootstrapping approach [*Efron*, 1979] was used to resample the various available parameter sets from each donor catchment. In the case of regionalization, it is impossible to know how the different methods fare when differing parameter sets are used. In the equifinality context, many acceptable parameter sets can exist and provide the same results in calibration and validation. It is therefore practically impossible to verify the extent at which equifinality can affect the regionalization strategies' performance. However, bootstrapping does allow estimating the added uncertainty under these circumstances. In this study, the observed variables are the calibrated parameter sets which are known to exhibit equifinality to a certain degree. Then a random parameter set (from the 10 available) for each donor catchment was taken with replacement, and the NSE metric was measured. This was repeated for all donors and all basins, allowing the computation of a Success Rate (SR). The entire process was repeated 1000 times which resulted in a distribution of the SR metric. Confidence intervals were then taken on the bootstrap distribution, which should be similar to the real (but unknowable) population distribution. The Bias Corrected and Accelerated (BCa) [*Efron*, 1987; *DiCiccio and Romano*, 1995] method was used to adjust bias and skewness in the distribution for more precise confidence interval estimations. A comprehensive explanation as well as a complete test-case pertaining to bootstrapping and confidence intervals is available in *Ebtehaj et al.* [2010].

6.4.3 Generalities common to all regionalization methods

Some aspects of the methodology were common throughout the study. They are listed here to avoid repeating them for each regionalization scheme.

First, all the strategies were analyzed using all of the 268 available basins. Then, the catchments whose mean calibration NSE values were less than 0.7 were discarded from the list of possible donor basins. However, they were used for the validation, seeing as in the real world it would be impossible to determine *a priori* which basins would perform well as they would be ungauged. The success rate thus includes the basins for which the best model parameter set found is poor (which we will define as "bad basins") as well for this part of the study. This is similar to the approach used by *Oudin et al.* [2008]. It also allows studying the effect of keeping bad catchments during the parameter transfer process, as the added diversity of the bad catchments could result in an overall increase in performance.

Second, the success rate (SR) was defined as the total number of acceptable predictions divided by the total number of predictions. A success rate of 0.4 would indicate that the model was successfully parameterized on 40% of the basins using the regionalization scheme. In this study, a successful parameterization requires that the validation NSE be at least 85% of the mean calibration NSE for the pseudo-ungauged catchment. Figure 6.2 shows an example of the effect of using a threshold rather than using a fixed value to determine the success criterion.

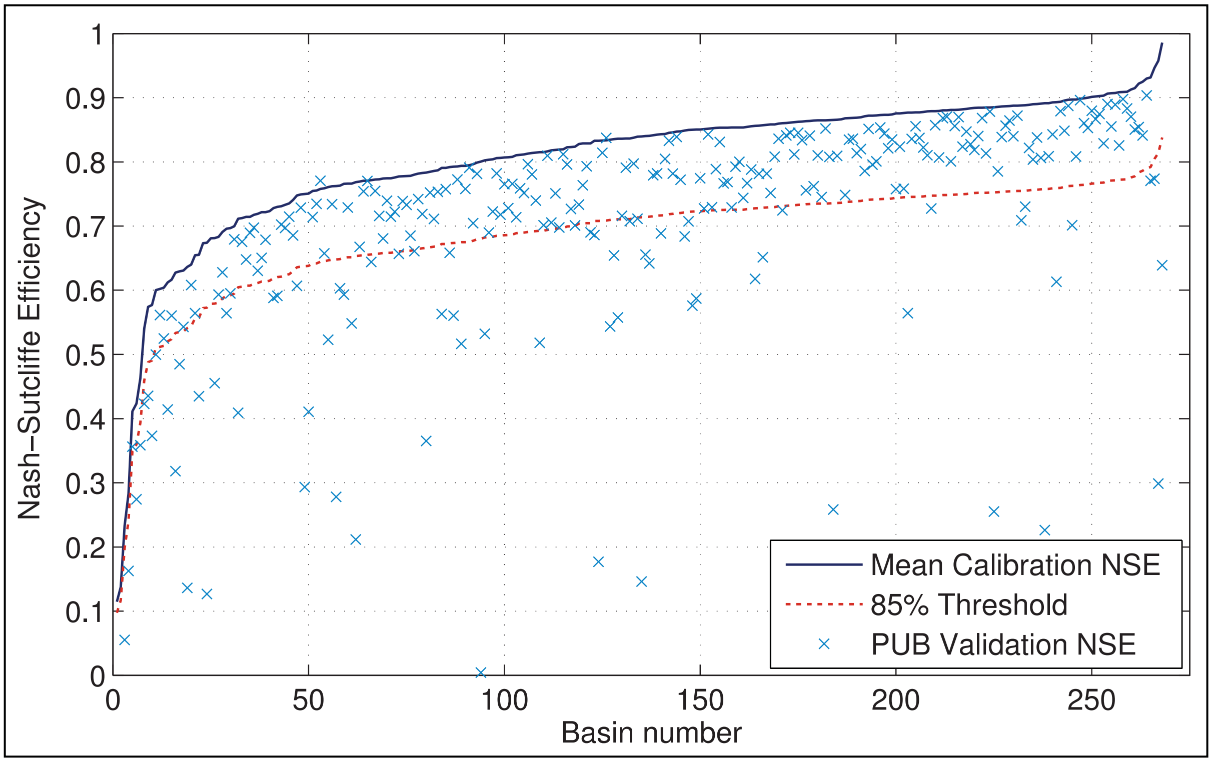

Figure 6.2 Visual representation of the 85% success rate threshold

For some catchments, it is easy to attain a NSE of 0.7 for example, while for others it would be almost impossible as the calibration NSE is lower than 0.7. Instead, the 85% threshold gives more latitude to the higher scoring basins (so a NSE of 0.85 is still considered acceptable even if the calibration NSE is 1) and less latitude to the lower scoring basins. The 85% limit was based on the 0.7 NSE value. If the validation NSE is higher than 0.6 (which corresponds roughly to 85% of 0.7), then the trial is considered a success. It can be seen that increasing the threshold would lead to more rejections, thus reducing the success rate. The effect of changing the threshold on the success rate for a given case is shown in figure 6.3.

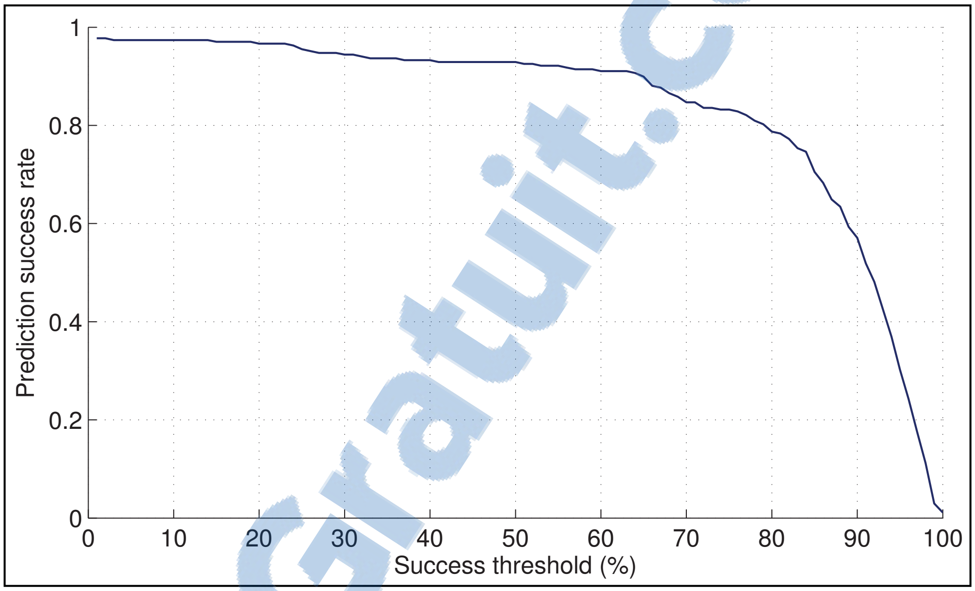

Figure 6.3 Effect of success threshold on the prediction success rate

While most other research papers on the subject use NSE values, its use here would have hidden information which was of critical value. The success rate gives the total number of successful predictions rather than the average value of the various predictions. In the bootstrapping methods, an increase in mean NSE does not automatically translate to using a better method. For example, the mean NSE can increase by improving on the bad catchments and worsening on the good ones, with the outcome being a higher mean NSE but fewer solid predictions. If it were possible to show the distribution of NSE values for all catchments and for all bootstrapping runs, it would have been possible to have this information available, but the success rate aggregates it into one simple and easy to interpret variable.

Third, all the methods were tested and statistically analysed through a bootstrapping method. The donor parameter sets were selected randomly from the 10 possible sets for each basin. The regionalization method's success rate was computed 1000 times per bootstrap run. This allows studying the effect of parameter set selection on the overall performance of the regionalization schemes. A value of 1000 was selected as it allowed a thorough exploration of the decision space while keeping computing costs reasonable. In essence, the regionalization approaches are run 1000 times each, and within each run, the donor parameter sets are selected randomly from the 10 calibrated sets for each basin. This allows exploring the effect of parameter set selection on the regionalization strategies performance. Finally, for the spatial proximity and physical similarity methods, both the arithmetic mean and the inverse distance weighting approaches were used to average multiple donor catchment outputs. For the spatial proximity approach, the physical distance was used for the weighting, whereas for the physical similarity approach, the Similarity Index Euclidian distance was used as the weighting metric.

6.4.4 Multiple linear regression regionalization method

The multiple linear regression regionalization approach was used using all available CDs, as shown in table 6.1. The regression models were constructed using all but one of the available basins (either 267 or 236, depending on if only the good basins were selected), leaving one out as ungauged. The parameters were then estimated at the pseudo-ungauged site, and the hydrological models simulated the streamflow using these parameters. The NSE was finally computed between the observed daily discharge and the daily simulated flow for the pseudo-ungauged basin. The entire process is repeated with a different pseudo-ungauged catchment until all catchments have been considered ungauged.

However, one question remains: Which parameter set should be used for the establishment of the regression models? To show the uncertainty the parameter set selection can induce, the regression models were built using one of the 10 calibration sets at random for each basin. Even with only 10 calibrated parameter sets per catchment, the bootstrapping method will sample 1000 unique combinations to show the uncertainty caused by the parameter set selection in building the regression models.

6.4.5 Physical similarity regionalization method

The physical similarity method was used with different combinations of CDs listed in table 6.1. For each trial, 1000 bootstrap runs were performed to analyze the methods performance with the particular CDs. In all cases, both parameter-averaging and model output-averaging were tested and analyzed. Up to 15 donor catchments were used for each step as experimentation showed that performance drops after 10 or 12 donor basins. Using 15 donors allows finding the "optimal" number of donor catchments to use, as will be discussed further. Some studies (such as *Oudin et al.* [2008]) use the rank of each CD to compute the similarity index. However, since the basins are very heterogeneous, a normalized sum of absolute values is used instead, as seen in equation 6.2. A comparison between an "optimal" selection of CDs and the use of all available CDs for physical similarity regionalization was performed.

6.4.6 Spatial proximity regionalization method

For the spatial proximity approach, the only criterion was to use distance between the pseudo-ungauged catchment and other catchments centroids. The donor catchment parameter sets were then directly transferred to the ungauged site, with the same bootstrapping method and cross-validation used in the previous sections. Furthermore, the parameter-averaging and output-averaging methods were also compared. Up to 15 donor catchments were used to allow finding an optimal number of donor catchments to use.

6.4.7 Regression-augmented approach

The method we call "regression-augmented" is based on the fact that the regression models are often poor, but sometimes a parameter has a relatively high coefficient of determination ($R^2 > 0.5$). The idea is to use a physical similarity or spatial proximity method to find a donor catchment which will transfer most parameters, and replace the parameters which have a high coefficient of determination with those from the regression model. The results should be better than the regression approach used alone, but it is unknown whether it can be better than the spatial proximity or physical similarity used by themselves also. While the similarity and proximity approaches transfer entire parameter sets, the equifinality problem may result in the transferred sets being non-behavioral on the ungauged catchment. Therefore in this method, the parameter set is initially set through the similarity or proximity methods, and is then "upgraded" using the regression strategy. The bootstrap method was used as was the leave-one-out cross-validation method. Only the good basins (NSE>=0.7) were used for this analysis.

6.5 Results

The results are presented in two sections. First, the results for each method are independently shown, and second, a comparison between the methods is made and the effect of parameter set selection is analyzed.

6.5.1 Regression-based approach

The regression-based approach was used for all parameters, even though for most parameters the coefficient of determination is low (below 0.5). In fact, depending on the parameter sets selected during the regression model building, only one to four parameters possessed a R^2 greater than 0.5 when the poor basins (NSE < 0.7) were included. When they were excluded, the number of parameters whose R^2 value was over 0.5 was only marginally higher. This phenomenon has been reported before in various studies [Seibert, 1999; Merz and Blöschl, 2004; Lee et al., 2005] and is expected when there is a lot of parameter interdependence.

The success rates after 1000 bootstrapping runs for the good basins had minimum, median and maximum values of 0.296, 0.377 and 0.444; whereas when all the basins were used, these success rates dropped to 0.291, 0.362 and 0.418 respectively. The median NSE value for all bootstrapping runs was 0.66 when only the good basins were used, whereas the NSE dropped to 0.65 using all basins. The slight reduction in NSE explains the small differences in SR rates as well.

An analysis of the 1000 bootstrapped runs shows that four parameters had R^2 values over 0.5 at least once.

1- Parameter 21 (which controls the surface hydrograph shape) in 100% of the runs;

2- Parameter 19 (subsurface reservoir emptying rate) in 10% of the runs when the good catchments only are used. This number falls to 1% when all catchments are used;

3- Parameters 15 and 20 (Surface runoff fraction and peak time of the unit hydrograph, respectively), in 0.8% of the runs each, and only when the good basins are used.

The other parameters are not correlated highly enough to the catchment descriptors to be adequately estimated, which explains the methods poor performance. Perhaps using other catchment descriptors not available in this study could increase the methods performance.

6.5.2 Physical similarity approach

For the physical similarity approach, different catchment descriptor combinations were used to determine the most influential ones. The descriptors were tested one at a time to find the most advantageous one. Then, a second one was added and the process repeated to find which one was the second best and so on, until the entire list was completed. Ten iterations were conducted for each trial to eliminate any bias caused by selecting poor donor parameter sets.

Table 6.2 lists the order in which the parameters were added to the catchment descriptor vector when all catchments were used. Note that all the basins were used for validation. As will be shown further, approximately 5 donor catchments were found to maximize the mean success rate. Thus, table 6.2 presents the values for 5 donor catchments. It is also important to note that the model output averaging method was used to compile this list given its significantly better performance, as will also be shown further. The p-value in table 6.2 was computed between the Success rates using the previous CD set and the set with the added CD.

Table 6.2 Catchment descriptors by order of importance and the bootstrapping results for success rates and NSE values, using 5 donor basins with model output averaging

Catchment descriptor	NSE med	SR min	SR med	SR max	p-value
Land Cover – Water	0.6759	.4179	0.444	0.4515	-
Latitude	0.7215	0.597	0.6213	0.6343	0
Mean annual precipitation	0.7399	0.6754	0.6828	0.6903	0.000001
Longitude	0.7421	0.6940	0.7090	0.7313	0.00006

Figure 6.4 shows the four variants applied to the physical similarity method while using the four CDs. It presents the Success Rate when all the basins are used and when only the good basins are used, and in each case, using model output averaging and parameter averaging approaches. In all cases, multiple donors were averaged using the arithmetic mean.

It can be seen in figure 6.4 that parameter averaging is worse than model output averaging for the physical similarity approach. In fact, model output averaging is able to increase performance by adding donor catchments, whereas the opposite is true for parameter averaging. Additionally, looking at the spread of success rate values, it can be seen that the uncertainty due to parameter set selection is higher for parameter averaging than for model output averaging. The uncertainty decreases when adding donor basins for the model output averaging, while there is no dependence in the case of parameter averaging. The median and mean NSE values were computed for the 1-8 donor tests and are presented in table 6.3. A Kruskal-wallis [*Kruskal and Wallis*, 1952] test shows that 5-10 donor basins is the best when model output averaging is used, and is statistically significantly better than (4 or less) and (11 or more) donor basins.

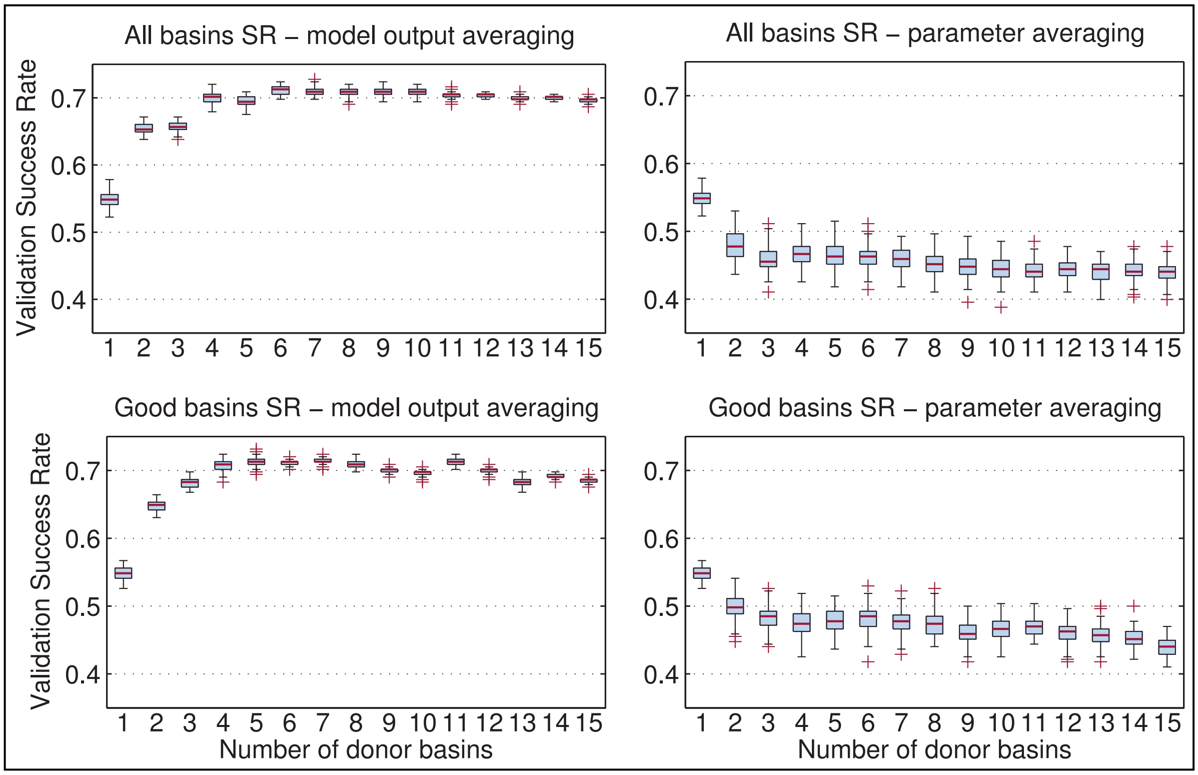

Figure 6.4 Bootstrapped success rates for four variants of the physical similarity regionalization scheme using 4 catchment descriptors. The top and bottom figures are respectively for all basins and only good basins. The left and right figures are respectively for model output averaging and parameter averaging. The box and whisker plots show the 25th and 75th percentiles (box edges) and the line in the center of the boxes represents the median value. The top and bottom whiskers represent the most extreme value within ±2.7 standard deviations. Outliers are plotted individually when they are outside of these values

A Wilcoxon rank-sum test [*Wilcoxon*, 1945] was used to compare the success rate while using all the basins to the success rate while using only the good basins. The test was performed 15 times, once for every number of donor catchments. The results indicate that the removal of bad basins is advantageous when 3 to 5 donors are used. When 6 or more donors are used, there is no statistical difference between both groups.

The entire process was repeated while using all available CDs, and the results are shown in figure 6.5. In all cases, multiple donors were averaged using the arithmetic mean.

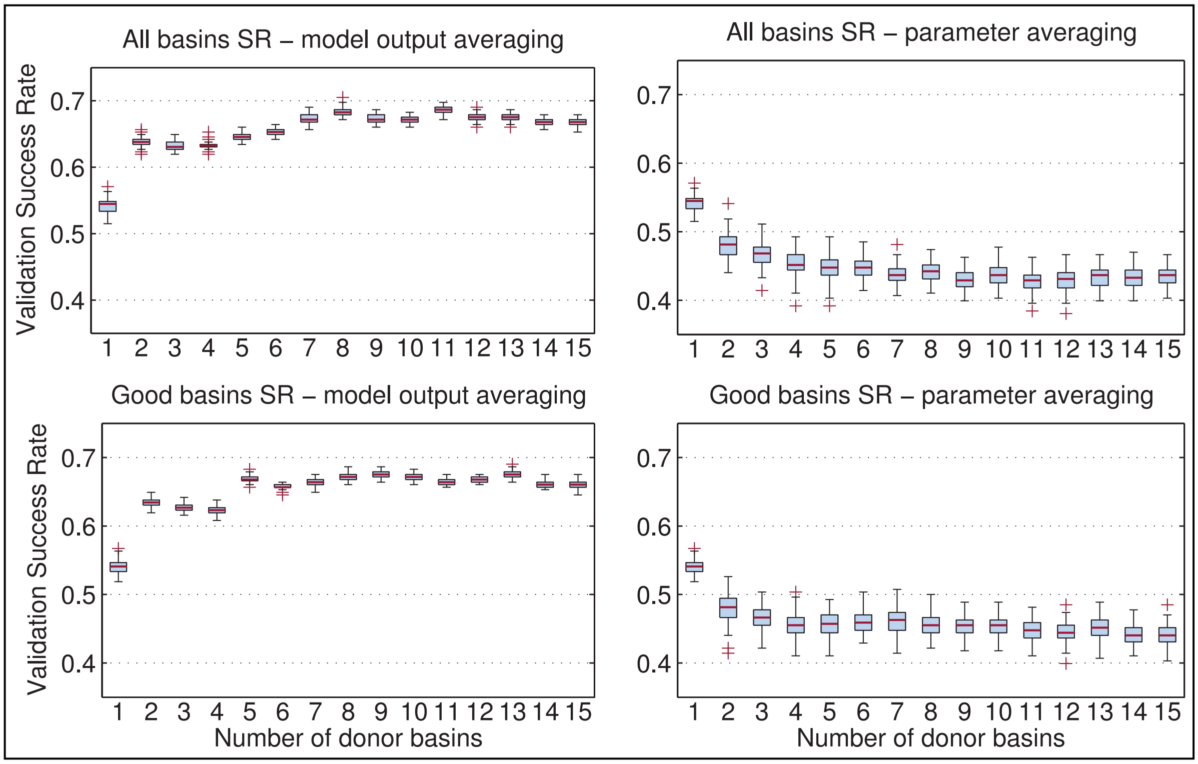

Figure 6.5 Bootstrapped success rates for four variants of the physical similarity regionalization scheme using all available catchment descriptors. The top and bottom figures are respectively for all basins and only good basins. The left and right figures are respectively for model output averaging and parameter averaging

The same general pattern can be seen as in figure 6.4, but it is clear that the Success Rates are lower using all CDs. When comparing both best case scenarios (good basins only with model output averaging), the Wilcoxon rank-sum test indicates that for every number of donor basins, the four-CD version is significantly better than the version with all CDs.

6.5.3 Spatial proximity

The spatial proximity method was performed similarly to the physical similarity method, except that the physical centroid-to-centroid distance was used instead of the similarity distance. Figure 6.6 shows the Success Rate for the model output averaging and parameter averaging strategies when all the basins are used and when only the good basins are used. In all cases, multiple donors were averaged using the arithmetic mean.

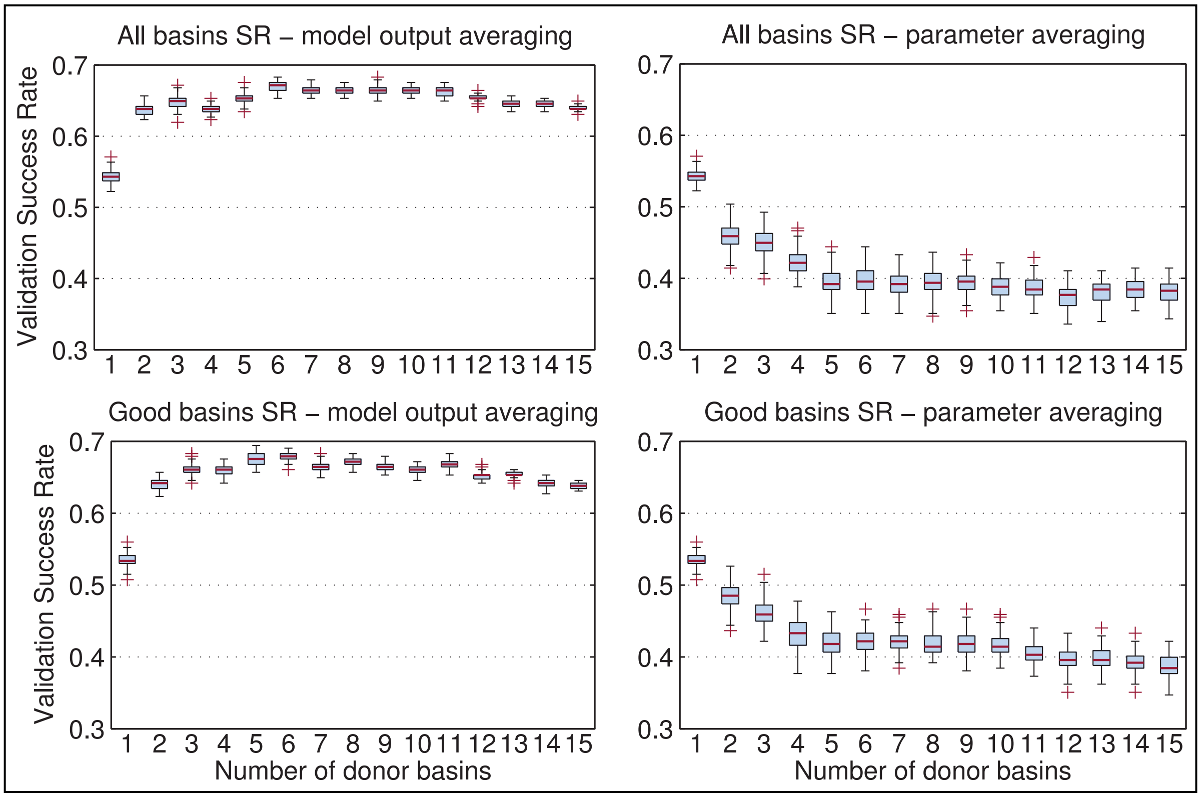

Figure 6.6 Bootstrapped success rates for four variants of the spatial proximity regionalization scheme. The top and bottom figures are respectively for all basins and only good basins. The left and right figures are respectively for model output averaging and parameter averaging

The analysis of figure 6.6 is essentially the same as for figures 6.4 and 6.5: Model output averaging is significantly better and produces much less uncertainty than parameter averaging. This is unavoidable since the information contained in an entire parameter set is lost when the parameters are averaged. Furthermore, the equifinality in between parameter sets guarantees that at least some parameters will differ greatly between donors and will reduce the overall performance in prediction mode. Statistics on the NSE results for the spatial proximity method are presented in table 6.3.

A Kruskal-Wallis test shows that 5-8 donor basins is the best when model output averaging is used, and is statistically significantly better than (4 or less) and (9 or more) donor basins. This could be due to the fact that at a small distance, the basin similarity is close enough the warrant a successful parameter transfer. As distance is increased, the similarity diminishes, and the performance drops. Then, a Wilcoxon rank-sum test was used to compare the success rate while using all the basins to the success rate while using only the good basins. The test was performed 15 times, once for every number of donor catchments. The results indicate that the removal of bad basins is advantageous when 2 to 6 donors are used. When 7 or more donors are used, there is no statistical difference between both groups.

6.5.4 Inverse distance weighting

The inverse weighting of the donor catchments for the spatial proximity and physical similarity methods was used for the model output averaging approach using only good catchments. Figure 6.7 shows the success rates for the spatial proximity and physical similarity methods, both using simple arithmetic mean of donor model outputs and inverse distance weighting (IDW) of the model outputs.

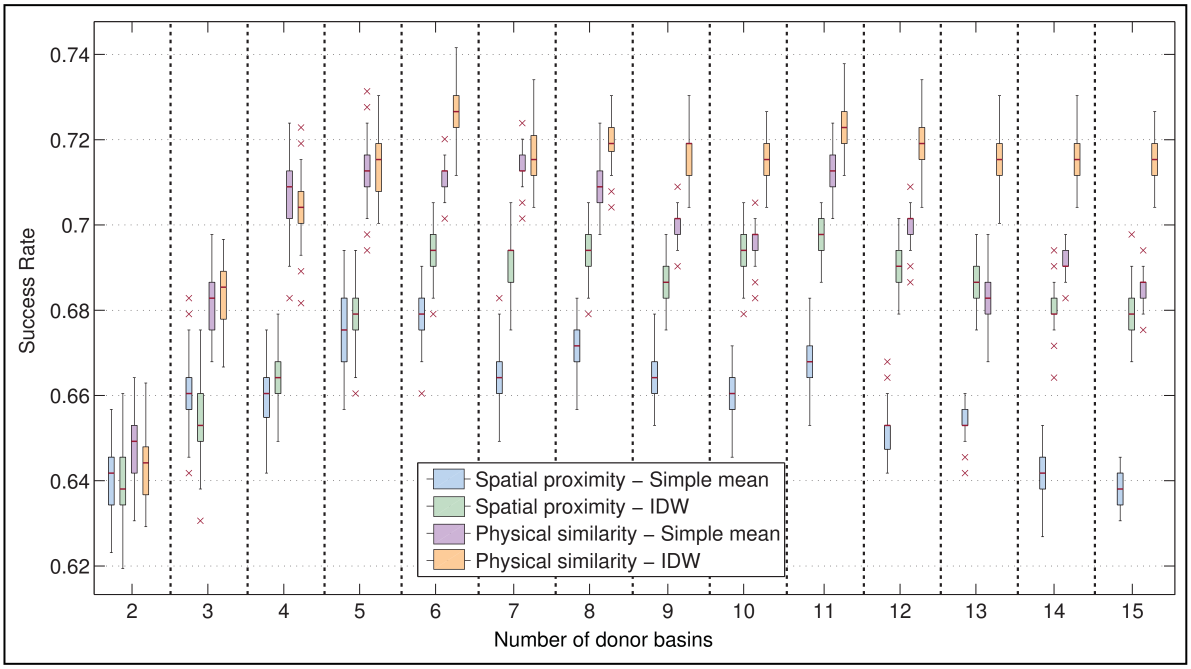

Figure 6.7 Bootstrapped success rates for the similarity and proximity methods using simple mean and inverse distance weighting (IDW) averaging of multiple donor catchment model outputs. Only good catchments are used

In all cases, physical similarity is still the best option for regionalization in this study. However, IDW approaches increase performance significantly for both methods, especially when a sizeable number of donors is used. The weighting scheme minimizes the farthest donors so their negative impact is reduced. This means the IDW approach allows good results even if too many donors are used compared to the simple mean method. The comparison between the standard and IDW variants in terms of NSE values is presented in table 6.3.

6.5.5 Regression-augmented

The first regression-augmented approach used the spatial proximity strategy as a starting point. Since model output averaging was significantly better than parameter averaging for all methods to date, it was the only variant tested for both regression-augmented tests. Furthermore, the IDW donor averaging was used as previous results showed its performance was better. Figure 6.8 shows a comparison between the best spatial proximity case (good catchments only with model output averaging, IDW donor averaging) as well as the regression-augmented spatial proximity approach with the same conditions. The same two cases are also seen with the physical similarity method. In each case, the results are shown for when 2-15 donor catchments are used. The single donor case was omitted since the results are poorer as seen in figure 6.5.

It can be seen that for 2-10 donor catchments, the regression-augmented approach increases the predictive skill on the ungauged basins, especially for the physical similarity approach. The NSE values presented in table 6.3 also show the same trend, although they are not as evident as seen in figure 6.8. The Wilcoxon tests show that for 3-10 donors, the regression-based approach is significantly better than the standalone spatial proximity method. The same holds true for the physical similarity method for 2-11 donors.

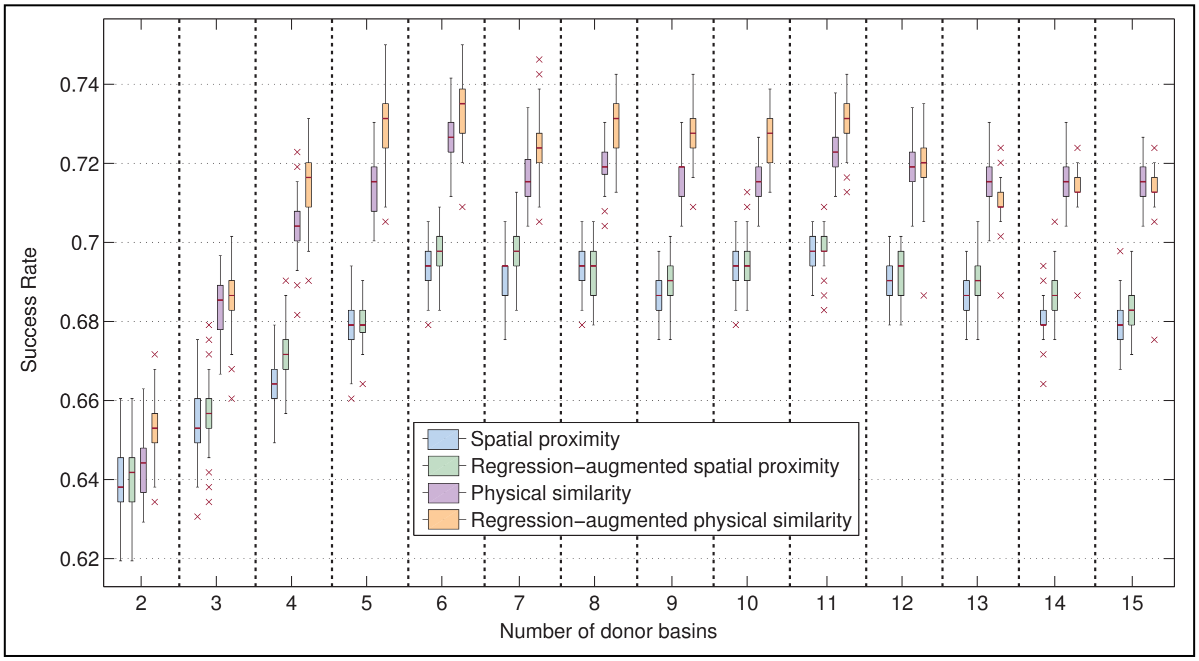

Figure 6.8 Standalone similarity and proximity approaches vs. their regression–augmented counterparts for 2-15 donor catchments

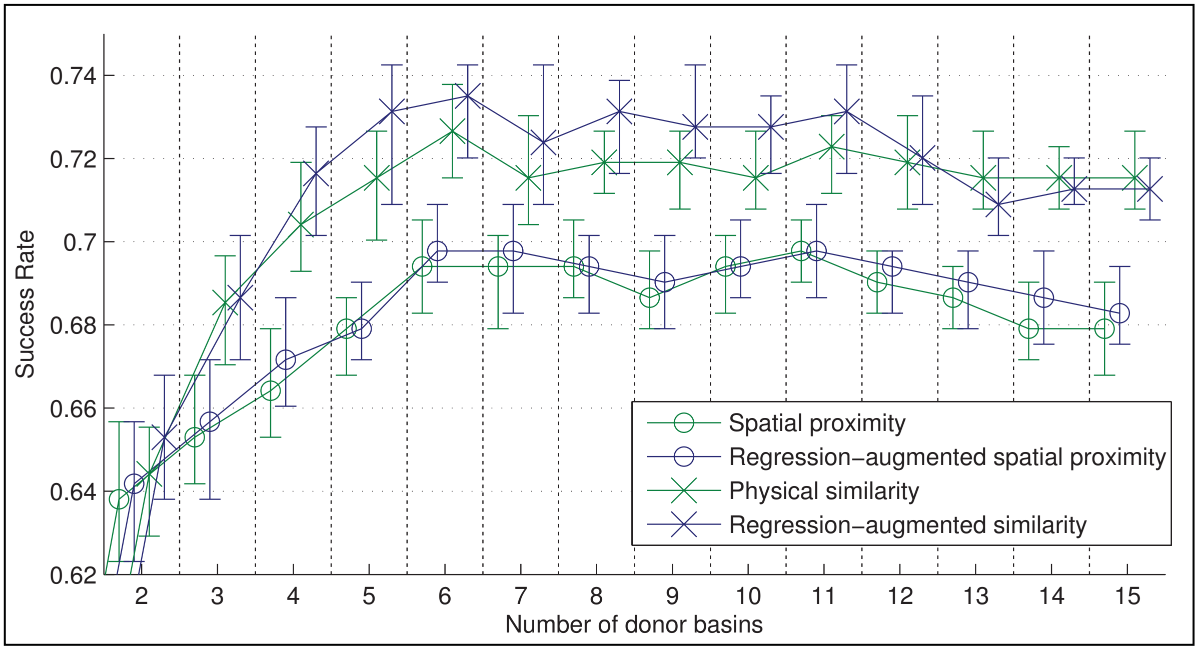

Figure 6.9 95% Confidence interval on Success Rates for Standalone similarity and proximity approaches vs. their regression–augmented counterparts for 2-15 donor catchments

However, one very apparent disadvantage of this method is that the uncertainty is much larger with the regression-augmented approach and some of the results are noticeably worse than the standalone option. This can be seen in figure 6.9, which shows the 95% confidence interval using the BCa method for 2-15 donor basins.

Of all the methods analyzed in this study, the regression-augmented physical similarity strategy using IDW for donor averaging is the one that procures the best results, especially when using 5-8 donors. In this case, it is expected that there should be more uncertainty as the regression model modifies the parameter set and breaks its unity, thus leaving it less coherent. The 5% confidence intervals (2.5% at each tail) are larger for the similarity methods, and the regression-augmented approaches generally have higher uncertainty than their traditional counterparts.

The hypothesis behind the fact that the regression augmented versions of the algorithms outperform their classical counterparts is that a parameter which is strongly correlated to catchment descriptors must increase the performance in two ways. First, the HSAMI model is known to be robust to different parameter sets. For example, taking a calibrated parameter set from a catchment to simulate streamflow on another basin will not produce terrible streamflows in HSAMI. The performance will drop, evidently, but a parameter set transposition will ususally be better than an NSE of 0. The parameters modified by the regression algorithm, then, must make the simulation better. Also, the parameter which is the most often found as being strongly correlated to the catchment descriptors is related to the surface unit hydrograph shape. This strongly influences the peak floods and their timing, therefore a better approximation of its value could generate better results. This can be seen by comparing the relative differences between the proximity methods and the similarity methods. For the proximity methods, the regression-augmented version increases the performance marginally. For the similarity method, which is based on catchment similarity and descriptors, the gain is more substantial. It is therefore expected that the added regression-based parameter would increase the similarity method's performance even further.

6.5.6 Inter-method comparison

In this section, the three classical methods are compared using their most advantageous setups. This includes using model output averaging for the physical similarity and spatial proximity methods, as well as using only the good catchments as donors.

As was shown in section 6.5.1, the regression method cannot compete satisfactorily with the other two methods. This can be attributed to the very low coefficients of determination throughout the analysis. As seen in figures 6.4 and 6.6, using a single donor basin for the physical similarity and spatial proximity methods is not worthwhile, as the success rate is approximately 20% lower than when a second donor is added.

It is clear from figures 6.7 to 6.9 that the physical similarity regionalization strategy is better than the spatial proximity as it outperforms it for any number of donors. Therefore, the physical similarity method is recommended if it is possible to obtain CDs for the series of catchments used for regionalization. Furthermore, the regression-augmented version of the physical similarity can outperform its standalone counterpart at the expense of more uncertainty, as shown in figures 6.8 and 6.9. For applications in real ungauged basins, depending on the allotted time and resources available, it could be possible to perform multiple regionalization runs using the regression-augmented physical similarity approach with multiple calibrated parameter sets for each donor basin. Then, to mitigate the added uncertainty of the method, it could be possible to select a median hydrograph amongst the model output averaged hydrographs. This would eliminate the outliers, further increasing the probability of a successful prediction.

6.5.7 Success Rate vs. Nash-Sutcliffe Efficiency

As most regionalization studies use the NSE metric to compare results, the NSE statistics obtained in this study are shown in table 6.3. Results are shown for 1-8 donors as the results do not vary much past this point.

Table 6.3 NSE statistics for the median and mean NSE values in cross-validation for 1-8 donors

	Median NSE (Mean NSE)							
Donor basins	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**
Calibration (control)	0.840 (0.804)	---	---	---	---	---	---	---
Regression(1)	0.662 (0.631)	---	---	---	---	---	---	---
Parameter averaging – All basins								
Similarity	0.693 (0.579)	0.686 (0.617)	0.684 (0.620)	0.691 (0.622)	0.698 (0.626)	0.696 (0.625)	0.698 (0.622)	0.684 (0.623)
Proximity	0.699 (0.576)	0.702 (0.605)	0.686 (0.601)	0.671 (0.595)	0.677 (0.588)	0.675 (0.592)	0.673 (0.585)	0.676 (0.584)
Model output averaging – All basins								
Similarity	0.693 (0.579)	0.727 (0.645)	0.729 (0.658)	0.735 (0.662)	0.737 (0.670)	0.742 (0.671)	0.748 (0.668)	0.746 (0.669)
Proximity	0.699 (0.576)	0.729 (0.635)	0.731 (0.635)	0.731 (0.635)	0.730 (0.632)	0.738 (0.639)	0.734 (0.634)	0.736 (0.634)
Parameter averaging – Only good donors								
Similarity	0.693 (0.579)	0.686 (0.617)	0.684 (0.620)	0.691 (0.622)	0.698 (0.626)	0.696 (0.625)	0.698 (0.622)	0.684 (0.623)
Proximity	0.699 (0.571)	0.702 (0.617)	0.692 (0.600)	0.685 (0.595)	0.688 (0.590)	0.684 (0.590)	0.678 (0.587)	0.683 (0.588)
Model output averaging – Only good donors								
Similarity	0.695 (0.573)	0.732 (0.650)	0.736 (0.662)	0.741 (0.663)	0.746 (0.671)	0.748 (0.667)	0.747 (0.665)	0.748 (0.664)
Proximity	0.699 (0.571)	0.734 (0.645)	0.733 (0.632)	0.735 (0.635)	0.736 (0.632)	0.739 (0.636)	0.741 (0.634)	0.743 (0.636)
Similarity IDW	0.692 (0.574)	0.733 (0.647)	0.740 (0.665)	**0.744** **(0.668)**	**0.747** **(0.674)**	**0.749** **(0.673)**	**0.750** **(0.672)**	0.751 (0.672)
Proximity IDW	**0.702** **(0.570)**	0.734 (0.641)	0.732 (0.635)	0.738 (0.640)	0.745 (0.639)	0.744 (0.644)	0.747 (0.644)	0.749 (0.646)
RA-Simil. IDW	0.699 (0.575)	**0.736** **(0.647)**	**0.742** **(0.664)**	**0.745** **(0.667)**	**0.747** **(0.674)**	**0.750** **(0.672)**	0.749 (0.672)	**0.753** **(0.672)**
RA-Proxim. IDW	**0.702** **(0.573)**	0.731 (0.643)	0.734 (0.636)	0.742 (0.643)	0.747 (0.641)	0.746 (0.646)	0.745 (0.646)	0.748 (0.647)

(1) The regression method results presented here were computed using the good basins only to define the regression models. By definition, there are no donor basins for the regression method as they all contribute to the regression model.

Since the NSE metric depends on the selected parameter sets during the regionalization and leave-one-out cross-validation, the statistics are computed on the median NSE value of the 1000 runs. The 268 median NSE values were then analysed for the selected number of donors. Table 6.3 shows the median and mean NSE values resulting from this analysis. Note that negative NSE values were set to 0 in this table to remove large negatives which would weigh too much in the mean calculation.

It can be seen that the difference in NSE values between the various regionalization approaches is smaller than the difference in SR values presented previously. This indicates that the distribution of the NSE values is different from one method to another since one can perform well on more catchments (higher SR) while the other method might have a lower SR but higher NSE values on the successfully simulated basins. It can also be seen that the highest scoring methods (in bold font) are still the IDW similarity methods, with the regression-augmented variant being marginally better than its standard counterpart.

6.5.8 Hydrograph analysis

The physical similarity and spatial proximity regionalization methods were analyzed to get a better understanding of how their simulated hydrographs compare to one another. Figures 6.10 (physical similarity with one donor), 6.11 (physical similarity with 5 donors) and 6.12 (spatial proximity with 5 donors) show the hydrographs for the Outardes river basin for the year 1981 under different conditions.

The Outardes river basin is located in central Québec, Canada, and has an area of 17119 km^2. It is used for hydropower generation and it drains into the St-Lawrence River. In figures 6.10-6.12, only one basin is shown (and for a single year) but the same conclusions are found with the other basins / years. Figures 6.10-6.12 show the hydrographs using 1000 bootstraps and 1 to 5 donor basins.

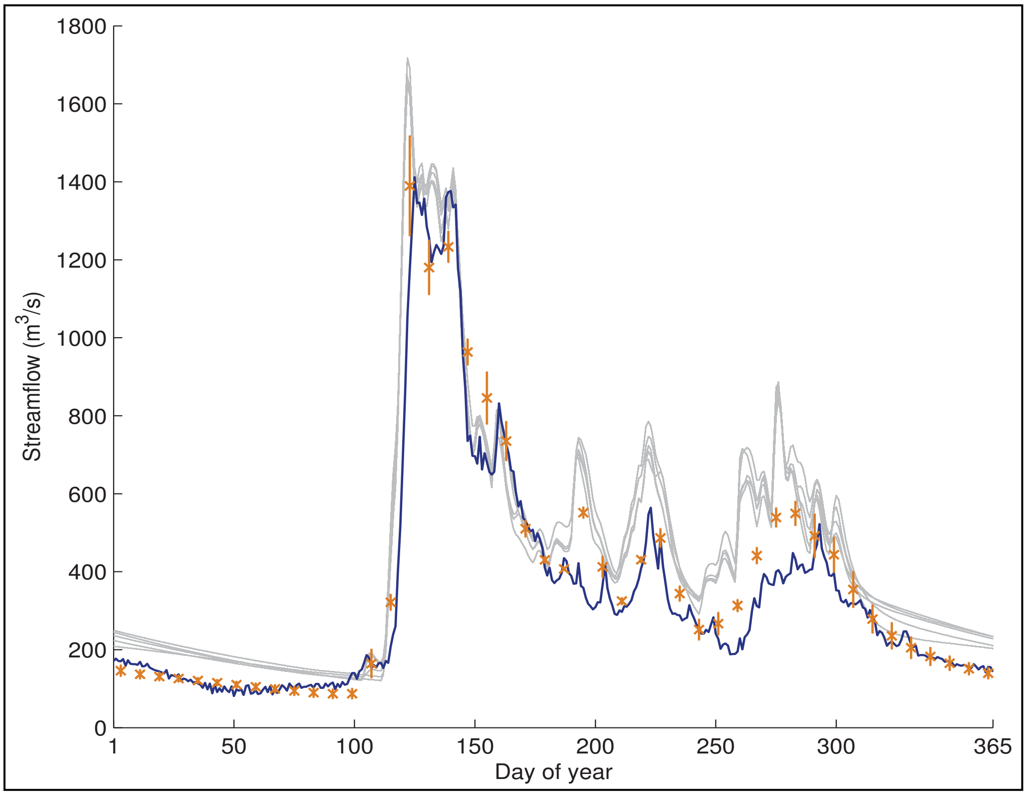

Figure 6.10 Observed (blue) and simulated (gray) hydrographs for the year 1981 using the physical similarity method and one donor basin. The gray lines represent the regionalization outcomes from the multiple bootstrapped runs and the error bars represent the variability in the hydrographs when the model is run using the 10 calibrated parameter sets on the calibration period

From figures 6.10-6.12, it is clear that using a single donor increases the spread and error in the predictions. It can also be seen that the spatial proximity method produces hydrographs that are slightly less precise than the physical similarity method, especially for the peak flood values. For low-flows, both methods are equally consistent in approximating observed discharge values. However, in the single donor case the error is much larger. This could indicate that the number of donors is more important than the selected regionalization method, as seen in figures 6.4-6.8.

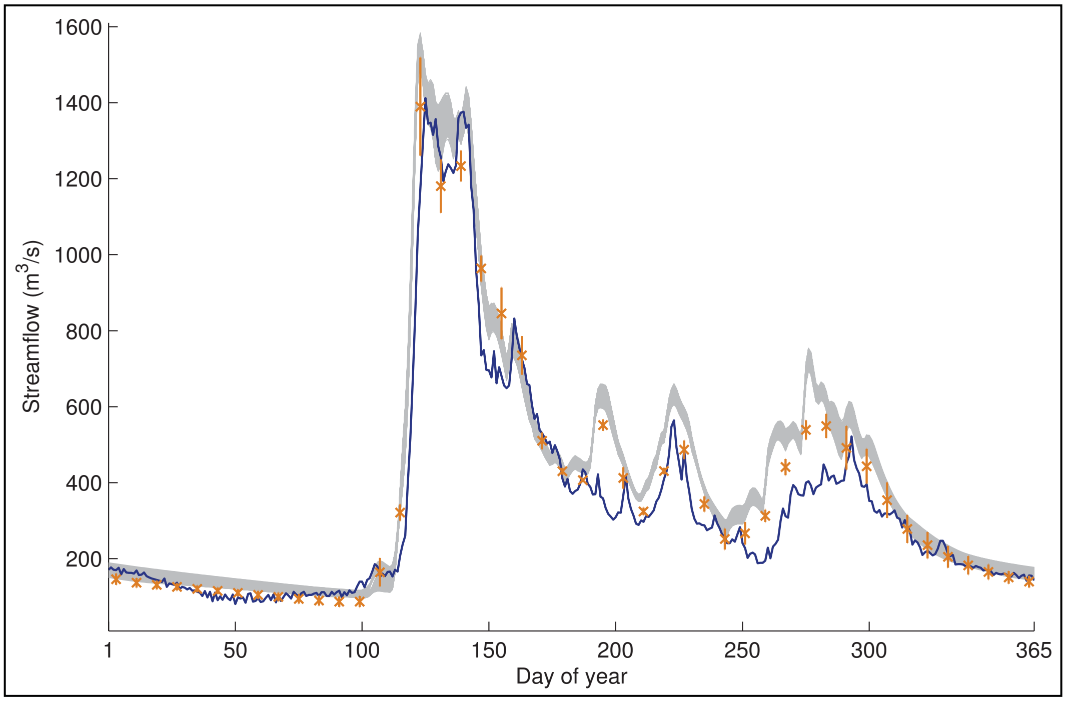

Figure 6.11 Observed (blue) and simulated (gray) hydrographs for the year 1981 using the physical similarity method and five donor basins

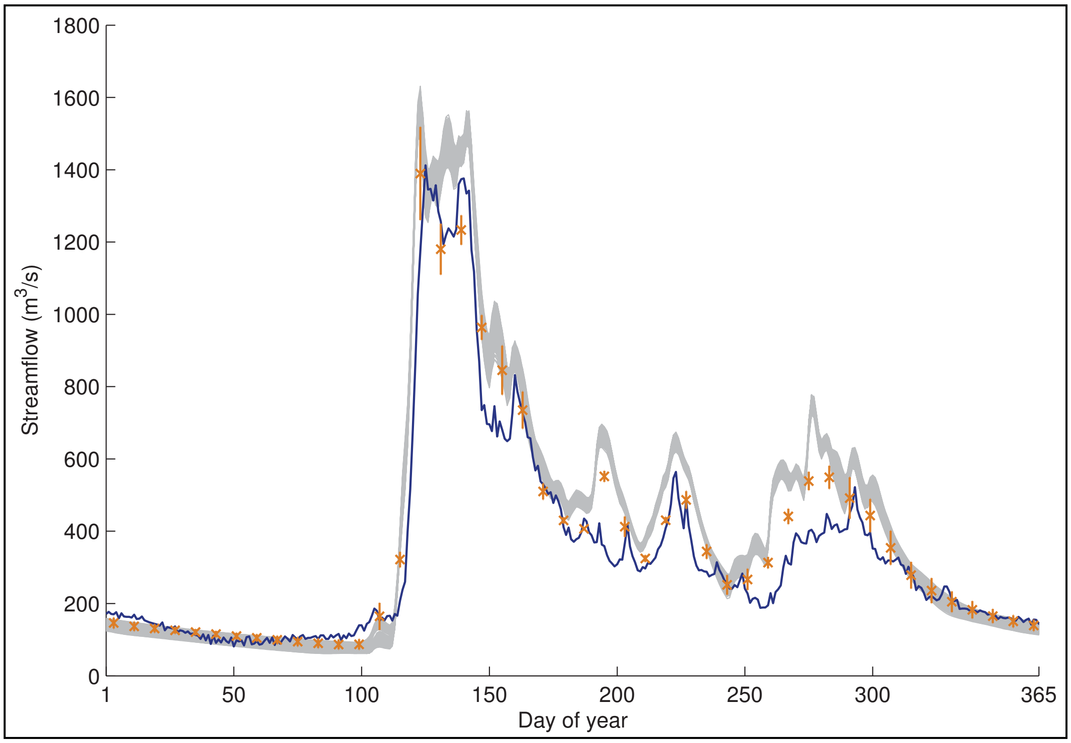

Figure 6.12 Observed (blue) and simulated (gray) hydrographs for the year 1981 using the spatial proximity method and five donor basins

6.6 Discussion

6.6.1 Number of donor catchments

The bell-shaped performance curve in figures 6.4 to 6.9 could be explained by the fact that there were only 31 bad catchments on a total of 268 available. When many donors are considered, the poor basins' effect is diluted enough to not influence the end result, especially in the case of IDW. It is probable that if there had been a higher poor/total ratio, the results would indicate that the success rate would be higher for the good basins versus all the basins. After the optimal number of donors, the curve descends principally because the added basins are either too far (thus less homogeneous) or too dissimilar. In later steps, it would be interesting to purposely alter some of the good basins datasets to make them artificially poor. By altering the good/poor ratio, it could be possible to detect the effect of the ratio on the optimal number of donor catchments for each method.

Next, it was shown that in all cases, success rates were higher when only the good basins were used to build the regionalization models. Perhaps this is due to bad climate or hydrometric data contaminating the bad catchments calibration parameters. When bad basins are used, the parameter set may be completely unrelated to the basin attributes, thus cascading the errors in the regionalization scheme. Although bad data is the primary source of poor performance, another source of error could be the calibration algorithm itself, but the use of 10 parameter sets mitigates this somewhat. It is thus recommended to remove all bad basins for the model building, but keeping them for the cross-validation to verify the predictive ability of the regionalization model.

In this study, there are 122 nested basins (nested in 74 larger basins). We looked at the regionalization methods performances when a single donor was used (its parent). Overall, the results are surprising, in that they offer slightly worse scores than when the most similar basin is used. This is opposite of previous findings by *Parajka et al*. [2005]. In their case, the performance was slightly better using the nested basins, although they conclude that the regionalization methods' performance is not all attributable to the fact that the basins are nested. They also state that there must be trans-boundary properties that can allow good regionalization on ungauged basins. In this study, the basin fraction that is water is probably the culprit as the other catchment descriptors should be similar (latitude, longitude and mean annual precipitation). The results show that the spatial proximity method is not affected by it being nested or not. When more donors are used, the errors seem to cancel out, which stresses the need to use more than a single donor. In this regard, nested basins should not necessarily be viewed as better sources of information than standalone basins.

This raises an interesting point. Oftentimes, distributed models are used to estimate flows at ungauged sites within a gauged catchment. This should be equivalent to using the same parameter set for the nested basin than the larger one. Perhaps it should be undertaken to verify that this approach is actually as efficient as using multiple donor basins and averaging the flows at the nested ungauged site.

6.6.2 Regionalization methods analysis

For the regression based approach, the fact that only 1-3 parameters at any time had coefficients of determination greater than 0.5 shows that the parameters are for the most part uncorrelated to the catchment characteristics. Consequently, most parameters are estimated with very poor confidence and are little better than random. It should therefore be expected that in the case of the HSAMI model, with its high dimensionality and parameter interdependence, that the multiple regression method was not as successful as the other methods. Nonetheless, it was shown to be beneficial to add the moderately correlated parameters to the donor catchment parameter sets from the two other approaches. This indicates that for a model with less interdependence, the regression-based approach would fare better. It would be possible to use various versions of the same model with a decreasing number of parameters to test this hypothesis. However, the reduction of the parameter space while maintaining performance is not an easy task.

Another problem which was found throughout this work is that the data availability periods were different for each catchment. Some had a full 30 years of data, while others had only one or two years. The majority had between 10 and 20 years. This is an important aspect especially for the spatial proximity approach. With this regionalization strategy, it could be possible that a donor basin was calibrated on a short, wet period while the pseudo-ungauged basin's dataset indicates a dry period. Perhaps the transferred parameter sets would have been adequate given similar climate data, but in this theoretical case, they would probably underestimate the methods performance. It would be a certain advantage to have long and contiguous time series for calibration of donor catchments to offset the possibility of disparate climates affecting the spatial proximity regionalization schemes outcome.

A key point for the physical similarity method was the selection of catchment descriptors to use to determine the similarity index. The results of the CD selection experiment shown in table 6.2 illustrate that only 4 CDs are necessary to optimize the regionalization method's performance. The next CD (Land Cover - Grass) in the list was a special case, as when it was used independently, it scored worse as a CD than if the donor catchments were chosen randomly. When added to the current vector, the regionalization approach performance dropped significantly (p-value = 0.003). As additional CDs were added, the performance incrementally dropped by small values. The physical similarity method was thus implemented with only the four first CDs: Fraction of land cover that is water, latitude, mean annual precipitation and longitude. Since the latitude and longitude define the spatial coordinates of the basins, these CDs hybridize the physical similarity strategy with the spatial proximity method. This result confirms the conclusions brought upon by *Oudin et al.* [2008] and *Samuel et al*. [2011], which show that combining physical similarity with spatial proximity methods provides better results.

The linear approach used does not take into account combinations of higher order. For example, it could be argued that basin area should be an important catchment descriptor for predicting streamflow, but in the present study it is not used as it is detrimental to the methods performance. However, it could be possible to increase the regionalization performance by adding two or more descriptors which, when taken alone, have a negative impact. The first reason this strategy was chosen is because of time constraints. It could be possible to cycle through the possible combinations and perhaps find some that further increase the approach's performance. However, we withheld from doing so as the added benefit was marginal. Already, when using all available CDs the physical similarity method outperformed the others. When the four CDs were used as per the method described in this paper, the performance increased again. It is quite possible that the optimal CD set was not used, but it is unlikely that it would change the study's conclusions. It could also be helpful, to further increase the method's performance and robustness, to use more catchment descriptors (such as drainage density, soil properties if available, and more climate indicators). In any case, further research into the CD selection would probably yield better success rates and allow for better real world applications. It would also be pertinent to pursue research efforts in analyzing and ameliorating the similarity index itself. The current approach could be biased by outlier values since no weighting is applied to important parameters. Care has been taken to avoid such mishaps, but a more robust methodology could negate these shortcomings.

Furthermore, the fact that latitude and longitude are selected in the top performing group of CDs reveals that spatial proximity is important in regionalizing model parameters. It can be argued that latitude is related to some parameters such as snowmelt and evapotranspiration, and its presence in the list is not odd. Longitude, however, should not be correlated directly to any parameters and stands out in the best CD list. The most probable explanation is that longitude is acting as a proxy for catchment attributes not directly used in this study, such as soil properties (type of soil, drainage capacity, depth) which are not available in the region of interest. It is expected that the nearest basins will have similar soil properties and therefore be hydrologically similar. This observation partly explains why a combination of physical similarity and spatial proximity has been shown to perform well in the past. If all the appropriate catchment descriptors were available, then the spatial advantage would probably be less important.

A post-hoc analysis was performed to try and correlate methods success rates with basin attributes. However, no correlation greater than 0.3 was found, and most were below 0.2.

This low score shows that it is impossible, with the catchment descriptors at hand, to select a regionalization strategy as a function of basin attributes. If the success rate or absolute NSE value were highly correlated to at least one of the catchment descriptors, it would have been possible to suggest a strategy or another to increase the odds of a successful prediction on the ungauged catchment. Unfortunately, this work shows no sign of this being the case.

6.6.3 Comparison with other studies

To better understand where the results lie in the midst of the wealth of literature, a comparison was made with the comparative assessment of predictions in ungauged basins by *Parajka et al.* [2013], in which regionalization methods performances are analyzed through various means.

First, the cold climate in the northern parts of Quebec and the cold/humid climate in the south increase the odds of seeing good performance, as does the large dataset. This is consistent with values found in the literature. Our results show median NSE values ranging from 0.73 for the regression method to 0.75 for the physical similarity methods, which places high for cold climates and average-high for humid climates. The aridity index shows that the climate is humid since it is inferior to 1 across all basins. Also, our results are on par as to the relative performances of the different methods. In studies where multiple methods are compared, parameter regression consistently scores lower, while similarity is equivalent or better than spatial proximity. In the case of the proposed regression-augmented approaches, they outperform their classical counterparts for mean NSE and SR values and should be taken into account in future comparative studies.

In this study, similarity approaches perform better than spatial proximity methods. However, this is usually not the norm for datasets with a low density of gauging stations such as in Quebec. For example, one study used 320 catchments in Austria with good results using the similarity approach. However the median basin area was 196 km^2, whereas it is of 2532 km^2 in this study. Our largest basin has an area equivalent to 82% of that of Austria. Therefore it is expected that the gauging network is much less dense. Nonetheless, the similarity approach used performs well, although it must be restated that the catchment descriptors using latitude and longitude make it a hybrid between traditional spatial and proximity methods.

In a comparable study, *Samuel et al.* [2011] show that on 94 basins in Ontario, Canada, the physical similarity method combined with the IDW approach (which was used in this study) was the best. Close behind were the spatial proximity methods, and last was the regression method. The results are similar, possibly due to the fact that the catchments share similar geophysical and hydrometeorological characteristics.

6.6.4 Parameter set selection uncertainty

As was seen in figures 6.4 to 6.9, parameter set uncertainty is a factor and does influence the regionalization strategies performance. In some cases it is quite important, such as in the regression-based approaches and in all cases where parameter averaging of donor sets is used. This should come as no surprise since the hydrological model used in this study has many parameters which are interdependent, and is subject to the effects of equifinality. Thus the methods that emphasize the individual parameters instead of the model output (or complete parameter sets) are prime candidates for generating uncertainty. Their values are uncertain to begin with (due to equifinality), but they are then further denatured by either averaging them (which eliminates any unity the parameter set might of had) or estimating them using a poor regression model. As was shown, only 1-3 parameters had a coefficient of determination that was higher than 0.5 for any given run. The highest R^2 value recorded during all the tests in this study (including the bootstrapping) was 0.71 for a single parameter. It is effectively normal to witness the kinds of uncertainty observed in this work using parameter based methods under these circumstances.

On the other hand, model output methods see much less uncertainty. As the parameter sets are kept intact, they generate hydrographs that are at least in some way coherent. The averaging of these hydrographs is much less damaging than the averaging of parameters, and is in fact a way of mitigating possible errors. It is the premise used in such disciplines as ensemble streamflow prediction. Therefore, the little uncertainty that is seen for the model output averaging stems from the equifinality problem. Had all 10 independent parameter sets been equal for each basin during the initial calibration step of this study (completely eliminating equifinality), the uncertainty at this level would have been nil. In light of our results, it is safe to say that model output averaging techniques should be used during regionalization studies as they contribute the least amount of uncertainty on the overall results. This information can be interpreted as a positive sign that hydrological models with many parameters can be used for regionalization project since the model output is not dependent on the individual parameters themselves but on the entire parameter set.

6.6.5 Type I errors in hypothesis testing

In most of the method comparisons, multiple Wilcoxon rank-sum tests were performed to reject or keep the null hypothesis that the two compared methods were identical, i.e. to assert statistical significance. One problem associated with the use of multiple tests is the possibility of a type I error, which is when the null hypothesis is true but is rejected. In this study, it is quite possible, given the number of tests, that one or more type I errors have been committed without our knowledge. However, the conclusions would not differ very much as the Wilcoxon tests showed similar patterns of statistical significance throughout the paper (for the number of donor catchments to use for example). In this regard, the type I errors can be neglected.

6.7 Conclusions

This paper provides a new analysis of the three most common parameter regionalization schemes for the HSAMI hydrologic model on 268 basins in the province of Québec, Canada. The analysis reveals that for this region and this dataset, the physical similarity approach has the highest success rate, followed closely by the spatial proximity method. When possible, multiple donors should be used and their respective outputs averaged according to an inverse distance weighting scheme. It is shown that the multiple linear regression approach is the worst as the parameters are in most part not correlated to the catchment attributes. However, the use of a regression-augmented physical similarity approach improved the results significantly, at the expense of added uncertainty.

It was also shown that parameter set selection plays a small role in total uncertainty when using model output averaging, while the uncertainty jumps when using parameter averaging of multiple donor parameter sets. The bootstrapping method allowed to quantify the uncertainty associated with equifinality and with the effort required to produce good parameter sets.

Finally, it was shown that the selection of catchment descriptors for the physical similarity method is important and must not be taken lightly, as different combinations can drastically increase or decrease the regionalization schemes performance. In this study, it was shown that four CDs were optimal, however more research in this area is still required.

Future work should focus on same-period datasets to eliminate invalid donor sets, on CD selection methods, on parameter reduction strategies for regression-based approaches and on the bad basin ratio effect on the number of donor catchments.

6.8 Acknowledgments

The authors would like to thank Catherine Guay from IREQ for making available the CQ2 database, and Kenjy Demeester from École de technologie supérieure for providing the land cover data for the project. We would also like to thank Dr. Juraj Parajka and the anonymous reviewers for their constructive comments which contributed to improving the overall quality of this article.

6.9 References

Arsenault, R., A. Poulin, P. Côté, and F. Brissette (2013), A comparison of stochastic optimization algorithms in hydrological model calibration, J. Hydrol. Eng., doi:10.1061/(ASCE)HE.1943-5584.0000938.

Bao, Z., J. Zhang, J. Liu, G. Fu, G. Wang, R. He, X. Yan, J. Jin, and H. Liu (2012), Comparison of regionalization approaches based on regression and similarity for predictions in ungauged catchments under multiple hydro-climatic conditions, J. Hydrol., 466–467, 37-46 doi:10.1016/j.jhydrol.2012.07.048.

Bardossy, A. (2007), Calibration of hydrological model parameters for ungauged catchments, Hydrol. Earth Syst. Sci., 11, 703–710.

Beven, K. (2006), A manifesto for the equifinality thesis, J. Hydrol, 320, 18–36.

Burn, D.H., and D.B. Boorman (1993), Estimation of hydrological parameters at ungauged catchments, J. Hydrol, 143(3–4), 429-454.

Chen, J., F. P. Brissette, A. Poulin, and R. Leconte (2011), Overall uncertainty study of the hydrological impacts of climate change for a Canadian watershed, Water Resour. Res., 47, W12509, doi:10.1029/2011WR010602

Chen, J., F. P. Brissette, D. Chaumont, and M. Braun (2013), Finding appropriate bias correction methods in downscaling precipitation for hydrologic impact studies over North America, Water Resour. Res., 49, 4187-4205, doi:10.1002/wrcr.20331.

DiCiccio, T.J. and J.P. Romano (1995), On bootstrap procedures for second-order accurate confidence limits in parametric models, Statist. Sinica, 5, 141-160.

Ebtehaj, M., H. Moradkhani, and H. V. Gupta (2010), Improving robustness of hydrologic parameter estimation by the use of moving block bootstrap resampling, Water Resour. Res., 46, W07515, doi:10.1029/2009WR007981.

Efron, B. (1979), Bootstrap methods: Another look at jackknife, Ann. Stat., 7, 1–26.

Efron, B. (1987), Better bootstrap confidence intervals, J. Am. Stat. Assoc., 82(397), 171–200.

Fortin, V. (2000), Le modèle météo-apport HSAMI: historique, théorie et application, Varennes: Institut de Recherche d'Hydro-Québec, 68 p.

Hansen, N. and A. Ostermeier (1996), Adapting arbitrary normal mutation distributions in evolution strategies: The covariance matrix adaptation, In Proceedings of the 1996 IEEE International Conference on Evolutionary Computation, 312-317.

Hansen, N., and A. Ostermeier (2001), Completely derandomized self-adaptation in evolution strategies, Evolutionary Computation, 9(2), 159-195.

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

Hutchinson, M. F., D.W. McKenney, K. Lawrence, J. H. Pedlar, R. F. Hopkinson, E. Milewska, and P. Papadopol, (2009), Development and testing of Canada-wide interpolated spatial models of daily minimum-maximum temperature and precipitation for 1961-2003, J. Appl. Meteor. Climatol., 48, 725-741.

Kruskal, W. H. and W. A.Wallis (1952), Use of ranks in one-criterion variance analysis, J. Amer. Statist. Assn., 47, 583-621.

Lee H., N. R. McIntyre, H. S. Wheater, and A. R. Young (2006), Predicting runoff in ungauged UK catchments, Proceedings of the Institution of Civil Engineers. Water Management 159(2): 129–138.

McIntyre, N., H. Lee, H. Wheater, A. Young, and T. Wagener (2005), Ensemble predictions of runoff in ungauged catchments, Water Resour. Res., 41, W12434, doi:10.1029/2005WR004289.

Merz, R., and G. Blöschl (2004), Regionalization of catchment model parameters, J. Hydrol, 287, 95–123.

Minville, M., F. Brissette, and R. Leconte (2008), Uncertainty of the impact of climate change on the hydrology of a Nordic watershed, J. Hydrol., 358(1-2), 70-83.

Minville, M., F. Brissette, S. Krau, and R. Leconte (2009), Adaptation to climate change in the management of a Canadian water resources system, Water Resour. Manage., 23(14), 2965-2986.

Minville, M., S. Krau, F. Brissette, and R. Leconte (2010), Behaviour and performance of a water resource system in Québec (Canada) under adapted operating policies in a climate change context, Water Resour. Manage., 24, 1333–1352.

Oudin, L., V. Andréassian, C. Perrin, C. Michel, and N. Le Moine (2008), Spatial proximity, physical similarity, regression and ungaged catchments: A comparison of regionalization approaches based on 913 French catchments. Water Resour. Res., 44, W03413. doi:10.1029/2007WR006240.

Parajka, J., A. Viglione, M. Rogger, J. L. Salinas, M. Sivapalan, and G. Blöschl (2013), Comparative assessment of predictions in ungauged basins – Part 1: Runoff hydrograph studies, Hydrol. Earth Syst. Sci. Discuss., 10, 375-409, doi:10.5194/hessd-10-375-2013.

Parajka, J., R. Merz, and G. Blöschl (2005), A comparison of regionalisation methods for catchment model parameters, Hydrol. Earth Syst. Sci., 9, 157–171.

Poulin, A., F. Brissette, R. Leconte, R. Arsenault, and J. Malo (2011), Uncertainty of hydrological modelling in climate change impact studies, J. Hydrol, 409(3–4), 626–636.

Razavi, T., and P. Coulibaly (2013), Streamflow prediction in ungauged basins: Review of regionalization methods, J. Hydrol. Eng., 18(8), 958–975.

Samuel, J., P. Coulibaly, and R. Metcalfe (2011), Estimation of continuous streamflow in Ontario ungauged basins: Comparison of regionalization methods, J. Hydrol. Eng., 16(5), 447–459.

Seibert, J. (1999), Regionalisation of parameters for a conceptual rainfall runoff model, Agric. For. Meteorol., 98-99(31), 279–293.

Sivapalan, M., K. Takeuchi, S. W. Franks, V. K. Gupta, H. Karambiri, V. Lakshmi, X. Liang, J. J. McDonnell, E. M. Mendiondo, P. E. O'Connell, T. Oki, J. W. Pomeroy, D. Schertzer, S. Uhlenbrook, and E. Zehe (2003), IAHS Decade on Predictions in Ungauged Basins (PUB), 2003-2012: Shaping an exciting future for the hydrological sciences, Hydrolog. Sci. J., 48, 857-880.

Wagener, T., and H. S. Wheater (2006), Parameter estimation and regionalisation for continuous rainfall-runoff models including uncertainty, J. Hydrol., 320, 132–154.

Wilcoxon, F. (1945), Individual comparisons by ranking methods, Biometrics, 1(6), 80–83.

Yadav, M., T. Wagener, and H. Gupta (2007), Regionalization of constraints on expected watershed response behavior for improved predictions in ungauged basins, Adv. Water Resour., 30, 1756–1774.

Zhang, Y., and F. H. S. Chiew (2009), Relative merits of different methods for runoff predictions in ungauged catchments, Water Resour. Res., 45, W07412, doi:10.1029/2008WR007504.

CHAPITRE 7

ARTICLE 5 : MULTI-MODEL AVERAGING FOR CONTINUOUS STREAMFLOW PREDICTION IN UNGAUGED BASINS

Richard Arsenault[1] et François Brissette[1]

[1] Département de Génie de la Construction, École de technologie supérieure,

1100 Notre-Dame Ouest, Montréal, Québec, Canada H3C 1K3.

Article soumis à la revue « Hydrological Sciences Journal » en janvier 2015.

Abstract

This paper assesses the possibility of using multi-model averaging techniques for continuous streamflow prediction in ungauged basins. Three hydrological models were calibrated on the Nash-Sutcliffe Efficiency metric and were used as members of 8 multi-model averaging schemes. The averaging methods were tested on 267 catchments in the province of Québec, Canada, in a leave-one-out cross-validation approach. It was found that the best hydrological model was practically always better than the others used individually or in a multi-model framework, thus no averaging scheme performed statistically better than the best single member. It was also found that the robustness and adaptability of the models were highly influential on the models' performance in cross-verification. The results show that multi-model averaging techniques are not necessarily suited for regionalization applications, and that models selected in such studies must be chosen carefully as to not be too heterogeneous.

Keywords: multi-model; model averaging; regionalization; streamflow prediction; PUB; physical similarity

7.1 Introduction

The science of predicting continuous streamflow time series in ungauged basins has progressed in the past few years, especially since the IAHS issued the 2003-2012 decade on prediction in ungauged basins (Sivapalan et al. 2003). Parajka et al. (2013) and Razavi and Coulibaly (2013) have published comprehensive reviews of the many attempts and breakthroughs made thus far, and Hrachowitz et al. (2013) show which difficulties persist in this ever-evolving aspect of hydrology. As the term “regionalization” has taken different meanings during these years (He et al. 2011), it should be noted that in this paper, regionalization refers to the art predicting streamflow values on ungauged basins using models calibrated on other, gauged basins. The body of literature is well established in single model regionalization and a few methods have been used extensively such as the spatial proximity or physical similarity methods (Merz and Blöschl 2004, McIntyre et al. 2005, Parajka et al. 2005, Bardossy 2007, Oudin et al. 2008, Zhang and Chiew 2009). The reader is invited to consult any of these works for details on the inner workings of the aforementioned strategies.

7.1.1 Multi-model averaging

In other subsets of hydrology, such as in model parameter calibration, precipitation forecasting and flood forecasting, multi-model averaging has been used extensively in the past years (Shamseldin et al. 1997, Ajami et al. 2006, Diks and Vrugt 2010). The body of literature suggests that the model averaging techniques make the best use of the information provided by each model in the group, thus reducing uncertainty and model error while improving on performance. The first noteworthy case of multi-model averaging for rainfall-runoff modelling was proposed by Shamseldin et al. (1997). They showed that the Weighted Average Method (WAM) and Neural Network Method (NMM) produced better results than the Simple Average Method (SAM), which is a simple arithmetic mean of the multiple model outputs.

Other multi-model approaches have been proposed by Ajami et al. (2006). They compared the SAM and WAM methods to the Multi-Model Super Ensemble (MMSE) and Modified MMSE (M3SE) methods using the Distributed Model Intercomparison Project Results (Smith et al. 2004). These methods include bias correction and variance reduction to further improve simulation quality. However these methods cannot be used in regionalization as they require knowing the measured streamflow timeseries.

Arsenault et al. (2014b) compared 9 multi-model averaging schemes using 421 catchments from the MOPEX database (Duan et al. 2006). The authors use the same hydrological models as in the present study and they conclude that multi-model averaging increases prediction skill better than any single model. They also find that the popular Bayesian Model Averaging method (BMA) (Raftery and Zheng 2003, Raftery et al. 2005, Neumann 2003, Vrugt and Robinson 2007) performs well but is not as robust as others, and is costly in terms of required computing power. They conclude that the Unconstrained Granger-Ramanathan variant C is as good as BMA but is much quicker to implement and is more robust, which seconds Diks and Vrugt's (2010) original findings.

The averaging aspect is quite well understood and promising for use on a single basin. For example, estimating streamflow during calibration using multi-model averaging and then applying to validation is common and has been shown to be efficient. However, in regionalization projects, the streamflow must be predicted on a different basin. In this case, the weights are not guaranteed to be good or even acceptable. In the case of ungauged catchments, this is a problem which cannot be avoided.

7.1.2 Multi-model averaging in regionalization

In this paper, the model averaging methods will be used as tools to help predict streamflow in ungauged basins. While multi-model approaches have been popular with hydrologists in general, regionalization studies have not used them quite as often. One major problem is the need to calibrate the weights of the ensemble members. By definition, it is impossible to do so in ungauged basins. However the weights can be determined based on similar donor catchments and then transferred to the ungauged site. McIntyre et al. (2005) were the first to use multi-model averaging in a regionalization context. They showed that ensemble and similarity weighed averaging (SWA) was significantly better than individual model regionalization on 127 catchments in the UK. Goswami et al. (2007) also tested multi-model averaging over 12 catchments in France, and concluded that the method performs better than any single model in calibration, but loses its advantage in validation. Viney et al. (2009) used five lumped rainfall-runoff models on 240 Australian catchments in a multi-model, multi-donor regionalization framework. They showed that a weighted average of the five models is better than unweighted averaging during calibration, but not in validation. They also find that multi-donor ensembles using the five-model averaging approach is better than the single-donor approach. They conclude that the best results are obtained using weighted multi-model and weighted multi-donor methods combined.

Previous studies do not all agree on the methods to be used or expected results, and some have used relatively limited datasets to validate their approach. This study will use three models and eight model averaging methods to widen the range of possible outcomes. The scope of the trials will help in understanding and estimating the usefulness of model averaging techniques in continuous streamflow prediction.

7.1.3 Averaging methods description

Eight multi-model averaging methods were selected in this work in an attempt to maximize prediction skill.

7.1.3.1 Simple Average Method (SAM)

SAM is the simplest of the tested methods and will be the benchmark by which others are compared. This method is used to determine if simple averaging can perform better than using a single model. The simulated flows from the different models are simply averaged with this method. No weights must be computed as they are de facto equal to the inverse of the number of models.

7.1.3.2 Unconstrained and Constrained Granger-Ramanathan Averaging (UGRA, CGRA)

The Unconstrained Granger-Ramanathan Averaging method (Granger and Ramanathan 1984) is a simple method that minimizes the RMSE between the simulated and observed variables. As the name implies, the weights are unconstrained. There is no bias correction mechanism, which makes it a candidate for regionalization purposes. The Constrained Granger Ramanathan Averaging method is the same as the UGRA method except that the weights must sum to 1.

7.1.3.3 Bates Granger Averaging (BGA)

The Bates-Granger averaging method (Bates and Granger 1969) aims to reduce the RMSE of the combined forecast, under the presumption that the streamflow values are unbiased and that the inter-member errors are uncorrelated. Each member's weight is estimated using $1/\sigma^2{}_i$ where $\sigma^2{}_i$ is the member's estimated variance.

7.1.3.4 Shuffled Complex Averaging (SCA)

The Shuffled Complex Averaging method uses a stochastic optimization algorithm (SCE-UA) (Duan et al. 1992) to optimize weights that maximize the Nash-Sutcliffe efficiency between the observed data and the resulting weighted average streamflow series. Boundary values of [-1:1] were set to limit the range of values the weights can take. This method is based purely on trial and error and is not based on mathematical hypotheses such as bias correction or assumptions such as the absence of correlation between the input members. The details of this method can be found in Arsenault et al. (2014b), in which the SCA method was found to be the best along with the Unconstrained Granger-Ramanathan (UGRA) averaging method.

7.1.3.5 Akaike and Bayes Information Criterion Averaging (AICA, BICA)

The Akaike and Bayes Information Criterion Averaging methods (Akaike 1974, Buckland et al. 1997, Burnham and Anderson 2002, Hansen, 2008) estimate the likelihood of each member using an average of the log of the error variance of all of the members, to which a penalty term is added for each member. In the AICA method, the penalty term is equal to twice the amount of configurable parameters during the calibration process. For BICA, the amount of configurable parameters multiplied by the natural log of the amount of time steps in the calibration period is used instead.

7.1.3.6 Neural network Method (NNM)

The NNM uses a multi-layer feedforward neural network comprising of 3 layers: The input layer, the output layer, and a central layer called the hidden layer. Each layer has a number of neurons where information is processed. The input layer has one neuron per hydrological model estimated streamflow series, the output layer has only one neuron (the estimated streamflow) and the hidden layer has a user-defined number of neurons. The higher the number of neurons, the better the fit. However, when taken into validation mode, overfitting issues arise if there are too many neurons in the hidden layer. Therefore, keeping the number to a minimum is preferable. In the present study, it was found that the optimal number of neurons in the hidden layer was 3. The different layers are linked together using transfer functions. Input neurons are transferred to the hidden layer neurons using these transfer functions, which can take many shapes. Usually, a non-linear logistic function is used as the activation function (between the input and hidden neurons) and a linear transfer function is used between the hidden and output neurons. The neural network assigns weights to each transfer function to minimize the mean square error between the observed and predicted streamflow values. Many types of NNMs exist, and different approaches using NMMs have been proposed, such as Ensemble NNMs and Non-Linear NN ensemble means (Krasnopolsky and Lin 2012). The reader is referred to Shamseldin et al. (1997) for more information on the mathematics and applications of NNM.

7.2 Models, study area and data

This section first introduces the hydrological models used in this paper, and then describes the study area and the data for each of the 267 basins.

7.2.1 Hydrological models

Three models of varying complexity were used during this study, with free parameters ranging from 10 for MOHYSE to 23 for HSAMI. All three are lumped rainfall-runoff models.

7.2.1.1 HSAMI

The HSAMI model (Fortin 2000; Minville et al. 2008, 2009, 2010, Poulin et al. 2011, Arsenault et al. 2013) has been used by *Hydro-Quebec* for over two decades to forecast daily flows on more than 100 basins over the province of Quebec. Runoff is generated by surface, unsaturated and saturated zone reservoirs through two unit hydrographs: one for surface and another for intermediate (soil water) reservoir unit hydrographs. The required inputs are spatially averaged maximum and minimum temperatures, liquid and solid precipitation. The model has 23 calibration parameters, all of which were used for this study.

7.2.1.2 MOHYSE

MOHYSE is a simple model that was first developed for academic purposes (Fortin and Turcotte 2007). Since then, the model has been used in research applications (e.g. Velazquez et al. 2010). MOHYSE is specifically built to handle Nordic watersheds and has a custom snow accumulation and melt as well as potential evapotranspiration (PET) modules. The required input data are mean daily temperatures, total daily rainfall depth and total daily snow depth (expressed as water equivalent). Ten (10) parameters need to be calibrated.

7.2.1.3 HMETS

HMETS is a model that uses two reservoirs for the vadose and phreatic zones (Chen et al. 2011). HMETS is a Matlab based model which has 21 parameters. The model requires the area of the watershed and the latitude and longitude of the centroid of the basin area as physiographic information. The minimum and maximum temperatures as well as snow and rain are also required as meteorological inputs. HMETS' structure resembles that of HSAMI as it accounts for snow accumulation, snowmelt and evapotranspiration using the hydrometeorological data available to simulate the streamflow at the outlet. It was fitted with more complex snowmelt and evapotranspiration models than HSAMI, which could improve simulations in the study area.

7.2.2 Study area

The study area consists of 267 basins covering the province of Québec, Canada. Figure 7.1 shows the study area and the basin locations.

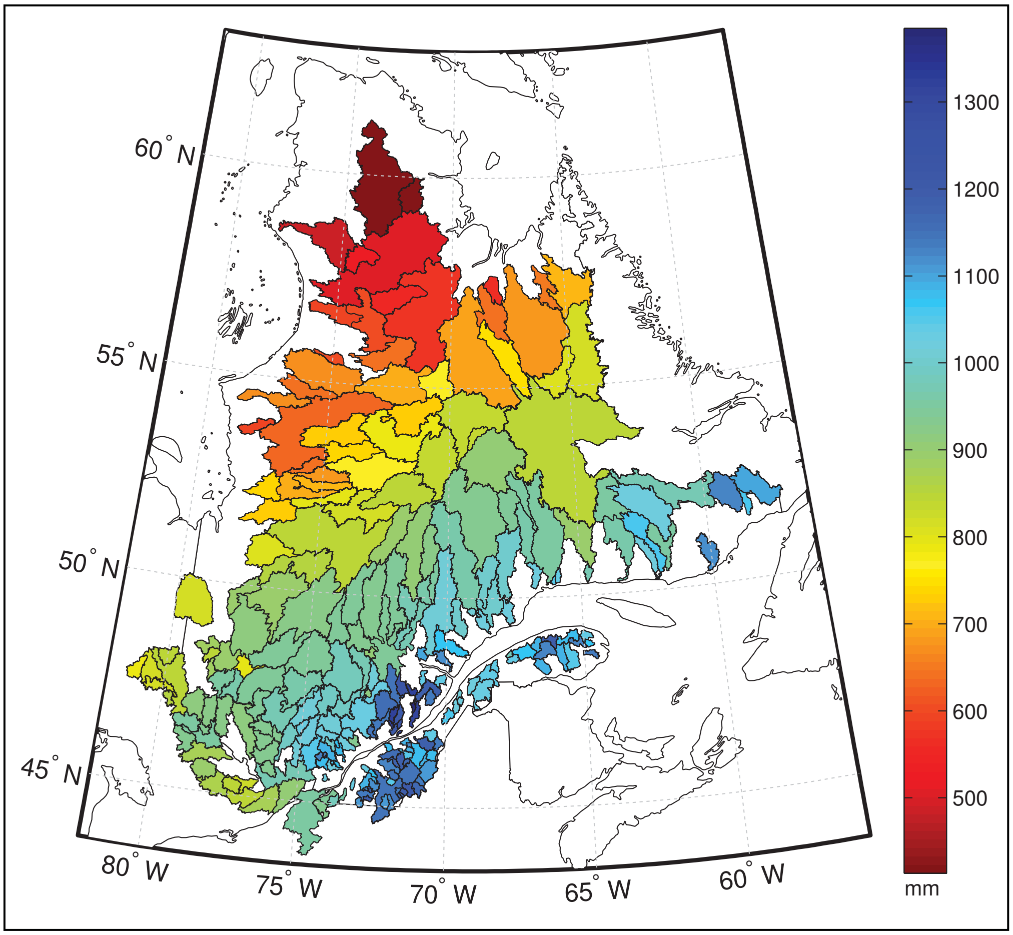

Figure 7.1 Catchment locations in the province of Québec used in this study

Some basins are nested within others which are included in the study. The basins range in size from 30 to 69191 square kilometres, and cover most of the province of Québec with a total area of 1.6 million square kilometers. A list of 12 catchment descriptors was used in this study according to the compilation by He et al. (2011). Some descriptors, such as soil properties, were not used in this study due to limited availability. The ones that were selected, as well as their statistics, are presented in table 7.1.

Table 7.1 Statistics of catchment descriptors used in this study

Catchment descriptors	Maximum	Minimum	Average
Area (km²)	69191	30	6832
Slope (%)	51.9	1.1	10.7
Elevation (m)	916	52	383
Land Cover - Crop (%)	83.1	0	8.7
Land Cover – Forest (%)	96	0	65.2
Land Cover - Grass (%)	65.5	0	13.6
Land Cover - Urban (%)	16.4	0	1.2
Land Cover - Water (%)	35.6	0	9.3
Land Cover - Wetlands (%)	17.1	0	1.2
Mean annual precipitation (mm)	1412	413	965
Longitude (degrees)	-57.9	-81	-72
Latitude (degrees)	59.9	44.5	49
Aridity index	0.99	0.31	0.61

7.2.3 Meteorological and hydrological datasets

The hydrometric data were obtained through a partnership between various province and industry partners who combined their hydrometric data into a single database. The observed climate data were substituted by the Canadian National Land and Water Information Service (NLWIS) 10km gridded dataset (Hutchinson et al. 2009). This choice was made since many catchments have no weather stations within their boundaries, but all the catchments in this study contained at least one NLWIS climate data point. The NLWIS climate dataset was shown to be a good replacement for missing observed data in hydrological applications (Chen et al. 2013), although it still suffers from the lack of observational data in the northernmost points in the study area.

7.3 Methodology

The methodology can be broken down into four main sections: The model calibration approach, the donor basin selection scheme (the regionalization method), the model averaging strategies and the multi-donor aggregation step.

7.3.1 Model calibration

The first step in this study was to calibrate all the models on all the catchments to obtain parameter sets to be transferred to the ungauged sites. All calibrations for the HSAMI and HMETS models were performed using the Covariance-Matrix Adaptation Evolution Strategy (CMAES) (Hansen and Ostermeier 1996, 2001). CMAES is an evolutionary algorithm for difficult problems, such as those with non-linear, non-convex and non-smooth fitness landscapes. It is an iterative second order method that estimates the positive definite matrix but is free of derivability requirements to estimate gradients. It was shown to outperform other algorithms in calibration for these models (Arsenault et al. 2014a). Following the same methodology, it was determined that the SCE-UA algorithm (Duan et al. 1992, 1993, 1994) was the better choice for the MOHYSE model. The models were calibrated using the Nash-Sutcliffe Efficiency metric as the objective function (Nash and Sutcliffe 1970), which is arguably the most common goodness-of-fit metric in hydrology. The calibration was performed on the first half of the available data.

Lower-scoring basins in calibration are sometimes discarded at this stage in regionalization studies; however they were not removed in this work in order to keep as much information as possible for the regionalization strategies under a multi-model averaging framework.

7.3.2 Donor basin selection scheme

As was shown in Zhang and Chiew (2009), a combination of physical similarity and spatial proximity may outperform both approaches taken individually. Therefore, a physical similarity method using spatial distance as one of the catchment characteristics was used.
The physical similarity approach uses catchment descriptors to rank the catchments in similarity to the ungauged one. The strategy involves transferring the parameter sets from the most similar catchments to the ungauged catchment for use in the hydrological models. The similarity between catchments was measured using the similarity index defined by Burn and Boorman (1993):

$$\Phi = \sum_{i=1}^{k} \frac{\left|X_i^G - X_i^U\right|}{\Delta X_i} \tag{7.1}$$

Where i is the catchment descriptor identifier, X^G is the catchment descriptor value for the gauged catchment, X^U is the catchment descriptor value at the ungauged catchment and $\varDelta X$ is the range of values taken by X^G in the dataset. The catchment that minimizes the difference in similarity index ϕ with the ungauged basin is used as the donor catchment. When multiple donors are used, they are selected in ascending order of similarity index value. The catchment descriptors were all used in preliminary testing in this work, but the results were not as good as with small subsets of the descriptors. A one-at-a-time approach allowed showing that only 4 descriptors were necessary to maximize the performance on almost all the catchments. These are the latitude, longitude, mean annual precipitation and fraction of land cover that is water. These were shown to be optimal or quasi-optimal for the three models by adding one descriptor at a time in descending order of performance increase (see Arsenault and Brissette 2014). In doing so, the similarity index is a hybrid of proximity and similarity metrics, thus making it an integrated similarity index. Adding more descriptors to this list only reduced the performance of the models.

It is important to note that all 267 available basins were used during the cross-validation phase as pseudo-ungauged targets, as it would be impossible in a real world scenario to know in advance if a basin would have good calibration efficiency metric values. Then, the catchments whose mean calibration NSE values were less than 0.7 were discarded from the list of possible donor basins. Therefore all basins are considered as ungauged in the cross-validation phase, however the basins which are poorly modelled are not considered as viable donors. This is similar to the approach used by Oudin et al. (2008) and Arsenault and Brissette (2014) which allows for more realistic simulation and validation results.

7.3.3 Model averaging strategies

The multi-model averaging step is the cornerstone of this project. The method will be detailed for one averaging scheme, but in the project the process was repeated for each of them. The steps are as follows:

1- Run the 3 hydrological models on the donor catchment and produce 3 hydrographs;
2- Apply the weighting schemes to the 3 hydrographs with the donor basin's observed hydrograph as the target. This will produce a set of weights W;
3- Run the hydrological models on the ungauged basin using the donor basin's parameter set for each model, resulting in 3 simulated hydrographs on the ungauged basin;
4- Apply the set of weights W to the 3 hydrographs generated in point 3. This produces an averaged hydrograph for the ungauged basin;
5- Compare the observed and averaged hydrographs on the ungauged basin, or use the averaged hydrograph in a multi-donor framework detailed below, as in Zelelew and Alfredsen (2014).

Multiple donors were used in this framework, so this step was repeated for each of the donor basins. Note that this procedure averages the discharge as simulated by the three models according to the weights that are determined on the donor catchment. Since the three parameter sets are transferred (one per model), it is assumed that the model structural error will be preserved at the target site. Therefore the weights are transferred from the donor to the target basin as-is. However, other methods of weighting the models have been proposed, such as in Reichl et al (2009) where prior belief in transferability is used instead of relying on the donor-calibrated weights. Moreover, different combinations of hydrological models were used to determine if any had more impact than any other.

7.3.4 Multi-donor averaging

Parajka et al. (2007), amongst others, showed that when multiple donors are used, inverse distance weighting (IDW) outperformed simple arithmetic averaging. Simple linear IDW will thus be used to predict streamflow at an ungauged site when multiple donors are selected. Simply put, the streamflow values produced with the multi-model averaging scheme from each donor were aggregated into a single multi-model, multi-donor streamflow time series. This average is then compared to the observed data to determine the efficiency metric and evaluate the multi-model averaging scheme performance. Furthermore, the distance measure is based not on the spatial distance, but on the physical similarity index distance, which happens to be heavily influenced by spatial distance. This double averaging approach has been shown to be effective in a study by Viney et al. (2009).

7.4 Results

7.4.1 Initial model calibration and weighting method evaluation

The hydrological model calibration process was performed on the three models with the NSE metric. Figure 7.2 shows the cumulative distribution function for the HSAMI, HMETS and MOHYSE hydrological models when calibrated on the NSE metric.

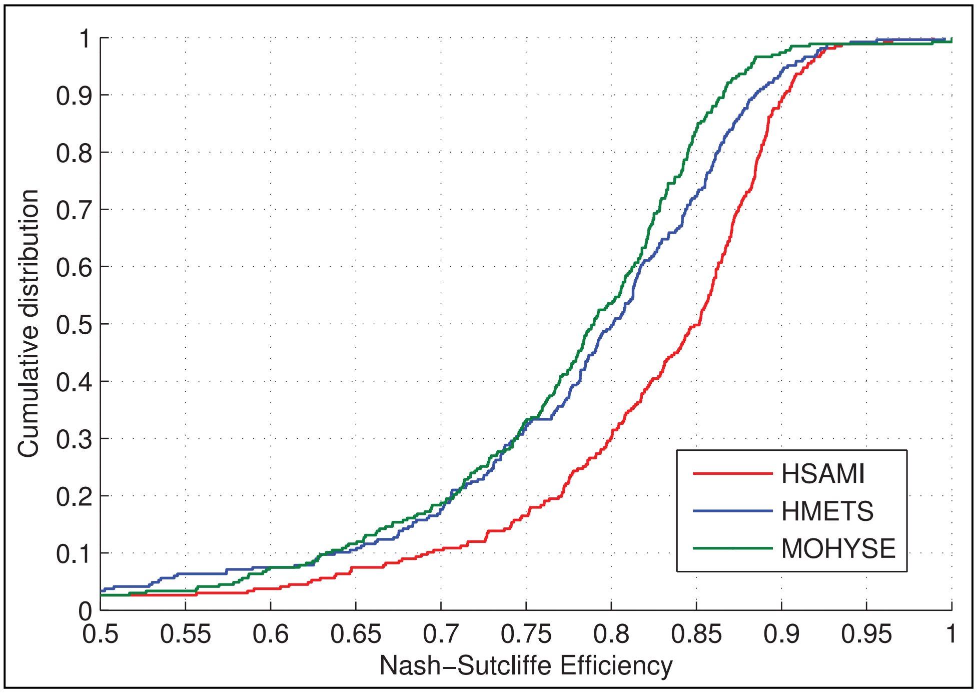

Figure 7.2 Cumulative distribution function of initial calibration performance of the three hydrological models calibrated on the NSE metric

Overall initial calibration of the three hydrological models revealed that the HSAMI model could adapt more easily than the other two models to the various basins in the database. The difference in NSE values between HSAMI and MOHYSE at the 50% probability level, for example, is of 0.07. Figure 7.2 starts at an NSE value of 0.5 since before that point, all three models are essentially the same.

The weighting methods were then evaluated locally on the gauged basins. The NSE of the best of the three models was pitted against the NSE obtained with the model averaging schemes. Figure 7.3 shows the results of this evaluation on the validation period, which was equal to the last half of the available data for each given site.

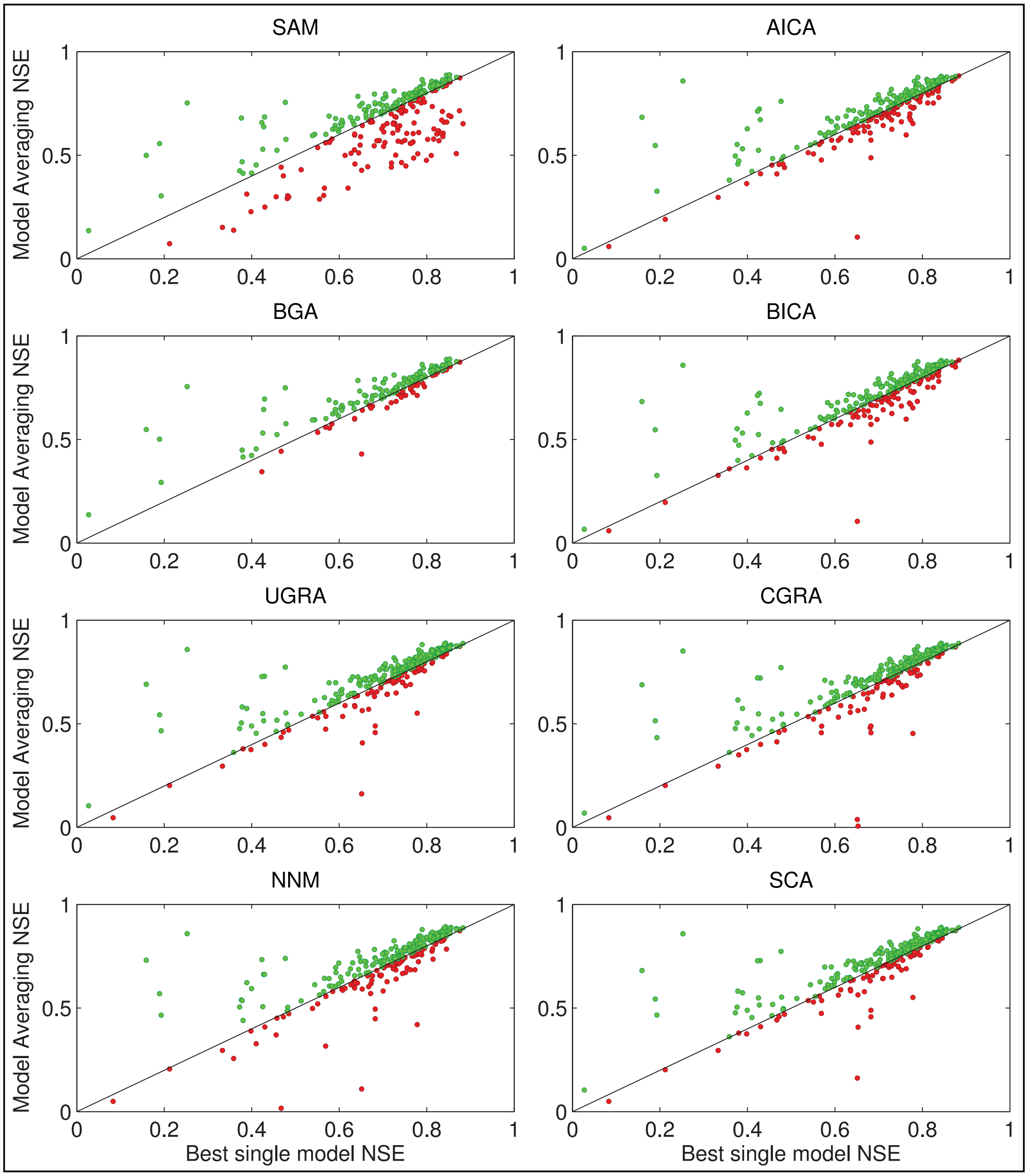

Figure 7.3 Best single model NSE and model averaging NSE in validation for the 8 averaging methods. The diagonal line represents the 1:1 ratio. Markers over (or to the left of) the line indicate basins where the model averaging methods were able to improve upon the best model's performance

It is clear from figure 7.3 that some of the model averaging methods are able to consistently equal or outperform the best individual model. This is consistent with the literature and is an expected result for local application on gauged basins, but serves as the comparison benchmark for testing in the regionalization mode.

7.4.2 Regionalization under the multi-model averaging framework

The performance of the multi-model averaging schemes in regionalization was measured by comparing their predictive skill to that of the hydrological models taken individually in a standard mono-model regionalization approach. Figure 7.4 shows the average NSE values of the 267 ungauged catchments for the 8 multi-model averaging schemes. Furthermore, the multi-donor aspect of the project is illustrated as up to 15 donors were used to maximize the NSE gain as is the case in mono-model regionalization. Finally, the individual performances of the three models in a mono-model framework were added to Figure 7.4 for ease of comparison with the multi-model averaging schemes.

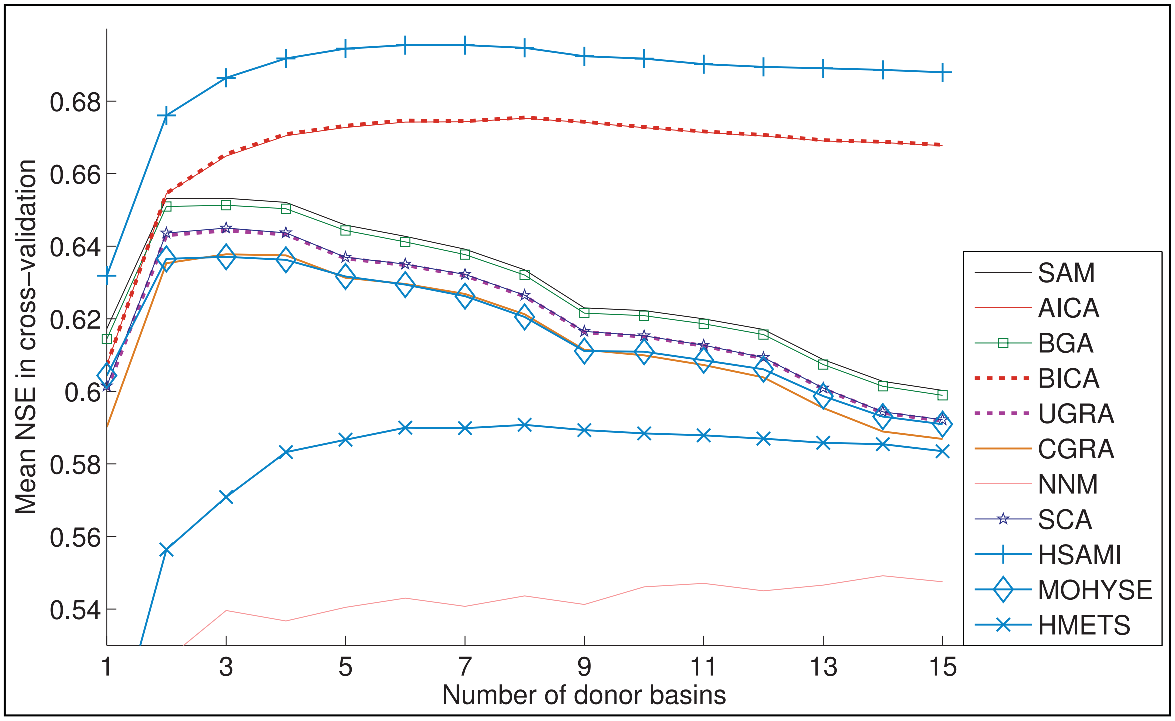

Figure 7.4 Mean NSE value in multi-model regionalization for a varying number of donor basins when the three-model ensemble is used

From figure 7.4 it is clear that the different model averaging schemes show diverse levels of success. The benchmark (SAM) outperforms all he methods except AICA and BICA, which are almost identical. The latter often find corner solutions, meaning that weights are attributed in a 0 or 1 fashion. The NNM method is largely the worst, and BGA, SCA, UGRA and CGRA are similar. This group behaves differently than what is usually seen in multi-donor regionalization. Indeed, the performance is maximal at 2 donors only, and sharply drops thereafter. AICA and BICA seem to have a sweet-spot at approximately 4 to 7 donors before slightly declining. However, and most importantly, no method was able to equal or beat the best single model (HSAMI) used alone in regionalization. Clearly the use of other models with poorer calibration NSE values is lowering the overall score.

Another test was performed by reiterating the method with different model ensemble members. Figure 7.5 shows the behaviour of the model averaging methods when all possible model combinations are used (3x single model, 3x 2 models and 1x 3 models). Each panel in figure 7.5 represents a different model averaging technique and each curve represents the mean regionalization NSE value for a given model combination. Note that for the NNM method, the HMETS-MOHYSE ensemble is not shown as its NSE values are too low to properly display.

Figure 7.5 also shows the effects of the different model weighting mechanisms. For instance, AICA and BICA offer the same performance as HSAMI when the HMETS model is not used. In this case, the MOHYSE model is never given a weight, leaving HSAMI as the only weighted model with a weight of unity. However, when the HMETS model is used with HSAMI, the performance drops uniformly according to the proportion of times the HMETS model is used in the weighting. Also, the performance of the HSAMI-HMETS ensemble is similar to that of the 3-model ensemble, thus confirming the relative uselessness of MOHYSE when HSAMI is present. This is expected from AICA and BICA since the algorithms strongly favour the best model and neglect the other members. In this case, the difference between HSAMI and MOHYSE in terms of relative performance in calibration (as shown in figure 7.2) is large enough as to render MOHYSE all but unused.

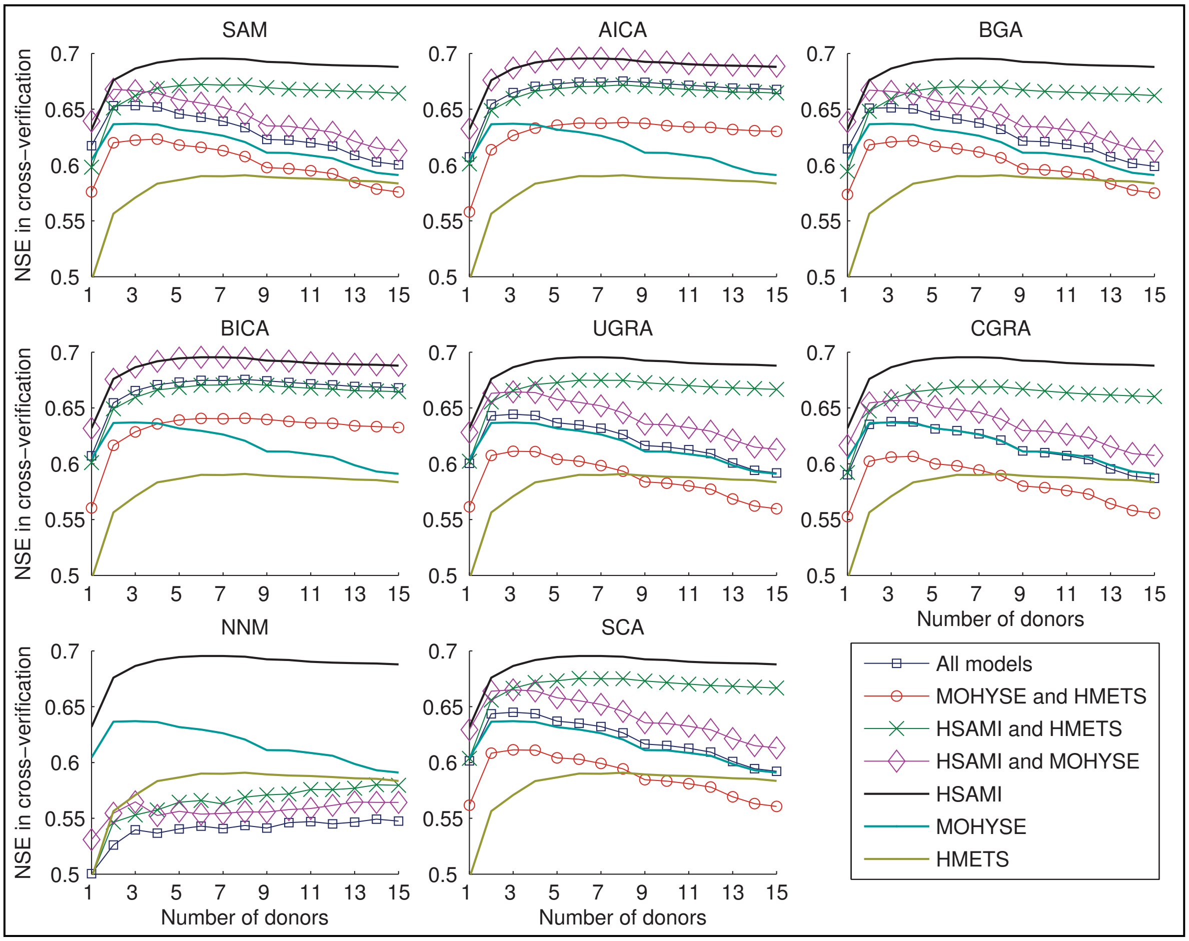

Figure 7.5 Mean NSE values in multi-model regionalization depending on the models included in the ensemble. Each panel presents the results for a specific model averaging method. Note than NNM does not show the MOHYSE-HMETS ensemble as the performance is too low to properly display

In the MOHYSE-HMETS ensemble, AICA and BICA are able to beat the individual members by selecting the best model for each case. The difference in performance between HMETS and MOHYSE in calibration was much smaller than with HSAMI, as was seen in figure 7.2. This shows that the level of similarity of the models is important in multi-model regionalization. For the other model averaging methods, the behaviour is different since the multiple models are often allocated non-negligible weights. It can be seen that the HSAMI-HMETS ensemble ranks highest and that the model average follows the same type of performance curve as HSAMI and HMETS. MOHYSE, on the other hand, has an optimum number of donors of 2 and performance drops quickly thereafter. It can be seen that when the

MOHYSE model is part of the ensemble, the model average performance follows the same type of downward trend, thus indicating that the MOHYSE model is often weighted. However, all the averaging methods except AICA and BICA are unable to perform at the same level as the single HSAMI model.

7.4.3 Weights distribution

In order to better understand the model averaging methods properties, the weights that are generated by each method were analyzed. Figure 7.6 shows the cumulative distribution function of the model weights for the 267 of the catchments with the 3-member ensemble. The X-axis is the value of the weight and the Y-axis is the frequency probability for that weight value. For example, for the BICA method, the HSAMI member has a weight of 0 for 30% of the basins, a weight between 0 and 1 for approximately 5% of the basins, and a weight of 1 for the remainder (65%). Note that the cumulative distribution function reorders the weight sets in increasing order, thus it is impossible to identify the weights for a given catchment from figure 7.6. For constrained methods with weights bounded from 0 to 1, the weights sum to unity, therefore when one member has a weight of 1 the others are necessarily set to zero.

It can be seen in figure 7.6 that the AICA and BICA methods favor corner solutions, in that they give no weight to the undesirable models and a weight of 1 to the best. The UGRA and CGRA methods distribute the weights fairly, but they are not bounded at 0 or 1. Many solutions therefore use negative and over-unity weights. The SCA method sees the same behavior even though the weights are bounded from -1 to 1. The fact that the UGRA and CGRA produce weights similar to SCA suggests that their weights do not require being very far out of the -1 to 1 range. It is also noteworthy that the SCA, UGRA and CGRA methods all attempt to minimize the square of model residuals between the averaged and observed flow, therefore their weights are expected to be similar.

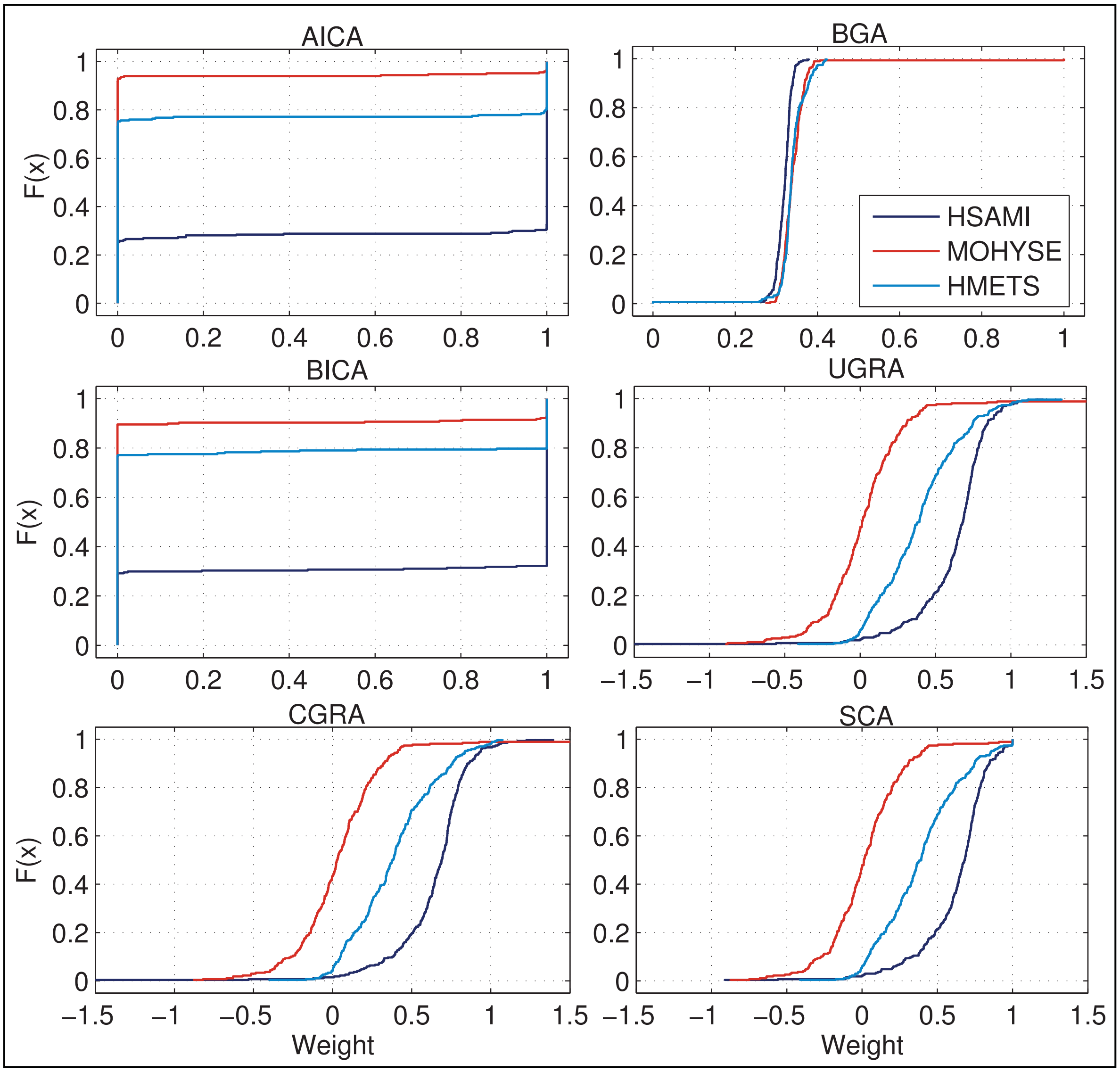

Figure 7.6 Cumulative distribution function of the model averaging methods' calculated weights for the 6 weighting schemes using the three models. SAM is not included as the weights are all set to one third, and NNM does not use weights but a neural network transfer function

7.5 Analysis and discussion

7.5.1 Overview of model averaging methods performances

Throughout this study, 8 model averaging methods were used. While most were able to perform well at some point and in specific conditions, the NNM was unable to compete with the others. Seeing as it does not use weights per se, but rather transfer functions, it is possible that the neural networks are well trained on the gauged basins but are unable to adequately use the inputs from other basins. NNMs have proven time and again that when they are used in the right conditions, they can be powerful tools, such as in classical rainfall-runoff prediction (Shamseldin et al. 1997, Krasnopolsky and Lin 2012). The number of neurons in the hidden layer was varied from 1 to 10, with 3 neurons returning the best results in validation. The project therefore used 3 neurons for all the tests. This could have biased the results somewhat, but it is doubtful that the end results would change.

Another visible trait is that AICA and BICA are able to handle the poorer HMETS model better than the other methods. In figures 7.3 and 7.4, they score the highest and maintain a good performance as the number of donors increases. This particularity is due to the way the AICA and BICA methods compute the weights, often attributing a weight of 1 to the best single model and a weight of 0 to the others, as shown in figure 7.6. This figure also shows that the other methods tend to set weights that can be negative or superior to 1. This could produce poor results if the hydrological models are not able to produce reasonable flows on the ungauged basin. For example the UGRA and CGRA methods behave in this manner. BGA, on the other hand, is constrained between 0 and 1, but put similar weights on the HMETS, MOHYSE and HSAMI models as opposed to AICA and BICA.

A notable find was that of the SCA method, which ranked similarly to the UGRA and CGRA methods in the performed tests. This method is a "brute force" method which optimizes the weights on the calibration period with a stochastic optimization algorithm. As it compares well with more mathematically sound methods, it could be interesting to use different algorithms and different objective functions during calibration of the weights to try to further improve its efficiency.

7.5.2 Multi-model averaging in regionalization

According to the results obtained herein, it would not be advisable to use the three same hydrological models in a regionalization context. The model averaging methods are unable to improve upon the best single member in regionalization, which implies that the donors are either too different from the ungauged basins (which limits model performance) or that the transferred weights are not adequate. Since the single HSAMI model is able to generate good results, it follows that the problem lies within the model averaging weights transfer. However, the model averaging methods are able to hedge against the use of a bad model by either ignoring the bad models completely (such as in AICA and BICA) or at least weighting them with other models, such as the CGRA and BGA. The results in figures 7.4 and 7.5 suggest that the most sensible approach would be to find and use at least 2, but preferably more, of the best possible models in calibration and to ensure that the selected models perform similarly. The large discrepancy between HSAMI and HMETS/MOHYSE made it difficult for the averaging methods to improve upon the best single model. Perhaps finding other models that are equivalent to HSAMI on the study area would allow for better transferability and more chance of success in regionalization. It also appears that the AICA and BICA methods would be the methods of choice in this case as they were able to select the best model in each case to improve upon the single-models. While this does not guarantee a better NSE value in regionalization, it does reduce the chance of using a model that fails on the ungauged site, therefore reducing some of the uncertainty. Finally, in the case at hand, the extra resources required to perform multi-model regionalization do not reap the benefits as expected, especially for the complex neural network method.

7.5.3 Model robustness

One of the fundamental aspects of regionalization is the hydrological model's robustness to different basin characteristics and datasets. The same hydrological models and most of the model averaging schemes used in this work were used in Arsenault et al. (2014b). In that paper, it was shown that when the model averaging takes place in simulation mode, the weighting schemes work very well and they almost always score better NSE values than any model taken individually in validation (as is shown in figure 7.3). The authors also go on to show that even hydrological models that perform poorly are used in the averaging schemes and they contribute to the increase in performance. However, in this current paper we show that in regionalization, models whose robustness is poor cannot be trusted as the transferred parameter sets can make the models produce unrealistic streamflow values which cannot be corrected by the model averaging schemes.

This has been noted in Viney et al. (2009), who state that "*relative calibration performances of different models in a donor catchment are not necessarily good indicators of how well the models will contribute to prediction in a neighbouring catchment*." Their conclusions are different from the ones presented in this study as they found that multi-model averaging did increase prediction skill in ungauged basins. However they used 5 models of similar complexity with more similar calibration objective function values than in this study. Furthermore, their weighting algorithm was based on the calibration skill rather than on the reduction of structural error and they used a different objective function. All this adds credibility to the necessity of using the right models to allow the weighting schemes to perform at their best, and also to review the weighting approach according to transferability of the model parameter sets. It is also possible that the climate in the Viney et al. (2009) paper was better suited for model averaging techniques as the warmer and drier conditions of Australia could lead to more uncertainty in the modeled flow, thus allowing room for improvement with the model averaging methods.

A test was devised to verify the robustness of the models in the study when subjected to the parameter transfer process. Each model was run on all the basins using the parameter sets of all the other basins. Therefore basin-1 was run with the parameters from basin-2 to basin-267, basin-2 using parameters from basin-3 to basin-267, and so on. The NSE values obtained were analysed with their cumulative distribution functions shown in figure 7.7.

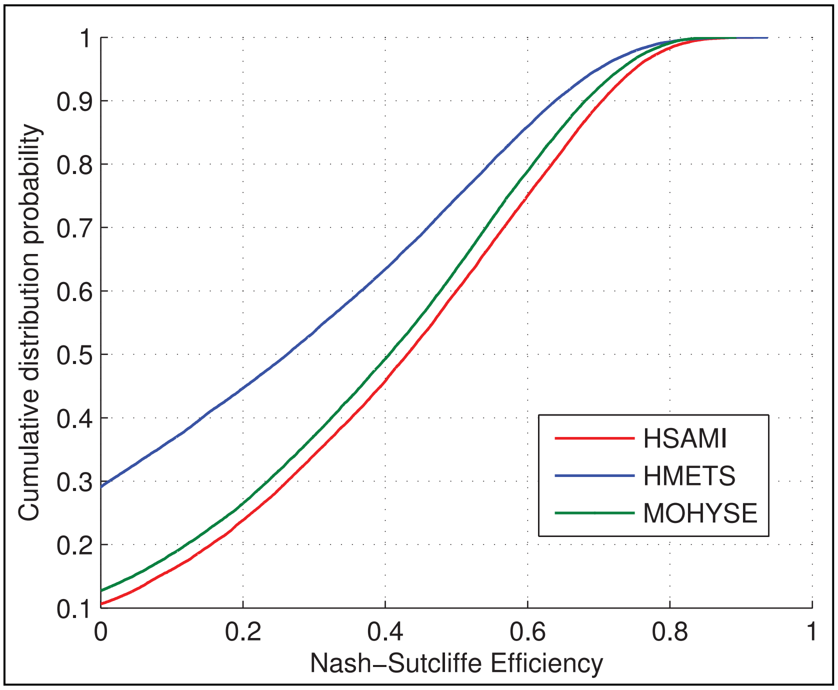

Figure 7.7 Cumulative distribution function of the hydrological models' performance when parameter sets are blindly transferred to another catchment. The CDFs contain all the possible donor-target combinations

The first obvious trait in figure 7.7 is that the HMETS model performs worse than taking the mean of the observed flows in approximately 30% of all cases (NSE<0). HSAMI and MOHYSE have no such problem, with approximately 11% and 14% of bad simulations respectively. The CDFs also show that HSAMI is more robust in that it can simulate better flows on randomly selected catchments compared to the other models, although MOHYSE is not very far behind.

The fact that HMETS is so much poorer than the other two models could explain why it is the worst performing single model in regionalization and why the multi-model regionalization scores are lower than when the HSAMI model is taken alone. If a method such as UGRA gives HMETS a non-negligible weight, it is possible that the predicted streamflow could be unrealistic in cross-validation due to HMETS' bad transferability properties. AICA and BICA are special cases in this regard, as they tend to give a weight of 1 to the best model and 0 to the others. In a case where HSAMI is the best model, the odds of the resulting average streamflow being adequate are good. However if the HMETS model is selected, then the end result could potentially be much poorer. Since HSAMI scores better than HMETS most of the time, AICA and BICA are generally the best methods in this scenario.

7.5.4 Multi-donor aspect

Using multiple donor basins has proved to be an excellent way to increase predictive skill in mono-model regionalization. Viney et al. (2009) used multi-donor averaging to further increase their gains on multi-model regionalization. In this work, multi-donor averaging was also found to be very effective in increasing the predictive skill of the multi-model approach. In figs 7.4 and 7.5 it is clear that using donor averaging improved performance for AICA and BICA, and even for the other methods, although the latter were generally poorer than the AICA and BICA methods in absolute terms. However, the donor averaging skill improvement was similar to that of the HSAMI model. Therefore it was impossible for AICA and BICA to surpass the HSAMI model. The optimal number of donors for the AICA and BICA were similar to the optimal number of donors for HSAMI, which was between 4 and 7. Nonetheless, the results show that the use of multiple donors consistently outperforms the single-donor approach. It is thus highly advisable to always use multiple donors when possible.

7.6 Conclusions

This study aimed at determining if multi-model averaging could be used to improve continuous streamflow prediction in ungauged basins. Eight model averaging methods were used in a multi-model, multi-donor regionalization framework based on physical similarity. It was shown that it is good practice to use multiple donors rather than a single donor, as is the case in mono-model regionalization. Every trial performed in this study showed marked gains when using multiple donors.

It was also noted that the performance of the model averaging methods is directly correlated to the robustness of the hydrological models. The HMETS model, when it was used by the model averaging schemes, contributed to lower the overall performance of the method. HSAMI, on the other hand, is more robust and it increased performance when it was used. Accordingly, weighting schemes performances were dependent on the models that were available. If there are non-robust models, simpler AICA and BICA are better than more complex methods. They also have the advantage of reducing the chance of failure if multiple good models are available in the ensemble. Overall, in this particular study, it was found that multi-model averaging was not able to consistently perform better than the best single model for regionalization purposes. More work is needed to better identify which hydrological models are the most robust as to better use the information they can gather on ungauged catchments. The results also show that good performance in model calibration is not a good indicator of regionalization skill in validation. The models can have good calibration skill but poor transferability; therefore they can decrease the overall performance. More research is needed in weighting models on ungauged basins.

7.7 Acknowledgments

The authors would like to thank Catherine Guay from IREQ for making the CQ2 measured streamflow database available and Kenjy Demeester from École de technologie supérieure for providing the land cover data for the project. We also thank the anonymous reviewers for their fruitful comments in improving this paper.

7.8 References

Ajami, N.K., Duan, Q., Gao, X., and Sorooshian, S., 2006. Multimodel combination techniques for analysis of hydrological simulations: application to Distributed Model Intercomparison Project results. Journal of Hydrometeorology. 7, 755–768.

Akaike, H., 1974. A new look at the statistical model identification. IEEE Transactions on Automatic Control, 19(6), 716-723.

Arsenault, R., Malo, J., Brissette, F., Minville, M. and Leconte, R., 2013. Structural and non-structural climate change adaptation strategies for the Péribonka water resource system. Water Resources Management, 27(7), 2075-2087. doi: 10.1007/s11269-013-0275-6.

Arsenault, R., and Brissette, F., 2014. Continuous streamflow prediction in ungauged basins: The effects of equifinality and parameter set selection on uncertainty in regionalization approaches. Water Resources Research, accepted July 8th 2014. doi: 10.1002/2013WR014898.

Arsenault, R., Poulin, A., Côté, P. and Brissette, F., 2014a. Comparison of stochastic optimization algorithms in hydrological model calibration. Journal of Hydrologic Engineering, 19(7), 1374-1384. doi: 10.1061/(ASCE)HE.1943-5584.0000938.

Arsenault, R., Gatien, P., Renaud, B., Brissette, F. and Martel, J.-L., 2014b. A comparative analysis of 9 multi-model averaging approaches in hydrological continuous streamflow prediction. Journal of Hydrology, (under review), 38p.

Bardossy, A., 2007. Calibration of hydrological model parameters for ungauged catchments. Hydrology and Earth System Sciences 11: 703–710.

Bates, J.M. and Granger, C.W.J., 1969. The Combination of Forecasts. Operational Research Quarterly, 20(4), 451-468.

Buckland, S.T., Burnham, K.P. and Augustin, N.H., 1997. Model Selection: An Integral Part of Inference. Biometrics, 53(2), 603-618.

Burn, D.H. and Boorman, D.B., 1993. Estimation of hydrological parameters at ungauged catchments. Journal of Hydrology, 143(3–4), 429-454.

Burnham, K.P. and Anderson, D.R., 2002. Model Selection and Multi Model Inference: A Practical Information-Theoretic Approach, Second Edition. United-States: Springer-Verlag, New-York. 487p.

Chen, J., Brissette, F.P., Poulin, A. and Leconte, R., 2011. Uncertainty of downscaling method in quantifying the impact of climate change on hydrology. Journal of Hydrology, 401(3-4), 190-202.

Chen, J., Brissette, F.P., Chaumont, D. and Braun, M., 2013. Finding appropriate bias correction methods in downscaling precipitation for hydrologic impact studies over North America, Water Resources Research, 49, doi:10.1002/wrcr.20331.

Diks, C.G.H. and Vrugt, J.A., 2010. Comparison of point forecast accuracy of model averaging methods in hydrologic applications. Stochastic Environmental Research and Risk Assessment, 24(6), 809-820.

Duan, Q., Sorooshian, S. and Gupta, V. K., 1992. Effective and efficient global optimization for conceptual rainfall runoff models. Water Resources Research, 24(7), 1163-1173.

Duan, Q., Sorooshian, S. and Gupta, V. K., 1993. A shuffled complex evolution approach for effective and efficient optimization. Journal of Optimization Theory and Applications, 76(3), 501-521.

Duan, Q., Sorooshian, S. and Gupta, V. K., 1994. Optimal use of the SCE-UA global optimization method for calibrating watershed models. Journal of Hydrology, 158, 265-284.

Duan, Q., Schaake, J., Andreassian, V., Franks, S., Gupta, H.V., Gusev, Y.M., Habets, F., Hall, A., Hay, L., Hogue, T.S., Huang, M., Leavesley, G., Liang, X., Nasonova, O.N., Noilhan, J., Oudin, L., Sorooshian, S., Wagener, T. and Wood, E.F., 2006. Model Parameter Estimation Experiment (MOPEX): An overview of science strategy and major results from the second and third workshops. Journal of Hydrology, 320, 3-17.

Fortin, V., 2000. Le modèle météo-apport HSAMI: historique, théorie et application, 68p. Institut de Recherche d'Hydro-Québec, Varennes, Canada.

Fortin, V. and Turcotte, R., 2007. Le modèle hydrologique MOHYSE (bases théoriques et manuel de l'usager). Note de cours pour SCA7420, Département des sciences de la terre et de l'atmosphère 17p. Université du Québec à Montréal, Montréal, Canada.

Goswami, M., O'Connor, K.M., and Bhattarai, K.P., 2007. Development of regionalisation procedures using a multi-model approach for flow simulation in an ungauged catchment. Journal of Hydrology, 333 (2–4), 517-531.

Granger, C.W.J. amd Ramanathan, R., 1984. Improved methods of combining forecasts. Journal of Forecasting, 3(2), 197-204.

Hansen B.E., 2008. Least-squares forecast averaging. Journal of Econometrics, 146(2), 342–350.

Hansen, N. and Ostermeier, A., 1996. Adapting arbitrary normal mutation distributions in evolution strategies: The covariance matrix adaptation. In Proceedings of the 1996 IEEE International Conference on Evolutionary Computation, pp. 312-317;

Hansen, N. and Ostermeier, A., 2001. Completely Derandomized Self-Adaptation in Evolution Strategies. Evolutionary Computation, 9(2), 159-195.

He, Y., Bárdossy, A., and Zehe, E., 2011. A review of regionalisation for continuous streamflow simulation, Hydrology and Earth System Sciences, 15, 3539-3553, doi:10.5194/hess-15-3539-2011

Hrachowitz, M., Savenije, H.H.G., Blöschl, G., McDonnell, J.J., Sivapalan, M., Pomeroy, J.W., Arheimer, B., Blume, T., Clark, M.P., Ehret, U., Fenicia, F., Freer, J.E., Gelfan, A., Gupta, H.V., Hughes, D.A., Hut, R.W., Montanari, A., Pande, S., Tetzlaff, D., Troch, P.A., Uhlenbrook, S., Wagener, T., Winsemius, H.C., Woods, R.A., Zehe, E., and Cudennec, C., 2013. A decade of Predictions in Ungauged Basins (PUB)—a review. Hydrological Sciences Journal, 58 (6), 1198–1255.

Hutchinson, M. F., McKenney, D.W., Lawrence, K., Pedlar, J. H., Hopkinson, R.F., Milewska, E., and Papadopol, P., 2009. Development and testing of Canada-wide interpolated spatial models of daily minimum-maximum temperature and precipitation for 1961-2003. Journal of Applied Meteorology and Climatology, 48, 725-741.

Krasnopolsky, V. M., and Lin, Y., 2012. A Neural Network Nonlinear Multimodel Ensemble to Improve Precipitation Forecasts over Continental US, Advances in Meteorology, 2012, Article ID 649450, 11p. doi:10.1155/2012/649450

McIntyre, N., Lee, H., Wheater, H., Young, A. and Wagener, T., 2005. Ensemble predictions of runoff in ungauged catchments. Water Resources Research, 41, W12434.

Merz, R., and Blöschl, G., 2004. Regionalization of catchment model parameters. Journal of Hydrology, 287, 95–123.

Minville, M., Brissette, F. and Leconte, R., 2008. Uncertainty of the impact of climate change on the hydrology of a Nordic watershed. Journal of Hydrology, 358(1-2): 70-83

Minville, M., Brissette, F., Krau, S. and Leconte, R., 2009. Adaptation to Climate Change in the Management of a Canadian Water-Resources System. Water Resources Management, 23(14): 2965-2986.

Minville, M., Krau, S., Brissette, F. and Leconte, R., 2010. Behaviour and Performance of a Water Resource System in Québec (Canada) Under Adapted Operating Policies in a Climate Change Context. Water Resources Management, 24, 1333–1352

Nash, J. E., and Sutcliffe, W. H., 1970. River flow forecasting through conceptual models: Part 1. A discussion of principles. Journal of Hydrology, 10(3), 282-290.

Neuman, S.P., 2003. Maximum likelihood Bayesian averaging of uncertain model predictions. Stochastic Environmental Research and Risk Assessment, 17(5), 291-305.

Oudin, L., Andréassian, V., Perrin, C., Michel, C. and Le Moine, N., 2008. Spatial proximity, physical similarity, regression and ungaged catchments: A comparison of regionalization approaches based on 913 French catchments. Water Resources Research, 44, W03413.

Parajka, J., Merz, R. and Blöschl, G., 2005. A comparison of regionalisation methods for catchment model parameters. Hydrology and Earth System Sciences, 9, 157–171.

Parajka, J., G. Blöschl, and R. Merz (2007), Regional calibration of catchment models: Potential for ungauged catchments, Water Resour. Res., 43, W06406, doi:10.1029/2006WR005271.

Parajka, J., Viglione, A., Rogger, M., Salinas, J.L., Sivapalan, M., and Blöschl, G., 2013. Comparative assessment of predictions in ungauged basins – Part 1: Runoff hydrograph studies, Hydrology and Earth System Sciences, 10, 375-409, doi:10.5194/hessd-10-375-2013.

Poulin, A., Brissette, F., Leconte, R., Arsenault, R. and Malo, J.S., 2011. Uncertainty of hydrological modelling in climate change impact studies in a Canadian, snow-dominated river basin. Journal of Hydrology, 409(3-4), 626-636. doi:10.1016/j.jhydrol.2011.08.057.

Raftery, A.E. and Zheng, Y., 2003. Discussion: Performance of Bayesian Model Averaging. Journal of the American Statistical Association, 98(464), 931-938.

Raftery A.E., Gneiting, T. and Bakabdaoui, F., 2005. Using Bayesian Model Averaging to Calibrate Forecast Ensembles. Monthly Weather Review, 133(5), 1155-1174.

Razavi, T. and Coulibaly, P., 2013. Streamflow Prediction in Ungauged Basins: Review of Regionalization Methods. Journal of Hydrologic Engineering, 18(8), 958–975.

Reichl, J. P. C., A. W. Western, N. R. McIntyre, and F. H. S. Chiew (2009), Optimization of a similarity measure for estimating ungauged streamflow, Water Resour. Res., 45, W10423, doi:10.1029/2008WR007248.

Shamseldin, A., O'Connor, K., and Liang, G., 1997 Methods for combining the output of different rainfall-runoff models, Journal of Hydrology, 197, 203– 229.

Sivapalan, M., Takeuchi, K., Franks, S.W., Gupta, V.K., Karambiri, H., Lakshmi, V., Liang, X., McDonnell, J.J., Mendiondo, E.M., O'Connell, P.E., Oki, T., Pomeroy, J.W., Schertzer, D., Uhlenbrook., S. and Zehe, E., 2003. IAHS Decade on Predictions in Ungauged Basins (PUB), 2003-2012: Shaping an exciting future for the hydrological sciences. Hydrological Sciences Journal 48: 857-880.

Smith, M. B., Seo, D.-J., Koren, V. I., Reed, S., Zhang, Z., Duan, Q., Moreda, F., and Cong, S., 2004. The distributed model intercomparison project (DMIP): Motivation and experiment design. Journal of Hydrology, 298, 4–26.

Velazquez, J.A., Anctil, F. and Perrin, C., 2010. Performance and reliability of multi-model hydrological ensemble simulations based on seventeen lumped models and a thousand catchments. Hydrology and Earth System Sciences, 14, 2303-2317.

Viney, N. R., Vaze, J., Chiew, F. H. S., Perraud, J.-M., Post, D. A., and Teng, J., 2009. Comparison of multi-model and multi-donor ensembles for regionalisation of runoff generation using five lumped rainfall-runoff models, 18th World IMACS Congress and MODSIM09 International Congress on Modelling and Simulation, Cairns, Australia.

Vrugt, J.A. and Robinson, B.A., 2007. Treatment of uncertainty using ensemble methods: Comparison of sequential data assimilation and Bayesian model averaging. Water Resources Research, 43, W01411, doi:10.1029/2005WR004838.

Zelelew, M.B., and Alfredsen, K., 2014. Transferability of hydrological model parameter spaces in the estimation of runoff in ungauged catchments. Hydrological Sciences Journal, 59 (8), 1470–1490. http://dx.doi.org/10.1080/02626667.2013.838003

Zhang, Y. and Chiew, F.H.S., 2009. Relative merits of different methods for runoff predictions in ungauged catchments. Water Resources Research, 45, W07412.

CHAPITRE 8

ARTICLE 6 : ANALYSIS OF CONTINUOUS STREAMFLOW REGIONALIZATION METHODS USING A REGIONAL CLIMATE MODEL ENVIRONMENT FRAMEWORK

Richard Arsenault[1] et François Brissette[1]

[1] Département de Génie de la Construction, École de technologie supérieure,

1100 Notre-Dame Ouest, Montréal, Québec, Canada H3C 1K3.

Article soumis à la revue « Hydrological Sciences Journal » en février 2015.

Abstract

The aim of this work was to analyze three common hydrological model parameter regionalization approaches (spatial proximity, physical similarity and multiple linear regression) and their limitations. To do so, 264 basins were modeled in a virtual-world setting, using a 15km resolution regional climate model to eliminate uncertainty due to measurement errors and missing data. This allowed analyzing the regionalization methods without the influence of uncertainty related to meteorological data quality and catchment descriptor estimates. The regionalization approaches were evaluated with a leave-one-out framework, effectively making 264 regionalization attempts. One to 10 donors were used during the process. The results were similar to those obtained in the real-world in a previous study, giving credence to the virtual world approach. It was shown that the physical similarity method outperforms the proximity approach and that averaging the outputs of multiple donors should be favoured. Inverse weighting distance should be preferred for the physical similarity method, while the simple arithmetic mean should be chosen for the spatial proximity method. It was also found that in many cases the best donor is neither the most similar nor the closest watershed to the ungauged site, indicating a need for better hydrologically relevant catchment descriptors. Furthermore, an analysis comparing the basins for which the regionalization methods worked the best and where they were the weakest was performed. It was found that the similarity distance between the donors and the ungauged sites was a strong indicator of regionalization skill. For the spatial proximity method, it was found that the closest donors worked well if they were similar, indicating that the proximity method is a good proxy only if there is reason to believe that the basins are similar. It was also shown that the ability to predict if a method will succeed or fail is limited by the quality of catchment descriptors and the inherent probabilistic nature of the problem. Finally, the virtual-world setting was shown to be a valuable tool for performing hydrological experiments which would otherwise be impossible to do. The unparalleled richness and quality of catchment descriptors from the virtual world was critical in assessing the reasons for regionalization methods' performance.

Keywords: prediction in ungauged basins, continuous streamflow, regional climate model, regionalisation, hydrological modelling

8.1 Introduction

Continuous streamflow prediction in ungauged basins (PUB) has been at the forefront of the hydrological sciences for decades, but has seen a revived interest after Sivapalan et al (2003) urged hydrologists to concentrate their efforts on the problem for the next 10 years. The community has seen non-negligible improvements during that time, but still today there exists an important gap in our ability to predict flows in ungauged locations (Wagener and Wheater 2006, Bao et al. 2012, Hrachowitz et al. 2013).

Various approaches have been tested, ranging from simply transposing the flows from an adjacent catchment and factoring for catchment size (McCuen and Levy 2000) to Monte-Carlo simulations with index-based constraints (Yadav et al. 2007). Each has their strengths and weaknesses, but the most promising and widely-used methods remain the hydrological model parameter regionalization approaches (Razavi and Coulibaly 2013). These methods can make use of multiple donor sites to extract the most information possible to increase the predictive skill on the ungauged basins.

Regionalization approaches can be categorized in three main classes. The first is spatial proximity, which aims to transfer the parameter set from the closest catchment and run the hydrological model on the ungauged site with these parameters (Vandewiele and Elias 1995). The distance is calculated as a Pythagorean distance between the latitude and longitude coordinates (Zelelew and Alfredsen 2014). The second, physical similarity, is analogous to the spatial proximity method except that the donor parameter set is selected from the most similar basin as calculated with a distance measure between catchment descriptors. The metric used in this study is taken from Burn and Boorman (1993), as shown in equation 8.1.

$$\Phi = \sum_{i=1}^{k} \frac{\left| X_i^G - X_i^U \right|}{\Delta X_i} \tag{8.1}$$

Where *i* is the catchment descriptor identifier, X^G is the catchment descriptor value for the gauged catchment, X^U is the catchment descriptor value at the ungauged catchment and ΔX is the range of possible values taken by X^G. The catchment that minimizes the difference in similarity index ϕ with the ungauged basin is used as the donor catchment (Bardossy 2007). Finally, the last category is the multiple linear regression approach. In this case, a regression model is built for each of the hydrological model parameters where catchment descriptors act as the model predictors. The regression model is then used to predict the parameter value at the ungauged site using its own observable catchment descriptors.

Past studies have shown that the multiple linear regression method is preferred only on arid and semi-arid basins (Parajka et al. 2013), whereas the physical similarity method is generally viewed as superior when a large number of donor basins and their catchment descriptors are available. The spatial proximity method is favoured for cases in which the catchment descriptors are lacking. Using the closest catchment as a donor implies assumptions about the soil and other physical characteristics being similar in the adjacent region (Shu and Burn 2003), McIntyre et al. 2005). Parajka et al. 2005, on the contrary, found that a proximity method based on parameter kriging slightly outperformed other regionalization methods. Later, Oudin et al. (2008), Zhang and Chiew (2009) and Samuel et al. (2011) showed that a combination of physical similarity and proximity outperformed the two independent approaches. This is understandable since the descriptors for which it is difficult to obtain measurements (soil or bedrock properties) should be more similar for adjacent catchments, all else being equal. In a previous work, Arsenault and Brissette (2014a) added the latitude and longitude of the basin centroid to the similarity distance calculation to integrate this hybrid method concept. Furthermore, in the same paper, a hybrid method was proposed in which the donated parameter sets were modified by the regression method, if and only if the regression model for the hydrologic model parameters was deemed as *good* (R^2>0.5). In such cases, the parameters for which the regression model performed well were replaced by the estimated parameter value. These new hybrid methods (regression-augmented proximity and regression-augmented similarity) outperformed their standalone counterparts, but to the cost of more uncertainty and less robustness. Arsenault and Brissette (2014a) also found the multiple linear regression method to vastly underperform when evaluated on 268 basins in Quebec, Canada.

Attempts to make use of multiple models in regionalization under a model averaging framework have also been proposed, with mitigated success. Goswami et al. (2007) showed that model averaging improved performance in calibration but the gains did not follow in validation mode. Viney et al. (2009) came to the same conclusion, but found that multi-donor averaging, rather than multi-model averaging, did increase performance significantly in validation. Arsenault and Brissette (2014c) showed that model robustness is a key factor in successfully predicting flows on ungauged basins in a multi-model averaging framework. The uncertainty on the limitations of the regionalization methods makes it difficult to effectively use the model averaging techniques.

One of the problems that is commonly expressed in trying to understand the limiting factors and the underlying mechanics of the regionalization methods (and why they sometimes either work well or fail miserably) is the quality of the climate and hydrometric data (Sellami et al. 2014) as well as the difficulty in getting consistent catchment descriptors. The uncertainty in the measurements is inherently reflected in the regionalization methods' performance. The meteorological observations are sparse, biased and are often riddled with missing data. The same is true for measured hydrometric time series. Furthermore, the meteorological data must often be homogenized at the catchment scale and lumped for modelling applications, which adds another layer of uncertainty. Another weakness related to observations pertains to the catchment descriptors, which are estimated on large scales but have a direct influence on hydrological response. Descriptors such as land cover use can change over time, whereas soil and bedrock properties are usually unknown or rough approximates.

The aim of this study is to use numerically generated data within a high-resolution (15km) reanalysis-driven regional climate model to analyze the regionalization methods and their limitations in an uncertainty-reduced framework. Additionally, in the proposed virtual-world setting, the catchment descriptors are numerous, perfectly well known and are directly linked to the hydrological response which is crucial for understanding the impact of such descriptors on the similarity-based regionalization methods.

8.2 Data and Methodology

8.2.1 Description of the virtual-world setting

The virtual-world setting is a numerical environment in which hydrological experiments can be performed with perfect knowledge of meteorological time series and physical basin characteristics. The virtual world consists in the combination of a Regional Climate Models (RCM) physical characteristics as well as its simulated data. The RCM used in this study is the Canadian RCM (CRCM, Caya and Laprise 1999) which was run on a 15km resolution grid allowing catchment-scale dynamics to be modelled. The physical processes and variables such as precipitation, temperature, radiation, wind, runoff depth and snow-water equivalent are all computed and archived at each grid point and for each time step. Since the processes rely on the CRCM characteristics (elevation, soil types and depths, land use, etc.) and its conservative laws of mass and energy balance, the CRCM offers a dense, coherent and complete database for a plethora of hydrologically relevant variables (Music & Caya, 2007, 2009; Music et al., 2009). The land and soil interactions are modelled with the Canadian Land Surface Scheme (CLASS, Verseghy et al. 1993). The virtual world results in variables that are physically coherent between themselves (e.g. precipitation, snowmelt and runoff) and with the CRCM characteristics (e.g. runoff, infiltration and soil type). Therefore the virtual-world setting offers a fertile ground for experimentation in a numerical environment free of data quality and quantity constraints of the real world. Regional climate models are becoming increasingly popular to perform hydrological experiments inside their virtual-world environments. Previous examples of such studies include Maraun (2012), Beauchamp et al. (2013), Arsenault and Brissette (2014b), Minville et al. (2014) and Lucas-Picher et al. (2015).

8.2.2 Meteorological data

The meteorological data used in the hydrological modelling aspect of this work was taken from the virtual-world setting.

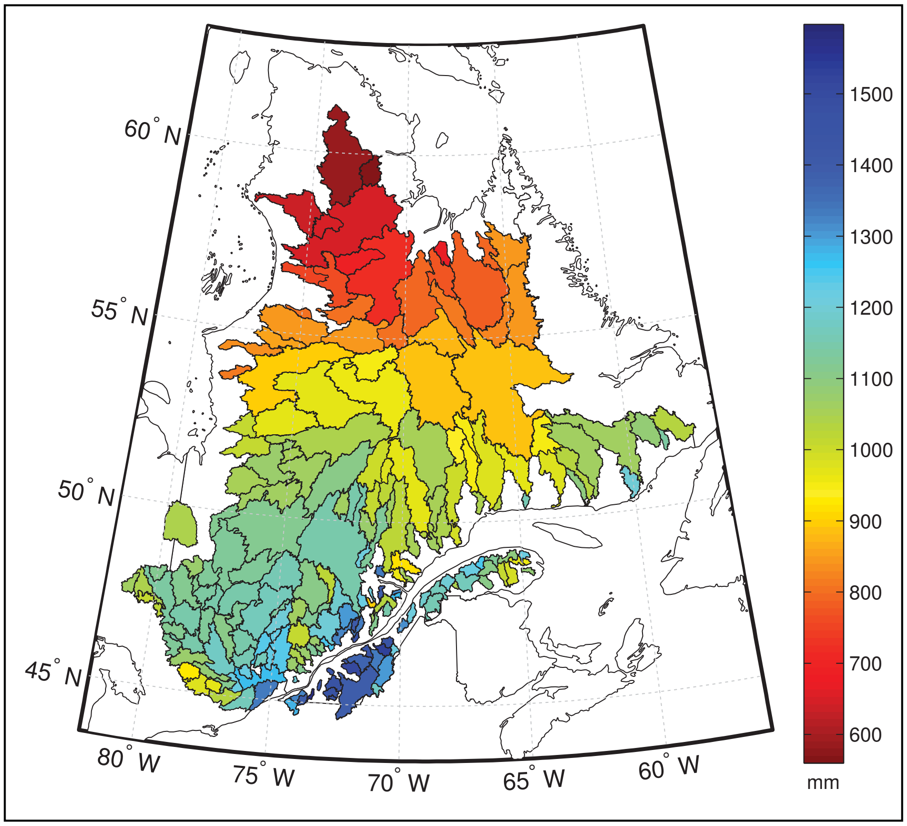

Figure 8.1 Geographical location of the 264 basins used in this study. The colors represent mean annual precipitation in the virtual world. There is a clear north-south gradient

Figure 8.1 shows the spatial representation of the 264 basins used in this study and their mean annual precipitation values in the virtual world. The basins were transposed in the virtual-world setting by taking the CRCM grid points that lied inside the real basins' boundaries. More precisely, basin-averaged daily maximum and minimum temperature, as well as daily precipitation (rain and snow), were introduced to the hydrological model for the study. There are no missing data in the entire simulation time span, which is 1961-2003.

8.2.3 Virtual-world setting hydrometric data

One of the problems encountered in the present study was the need for streamflow databases perfectly coherent with the climate data. This is required to isolate and analyse the regionalization methods' behaviours in a perfect context, which would then enable us to better understand the observed deviations. Since this situation is utopic in the real-world, a virtual copy of the study area was built using the CRCM as described in the previous section. As do other regional climate models, the CRCM generates runoff values and sub-surface water budgets for each grid node (each virtual climate station location). These runoffs are limited to each tile and are simply removed from the model after each time step as there is no runoff routing inside the CRCM. The solution to this caveat was to use an empirical flow routing scheme based on surface and sub-surface unit hydrographs to produce river flows at each of the basins outlets. Readers are referred to Arsenault and Brissette (2014b) for the complete methodology of the runoff routing scheme as it is out of the scope of this paper. It is important to note that this method produces routed flows that are coherent with the CRCM climate in terms of mass balance and approximate timing.

8.2.4 Catchment descriptors

Since the regionalization approaches were evaluated in a virtual-world setting, the catchment descriptors were readily available from the CRCM database. Table 8.1 shows the catchment descriptors used in this study as well as a few statistics describing their properties. Note that these properties are perfectly known within the virtual-world setting and that the governing physics are directly linked to them.

The slope has been omitted from this study because of the gridded nature of the basins in the CRCM environment. The slope would not reflect accurately on the hydrological processes due to the 15km resolution which is too coarse to estimate the slope. Also, many more catchment descriptors were available but were omitted due to them being perfectly correlated with other descriptors in table 8.1. For example, sand content was perfectly correlated with the porosity of the first soil layer.

Table 8.1 Catchment descriptors and basic statistics (minimum value, 25th, 50th and 75th percentiles and maximum value)

Catchment descriptor	**Units**	**Min**	**25th**	**50th**	**75th**	**Max**
Area	(km^2)	225	1350	3038	8888	72900
Elevation	(m)	37	300	373	452	864
Porosity of first layer of soil	---	0.37	0.37	0.37	0.39	0.49
Canopy rooting depth - coniferous	(m)	0.00	1.00	1.21	1.49	1.50
Canopy rooting depth - broadleaf	(m)	0.00	0.44	0.75	1.83	2.00
Canopy rooting depth - grass and swamp	(m)	0.00	0.02	0.08	0.14	1.20
Aerial fraction of canopy - coniferous	---	0.00	0.45	0.67	0.79	1.00
Aerial fraction of canopy - broadleaf	---	0.00	0.08	0.16	0.31	1.00
Aerial fraction of canopy - arable/crop	---	0.00	0.00	0.00	0.01	0.86
Aerial fraction of canopy - grass and swamp	---	0.00	0.00	0.04	0.12	0.83
Bedrock depth	(m)	0.10	1.85	3.02	3.02	3.02
Latitude	(degrees)	44.86	46.69	48.37	51.44	59.94
Longitude	(degrees)	-81.04	-75.01	-72.41	-69.66	-57.94
Snow duration	(days)	204	236	254	281	323
Aridity index (PET/P)	---	0.40	0.48	0.55	0.62	0.89
Actual ET /Precipitation	---	0.23	0.33	0.37	0.42	0.52
Mean annual precipitation	(mm)	560	1006	1095	1194	1598

8.2.5 HSAMI hydrological model

The HSAMI model (Fortin 2000; Minville 2008, 2009, 2010; Poulin et al. 2011) has been used by *Hydro-Québec* for over two decades to forecast daily flows on many basins over the province of Québec, Canada. It is a lumped conceptual model based on surface and underground reservoirs. It simulates the main processes of the hydrological cycle, such as evapotranspiration, vertical and horizontal runoffs, snowmelt and frost. Runoff is generated by surface, unsaturated and saturated zone reservoirs through two unit hydrographs: one for surface and another for intermediate (soilwater) reservoir unit hydrographs. The required inputs are spatially averaged maximum and minimum temperatures as well as liquid and solid precipitation depths. The model has up to 23 calibration parameters, all of which were used for this study.

8.2.6 Model Calibration

The HSAMI model was calibrated using the CMAES algorithm (Hansen and Ostermeier 1997, 2001) as it was shown that this algorithm outperformed other popular ones under the circumstances in this study (Arsenault et al. 2014). The Nash-Sutcliffe Efficiency (NSE) (Nash and Sutcliffe 1970) was selected for the calibration metric as it is generally accepted as an efficient measure of continuous streamflow simulation performance, even if it places more importance on peak floods. Also, the NSE is easily comparable throughout studies as it is the most widespread. The model was calibrated 10 times in order to evaluate the effects of equifinality in the regionalization environment. The 10 parameter sets could then be sampled randomly to estimate the sensitivity of the parameter set selection during regionalization following the work of Arsenault and Brissette (2014a). The HSAMI calibration results are presented in figure 8.2.

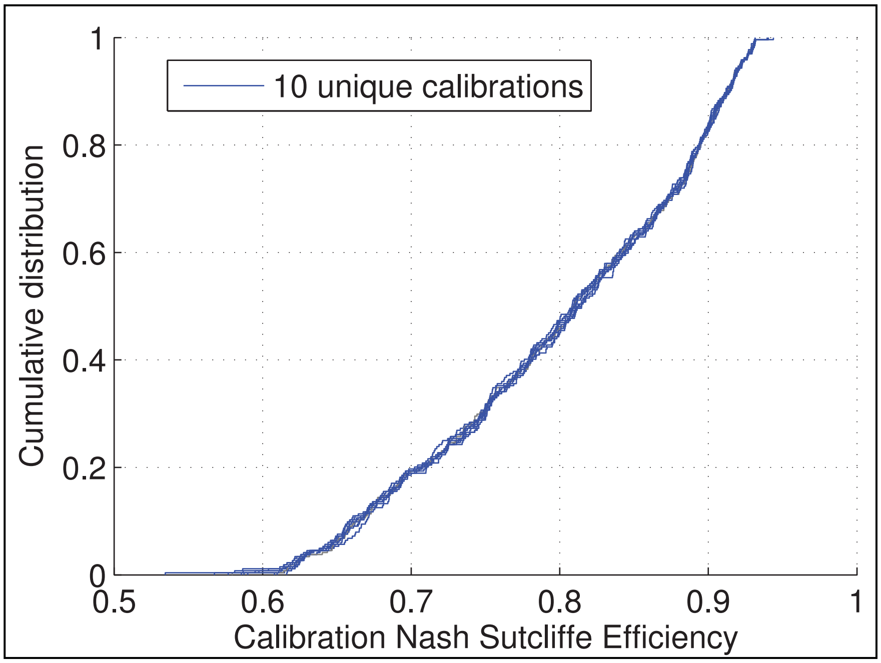

Figure 8.2 Cumulative distribution of the calibration NSE values for the 10 calibrations in the virtual world

The 10 calibrated model parameter sets show relatively good performance, with 80 percent of the catchments having a NSE value superior to 0.70. The tight spread also shows the extent of the equifinality (Beven 2006), in which the 10 different parameter sets allow for equivalent performance.

8.2.7 Regionalization methods

This study compares 3 regionalization methods: Multiple linear regression, spatial proximity and physical similarity. Furthermore, two hybrid methods (the regression-augmented similarity and proximity methods) developed previously (Arsenault and Brissette 2014) were tested. For all of these methods except for the multiple linear regression approach, it is possible to select more than one donor, usually in increasing order of distance from the ungauged basin. When multiple donor basins are selected, their parameter sets are transferred to the ungauged basin and the hydrological model is run with each set of these parameters. The resulting hydrographs are then averaged to generate a unique streamflow time series on the ungauged basin (Viney et al. 2009). With multiple donors, the weights of the donors can be considered equal using a simple mathematical average (SMA) or weighted according to the inverse of the distance (IDW). Another possible approach consists in averaging the parameter sets and modeling the end result. However, previous work has shown that averaging the parameter sets leads to poor performance, and thus this approach is not considered in this study (Oudin et al. 2008, Arsenault and Brissette 2014a).

8.2.8 Methodology

The entire study was performed in the virtual-world setting. This allowed controlling the climatic and hydrometric time series as well as the catchment descriptors to evaluate regionalization methods in an uncertainty-reduced environment. The first step was to calibrate the HSAMI model on the CRCM routed flows and virtual-world meteorological data. Then, each of the virtual-world basins was selected in turn to act as the ungauged basin. The pseudo-ungauged basin was removed from the contributing pool and the regionalization approaches were applied using one to ten donors in this leave-one-out framework. One hundred realisations were conducted in each case in order to sample the parameter sets produced under equifinality durning the model calibration process. A control group was also generated by selecting donors at random in order to evaluate the gain made by the regionalization methods versus a random selection. The Nash-Sutcliffe Efficiency metric was calculated on the ungauged basins to estimate the performance of the regionalization approaches.

Finally, groups of high (and low) performing basins were analyzed to identify key differences in order to try and understand the underlying mechanics using statistical tests. The main statistical test employed in this paper is the non-parametric Mann-Whitney (Wilcoxon 1945) test in which the null hypothesis is that the groups come from the same distribution.

8.3 Results

The main results of this study are summarized as NSE values at the ungauged sites obtained with the various regionalization approaches. The NSE metric was used since it is arguably the most widely known and used metric, which makes using it a requirement to compare results to other studies (Parajka et al. 2013). Figure 8.3 shows the NSE values for the regionalization methods using the IDW approach. The regression method is not donor dependent and thus is constant in the four panels for comparative purposes. The NSE value reported here is the median value of the 100 resamplings.

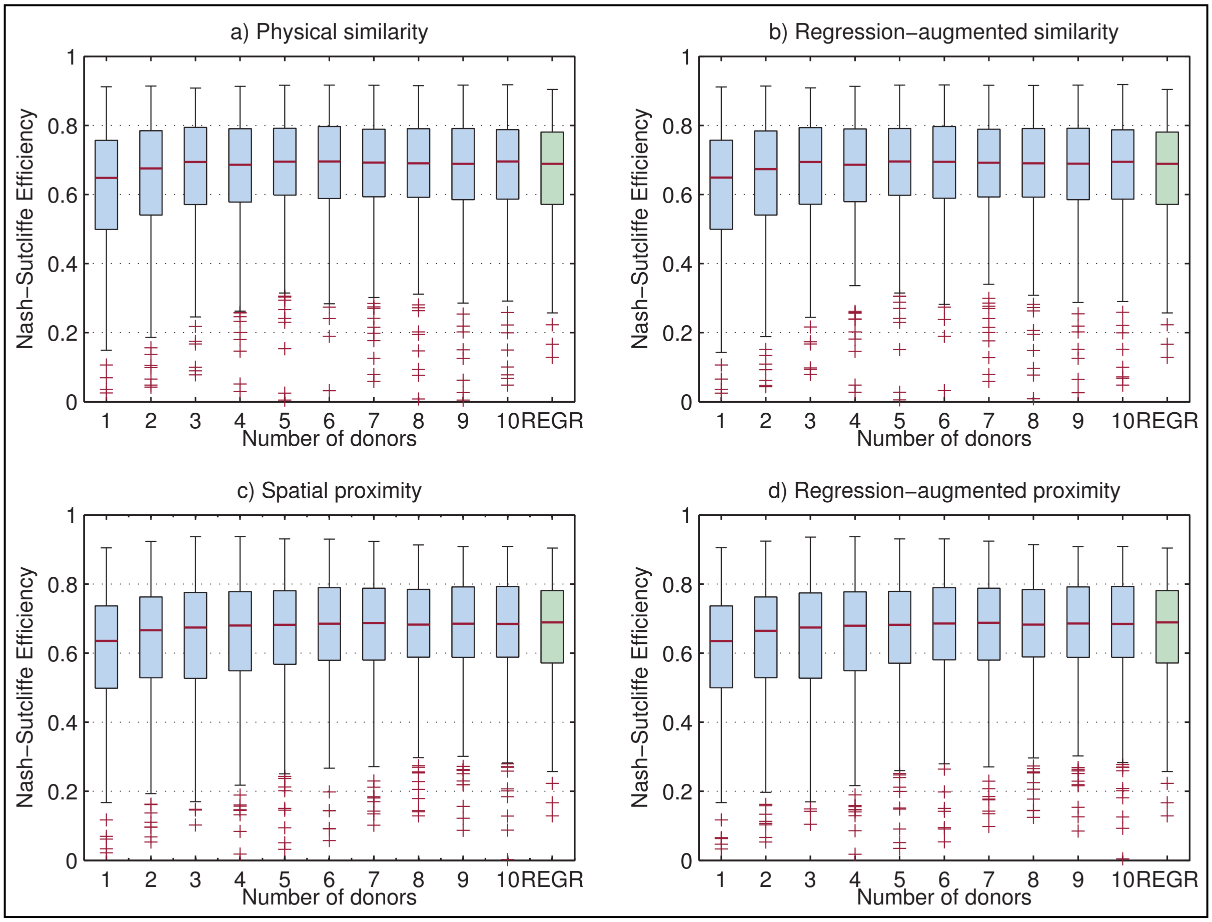

Figure 8.3 Median NSE values in regionalization on the 264 ungauged basins for 1 to 10 donor basins. The regression (REGR) results are constant in the four panels

The results show that the similarity methods are slightly better than the proximity methods; however the difference is not statistically significant as measured by a Kruskal-Wallis (Kruskal and Wallis 1952) test between each of the groups. Also, when multiple donors are used, the regionalization approaches tend to converge to a maximum limit that is slightly better than the multiple linear regression method, which is significant at the 95% confidence level.

The box-and-whisker plots in figure 8.3 are aggregates of large amounts of results and do not convey the entire picture. A second analysis was performed on the results, this time using a threshold method defined as the Success Rate (SR). The SR is the number of successful regionalization applications divided by the total number of trials. For example, 132 successful cases out of the 264 trials would lead to a 0.5 success rate (132/264). A "successful regionalization application" is defined to be a case in which the regionalized NSE in validation is equal or higher than 85 percent of the calibration NSE on the ungauged basin. This threshold was taken from previous work where it was found to be a good compromise (Arsenault and Brissette 2014a). A higher threshold (>0.9 for example) could leave little room for success, while a lower threshold (<0.8) can be too easy to attain.

The success rate allows for a better understanding of the regionalization methods performance as the validation results are compared to a baseline (the calibration NSE) rather than being independently evaluated. Furthermore, it allows comparing methods for consistency as a higher success rate means more catchments are adequately modelled.

The success rate was computed for each of the regionalization runs. The 100 iterations performed to estimate the effects of parameter set selection thus lead to 100 SR values for each regionalization approach. Figure 8.4 shows the SR results for 1, 5 and 10 donors. Note that the multiple linear regression method is identical in all cases as it uses all available information, making the donor basin concept irrelevant. Also note that the y-axis is different in panel a) than in panels b) and c) for display clarity.

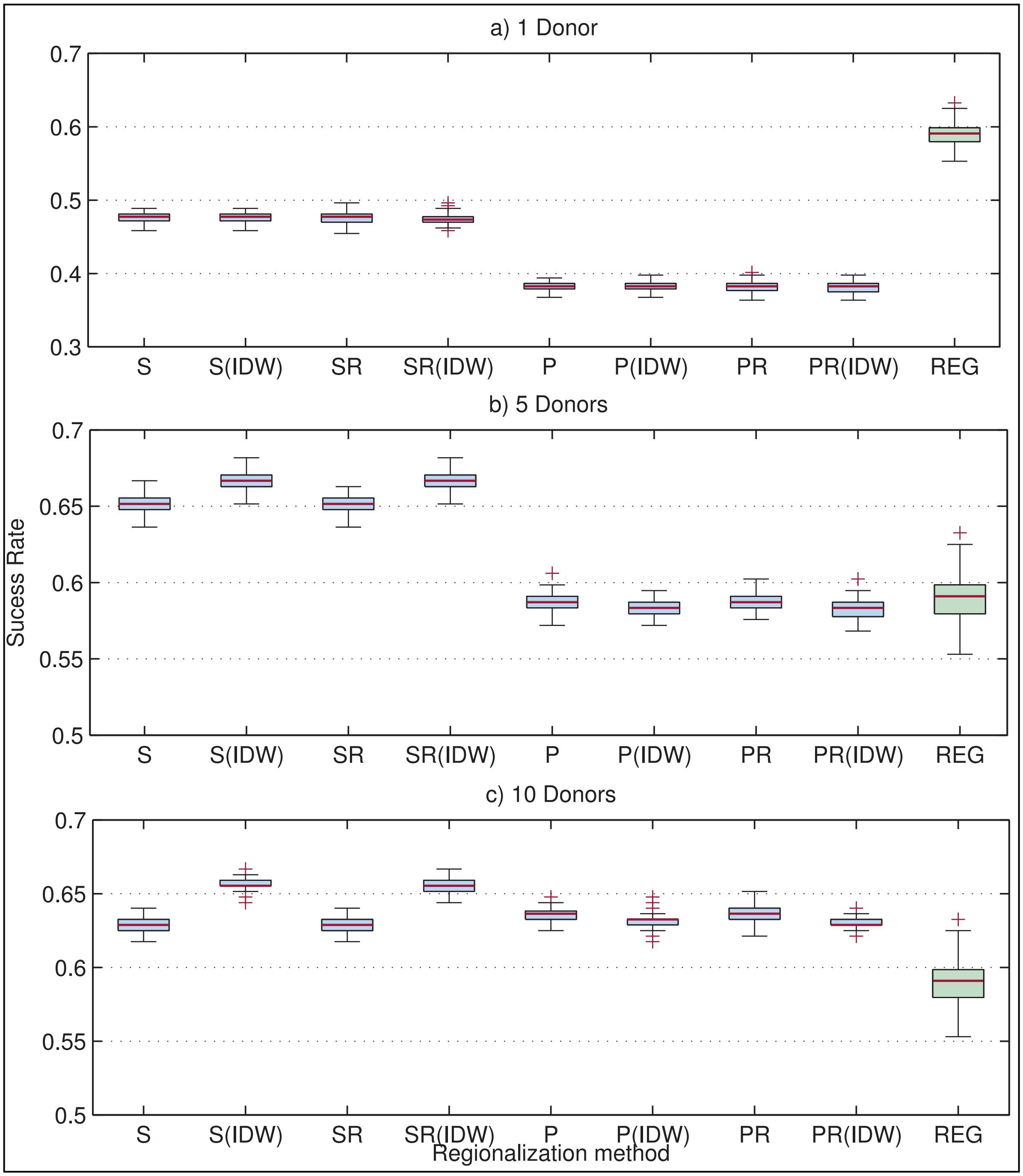

Figure 8.4 Success rates for the Similarity (S), Proximity (P), Regression-augmented similarity (SR) and proximity (PR) and Regression (REG) methods. IDW represents the inverse distance weighting of donor outputs, others are simple arithmetic mean. Panels a), b) and c) are for respectively 1, 5 and 10 donors

It can be seen that the regression method, which does not rely on donor parameter sets but on linear regression models based on the donor basins characteristics, performs well compared to the other methods. When a single donor is used with the other methods (figure 8.4 a), it outranks them completely. However, with more donors, the physical similarity and proximity methods are able to surpass the regression method.

For the similarity methods, the IDW averaging outperformed the SMA method, especially when more than 2 donors were used. This is to be expected if the donor basins are progressively less able to predict the target flows accurately, as the IDW method reduces the bad basins weights the further they are from the target basin. This is also consistent with the literature (see Oudin et al. 2008). As for the proximity-based methods, the simple mean seems to perform better than the inverse weighting with more donors (figure 8.4c). This suggests that the closest donors (centroid-to-centroid) are not necessarily the optimal donors for the ungauged basins, as using progressively farther basins as donors improves the performance more when they have larger weights, as opposed to the progressively lower weights for the IDW method. The discrepancy between the proximity and similarity methods' performance with SAM and IDW can explain the fact that the similarity methods are better than the proximity methods when fewer donors are used.

It is also noteworthy that the physical similarity method shows a slight performance drop between 5 and 10 donors. The IDW variants drop significantly less than the SAM variants, thus demonstrating another advantage of using IDW: performance drop caused by more donors than the optimum is mitigated by the progressively lower weights. As for the spatial proximity methods, they gain skill with added donors, enforcing the suspicion that the further donors are helping the performance rather than reducing it. However, it is clear that using multiple donors is required to extract the most information from the regionalisation methods. Therefore the first donor's performance is crucial for the regionalisation method's performance.

The regionalization methods' performance was then compared to randomly selected donors as a baseline. Results are shown in figure 8.5.

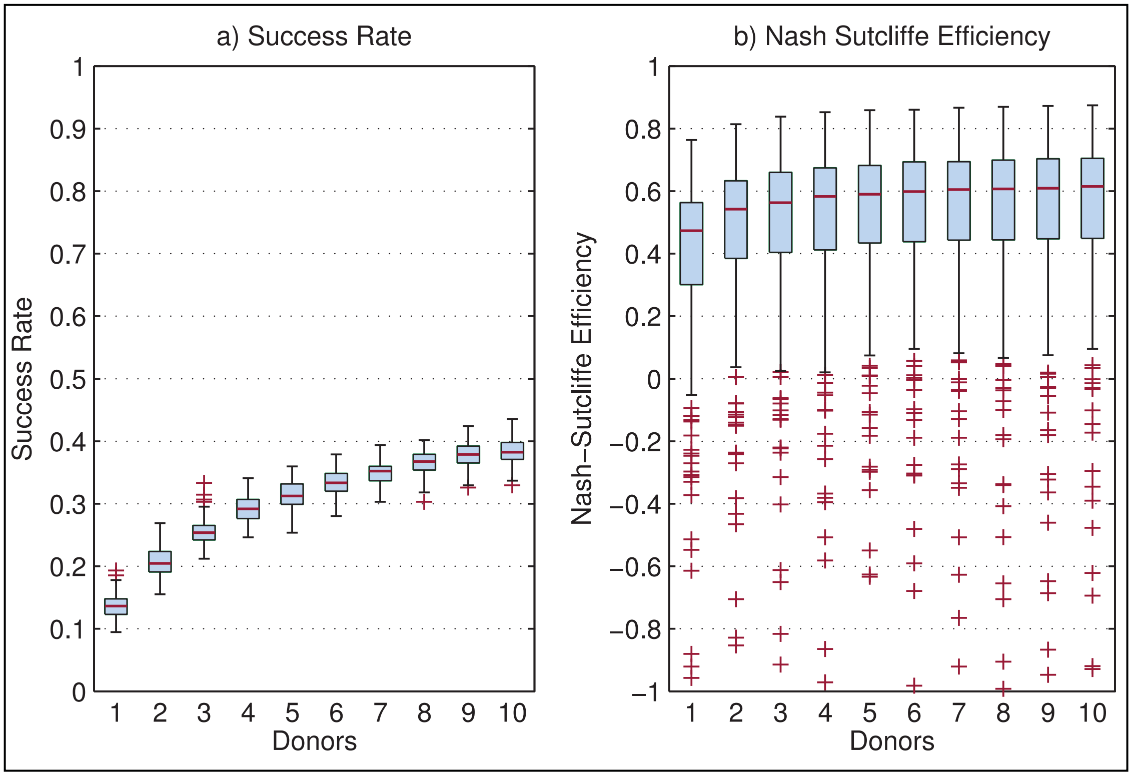

Figure 8.5 Success rate and Nash-Sutcliffe efficiency using random donors, for 1 to 10 donors

Several interesting observations can be made from figure 8.5. First, the SR climbs with added donors, even though the donors are randomly selected. This tends to prove the fact that model output averaging is absolutely necessary to get the most out of the regionalization methods. Indeed, the performance increase is attributed to the added value of multiple donors which, when averaged, cancel their respective errors out and produce a better simulation. Second, the NSE values lag the regionalization methods performance by approximately 0.10, which is significant. However this is more of a testament to the robustness of the HSAMI model, which is known to perform better than other models in such situations (Arsenault and Brissette 2014c). Finally, these results confirm that the random selection performs generally worse than the regionalization methods, meaning that the latter are able to make use of the available information as anticipated in the CRCM environment. The SR and NSE values are comparable to the results obtained in the real world (Arsenault and Brissette 2014a).

A final analysis of the results was undertaken to explore the relationship between the NSE value in regionalization and the calibration NSE value on the ungauged basin. The Nash Ratio (NR) was defined as the ratio of regionalization NSE to the calibration NSE at the ungauged site. A score of 1 would imply that the regionalization NSE is equivalent to the calibrated NSE. The success rate defined previously would include basins with a Nash Ratio at least equal to 0.85. Figure 8.6 shows the Nash Ratio and distance metric to the closest donor for each of the ungauged basins. The similarity method uses the similarity distance, while the proximity method uses the geographical distance.

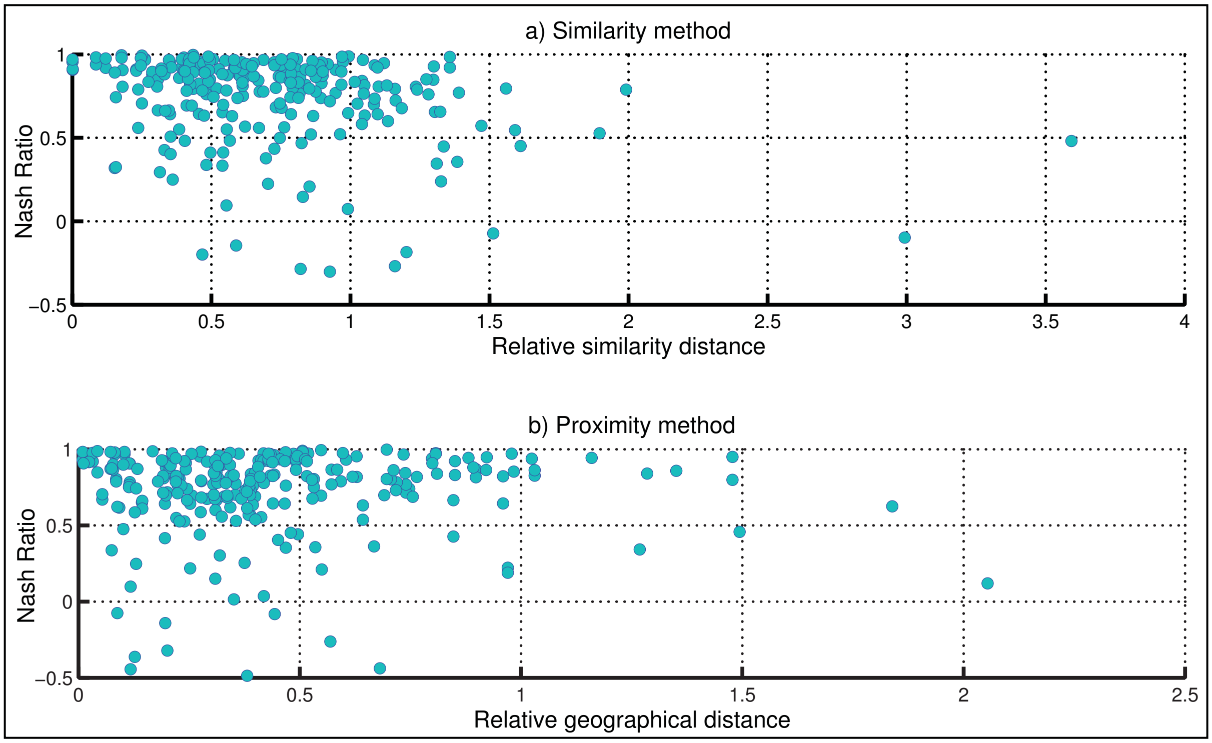

Figure 8.6 Nash Ratio in regionalization on 264 basins and their respective distance to their closest or most similar donor

From figure 8.6, it is clear that the ability to predict streamflow to ungauged sites is somewhat correlated to the distance to the first donor. It is also important to note that some catchments have poor Nash Ratios (<0.85) even though the distance to their closest donor is small. A test was devised to determine the sensitivity of the catchments to different donor sets in regionalization. Figure 8.7 presents the NSE values of 263 donor basins on nine randomly selected ungauged basins using the similarity approach.

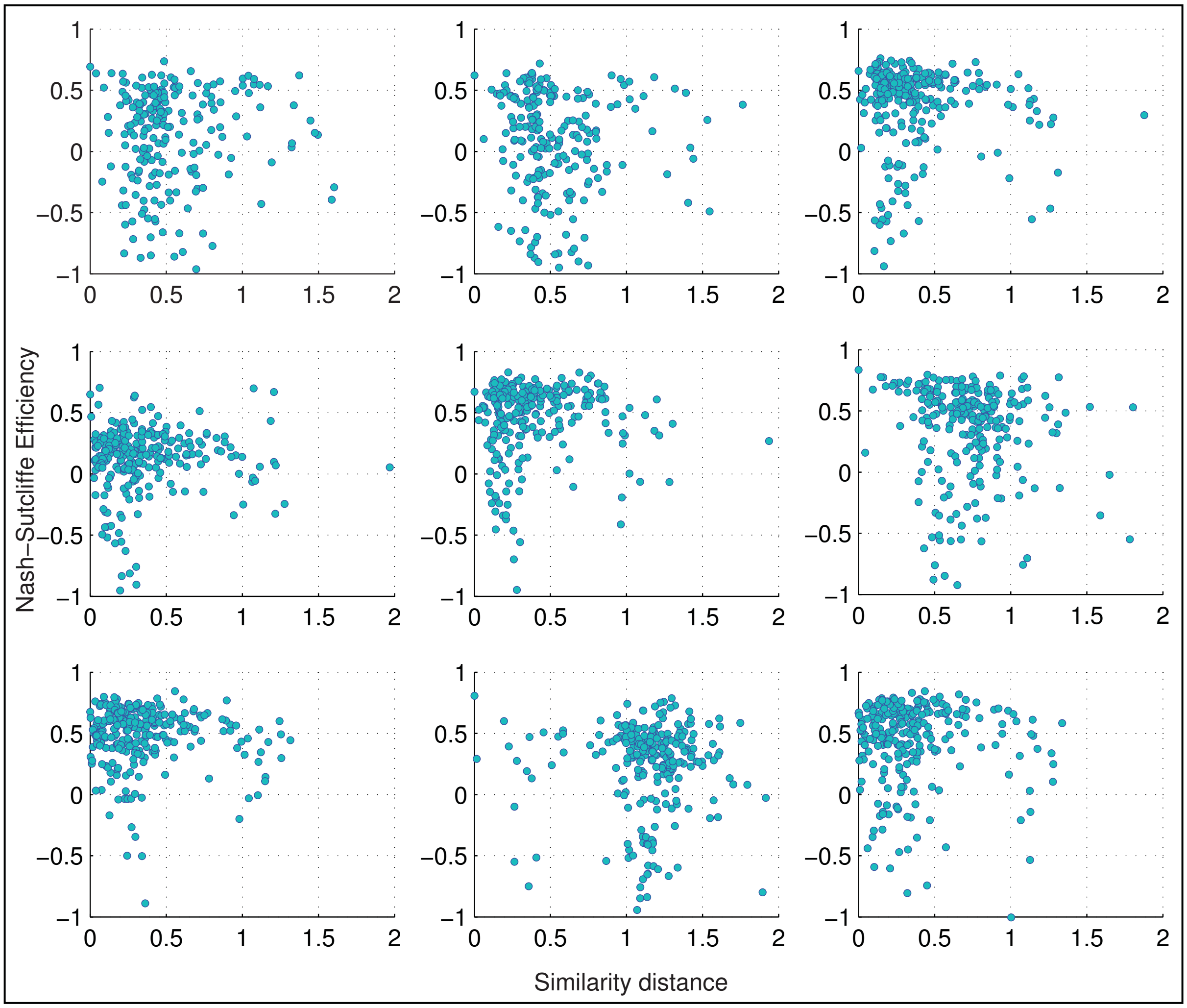

Figure 8.7 NSE value in regionalization when the 263 gauged basins are used as donors in the similarity regionalization approach. The NSE value is plotted against the similarity distance to the donor basins. Each panel represents a randomly selected ungauged basin

It can be seen in figure 8.7 that the similarity distance between donors and the ungauged basins does not seem to be the only determining factor in regionalization performance. In fact, in many cases, the best donor is far from being the most similar. Finally, the best donor for each catchment was found and the distance from this donor to the ungauged site was measured. Results are presented in figure 8.8.

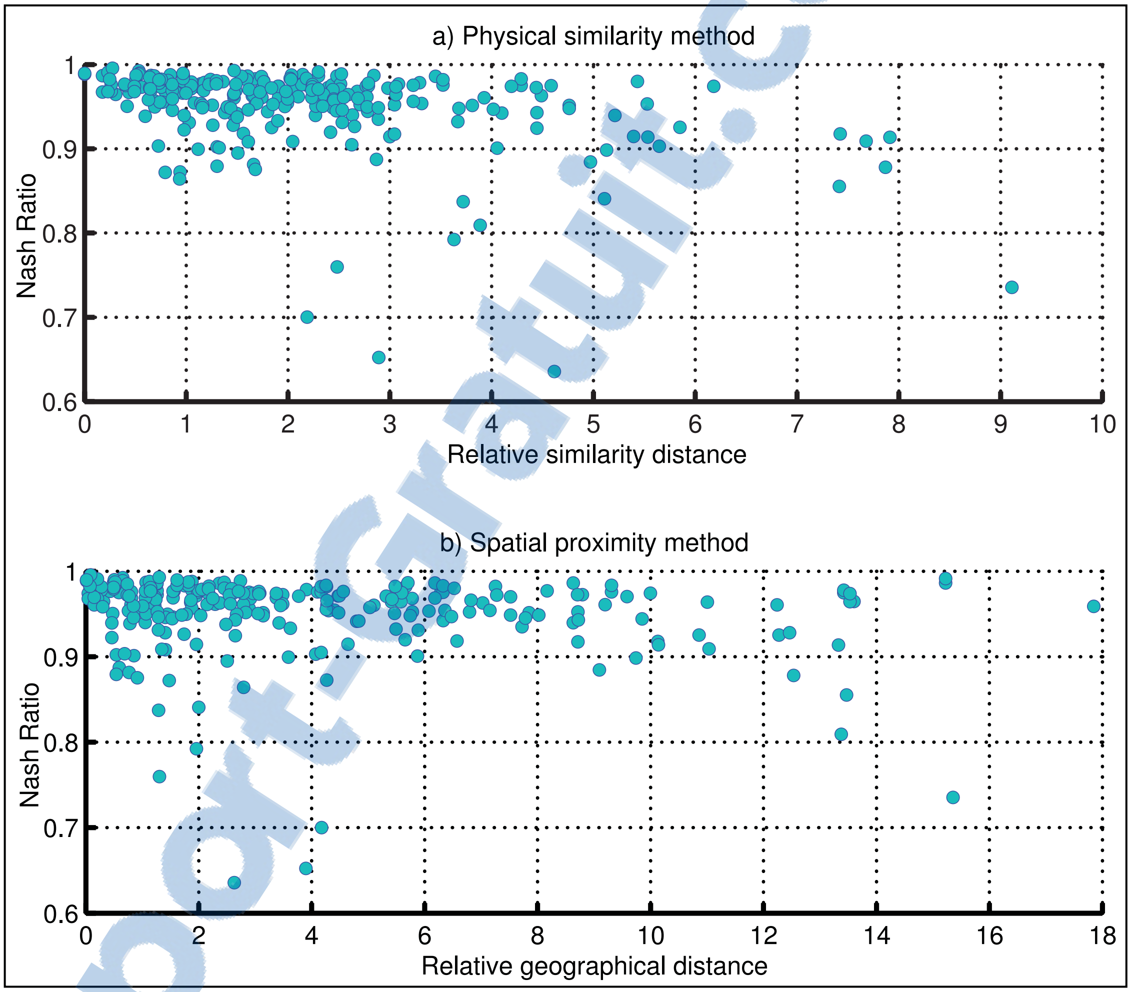

Figure 8.8 Nash Ratio in regionalization on 264 basins and their respective distance to their donor which returns the best NSE value

In figure 8.8, it can be seen that the best donors (those which result in the best regionalization performance as measured by the Nash Ratio) can be very dissimilar. However there is a slight downwards trend in performance with increasing distance, suggesting that distance does play a role in regionalization performance. Furthermore, it can be seen that for some basins, even the best donor results in poor Nash Ratios.

The next step was to compare the basins in which the regionalization methods performed well and those where they failed. To do so, the 50 catchments with the largest NR with their closest donor were grouped into the "good basin" group, whereas the 50 catchments with the lowest NR were grouped into the "bad basin" group. One set was generated for the similarity method and another for the proximity method. The number of basins per group (50) was selected to eliminate the 150 average performing basins to better distinguish differences between the groups, while keeping a large enough database to infer statistically significant findings. The first comparative analysis looked at the distance metrics between the good and bad basin sets. Figure 8.9 shows the distance to the 10 closest donors in each case.

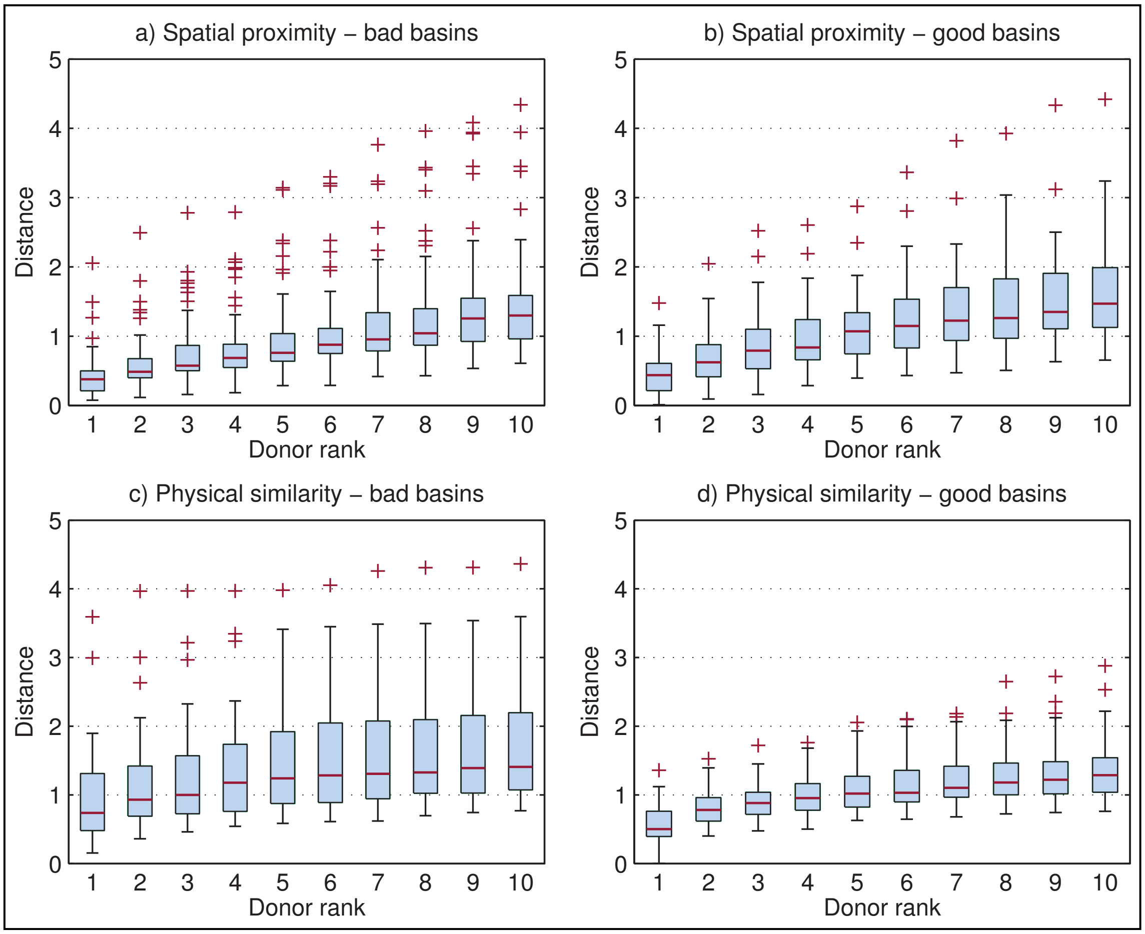

Figure 8.9 Comparison of distances to the ungauged basins' donors for the good basins group and the bad basins group. Panels a) and b) represent the proximity method and geographical distance, whereas panels c) and d) represent the similarity method and similarity distance

Two findings are immediately apparent in figure 8.9. The first is that the physical similarity method performs better when the donors are more similar. In cases where the most similar donor has a larger distance metric value, the probability of it being a bad basin for regionalization purposes increases.

Second, the distances are similar for both the good and bad basins for the proximity method. This is unsurprising since the donor catchments are only determined by their geographical location. Therefore any particular catchment descriptor which would allow selecting a better donor is not used. This indicates that the catchment descriptors play an important role in regionalization. In fact, basins are more likely to perform well if the closest basin is the most similar, as shown in figure 8.10.

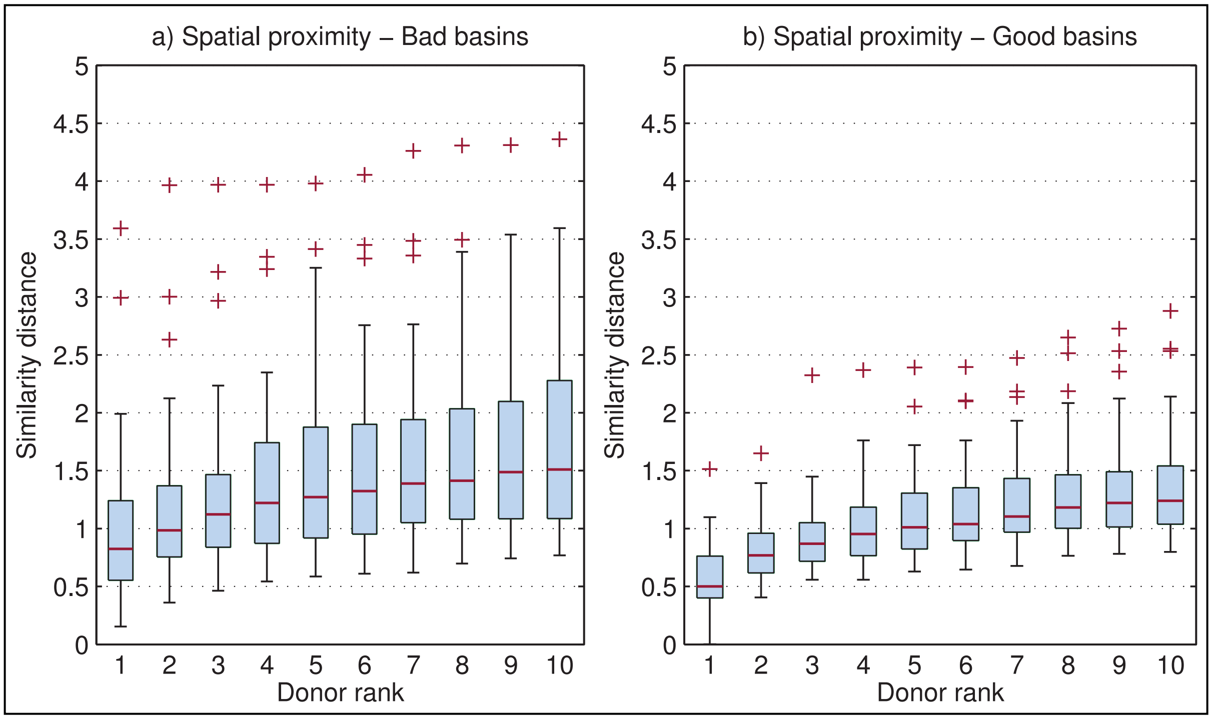

Figure 8.10 Physical similarity distance between good basins and bad basins using the spatial proximity approach

Here the similarity distance was measured between the good and bad basins as classified by the proximity method. It is clear that the proximity method works better when the donor basins are more similar to the ungauged target, which points to the physical similarity to be the underlying cause for the proximity method's success. This is the founding hypothesis of the proximity method. Figure 8.10 then confirms that spatial proximity methods can be useful if there is enough evidence pointing to the fact that the donor catchment is similar to the ungauged site.

The next step was to try and identify the most impactful catchment descriptors in an attempt to predict which ungauged catchments could be successfully regionalized. To do so, the similarity distances to the first donor (most similar basin for the physical similarity method and closest basin for the spatial proximity method) were analyzed by separating them into their individual descriptors. This allowed comparing the distances between each of the descriptors instead of the aggregated measure. Figure 8.11 presents the results for the good and bad basins for the proximity and similarity methods. Note that even for the spatial proximity method, the physical catchment descriptors were analyzed as was done in figure 8.10.

Figure 8.11 confirms the findings from figures 8.9 and 8.10, which are that the good basins are more similar to their donors than the bad basins are since the individual descriptor distances are smaller in almost all cases. However, figure 8.11 allows a deeper analysis in order to see which catchment descriptors are more important. For a few catchment descriptors (2, 3, 4, 6, 14, 15, 16 and 17 in particular), the difference is statistically significant. There is one caveat, which is the fact that there is no complete dissociation between the groups. In other words, the distributions for each of the catchment descriptors overlap each other in such a way that it is not possible to identify a threshold to predict if a regionalization method will perform well based on any individual catchment descriptor.

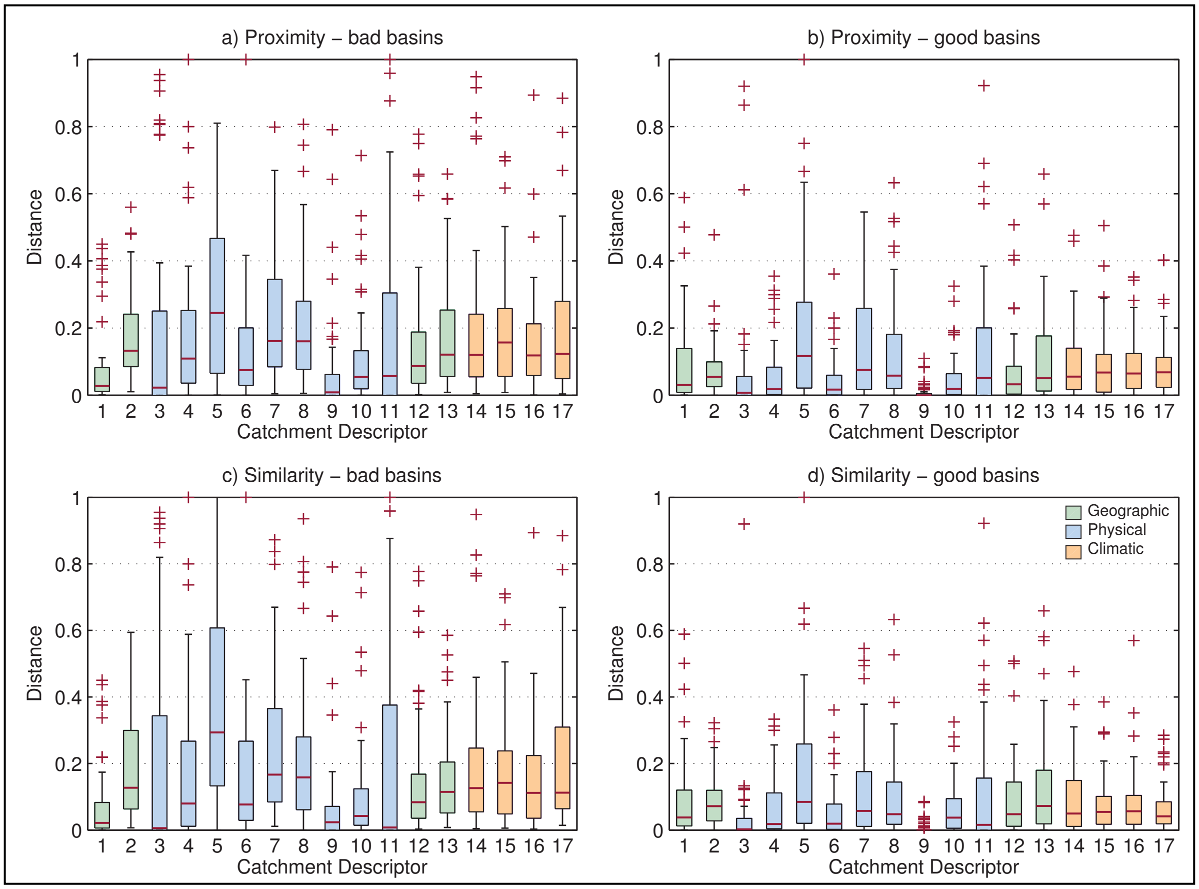

Figure 8.11 Breakdown of the similarity distance measure components to their individual catchment descriptor distances. Panels a) and b) respectively represent the bad and good basin sets for the proximity method, whereas panels c) and d) represent the bad and good basins for the similarity approach

A final test was performed to attempt to categorize the groups using a linear discriminant analysis (LDA) (McLachlan 1992). An LDA identifies the (n-1) dimensional surface (or hyperplane) that separates two groups the most efficiently based on the catchment descriptors. However, in this project, the LDA failed to produce an acceptable 16-dimension hyperplane. All possible combinations of catchment descriptors were used to identify a potentially optimal CD set which would allow efficient regionalization analysis. However the results show that in the best outcome, only 60% of cases could be accurately classified at the 95% confidence level. The groups were nonetheless separated by the best catchment descriptor hyperplane and the empirical probability of them being in the good groups was measured. Figure 8.12 shows the empirical probability that the basins in each group are a "good basin".

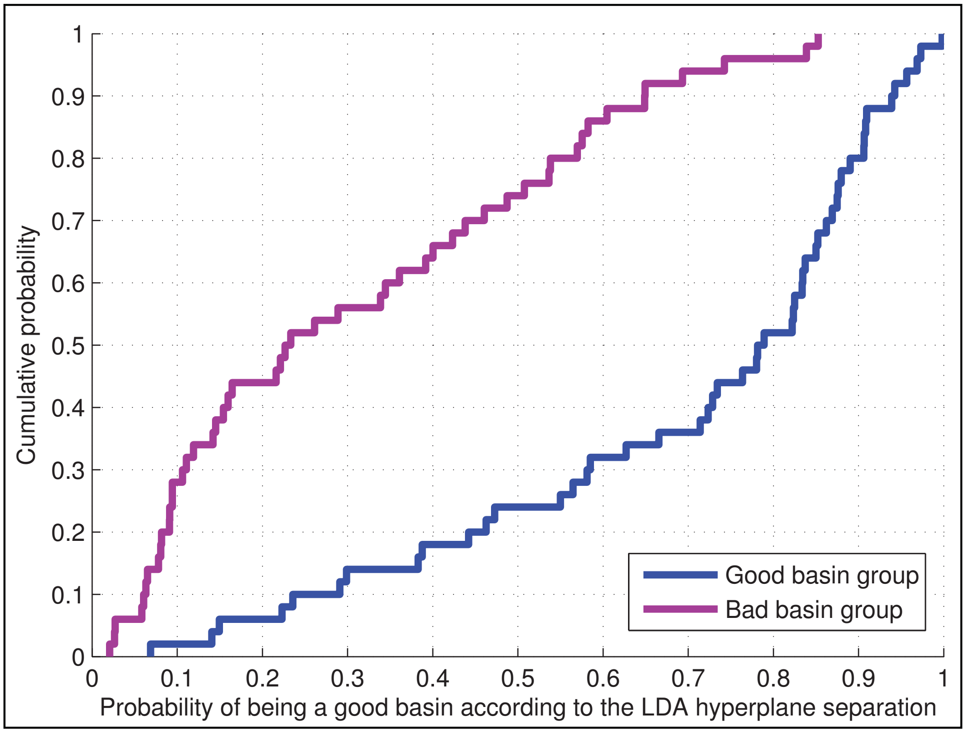

Figure 8.12 Empirical cumulative distribution of the probability that the basins are in the "good basins" group

From figure 8.12, it is clear that the best separation of the two distributions is still imperfect. A strong split would have the bad basin group entirely within the [0 0.5] range and the good basin group entirely within the [0.5 1] range. A perfect fit would have a probability of 0 for the 50 basins in the bad group and a probability of 1 for the 50 basins in the good group. In the case at hand, 20% of the good basins are more likely to be in the bad group after the LDA separation. The opposite is also true, with 30% of the bad basins more likely to be in the good group. These results are similar with the proximity method although the difference between the groups is much smaller.

From these results, it seems clear that the best possible separation cannot identify with certainty *a priori* which ungauged basins will positively respond to a regionalization application.

8.4 Analysis

8.4.1 Real world and CRCM environment

One of the most important aspects of this paper is the fact that the experiments were conducted in a virtual environment based on the CRCM's simulation data. The reasoning behind this was that the study area is very poorly characterized from meteorological and geophysical points of view. The sparseness of meteorological stations on the study area, combined with biases and missing data, makes it difficult to adequately analyze the regionalization methods due to the uncertainty in the input data. As for the geophysical data, soil depth and composition, hydraulic conductivity, bedrock fractures and ground cover, for example, are either approximated by coarse satellite data or simply unknown at the catchment scale. On the other hand, these properties are exactly prescribed in the CRCM's land surface scheme. Whether or not they accurately represent the real-world characteristics is irrelevant in this study, although they are derived from global maps. And since the runoff depths are perfectly correlated to the soil characteristics in the CRCM, the uncertainty due to catchment descriptors was eliminated.

One trade-off that must be done in the virtual-world is the runoff routing. Ideally, the regional climate model would produce streamflow values at basin outlets. However this is a difficult task due to scaling issues, namely the model's 15km resolution. Instead, a parameterized unit hydrograph system is used to convert runoff depths to outlet streamflow. This necessarily leads to a certain amount of filtering. However, the hydrologic model was able to perform well under calibration and validation with the routed hydrographs, thus ensuring that even if the hydrographs are slightly different in the CRCM environment than in the real world, the general hydrologic processes are physically coherent and representative of those in the real-world. This approach provides a dataset richness very difficult to attain in the real world, from both geophysical and hydrometeorological standpoints. Furthermore, it allows extending the time series to the length of the simulation, which was 43 years, as compared to the real-world time series which vary in length and are often much shorter. The uncertainty due to short observation periods and missing data are thus eliminated here also.

There are also limitations in using the CRCM environment for hydrological purposes. First, the routing can reduce the day-to-day variability due to short rainfall periods and low rainfall depth as the parameterization of the routing algorithm naturally favours the peak flows. For regionalization purposes, this is acceptable since the NSE metric also focuses on peak floods. More importantly, regionalization is useful to determine long-term hydrologic regimes, not exact daily values. Therefore the low hydrograph amplitude in short or small rainfall events is not necessarily critical as long as the mass balance is met over a medium duration window.

Second, the 15km resolution limits some of the hydrologic processes precision. For example, many short, high-intensity convective storms mainly occur under the 15km resolution. For distributed models, this could be problematic since the 15km grid would assume the storm cell covered the entire area (225 square kilometers). Similarly, urban watersheds are too small to be appropriately modeled in the CRCM's current resolution.

However, the conclusions in this study seem to demonstrate that the regionalization methods perform similarly in the uncertainty-reduced virtual-world and in the real world (see Arsenault and Brissette 2014a for the real-world analysis).

8.4.2 Analysis of the methods performance

The first results, seen in figure 8.3, show the Nash-Sutcliffe Efficiency for the regionalization approaches. At first glance, the methods seem to produce almost identical results, in which added donors contribute to improving the regionalization skill. A closer look reveals small differences between the proximity and similarity groups, with the similarity groups being slightly better than the proximity variants. However, the Success Rate metric in figure 8.4 showed that there are indeed larger differences hidden in the box-and-whisker plots. By categorizing the data in this manner, it is clear that the similarity method is more robust since it is able to maintain its performance on a larger number of catchments than the proximity method. From figures 8.3 and 8.4, it is also clear that the regression-augmented variations of the similarity and proximity methods did not contribute to improving simulation performance. This is largely due to the fact that there were only a few occasions where the regression model produced good coefficients of determination which warranted a parameter value modification.

Indeed, the regression method only found poor correlations between the catchment descriptors and the hydrological model parameters. This was expected since the hydrological model is known to be overparameterized and its parameters are interdependent, and is consistent with the literature (Seibert 1999; Merz and Blöschl 2004; Lee et al. 2005). The calibrated parameter sets can therefore take many values and still perform adequately due to equifinality. Furthermore, the relative homogeneity of the catchment descriptors in the virtual environment makes it difficult for the regression models to find strong correlations. This is why the regression method, while it outperformed the other methods using a single donor, was not investigated further. It would have been interesting to understand in which cases the regression method worked well if it had been one of the best methods, however from figure 8.4 it is clear that the similarity and proximity methods outperform it when multiple donors are used.

Another finding in figure 8.4 was that the Inverse Distance Weighting (IDW) was beneficial to the similarity method but detrimental to the proximity method. The reasoning was that the similarity method had better first donors than the proximity method, which makes the similarity method start strong and stay ahead as the added donors are weighted progressively lower. In the case of the proximity method, the main driver of performance increase is the fact that the donors further from the ungauged basin can be more similar than the closest ones. Thus some of the far donors bring better information in the flow averaging. This dynamic favors the unweighted mean as the IDW strongly favours the closest donor, which we showed to be problematic for the proximity method in figures 8.9 and 8.10. Furthermore, as was seen in figure 8.5, the simple fact of adding donors improves regionalization skill through the model averaging concept (Diks and Vrugt 2010). The added value of the progressively more distant donors should be taken advantage of with the simple arithmetic mean of the model outputs for the proximity method.

Finally, the choice of climatic descriptors was made to include the most common for comparative reasons, such as mean annual precipitation and aridity index (He et al 2011). Furthermore, all available physical characteristics were selected. However, certain properties were perfectly correlated with one another. For example, rooting depth of coniferous trees was perfectly correlated with soil depth. This appears to be because CRCM soil and ground cover databases were approximated and any unknowns were linked to other descriptors. The perfectly correlated variables were removed from the study to reduce the problem dimensionality. In the end, only 17 descriptors remained from the original list of 35.

Interestingly, the latitude and longitude descriptors (number 12 and 13 in figure 8.11) are not determining factors for the similarity method, and the difference is minor for the proximity method. The latter is expected as the physical distance was the main criteria to separate the groups. For the similarity method, however, the current literature is practically unanimous in that a combination of proximity and similarity is ideal (Zhang and Chiew 2009). Perhaps their weight is diluted amongst the 15 other descriptors and thus are not as meaningful. For example, in previous work (Arsenault and Brissette 2014a) only four descriptors were necessary to optimize the regionalization performance, with the latitude and longitude being part of the selection. Also, from figure 8.11, it can be seen that the most critical descriptors are the mean annual precipitation, aridity index, actual evapotranspiration to precipitation ratio, elevation, soil porosity and canopy fraction of different canopy types. Rooting depths were not as important in the differentiation between the good and bad groups, likely because they are vaguely estimated from other descriptors and thus are not as strongly linked to the real-world values.

8.4.3 Evaluation metrics and donor quality analysis

A few points are of interest regarding the selected analysis metrics. The Nash-Sutcliffe Efficiency was used in this paper primarily for ease of understanding. Granted, it is generally regarded as a good metric but with the obvious caveat of strongly weighting the high flows. Nevertheless, it is evident that other metrics could have performed differently under this framework depending on the ultimate goal of the end-user. However the general conclusions should remain valid, which are that the similarity methods should be favoured over the proximity methods unless the required data is unavailable.

The Success Rate (SR) was defined as a threshold value to discriminate the basins for which the regionalization approaches performed well from those where the methods failed. It permits a rough estimation of the methods' robustness, which the NSE value alone cannot. The SR is also a good aggregator of the 100 resampling results since it uses all available information rather than taking the median value of the 100 iterations. If a single iteration had been made (a single calibration parameter set per catchment) then the SR would have been a single value. This would have been problematic since it would not show the distribution of values due to equifinality. With 100 iterations, the distribution can be estimated and the results show that the spread is not particularly large. This confirms that the methods were minimally affected by the model parameter equifinality.

The Nash ratio allowed representing the relative ability of the methods to predict streamflow on ungauged basins with either the closest donor (figure 8.6) or the best possible donor (figure 8.8). From figure 8.6, it was shown that the geographically closest donors did not perform as well as the most similar donors. Furthermore, figure 8.8 showed that the best donors were at times very dissimilar or very far geographically. This perplexing result was not expected, as the rationale behind the regionalization methods is that the most similar basins should also react similarly in the hydrological sense. Figure 8.13 pits the best donor distance against the closest donor distance for the 264 ungauged basins for the similarity and proximity methods.

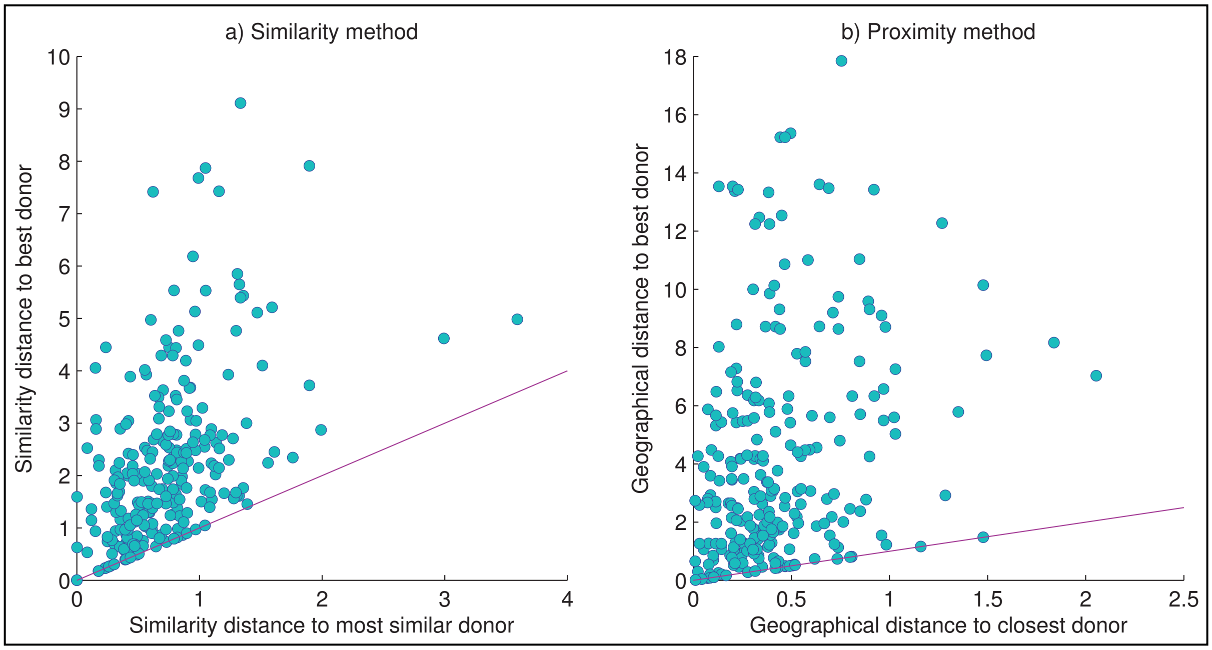

Figure 8.13 Comparison between the best donor distances and the closest or most similar donor distances. Markers on the 1:1 line represent basins whose closest/most similar donor is also the best donor.

It seems evident from figure 8.13 that some closest donors are also the best donors, which is what the regionalization methods are built to achieve. However, in many cases the best donor is neither the closest nor most similar, thus indicating a weakness in the donor selection method. This is what the LDA analysis and figure 8.12 attempted to resolve, with inconclusive results.

8.4.4 Predicting probability of success

In this study, 17 catchment descriptors were analyzed and compared for the good and bad basin groups in figure 8.11. The figure shows quite clearly that some descriptors are significantly less distant for the good groups than for the bad groups. Although the differences are almost all statistically significant according to Mann-Whitney tests, there is no foolproof method to predict if a basin will be in the good group or bad group *a priori*. As was shown in the LDA analysis and figure 8.12, it is possible to estimate if the regionalization will be fruitful on an ungauged basin based on its catchment descriptors and those of its donor basin. However, there remains a fair chance of failure due to the probabilistic nature of the problem and the incomplete separation of the good and bad basin groups. Even when the best catchment descriptors are selected for this means, the distributions of good and bad basins overlap.

In a real world application, the uncertainty regarding the catchment descriptors would likely add to the noise and make the separation of good and bad basins more difficult still. The main takeaway from this aspect of the analysis is that even in the near perfect conditions of the virtual world, there remains much doubt in the ability to predict if a regionalization attempt will be successful. The application of regionalization methods, therefore, must be made while knowing that there is a small yet persistent risk that the generated streamflow will be way off target. It is possible to mitigate this risk, however, by using multiple donors. Doing so was shown to improve the simulation skill by averaging the modeled outflows, as was previously reported by Oudin et al. (2008), amongst others.

A final note of importance must be made regarding an alternative form of regionalization in which multiple catchments are calibrated simultaneously with the intent of having a unique parameter set for a given region (Ricard et al. 2013). An ungauged basin in the region of interest is then modeled with the area's parameter set. Our work has shown that the proximity measure fails to consider some important aspects regarding catchment similarity. The lower performance of the proximity method suggests that the similarity measure should be used when producing parameter sets for common regions if the information is available.

8.5 Conclusion

The aim of this work was to analyze the main hydrological model parameter regionalization approaches and their limitations. The CRCM virtual environment was used to control the catchment descriptors, weather and runoff time series. The virtual world made it possible to explore the regionalization methods' limitations under a reduced uncertainty framework. The first main result was that the regionalization methods performances were very similar to their real world counterparts, lending credibility to the numerical environment approach. While not recommended for all types of hydrology research, this is one application that was well suited for the CRCM environment. Second, it was found that the similarity methods outperformed the proximity methods and should be preferred if the available data permits it. It was demonstrated that the proximity methods work well mainly if the closest donor is also very similar. In both cases, the use of multiple donors was shown to improve performance significantly, with the inverse distance weighting being the best approach with the similarity method and the simple arithmetic mean approach of the proximity method. However, counter-intuitively, it was shown that in many cases the best donor was not the closest or the most similar, but a distant, dissimilar basin. This suggests that there are other descriptors than those available in the CRCM which could potentially improve overall performance, although this hypothesis remains to be validated. Finally, it was shown that the similarity distance between certain catchment descriptors can help predict if a regionalization method will succeed or fail, although it is not currently possible to do so with complete certainty. The probabilistic nature of the prediction on ungauged sites problem seems to be impossible to overcome. Future work should assess if Regional Climate Models could allow improving upon existing knowledge once their resolutions improve and their databases are refined.

8.6 Acknowledgements

The authors would like to thank Blaise Gauvin-St-Denis and the simulations group from the Ouranos consortium on climate change for providing the Canadian Regional Climate Model data (CRCM version 4.2.4 driven by ERA-40 reanalysis over the Quebec domain) required for this project. Thanks also to Catherine Guay from the Institut de Recherche d'Hydro-Québec (IREQ) for providing the CQ2 hydrometric database.

8.7 References

Arsenault, R., Poulin, A., Côté, P. and Brissette, F., 2014. Comparison of stochastic optimization algorithms in hydrological model calibration. Journal of Hydrologic Engineering, 19(7), 1374-1384. doi: 10.1061/(ASCE)HE.1943-5584.0000938.

Arsenault, R., and Brissette, F., 2014a. Continuous streamflow prediction in ungauged basins: The effects of equifinality and parameter set selection on uncertainty in regionalization approaches. Water Resources Research, 50(7), doi: 10.1002/2013WR014898.

Arsenault, R. and Brissette, F. 2014b. Determining the Optimal Spatial Distribution of Weather Station Networks for Hydrological Modeling Purposes Using RCM Datasets: An Experimental Approach. J. Hydrometeor, 15, 517-526. doi: http://dx.doi.org/10.1175/JHM-D-13-088.1

Arsenault, R. and Brissette, F. 2014c. Multi-model averaging for continuous streamflow prediction in ungauged basins. Hydrological Sciences Journal, Under re-review, 35p.

Bao, Z., J. Zhang, J. Liu, G. Fu, G. Wang, R. He, X. Yan, J. Jin, and H. Liu (2012), Comparison of regionalization approaches based on regression and similarity for predictions in ungauged catchments under multiple hydro-climatic conditions, J. Hydrol., 466–467, 37-46 doi:10.1016/j.jhydrol.2012.07.048.

Bardossy, A. (2007), Calibration of hydrological model parameters for ungauged catchments, Hydrol. Earth Syst. Sci., 11, 703–710.

Beauchamp, J., Leconte, R., Trudel, M. and Brissette, F. (2013) Estimation of the summer-fall PMP and PMF of a northern watershed under a changed climate. Water Resources Research, 49(6), 3852-3862 DOI: 10.1002/wrcr.20336

Beven, K. (2006), A manifesto for the equifinality thesis, J. Hydrol, 320, 18–36.

Burn, D.H. and Boorman, D.B., 1993. Estimation of hydrological parameters at ungauged catchments. Journal of Hydrology, 143(3–4), 429-454.

Caya, D., and Laprise, R. (1999) A semi-implicit semi-Lagrangian regional climate model: The Canadian RCM. Mon. Wea. Rev. 127 (3), 341-362.

Diks, C.G.H. and Vrugt, J.A., 2010. Comparison of point forecast accuracy of model averaging methods in hydrologic applications. Stoch. Env. Res. Risk A. 24(6), 809-820.

Fortin, V., 2000. Le modèle météo-apport HSAMI: historique, théorie et application, 68p. Institut de Recherche d’Hydro-Québec, Varennes, Canada.

Goswami, M., O’Connor, K.M., and Bhattarai, K.P., 2007. Development of regionalisation procedures using a multi-model approach for flow simulation in an ungauged catchment. Journal of Hydrology, 333 (2–4), 517-531.

Hansen, N. and Ostermeier, A., 1996. Adapting arbitrary normal mutation distributions in evolution strategies: The covariance matrix adaptation. In Proceedings of the 1996 IEEE International Conference on Evolutionary Computation, pp. 312-317;

Hansen, N. and Ostermeier, A., 2001. Completely Derandomized Self-Adaptation in Evolution Strategies. Evolutionary Computation, 9(2), 159-195.

He, Y., Bárdossy, A., and Zehe, E., 2011. A review of regionalisation for continuous streamflow simulation, Hydrology and Earth System Sciences, 15, 3539-3553, doi:10.5194/hess-15-3539-2011

Hrachowitz, M., Savenije, H.H.G., Blöschl, G., McDonnell, J.J., Sivapalan, M., Pomeroy, J.W., Arheimer, B., Blume, T., Clark, M.P., Ehret, U., Fenicia, F., Freer, J.E., Gelfan, A., Gupta, H.V., Hughes, D.A., Hut, R.W., Montanari, A., Pande, S., Tetzlaff, D., Troch, P.A., Uhlenbrook, S., Wagener, T., Winsemius, H.C., Woods, R.A., Zehe, E., and Cudennec, C., 2013. A decade of Predictions in Ungauged Basins (PUB)—a review. Hydrological Sciences Journal, 58 (6), 1198–1255.

Kruskal, W. H. and W. A.Wallis (1952), Use of ranks in one-criterion variance analysis, J. Amer. Statist. Assn., 47, 583-621.

Lee H., N. R. McIntyre, H. S. Wheater, and A. R. Young (2006), Predicting runoff in ungauged UK catchments, Proceedings of the Institution of Civil Engineers. Water Management 159(2): 129–138

Lucas-Picher, P., Riboust, P., Somot, S., and Laprise, R. 2015: Reconstruction of the Spring 2011 Richelieu River Flood by Two Regional Climate Models and a Hydrological Model. J. Hydrometeor, 16, 36–54. doi: http://dx.doi.org/10.1175/JHM-D-14-0116.1

Maraun, D. 2012. Nonstationarities of regional climate model biases in European seasonal mean temperature and precipitation sums, Geophys. Res. Lett., 39, L06706, doi:10.1029/2012GL051210.

McCuen, R. H. , and Levy, B. S. (2000). Evaluation of peak discharge transposition. J. Hydrologic Eng., 5 (3), 278–289.

McIntyre, N., Lee, H., Wheater, H., Young, A. and Wagener, T., 2005. Ensemble predictions of runoff in ungauged catchments. Water Resources Research, 41, W12434.

McLachlan, Geoffrey J. 1992. Discriminant analysis and statistical pattern recognition.Wiley Series in Probability and Mathematical Statistics: Applied Probability and Statistics. Wiley-Interscience. John Wiley & Sons, Inc., New York. 526 pp. ISBN: 0-471-61531-5

Merz, R., and Blöschl, G., 2004. Regionalization of catchment model parameters. Journal of Hydrology, 287, 95–123.

Minville, M., Cartier, D., Guay, C., Leclaire, L.-A., Audet,C., Le Digabel, S., and Merleau, J. (2014),Improving process representation in conceptual hydrological model calibration using climate simulations, Water Resour. Res., 50, 5044–5073, doi:10.1002/2013WR013857.

Minville, M., Brissette, F. and Leconte, R., 2008. Uncertainty of the impact of climate change on the hydrology of a Nordic watershed. Journal of Hydrology, 358(1-2): 70-83

Minville, M., Brissette, F., Krau, S. and Leconte, R., 2009. Adaptation to Climate Change in the Management of a Canadian Water-Resources System. Water Resources Management, 23(14): 2965-2986.

Minville, M., Krau, S., Brissette, F. and Leconte, R., 2010. Behaviour and Performance of a Water Resource System in Québec (Canada) Under Adapted Operating Policies in a Climate Change Context. Water Resources Management, 24, 1333–1352

Music, B., A. Frigon, M. Slivitzky, A. Musy, D. Caya, and R. Roy, 2009: Runoff modelling within the Canadian Regional Climate Model (CRCM): analysis over the Quebec/Labrador watersheds. In: New Approaches to Hydrological Prediction in Data Sparse Regions (Proc. of Symposium HS.2 at the Joint IAHS & IAH Convention, Hyderabad, India, September 2009). International Association of Hydrological Sciences (IAHS) Red Book Series Publ. 333, 183-194.

Music, B. and D. Caya, 2009: Investigation of the sensitivity of the water cycle components simulated by the Canadian Regional Climate Model (CRCM) to the land surface parameterization, the lateral boundary data and the internal variability. Journal of Hydrometeorology, 10, 3–21.

Music, B., and D. Caya, 2007: Evaluation of the hydrological cycle over the Mississippi River basin as simulated by the Canadian RCM (CRCM). Journal of Hydrometeorology, 8: 969–988.

Nash, J. E., and Sutcliffe, W. H., 1970. River flow forecasting through conceptual models: Part 1. A discussion of principles. Journal of Hydrology, 10(3), 282-290.

Oudin, L., Andréassian, V., Perrin, C., Michel, C. and Le Moine, N., 2008. Spatial proximity, physical similarity, regression and ungaged catchments: A comparison of regionalization approaches based on 913 French catchments. Water Resources Research, 44, W03413.

Parajka, J., Merz, R. and Blöschl, G., 2005. A comparison of regionalisation methods for catchment model parameters. Hydrology and Earth System Sciences, 9, 157–171.

Parajka, J., Viglione, A., Rogger, M., Salinas, J.L., Sivapalan, M., and Blöschl, G., 2013. Comparative assessment of predictions in ungauged basins – Part 1: Runoff hydrograph studies, Hydrology and Earth System Sciences, 10, 375-409, doi:10.5194/hessd-10-375-2013.

Poulin, A., Brissette, F., Leconte, R., Arsenault, R. and Malo, J.S., 2011. Uncertainty of hydrological modelling in climate change impact studies in a Canadian, snow-dominated river basin. Journal of Hydrology, 409(3-4), 626-636. doi:10.1016/j.jhydrol.2011.08.057.

Razavi, T. and Coulibaly, P., 2013. Streamflow Prediction in Ungauged Basins: Review of Regionalization Methods. Journal of Hydrologic Engineering, 18(8), 958–975.

Ricard, S., Bourdillon, R., Roussel, D., and Turcotte, R. (2013). Global Calibration of Distributed Hydrological Models for Large-Scale Applications. J. Hydrol. Eng., 18(6), 719–721.

Samuel, J., P. Coulibaly, and R. Metcalfe (2011), Estimation of continuous streamflow in Ontario ungauged basins: Comparison of regionalization methods, J. Hydrol. Eng., 16(5), 447–459.

Seibert, J. (1999), Regionalisation of parameters for a conceptual rainfall runoff model, Agric. For. Meteorol., 98-99(31), 279–293.

Sellami, H., I. La Jeunesse, S. Benabdallah, N. Baghdadi, M. Vanclooster. Uncertainty analysis in model parameters regionalization: a case study involving the SWAT model in Mediterranean catchments (Southern France). Hydrology and Earth System Sciences, 2014, p. 2393 - p. 2413

Shu, C. and Burn, D.H.: Spatial patterns of homogeneous pooling groups for flood frequency analysis, Hydrol. Sci. J., 48(4), 601-618, DOI: 10.1623/hysj.48.4.601.51417, 2003.

Sivapalan, M., Takeuchi, K., Franks, S.W., Gupta, V.K., Karambiri, H., Lakshmi, V., Liang, X., McDonnell, J.J., Mendiondo, E.M., O'Connell, P.E., Oki, T., Pomeroy, J.W., Schertzer, D., Uhlenbrook., S. and Zehe, E., 2003. IAHS Decade on Predictions in Ungauged Basins (PUB), 2003-2012: Shaping an exciting future for the hydrological sciences. Hydrological Sciences Journal 48: 857-880.

Vandewiele, G. L. et Elias, A. (1995) Monthly water balance of ungauged catchments obtained by geographical regionalization. Journal of Hydrology, 170(1-4), 277–291.

Verseghy, D. L., McFarlane, N. A. and Lazare, M. (1993), Class—A Canadian land surface scheme for GCMS, II. Vegetation model and coupled runs. Int. J. Climatol., 13: 347–370. doi: 10.1002/joc.3370130402

Viney, N. R., Vaze, J., Chiew, F. H. S., Perraud, J.-M., Post, D. A., and Teng, J., 2009. Comparison of multi-model and multi-donor ensembles for regionalisation of runoff generation using five lumped rainfall-runoff models, 18th World IMACS Congress and MODSIM09 International Congress on Modelling and Simulation, Cairns, Australia.

Wagener, T., and H. S. Wheater (2006), Parameter estimation and regionalisation for continuous rainfall-runoff models including uncertainty, J. Hydrol., 320, 132–154.

Wilcoxon, F. (1945). Individual Comparisons by Ranking Methods. Biometrics, 1(6), 80–83.

Yadav, M., T. Wagener, and H. Gupta (2007), Regionalization of constraints on expected watershed response behavior for improved predictions in ungauged basins, Adv. Water Resour., 30, 1756–1774.

Zelelew, M.B., and Alfredsen, K., 2014. Transferability of hydrological model parameter spaces in the estimation of runoff in ungauged catchments. Hydrological Sciences Journal, 59 (8), 1470–1490. http://dx.doi.org/10.1080/02626667.2013.838003

Zhang, Y. and Chiew, F.H.S., 2009. Relative merits of different methods for runoff predictions in ungauged catchments. Water Resources Research, 45, W07412.

CHAPITRE 9

ARTICLE 7 : PARAMETER DIMENSIONALITY REDUCTION OF A CONCEPTUAL MODEL FOR STREAMFLOW PREDICTION IN UNGAUGED BASINS

Richard Arsenault[1], Dominique Poissant[1] et François Brissette[1]

[1] Département de Génie de la Construction, École de technologie supérieure,

1100 Notre-Dame Ouest, Montréal, Québec, Canada H3C 1K3.

Article soumis à la revue « Advances in Water Resources » en février 2015.

Abstract

Continuous streamflow prediction in ungauged basins is one of the most complex challenges in hydrology due to the unavailability of data and limited model ability to simulate hydrological processes. Parametric equifinality also poses a challenge as different parameter sets acting similarly on a given catchment may perform otherwise when applied to an ungauged basin. Model parsimony can reduce the extent of equifinality and could allow better identification of model parameters for regionalization applications. This paper evaluates five regionalization methods when applied in a parameter reduction framework on 267 catchments in the province of Quebec, Canada. The Sobol' variance-based sensitivity analysis is used to rank the model parameters by their influence on the model results, including the parameter cross-correlations. The parameters are fixed to a priori defined values and the regionalization approaches are re-evaluated for each new fixed parameter. The reduction in parameter correlations is shown to improve parameter identifiability, both for the model calibration step and for the regression model predictions based on physical catchment descriptors. However, this improvement is found to be minimal and is not transposed in the regionalization mode. Furthermore, all the tested methods performed worse with less free parameters. It is shown that for the model in this study, 8 of 23 parameters can be fixed with no loss in performance in regionalization, and 11 of 23 can be fixed with minimal loss. Uncertainty is shown to be minimally reduced with fewer parameters as opposed to model robustness which increased with the parsimonious versions of the model. The main conclusions are that conceptual models do not represent physical processes sufficiently well to warrant parameter reduction for physics-based regionalization methods and, most importantly, catchment descriptors do not represent the relevant hydrological processes sufficiently well. Finally, the reduction in parameter correlations achieved with Sobol' sensitivity analysis did not translate into a better performance in regionalization approaches.

Keywords: Parameter reduction, Sobol' sensitivity analysis, regionalization, parameter identifiability, prediction in ungauged basins

9.1 Introduction

Hydrological models are the main tools used to simulate streamflow on water basins. Typically, model parameters are calibrated as to tweak the model response as closely as possible to a measured streamflow time series. The model is then validated on an independent period, and if the results are satisfactory, the model can be used to simulate streamflow on past or future periods where hydrometric data are missing or unavailable. However, we there is often a need to estimate the hydrologic response in areas where there are no gauging stations. In such cases, the approach described above cannot be applied as there is no data to perform the calibration step.

A few methods were developed to counter this problem by using the available information in neighbouring catchments. The first proposed method relied on parametric regression, which entails building a regression model between the gauged basins' model parameters and their physical attributes and then estimating the parameters at the ungauged site based on its physical properties (Sefton and Howart 1998, Seibert 1999). Multiple linear regression methods were shown to be adequate in some cases, such as in drier climates, but are usually outperformed by other approaches in humid climates (Parajka et al. 2013). Among these, the spatial proximity method transfers entire parameter sets at one, as the closest gauged catchment "donates" its parameter set to the hydrological model on the ungauged basin. This method wagers that the hydrological characteristics of neighbouring catchments are similar enough to warrant the use of the same parameter set (Vandewiele and Elias 1995, Merz and Blöschl 2004, Parajka et al. 2005). A similar approach, the physical similarity method, uses the same logic except the donor basin is selected according to its physical similarity to the ungauged basin. The idea is that the more catchments share similar characteristics, the more they should share similar hydrological behavior. (McIntyre et al. 2005, Bardossy 2007, Samuel et al. 2011, Bao et al. 2012). However, Oudin et al. (2010) showed that the assumption of hydrologic and physical similarity being correlated is not always adequate and that our limited knowledge of hydrologic processes at the fundamental level limits the number of useful catchment attributes. Nonetheless, the physical similarity method performed better than the others in some studies, especially when geographical information is added to the similarity measure (Zhang and Chiew 2009).

The decade on Prediction in Ungauged Basins (PUB) initiative (2003-2012) rekindled interest in the field and a few innovations improved regionalization performance (Sivapalan et al. 2003). Razavi and Coulibaly (2013) and He et al. (2011) have made comprehensive reviews of the findings over this period. It has become quite clear that donor-based methods are to be preferred, and that spatial proximity outperforms physical similarity in areas with high-density gauging stations (Oudin et al 2008, Zhang and Chiew 2009, Parajka et al 2013).

Nonetheless, there remain many questions concerning the prediction of streamflow in ungauged basins. The hydrological models themselves are an integral part of the regionalization approach and they can be more or less suited to the PUB applications. They vary in complexity and in dimensionality, ranging from a few up to dozens of calibrated parameters. In conceptual models, there is often parameter correlation that can lead to identifiability problems, which in turn can complicate the model parameter / catchment descriptor correlation for regionalization purposes (Wagener and Wheater 2006). This encompasses the equifinality problem, which is a result of multiple different parameter sets that behave similarly in calibration (Beven, 2006). Regionalization approaches use calibrated parameters on gauged catchments to predict flow at the ungauged sites, therefore parameter pairs that are optimal on one catchment may be inappropriate for the target basin.

A previous study (Arsenault and Brissette 2014) showed that the equifinality played a minor role in regionalization when the 23-parameter HSAMI hydrological model was used and it was found that the physical similarity method performed better than spatial proximity or regression methods. The fact that physical similarity outperformed spatial proximity suggests that the gauging network was not dense enough and that there exists a link between the catchment descriptors and the model parameter set as a whole. Lee et al. (2005) showed that certain model structures are more suitable for regionalization approaches given a set of catchments. However the poor regionalization performance with the multiple linear regression method shows that the model parameters, taken individually, are not correlated to the catchment descriptors. This in turn suggests that the regression models are unable to make use of the parameter correlations. One possible explanation is parametric equifinality, in which different parameter sets produce similar hydrographs on a given catchment. These equifinal parameter sets can produce different streamflow simulations once transferred to a target catchment through regionalization. Therefore, a reduction in the number of parameters could lead to fewer parameter correlations and possibly improve regionalization skill.

One of the approaches used in simplifying the parameter complexity in models is the Global Sensitivity Analysis (GSA). GSA methods rely on analyzing carefully sampled parameter sets from the parameter space and their impact on the objective function. The most widespread method, Sobol' sensitivity analysis, decomposes the variance of the parameters on the model performance and indicates how much each parameter contributes to the model variance. Tang et al. (2007) found the Sobol' method to be more robust and more efficient than three other sensitivity analysis methods for hydrological model parameters. From this, it is possible to fix parameters that contribute the least to the total variance and thus reduce complexity while maintaining good model performance, as suggested by Saltelli and Tarantola (2002). For example, Nossent et al. (2011) showed that it was possible to reduce the number of parameters of their SWAT model from 26 to 9 with little to no loss in model performance using the Sobol' method. Van Werkhoven et al. (2008) found that they could reduce the number of parameters of the Sacramento Soil moisture accounting model by 30-40% while maintaining high quality predictions. More recently, Zhang et al. (2013) used the Sobol' method to reduce the complexity of the SWAT model using 4 goodness-of-fit metrics and found that only a few parameters controlled much of the variance for the 4 metrics.

These methods have been used successfully in streamflow simulation on gauged sites. This paper attempts to study the impact of parameter correlation on regionalization performance. The next section specifies the paper objectives, followed by the methodology used in this study and the main results. Finally, the implications of these findings are discussed and paths for further research are recommended.

9.2 Scope and aims

The main objective of this study is to measure the effects of reducing the correlations on parameter identifiability and to assess their impacts on the regionalization approaches performance. Ultimately, this is done in the hope of improving the performance of the main regionalization methods (multiple linear regression, physical similarity and spatial proximity). The model parameters will be sequentially fixed in increasing order of total variance explanation, thus reducing the parameter space at each new fixed parameter. Finally, the effects of equifinality will be measured on progressively simpler versions of the model used in regionalization.

9.3 Study area and data

The study was performed on 267 basins covering the province of Québec, Canada. Figure 9.1 shows the study area and the basin locations. Some basins are nested within others and are included in the study.

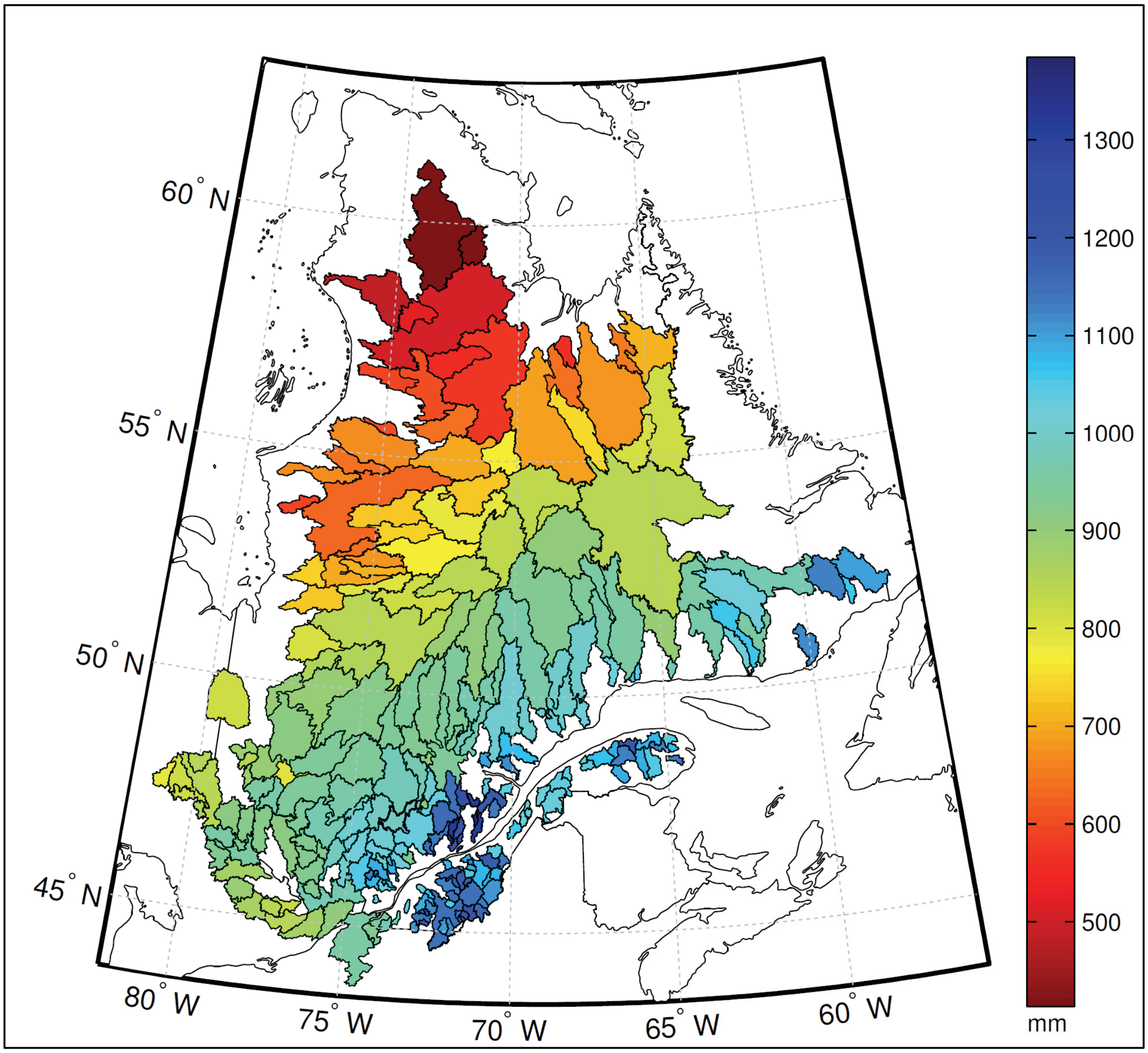

Figure 9.1 Sizes, locations and mean annual precipitation of the 267 basins in the study area, situated in the province of Quebec, Canada

The basins range from 30 to 69191 square kilometres in size, and cover most of the province of Quebec with a total area of 1.6 million square kilometers. A list of 13 of the catchment descriptors described in He et al. (2011) was used in this study. Most popular descriptors, taken from He et al. (2011), were used except for soil properties, which were not integrated in this study due to limited availability of reliable data. The descriptors that were selected and a few relevant statistics are presented in table 9.1.

Table 9.1 Statistics of catchment descriptors used in this study

Catchment descriptors	**Maximum**	**Minimum**	**Average**
Area (km²)	69191	30	6832
Aridity index	0.99	0.31	0.61
Elevation (m)	916	52	383
Land Cover - Crop (%)	83.1	0	8.7
Land Cover – Forest (%)	96	0	65.2
Land Cover - Grass (%)	65.5	0	13.6
Land Cover - Urban (%)	16.4	0	1.2
Land Cover - Water (%)	35.6	0	9.3
Land Cover - Wetlands (%)	17.1	0	1.2
Latitude (degrees)	59.9	44.5	49
Longitude (degrees)	-57.9	-81	-72
Mean annual precipitation (mm)	1412	413	965
Slope (%)	51.9	1.1	10.7

The Burn and Boorman (1993) approach was used to combine these 13 catchment characteristics into a single similarity index:

$$\Phi = \sum_{i=1}^{k} \frac{\left|X_i^G - X_i^U\right|}{\Delta X_i} \tag{9.1}$$

where *i* is the catchment descriptor identifier, X^G is the descriptor value for the gauged catchment, X^U is the descriptor value at the ungauged catchment and *ΔX* is the range of values taken by the respective X^G in the dataset. Note that the latitude and longitude are present in the similarity index, making it somewhat of an integrated similarity measure instead of a purely physical one. The latitude and longitude serve as proxies to unknown physical properties such as soil characteristics, which are assumed to be similar in adjacent regions.

9.3.1 Meteorological and hydrological datasets

The hydrometric data were obtained from the CQ2 database, which is a shared archive maintained and supplied by various province and industry partners who combined their hydrometric data for hydrological research. The observed climate data were substituted by the Canadian National Land and Water Information Service (NLWIS) 10 km gridded dataset (Hutchinson et al. 2009). This choice was made since many catchments have no weather stations within their boundaries, but the NLWIS dataset was dense enough that all the basins contained at least one grid point. The NLWIS climate dataset was shown to be a good replacement for missing observed data in hydrological applications (Chen et al. 2013), which is why it was selected as the preferred climate input. The NLWIS meteorological data was also used in another study which focused on multi-model averaging in regionalization on the same basins (Arsenault and Brissette 2015).

9.4 Methodology

9.4.1 Hydrological models

9.4.1.1 HSAMI

The HSAMI model (Fortin 2000) has been used by *Hydro-Quebec* for over two decades to forecast daily flows on more than 100 basins over the province of Quebec, Canada. It has been used extensively in research applications as well, such as in reservoir management (Minville et al. 2008, 2009, 2010) and climate change impact studies (Poulin et al. 2011, Arsenault et al. 2013). It simulates the entire hydrological cycle with a strong snow accumulation and melt model. Potential evapotranspiration is estimated using a proprietary formulation requiring only daily maximum and minimum temperatures. There are four interconnected reservoirs that contribute to the vertical water transfer balance: Snow on ground, surface runoff, saturated soil layer and unsaturated soil layer. The horizontal water transfer is based on two unit-hydrographs (one for surface runoff and one for underground runoff) and a linear reservoir. HSAMI requires spatially averaged maximum and minimum temperatures, liquid and solid precipitation and, if available, updated snow on ground depth.

The model has 23 adjustable parameters, all of which were initially calibrated in this study: 10 for the various production function processes, 5 for the horizontal transfer through reservoir-type soil layers, 2 for evapotranspiration and 6 for snow-related processes. The HSAMI model parameter descriptions are presented in table 9.2.

Table 9.2 Description of the HSAMI model parameters and process sub-models

Sub-model	**ID**	**Parameter description**	**Units**
Evapo-transpiration	1	Factor multiplying potential evapotranspiration (PET) for the estimation of summer real evapotranspiration (RET)	--
	2	Factor multiplying PET for estimating the RET in winter	--
Snowmelt	3	Snow melting rate during daytime. ΔT in Celsius is calculated as the difference between Tmax and parameter of Tmax threshold for snowmelt (parameter 5).	cm/Δ° C/day
	4	Snow melting rate during nighttime. ΔT in Celsius is calculated as the difference between parameter 5 and Tmin.	cm/Δ° C/day
	5	Tmax threshold for snowmelt	°C
	6	Tmin threshold for accelerated snowmelt	°C
	7	Reference temperature for calculating heat supplied by the rain to the snow cover	°C
	8	Empirical parameter used to connect the state variables describing snow cover and cumulated snowmelt to the proportion of the basin covered by snow	--
Surface runoff	9	Empirical parameter used to connect the state variables describing soil freezing and thawing to the proportion of snowmelt water flowing on the surface	--
	10	24-hour rainfall amount needed to generate 50% runoff with completely dry soil.	cm
	11	24-hour rainfall amount needed to generate 50% runoff with completely saturated soil.	cm
Vertical water transfer	12	Water amount in the unsaturated zone that cannot drain by gravity	cm
	13	Maximum water amount that can be contained in the unsaturated soil zone	cm
	14	Maximum water amount that can be contained in the aquifer before generating surface runoff	cm
	15	Proportion of surface water flowing through the intermediate hydrograph instead of moving through the soil column	--
	16	Proportion of soil water that is directed to the intermediate hydrograph when the unsaturated zone overflows	--
	17	Emptying rate of the unsaturated zone to the groundwater reservoir	Day^{-1}
	18	Emptying rate of the groundwater reservoir (base flow)	Day^{-1}
Horizontal water transfer	19	Emptying rate of the intermediate reservoir, through the intermediate hydrograph	Day^{-1}
	20	Time to peak for the surface unit hydrograph	Day
	21	Shape parameter of the surface hydrograph (using a gamma distribution function)	--
	22	Time to peak for the intermediate unit hydrograph	Day
	23	Shape parameter of the intermediate hydrograph (using a gamma function)	--

9.4.1.2 MOHYSE

MOHYSE is a simple model that was first developed for academic purposes (Fortin and Turcotte 2007). Since then, the model has been used in research applications (e.g. Velazquez et al. 2010). MOHYSE is specifically built to handle Nordic watersheds and has custom snow accumulation and melt as well as potential evapotranspiration (PET) modules. The required input data are mean daily temperatures, total daily rainfall depth and total daily snow depth (expressed as water equivalent). It has 10 free parameters which must be calibrated. The interest of using the MOHYSE model is that it was created with the intent of minimizing the number of parameters to improve their identifiability and reducing their cross-correlations. It was used for only a small part of this study, therefore its parameters are not shown here. Readers are invited to read Fortin and Turcotte (2007) for more details on MOHYSE and its parameters.

9.4.2 Model calibration

The first steps in this study were to set-up and calibrate the models on the 267 catchments to obtain parameter sets to be transferred to the pseudo-ungauged sites. The calibrations for the HSAMI model were performed using the Covariance-Matrix Adaptation Evolution Strategy (CMAES) (Hansen and Ostermeier 1996, 2001). CMAES is an evolutionary algorithm for difficult problems, such as those with non-linear, non-convex and non-smooth fitness landscapes. It was shown to outperform other algorithms in calibration for HSAMI (Arsenault et al. 2014). Using the approach in the previously mentioned study, it was determined that the Shuffled Complex Evolution – University of Arizona algorithm (SCEUA) (Duan et al. 1992, 1993, 1994) was the better choice for the simpler MOHYSE model. The hydrological models were calibrated using the Nash-Sutcliffe Efficiency metric on daily hydrographs as the objective function (Nash and Sutcliffe 1970), which is arguably the most common goodness-of-fit metric in hydrology. For each catchment, 10 calibrations were performed. This allowed sampling equifinal parameter sets to analyze parameter non-uniqueness in the regionalization approaches with reduced parameter space, all will be detailed further.

All the basins were kept in this study, including those with poor calibration NSE values. Lower-scoring basins in calibration were kept to determine the regionalization methods' abilities to predict streamflow on poorly modelled catchments. However, they do not contribute in the parameter identification process for the ungauged donors. They are effectively only used as target catchments, but never as donors.

9.4.3 Sobol' Global sensitivity analysis

The variance-based Sobol' sensitivity analysis method (Sobol', 1993) (hereafter referred to as the Sobol' method) was used to determine the relative importance of the model parameters according to their contribution to the total order variance. The total order variance, by definition, is equal to the first order variance of the parameter (the effect of the single parameter on the model response) added to any variance attributed to interactions with other parameters (Chen et al. 2015). Therefore the total order variance was used as it includes all multi-parameter effects in the modelling response. The Sobol' method was applied with 250000 sampled parameter sets following a Sobol' pseudo-random sequence and the respective objective function values when fed to the model. The sample size was selected as it consistently returned confidence intervals within 10% of the total order indices value for each parameter, which was considered reasonable. A larger sample would have given more precise estimations, but would have been more costly in computing resources. The objective function selected for the Sobol' analysis was a normalized Nash-Sutcliffe Efficiency metric shown in equation 9.2.

$$O.F. = \frac{1}{2 - NSE} \tag{9.2}$$

The transformation is necessary to prevent overweighting of parameters that can return very poor NSE values. For example, a parameter that could lead to a -100 NSE value would be weighted much more than a parameter that could lead to a -50 NSE, even in cases where the parameter is likely to have more impact in the optimal range. The transformation limits the range of the objective function to [0:1] and was shown to be a better method to identify parameter importance (Nossent and Bauwens 2012).

The Sobol' method was completed on each of the 267 catchments to verify the consistency of the parameter rankings. Figure 9.2 shows the distribution of the parameter rankings, in which each panel represents a ranking and the histogram values represent the number of basins in which the parameter on the x-axis is found to occupy the rank.

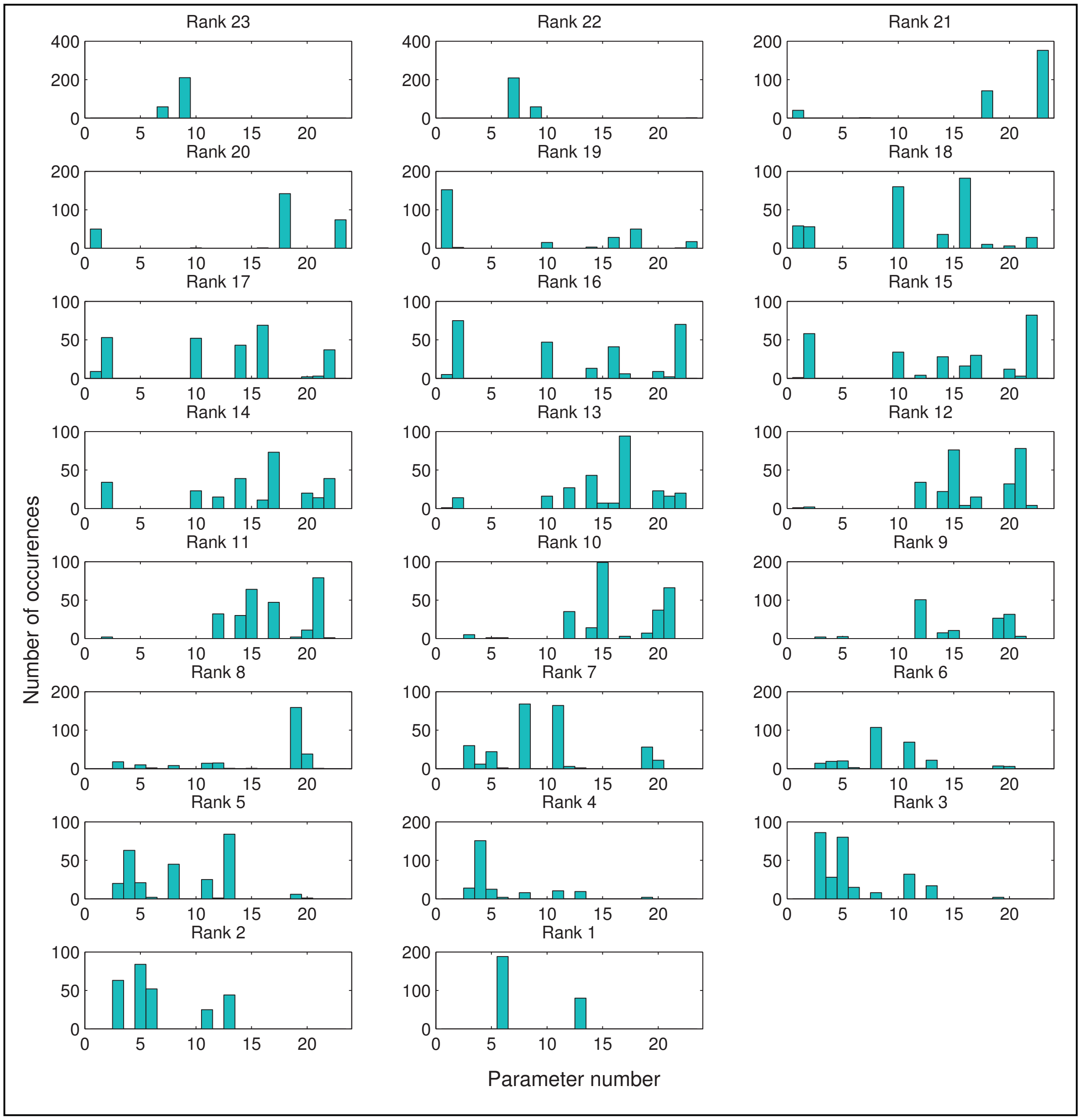

Figure 9.2 Parameters ranked from least to most influential according to their total order effects. The number of occurrences for each parameter is displayed for a given rank

For example, in panel 1 (rank 23, the least important parameter), parameter 9 was found to be the least important for 208 of the catchments, while parameter 7 was found to be the least important in 54 cases. In panel 2, parameters 7 and 9 are inverted. The two least influential parameters are therefore parameters 7 and 9 as they are the worst two for all but 5 basins. Table 9.3 shows the final parameter rankings in order of importance

Table 9.3 Sobol' sensitivity analysis results for the HSAMI Model

Parameter	Total order indices	Variance explained (%)	Cumulative variance exp. (%)
9	0.0001	0.004	0.004
7	0.0005	0.030	0.034
23	0.0009	0.054	0.088
18	0.0012	0.070	0.159
1	0.0017	0.095	0.253
2	0.0025	0.139	0.393
16	0.0027	0.152	0.544
12	0.0051	0.287	0.831
10	0.0051	0.289	1.120
17	0.0078	0.439	1.560
22	0.0081	0.457	2.017
15	0.0088	0.496	2.512
21	0.0092	0.520	3.032
14	0.0097	0.547	3.579
19	0.0258	1.462	5.041
11	0.0299	1.694	6.735
20	0.0368	2.084	8.818
13	0.0389	2.200	11.018
8	0.0530	2.999	14.017
4	0.1612	9.127	23.144
5	0.2755	15.597	38.741
3	0.3983	22.548	61.289
6	0.6838	38.711	100.000
Total	1.7663	100	

From table 9.3, it can be seen that 86% of the variance can be explained by the 4 most influential parameters and their interactions.

The results for the MOHYSE model are presented in table 9.4.

Table 9.4 Sobol' sensitivity analysis results for the MOHYSE Model

Parameter	Total order indices	Variance explained (%)	Cumulative variance exp. (%)
1	0.049	2.867	2.867
8	0.052	3.085	5.952
10	0.059	3.459	9.411
5	0.070	4.156	13.567
4	0.126	7.454	21.021
9	0.167	9.873	30.894
2	0.182	10.747	41.641
3	0.240	14.143	55.783
6	0.326	19.258	75.041
7	0.423	24.959	100
Total	1.694	100	

The results in table 9.4 show that the parameters have more uniformity in their variance explanation. 86% of the variance can be explained by the 6 most influential parameters, and the four remaining parameters seem to contribute to non-negligible amounts of the total order variance compared to the HSAMI model.

9.4.4 Sequential model parameter fixing and recalibration

The model parameter fixing required an a priori value to be set to the fixed parameter. To do so, the median parameter value of the calibration dataset was chosen. This ensures a fair way to treat unknown parameter values in the conceptual model. Of course for certain catchments the parameter selection will be detrimental, however in the case of non-important parameters, the effects should be negligible. The use of this method generates a small bias as all the catchments are selected at this stage, even though each of the catchments will later be treated as ungauged. However, the effects on the median parameter value are expected to be minute. At each step, the least important parameter remaining was fixed to its median calibrated value. The model was then recalibrated 10 times to generate new sets of parameters that are independent of the fixed parameter. The process is repeated until all but the last parameters are fixed. The aim of this method was to force the model into progressively more constrained parameter space to measure the effects on model robustness in regionalization applications. This also forces the reduction of model parameter equifinality at each step.

9.4.5 Regionalization methods application

Five regionalization approaches were tested in this study (multiple linear regression, spatial proximity, physical similarity, regression-augmented spatial proximity and regression-augmented similarity). The regression-augmented similarity (similarity-regression) and proximity (proximity-regression) methods are modified versions of the regular spatial proximity and physical similarity methods. In these hybrid versions, a linear regression model is built to estimate each of the parameters. If the coefficient of determination is superior to 0.5, that parameter is replaced in the donor parameter set by the estimated value. Arsenault and Brissette (2014) showed that the regression-augmented similarity method performed better than its standalone version. In all cases, basins are only considered candidate donors if their calibration NSE exceeds 0.7. This is to eliminate the lower scoring basins from contributing to the regionalization approaches.

In each case, they were applied on the 23 versions of HSAMI model (with 1 to 23 calibrated parameters). Furthermore, up to 10 donor basins were selected and their resulting simulated flows averaged to improve regionalization performance. The averaging was performed by taking the simple mean of the resulting hydrographs on the target catchment, or by weighting each hydrograph by the inverse distance to the donor catchment (IDW) (see Oudin et al. 2008, Viney et al. 2009). The entire process was repeated 100 times, each time selecting one of the 10 calibrated parameter sets at random from the donor catchments for each regionalisation attempt. This allows sampling the parametric equifinality in regionalization. The regionalization methods' skill was evaluated with the NSE metric as well as a metric defined as the Nash Ratio (NR). The NR is defined as the ratio of the regionalized NSE to the calibration NSE. This allows normalizing the results and to focus on the regionalization methods' abilities rather than the hydrological model's ability to simulate the streamflow on the given catchments. A final metric defined as the Success Rate (SR) was used to analyze the regionalization methods performances under equifinality (Arsenault and Brissette 2014). The SR is computed as the fraction of catchments whose regionalization NR is equal or superior to 0.85.

9.5 Results

9.5.1 Model calibration performance

The calibration NSE values for the HSAMI model with an increasing number of fixed parameters are presented in figure 9.3. Each box-and-whisker plot contains the median calibration NSE for the 267 basins.

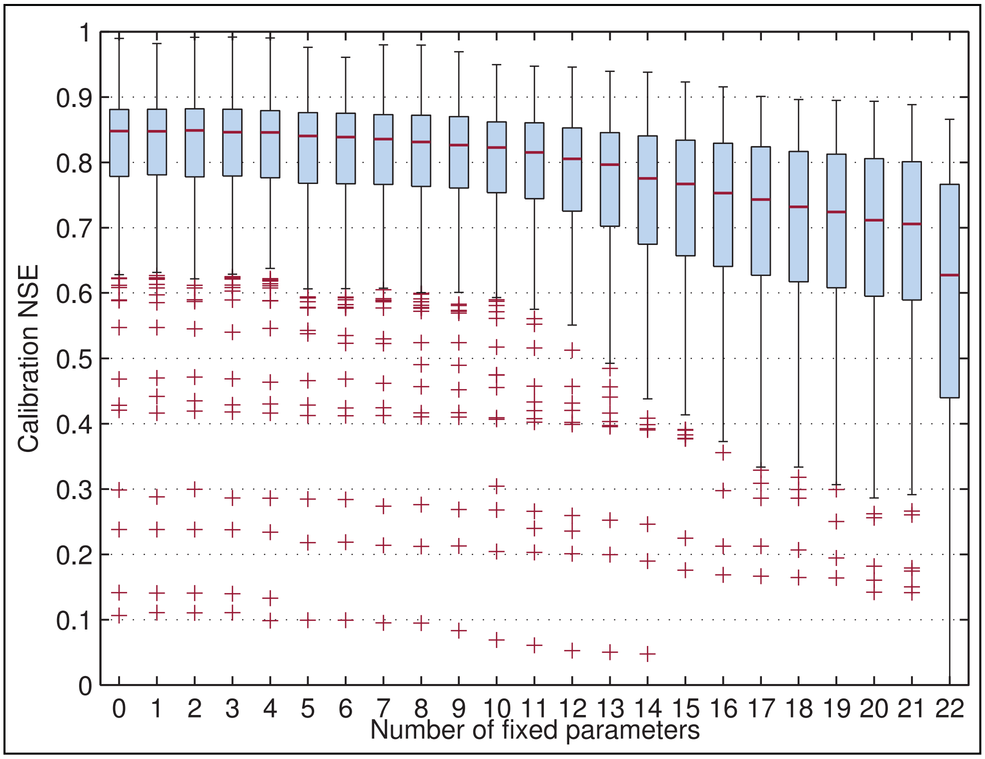

Figure 9.3 Calibration NSE for the HSAMI model with reducing number of free parameters

It can be seen in figure 9.3 that the first 4 or 5 parameters do not seem to contribute to the calibration skill, indicating that they could be removed without impacting the model performance. However, removing more parameters starts to noticeably degrade the NSE value.

The median calibration NSE results are also shown for the MOHYSE model in figure 9.4.

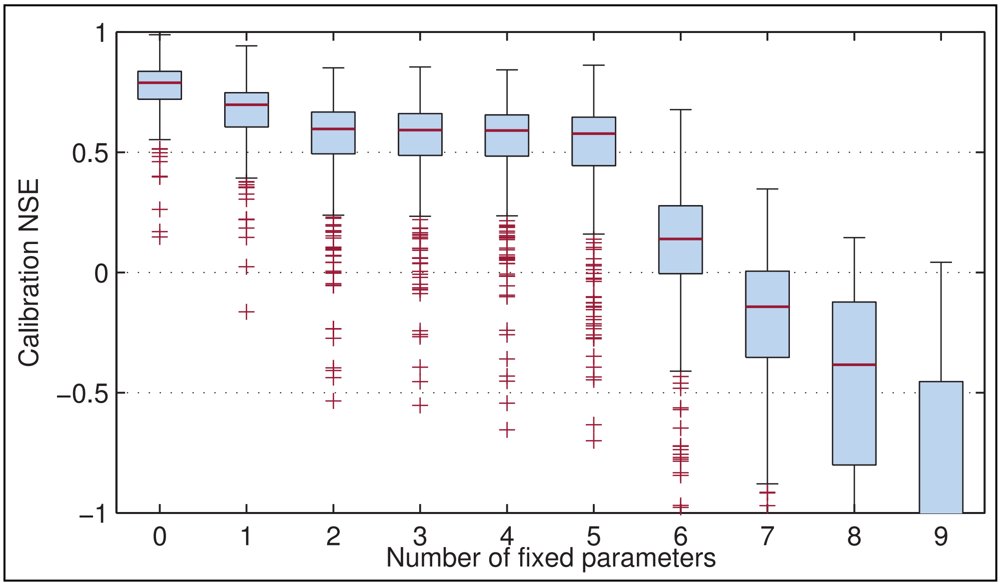

Figure 9.4 Calibration NSE for the MOHYSE model with reducing number of free parameters

Contrarily to the HSAMI model, MOHYSE is unable to cope with the fixing of its parameters in calibration. Even the least influential parameter, when fixed, produces an important loss in performance. This was predictable to some extent, as the least important parameter in MOHYSE explains 2.86% of the variance, whereas in HSAMI, the 19th fixed parameter explains the same amount of variance. This result goes against the recommendations of van Werkhoven et al. (2009) who recommend that parameters that explain less than 20% of the variance could be fixed. Also, the cumulative variance in HSAMI is approximately 2.5% when the first 12 parameters are accounted for. It is therefore not surprising that the MOHYSE model reacts in this manner since the model contains less parameters to begin with, thus they are expected to each bear more importance.

9.5.2 Regionalization application results

9.5.2.1 HSAMI Donor-based

The multi-donor approach was applied to all the regionalization methods except for the multiple regression method, which does not rely on donors. Results for the physical similarity method with inverse distance weighting (IDW) of the donor basins are presented in figure 9.5. Each panel contains the results for a given number of donors. The x-axis shows the number of fixed parameters in the HSAMI model. The NSE and NR values are the medians taken from the 100 regionalization iterations, for a total of 267 medians in each distribution. For the SR, each value corresponds to a metric related to a single iteration therefore the distributions represent all the project data.

The results in figure 9.5 show that the performance increases from 1 to 5 donors, then remains constant at 10 donors. The complete results (1 to 10 donors) are not shown here but the performance stops increasing after 5 donors. Figure 9.5 also shows that the overall performance of the physical similarity method is constant with up to 11 fixed parameters. When more parameters are fixed, the model starts losing its ability to adequately model flows, as it was seen in figure 9.3 (calibration). The effects of multi-donor averaging results were found to be similar for all the methods, with an optimum number of donors always found between 4 and 7.

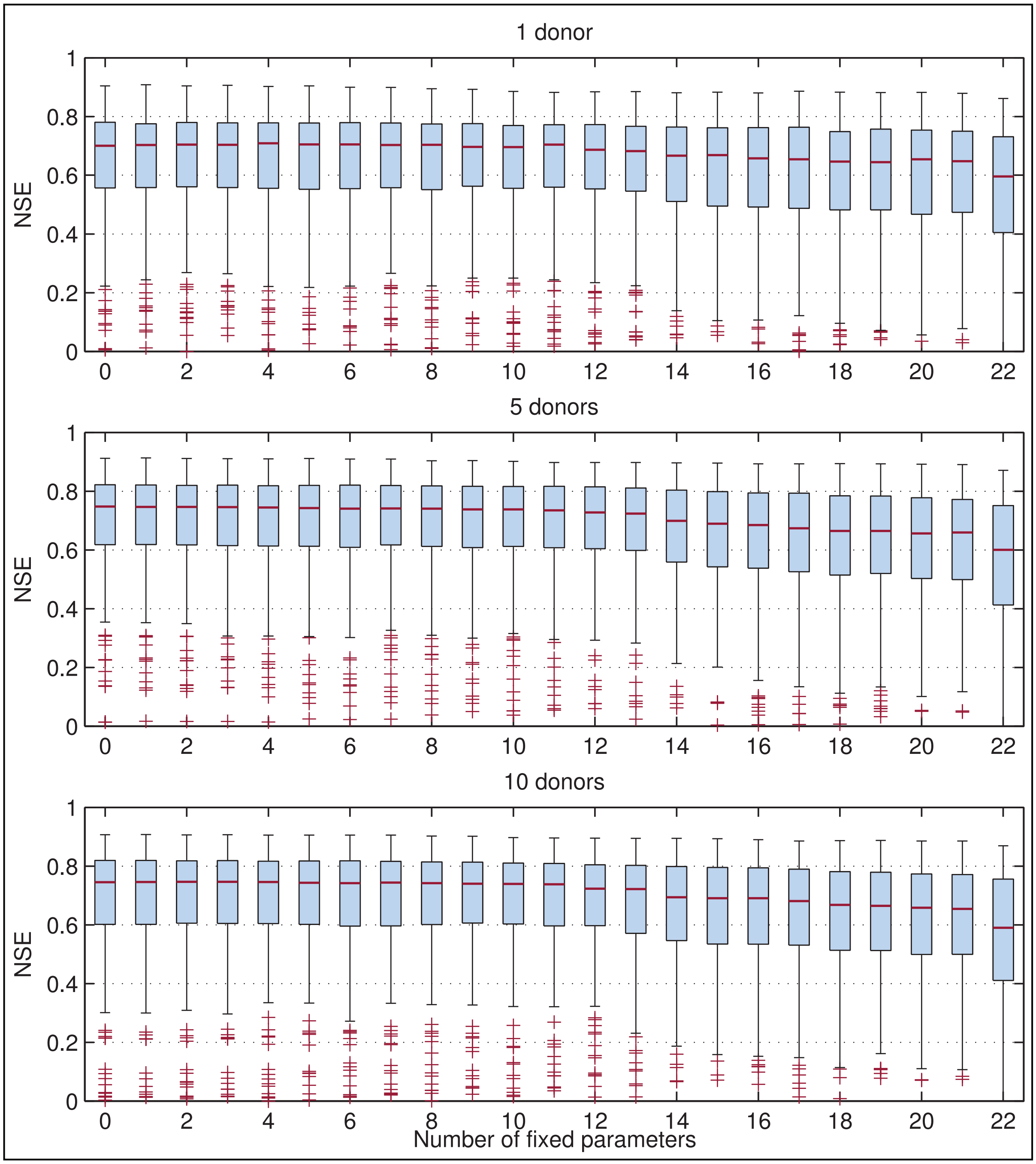

Figure 9.5 Nash-Sutcliffe Efficiency in regionalization using the physical similarity method with a reducing number of free parameters in the HSAMI model. Results for 1, 5 and 10 donor basins are shown

Figure 9.6 shows the results for all the methods when 5 donors are used. Once again, the NSE values are the medians taken from the 100 regionalization iterations.

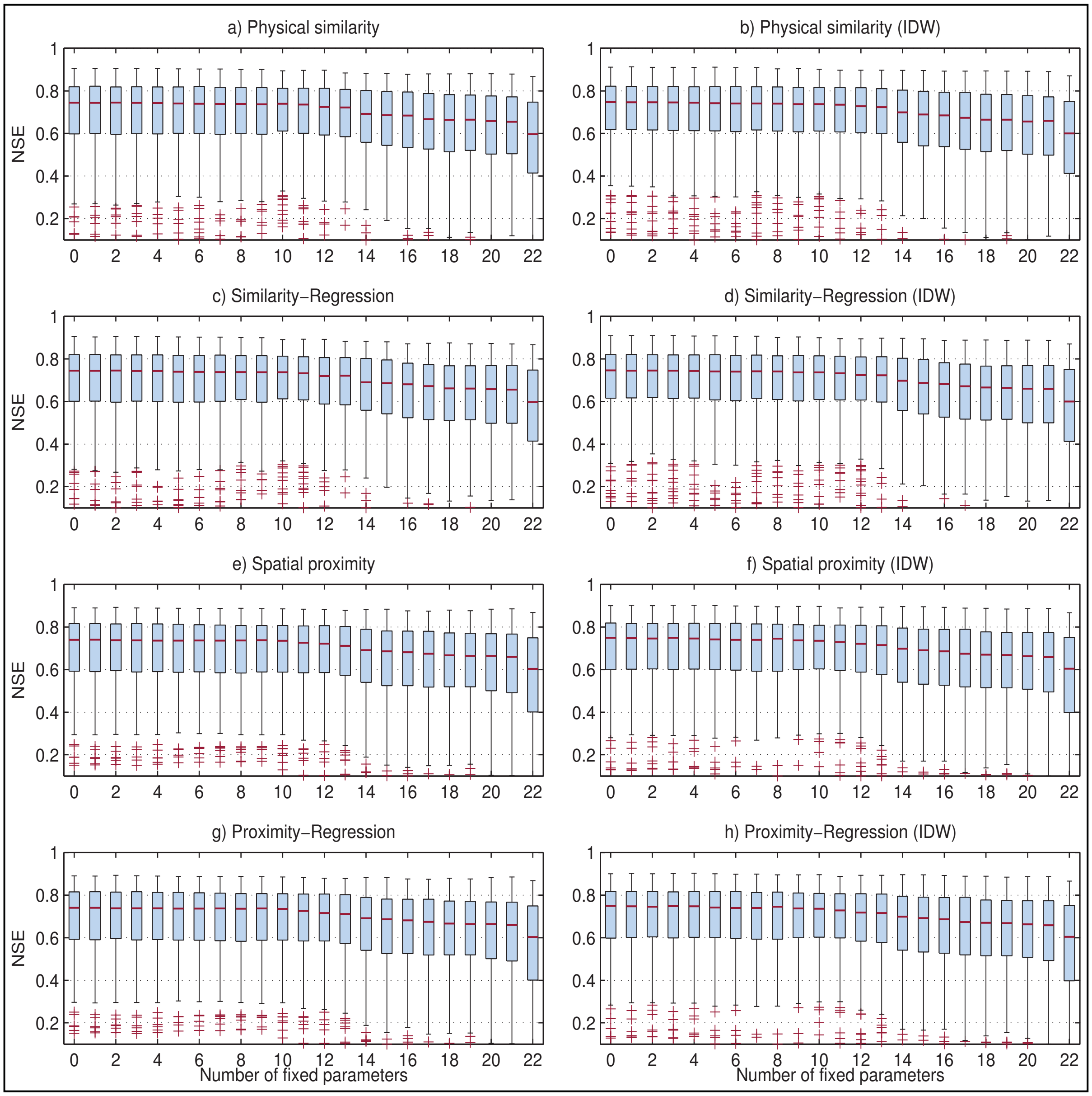

Figure 9.6 Nash-Sutcliffe Efficiency in regionalization using the eight donor-based methods (4 with simple mean and 4 with IDW) with a reducing number of free parameters in the HSAMI model. Results are displayed only for the 5-donor version

From figure 9.6, it seems that the regionalization methods all perform similarly when 5 donors are used. The NR was also evaluated to determine their performance independently from the model's ability to model streamflow on the basins. The NR results are not shown as they are similar for all methods, with the IDW versions slightly outperforming their simple mean counterparts. In all cases, the maximum number of removable parameters ranges from

10 to 11. Instead, the SR was evaluated over the 100 iterations to determine the percentage of catchments that are good regionalization targets. Figure 9.7 shows the SR with a diminishing free parameter dimensionality of the HSAMI model.

It can be seen that the SR is better for the spatial proximity method, which means more basins are successfully modelled. It is important to note that the important difference between the SR results and the NSE results are due to two factors. First, the SR is based on thresholds which amplify the differences due to minor differences around the threshold value. Second, the distribution in SRs uses all available information (100 iterations), whereas the NSE in figures 9.5 and 9.6 are the median value of the 100 iterations. This aggregation of the NSE values has the side-effect of neglecting the NSE spread, therefore eliminating the extreme extents of the results.

Also, compared to Arsenault and Brissette (2014), it is clear that the catchment descriptor (CD) selection here does worse than with the 4 CDs taken in the other paper. The 23-parameter version of HSAMI is better in the other paper in all aspects. The CD selection is the only difference between the two versions. Furthermore, the spatial proximity method performs better than physical similarity in the current study, whereas the opposite was true in the Arsenault and Brissette (2014) paper. This suggests that the latitude and longitude, which were part of the 4 CDs in the other paper, are now too diluted in the similarity index to bear any importance in the donor selection process. Alternatively, the added catchment descriptors could be too homogeneous and could not be discriminating enough to select the most suitable donors. Finally, it is important to see that to keep the SR intact, a maximum of 8 parameters should be fixed.

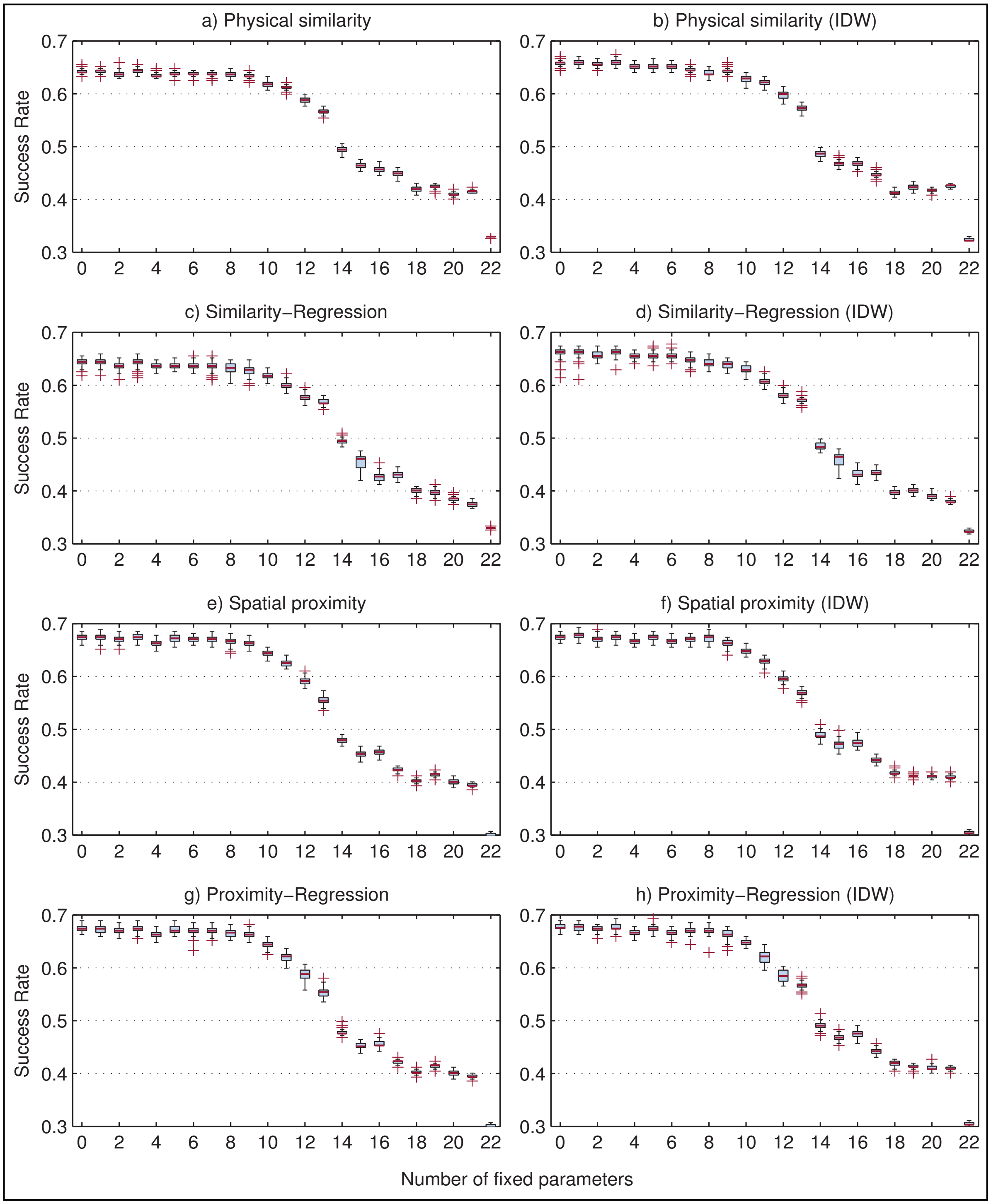

Figure 9.7 Success Rate in regionalization using the eight donor-based methods (4 with simple mean and 4 with IDW) with a reducing number of free parameters in the HSAMI model. Results are displayed only for the 5-donor version

9.5.2.2 MOHYSE donor-based

For MOHYSE, the regionalization methods displayed approximately identical behaviours. For this reason, only the results for the physical similarity method with inverse distance weighting are shown in figure 9.8.

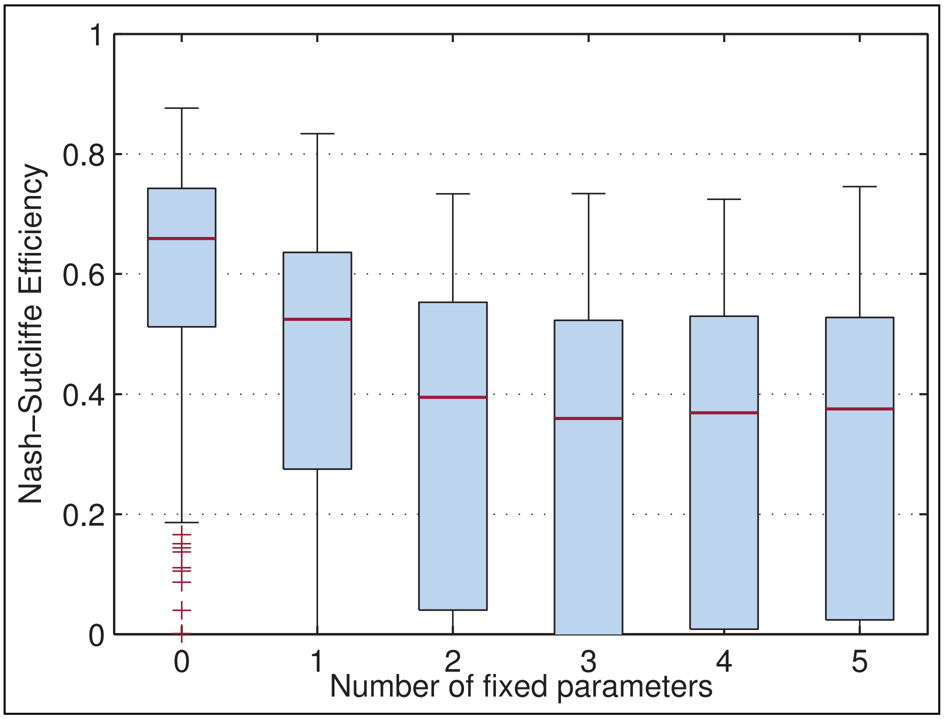

Figure 9.8 Nash-Sutcliffe efficiency in regionalization of the MOHYSE model using the physical similarity method with reducing number of free parameters and 5 donors. Only 5 parameters could be fixed before no more catchments were eligible to be donors.

However, the fixing of parameters was immediately detrimental to the regionalization skill. The first, least influential fixed parameter caused a decrease of 0.1 in NSE. It must be noted that figure 9.8 shows results only for up to 5 fixed parameters even though the MOHYSE model has 10 calibrated parameters. This is because basins with NSE values lower than 0.7 do not contribute as donors during regionalization, and with more than 5 fixed parameters, there remain no basins to act as donors, as can be seen in figure 9.4. For this reason, the MOHYSE model will not be evaluated further in this study. It is clear that the MOHYSE model, with its simpler structure, requires its 10 parameters for regionalization and should not be tampered with.

9.5.2.3 Regression based

The results for the multiple linear regression method are shown in figure 9.9. The panels respectively represent the Nash-Sutcliffe efficiency, the Nash Ratio and the Success Rate for the 23 versions of HSAMI.

The regression method underperforms the donor-based methods by approximately 0.05 in NSE values, except when more than 20 parameters are fixed. With more fixed parameters, the donor-based methods see a larger drop in performance than the regression method, thus eliminating the gap between the two methods. The SR, however, shows larger discrepancies between the regression method and the others. Even in the best case, flows are successfully predicted in less than 50% of the catchments. Figure 9.9 also shows the effects of reducing parameter correlation on their identifiability. The reduction in parameter dimensionality and its associated loss in model performance are compensated by the increase in parameter identifiability for the regression method. This can be seen for the 5th and 10th fixed parameters in particular, for which there is a sudden increase in performance and slow degradation thereafter. The regression methods performance depends on the regression model's ability to use the predictors (catchment descriptors) to estimate the parameter values.

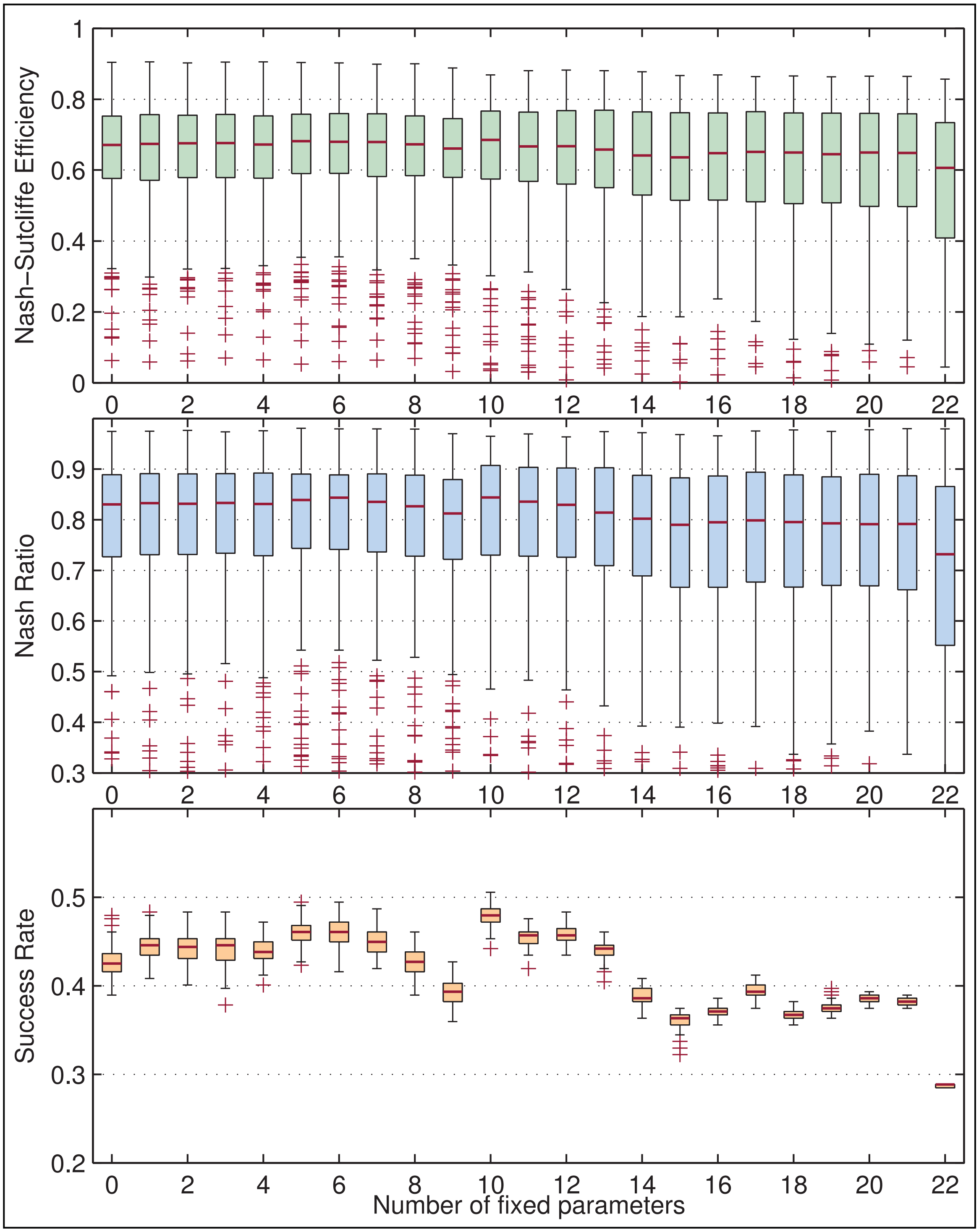

Figure 9.9 Nash-Sutcliffe Efficiency, Nash Ratio and Success Rate for the multiple linear regression method and a reducing number of parameters for the HSAMI model

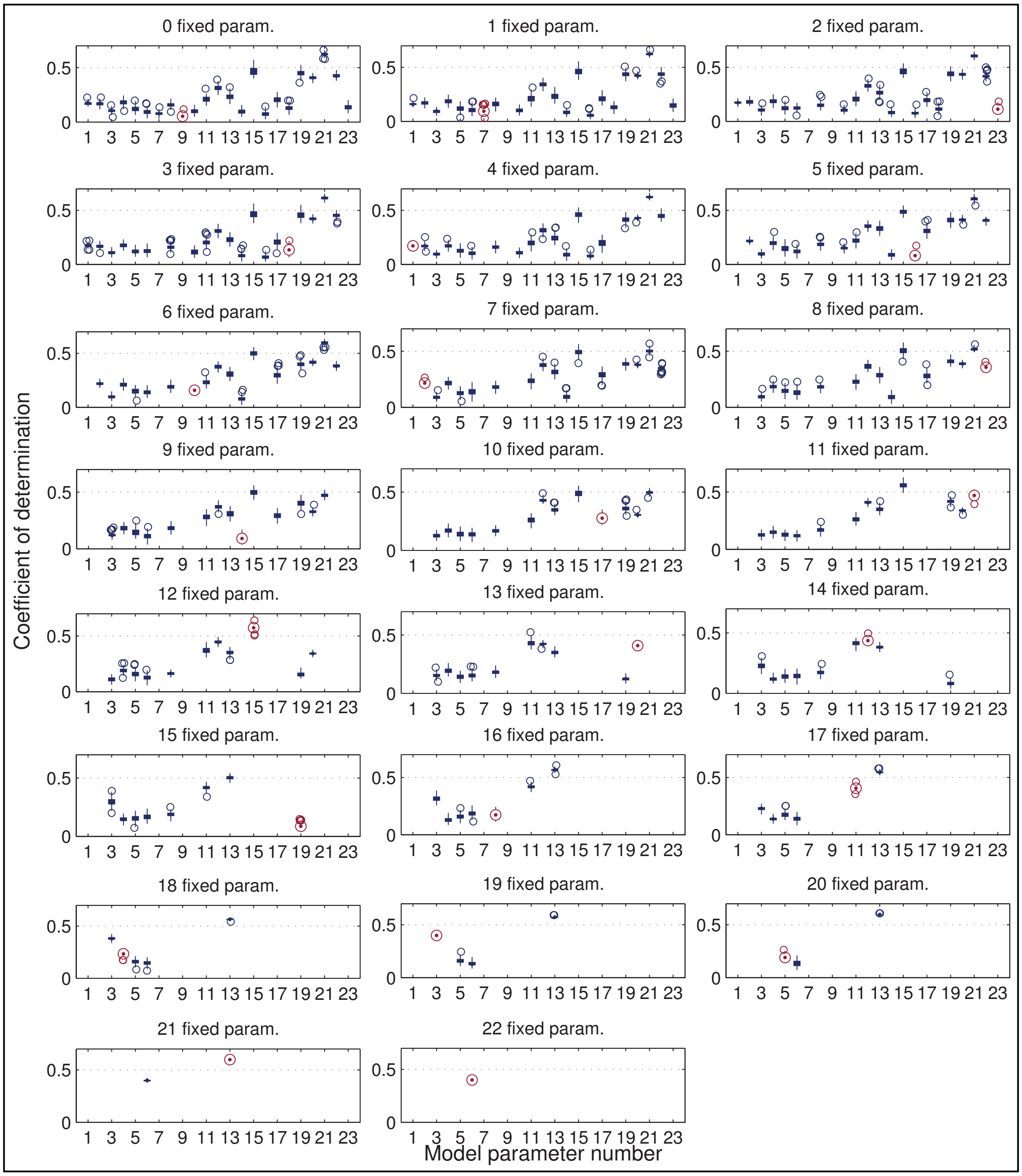

Figure 9.10 Regression models' coefficients of determination with varying number of fixed parameters. Each panel represents the HSAMI model with a different number of free parameters. Each box-and-whisker plot represents the distributions of the 100 iterations with randomly selected parameter sets. The highlighted parameters are the next ones to be fixed according to the Sobol' sensitivity analysis.

Figure 9.10 shows the regression method's coefficients of determination for each of the variable models' parameters. Each of the panels represents one version of HSAMI (with a given number of fixed parameters). The red color indicates that the parameter is the next to be assigned a fixed value.

Three main points can be taken from figure 9.10. First, only a few parameters have good identifiability with the regression model ($R^2 > 0.5$). Second, the order of parameter fixing is not dependent on the R^2 score. Finally, the fixing of parameters does influence the R^2 of the variable parameters in the parameter-reduced model. For example, parameter 6 has relatively low coefficients of determination (lower than 0.2) until parameter 5 is fixed, at which point its R^2 value jumps to 0.4. Parameter 15 also displays the same trend, and becomes fixed while it is at its highest R^2 level. This shows that the importance of a parameter is not related to the correlation between that parameter and the predictors for regionalization purposes. Also, parameter 21 was the only one displaying a good R^2 coefficient (R^2>0.5) when all the parameters were free, which matches the findings in Arsenault and Brissette (2014) for the regression-augmented methods. However the high-R^2-scoring parameters evolve with the number of fixed parameters, with parameters 13 and 15 briefly becoming the most predictable (with the strongest correlation between the parameter values and the catchment descriptors). Nonetheless, most of the parameters never enter this range.

9.5.3 Robustness evaluation

A test was devised to detect any difference in model robustness due to parameter fixing. In this test, the model was run on each catchment with parameters donated from the other 266 catchments. This produces 266 individual streamflow values for the selected pseudo-ungauged basin. The NSE is then computed between these flows and the observed flows on the target catchment. The process is repeated for all the catchments, thus resulting in 267 catchments x 266 donors per catchment = 71022 NSE values. The procedure was reiterated for the 23 versions of the HSAMI model. The results of this test are presented in figure 9.11.The idea behind this test is to estimate the regionalization skill probability if a random basin were selected as a donor. Note that the outliers are not shown in figure 9.11, however they represent less than 0.7% of the data and they follow the same trend as the quartiles and median.

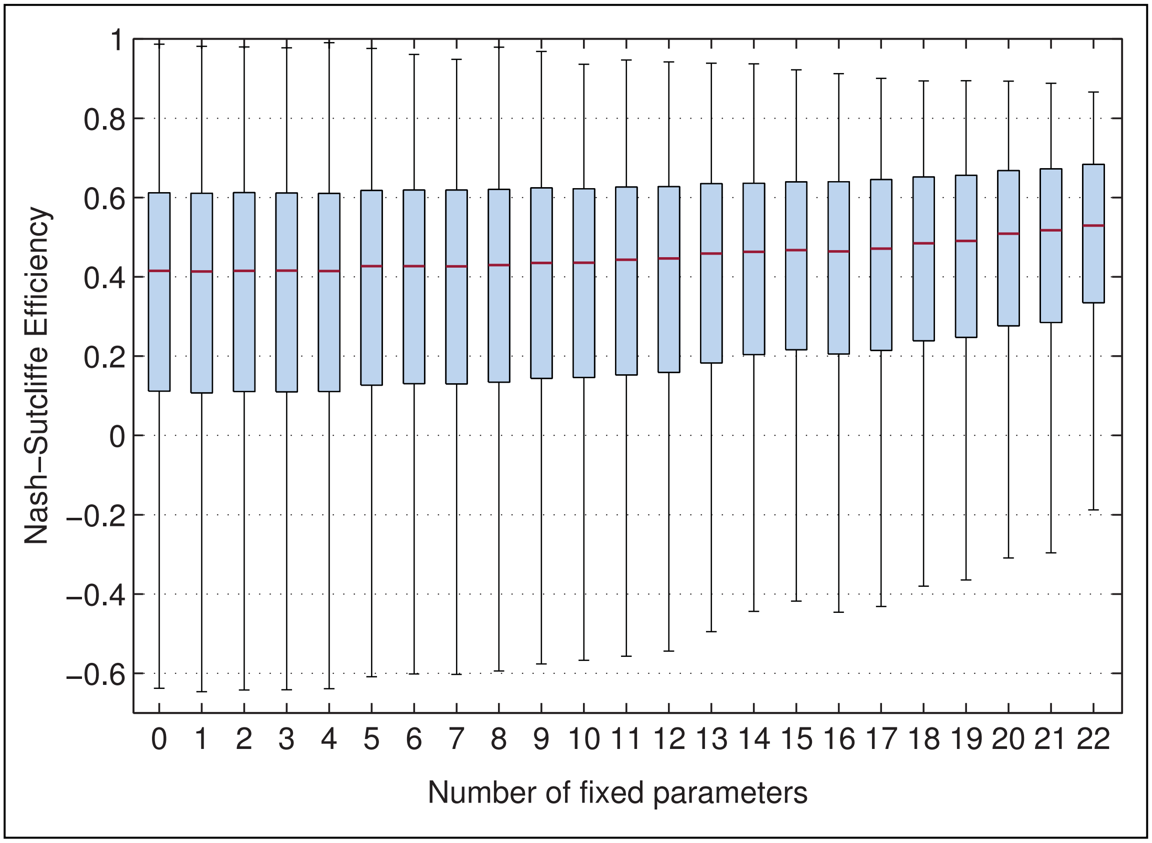

Figure 9.11 Nash-Sutcliffe Efficiency distributions when all basins are simulated with all of the 266 other parameter sets. The median and low-scoring values improve with fewer parameters at the expense of the best possible values. Each box-and-whisker plot represents the HSAMI model with a given number of fixed parameters.

Figure 9.11 shows that while the best attainable NSE value decreases along with a reduction in model dimensionality, the 25th, 50th and 75th percentiles actually improve. It is important to note that this method represents the NSE values when donors are randomly selected. When comparing the results in figure 9.11 to those in figure 9.6, it is clear that the regionalization methods are better than random selection of donors since their performance is similar to the best possible randomly selected donor, but that parsimonious models are indeed affected less by parameter correlations.

9.6 Discussion

9.6.1 Verification of the main hypothesis

The main hypothesis in this study was that reducing the number of parameters in lumped conceptual rainfall-runoff models could lead to better parameter identifiability. It was also sought to determine if regionalization performance could be improved by limiting the parameter correlations and increasing parameter identifiability. For HSAMI, The results showed that it is possible to fix up to 8 parameters with no loss in performance (and with up to 11 parameters with minimal loss in performance) in regionalization, indicating that the model is indeed overparameterized. The Sobol' method results showed that the variance explanation of these parameters is negligible. However, there was no improvement in regionalization when fixing any amount of parameters. For MOHYSE, it was shown that the model requires all its parameters since there is an important drop in calibration performance when the least influential parameter is fixed. This loss in performance was transposed in the regionalization mode. By definition, the MOHYSE model was designed to be parsimonious. The fact that no parameters are useless confirms that the model requires its 10 parameters.

The results with the donor-based approaches seem to invalidate the hypothesis because the reduction in parameter correlations did not allow better regionalization performance. However, when looking at the results for the multiple linear regression method (figures 9.9 and 9.10), it can be seen that there are elements that tend to support the hypothesis. In figure 9.9, the results show jumps in the NSE, NR and SR when the 5th and 10th parameters are fixed. The following decline (fixed parameters 11-15) is gradual and follows the calibration NSE trend. These jumps demonstrate that the fixing of a parameter can reduce (or eliminate) the correlations it has with other parameters that require calibration, thus improving parameter identifiability. Figure 9.10 shows the same conclusions in more detail. The most outstanding manifestation is found when the 20th parameter must be fixed. Indeed, when parameter 5 is fixed, the regression model for parameter 6 sees its coefficient of determination R^2 increase from 0.1 to 0.4. The same can be seen for other parameters, notably for parameter 15, whose R^2 statistic steadily increases until it is fixed and eliminated.

However, these improvements in parameter identifiability are not sufficient to overcome the loss in the model’s degrees of freedom, therefore resulting in a net loss in regionalization applications. The parameter/catchment descriptor correlations are limited by the quality of available catchment descriptors as well as the calibrated parameter values, however some of the most influential descriptors are unavailable in the study area (such as soil properties and hydraulic conductivity). This subject will be discussed in a further section.

9.6.2 Sobol’ sensitivity analysis

The Sobol’ sensitivity analysis was used in this paper in spite of its known drawbacks. The Sobol’ method is based on the strong hypothesis that the parameters are uncorrelated. However, this is known to be false for most hydrological models. An increasing number of studies are using the Sobol’ method for parameter reduction in hydrological models, and they show good results. In the past few years, mathematical researchers have started to address this issue with modifications to the Sobol’ method to make it compatible with correlated inputs (Chastaing et al. 2014). The main limitation of these methods is that they require less than 7 to 8 parameters to converge with reasonable computing power. Our 23 parameter version of the HSAMI model was simply too complicated to be successfully processed with these improved algorithms. The results with the classical Sobol’ analysis show that it does perform well nonetheless, however there is reason to believe that more precision and better overall results could be obtained if improved Sobol’ methods were to be applied on complex hydrological models and their parameters. In any case, the methodology’s robustness would lend credit and give more confidence in the results.

Another important aspect of the Sobol’ analysis was the selection of a transformed NSE metric to constrain the results between 0 and 1. Previous tests showed that non-important parameters could be falsely attributed high rankings if they lead to very bad NSE results. Parameter boundaries for calibration could be a source to this problem, as parameter values that result in very poor (or even impossible) model simulations would seem very important to Sobol’. Usually this is not a problem since the calibration algorithm selects parameter sets that generate acceptable flows, but the Sobol’ method samples uniformly across the parameter space without taking interdependence and correlations into account. Therefore, two parameters which are inversely proportional to one another would lead to good model performance when calibrated together, but the Sobol' method could sample a point where both parameters are at their maximum value. In such cases, the model could not generate acceptable flows and the very low NSE (sometimes negative infinity) would prompt the Sobol' method to label the parameters as very important, as they would explain a large part of the variance. By taking the $(2\text{-NSE})^{-1}$ metric, the bad results are given a more appropriate, less biased, comparative basis. Nonetheless, the regionalization and calibration aspects of this work were conducted with the classic NSE metric, with the transformed NSE metric being used only to determine the parameters' importance.

9.6.3 Parameter fixing

One of the main choices made in this paper was to set the fixed parameter values to the median calibration value. This has implications for the regionalization methods performances when an important number of parameters is fixed. By constraining the parameter space to a smaller sub-region, the parameter correlations are reduced. But at the same time, the performance gain usually brought upon by the synergetic parameter interdependence is lost. This was seen in the results, where influent parameters cause a drop in performance when they are fixed. For the least important parameters, the impact is more subtle since they can take any value without reducing the model performance in a measurable manner. This was also clearly demonstrated in the results section. In addition, the fact that all the catchments were used to calculate the median parameter value automatically added a bias in the results. In real applications, the calibrated parameter set would not be known for the ungauged basins, therefore the parameter fixing should not have taken them into account. However, this would have required 267 independent iterations of the paper, which would lead to almost identical results. Indeed, by taking the median, the effects of this choice become negligible, especially at this scale.

It is also important to note that the parameter fixing order was based on the total order sensitivity indices, which include the variance explanation of the parameter itself added to the variance caused by interactions with other parameters. Since the aim was to reduce parameter correlations, the total-order indices were selected as they contain information about the parameter dependencies as opposed to the first-order indices.

The re-calibration of the remaining parameters after each new parameter fixing was an essential step in making sure that the correlations with the fixed parameters were removed. In doing so, the idea was to allow the parameters to take values that would represent the physical processes better without the effects of interactions with the fixed parameters. It was shown that the calibration performance weakened with each new fixed parameter, which is expected due to the more limited degrees of freedom. However the regionalization performance did not suffer until at least 8 parameters were fixed. This clearly points to an overparameterization of the HSAMI model.

Finally, it is very important that model parameter reduction projects be reassessed at regular intervals. In some cases, a parameter could be used only on rare occasions (processes that occur after a certain flow threshold, or extreme events for example) and not be useful in the calibration period. In such cases, the parameter would have no effect on the model results. Therefore it would be considered as the least influential and could be fixed to a reasonable value. However, if an event happened which would warrant the use of the parameter, the Sobol' analysis and parameter fixing should be undertaken again with the new period and its extreme event. This would ensure that the model is not artificially limited by a fixed parameter. Regular re-calibrations and Sobol' analyses should be implemented to any project framework in this regard. However, the larger picture seems to be that for regionalization applications, fixing hydrological model parameters is not worthwhile. Lots of effort must be put in to reduce the number of parameters and finding adequate fixed parameter values, but the results show that there is no added value in doing so. In fact, the regionalization approaches perform either as well or better with more parameters. Perhaps the best strategy is to give the model more degrees of freedom to better integrate the parametric interdependence, which is linked to the hydrologically important catchment descriptors.

9.6.4 Regionalization performance

An important aspect driving the regionalization methods performance is the catchment descriptors selected in this study. In a previous paper, it was shown that the optimal set of catchment descriptors contained the latitude, longitude, mean annual precipitation and water fraction of land cover (Arsenault and Brissette 2014). In the present study, all the available physical and meteorological descriptors were taken and the results show that the methods performed worse overall. The reasoning was that limiting the number of descriptors would hinder the ability to detect correlations between the model parameters and the basins' physical characteristics. The fact that the regression models were mostly unable to provide good R^2 values shows that the descriptors were either inadequate estimators of hydrologic regime, that the linear regression model is inadequate or that the hydrological model is incapable of taking this information into account. Since it was showen that reducing parameter equifinality and improving identifiability can influence the regression model predictive skill, the evidence seems to indicate that the linear regression model is unable to find the more complex interactions. Also, hydrologic models seem too conceptual to make better use of the physical properties of the basins. The parameter values are too weakly influenced by the physical characteristics of the modelled basins, as was reported by Lee et al. (2006).

In figure 9.10, it is clear that the parameters which show higher R^2 values by the regression model are the parameters that control runoff horizontal and vertical transfer, which likely depend more on the basins' physical characteristics than their climate characteristics. The climate-based descriptors and some model parameters are strongly correlated, thus leading to situations as in figure 9.10 when 20-21 parameters are fixed. The R^2 improves for the remaining parameters in figure 9.10, but the overall model performance diminishes as seen in figures 9.5 and 9.6. This is the result of the parameter identifiability improvement (R^2 increases) as well as the generalized reduced model performance following parameter fixing (calibration NSE and regionalization NSE decrease). It also explains the poor NSE, NR and SR values for the multiple linear regression method, in that the drop in model performance is larger than the improvement in parameter identifiability. Furthermore, the increase in R^2 does not indicate a necessarily good fit as the R^2 never rises above 0.7.

In the case of the regression-augmented proximity and similarity methods, the results do not seem to show any improvements over their standard counterparts. In some cases (e.g. for 21 fixed parameters), the performance is lower than the standalone version. In these cases, it is hypothesized that the strong correlation between the remaining parameters is broken and the parameter values are replaced by incoherent predicted values. It is important to note that only one parameter (parameter 21) was consistently modified according to the regression model, as seen in figure 9.10. Within our previous work, it was exposed that there was a net improvement with these methods, whereas in this paper they are almost identical to the basic spatial proximity and physical similarity methods. The only difference in the 23-parameter version of HSAMI between both papers is that the catchment descriptors are not the same. This is further proof that catchment descriptors play a major role in the PUB problem.

It was also shown that multi-donor averaging is a necessity in regionalization. From figure 9.5, it can be seen that using 5 donors improves performance to a point where it is possible to fix 15 parameters and still obtain the same level of performance of the full 23 free parameters version of HSAMI. It was found that 5 to 6 donors were optimal for the present study, which is consistent with the body of literature. It is also clear from the literature that hydrological models with many parameters can easily be constrained by fixing parameters, as this paper reveals. What is not as obvious is the uncertainty reduction. It was expected that reducing the number of parameters would reduce model uncertainty in regionalization. However from figures 9.5, 9.6 and 9.7, it would seem that similar levels of uncertainty remain. The equifinality was already known to play a minimal role in regionalization (Arsenault and Brissette 2014), and this study confirms these results. The SR metric becomes progressively less dispersed as more parameters are fixed, as can be seen in figures 9.7 and 9.9, but the scales are relatively small and no significant trends to overall results could be explained by the equifinality, which seems to barely affect the results. Figure 9.11 shows that the robustness of the model is improved which is expected as the lower parametric dimensionality forces the model to respond similarly for the different catchments. This comes at the cost of lower performance of the best simulations.

Overall, the methods in this paper respond similarly to the fixing of parameters, without improving performance. In light of the results, it seems clear that conceptual models are able to correlate catchment descriptors to their parameters in complex, non-linear ways, but are limited in their ability of doing so because the hydrological processes and the catchment properties are ill-defined. Therefore removing correlated parameters does allow better parameter estimation, but the complex interactions between the parameters are what allow the models to perform under these circumstances. Parameter identifiability could be improved, as was shown in figure 9.10, but the parameters are not hydrologically meaningful by themselves. Therefore unless better catchment descriptors become available (most notably soil properties), there is no reason to believe that better parameter identifiability would lead to better regionalization performance (Oudin et al. 2010). Even then, some of the most correlated parameters were fixed early in the Sobol' analysis as they were not influential in the hydrological model. In fact, our conceptual models are not penalized by donor-based regionalization approaches even under the assumption of overparameterization as the effects of equifinality were shown to be relatively minor.

9.7 Conclusions

In this paper, the effects of parameter correlations and equifinality in regionalization were investigated. A parameter fixing framework was used to attempt to improve parameter identifiability and to assess its effects on regionalization performance using 5 regionalization methods on 267 catchments in the province of Quebec, Canada. The Sobol' sensitivity analysis identified and ranked model parameters according to their influence (both direct and combined with other parameters) on the model performance. The HSAMI model was shown to be overparameterized as up to 11 parameters could be fixed (out of 23) with little to no loss in regionalization skill. The simpler MOHYSE model, however, could not withstand the fixing of a single parameter. With both HSAMI and MOHYSE, the performance dropped significantly after fixing enough parameters to reduce the cumulative total variance by less than 3%, which corresponds to 12 parameters for HSAMI and 1 parameter for MOHYSE. The multiple linear regression method's analysis showed that there are correlations between model parameters and catchment descriptors; however these links are weak and convoluted. Reducing parameter correlations did improve parameter identifiability, but this did not transpire in the regionalization results. For the 5 regionalization methods tested in this paper, the results were never better than when all the parameters were free. The main conclusions are twofold. First, conceptual models do not represent the hydrological processes in a physically adequate manner, requiring intricate links between catchment descriptors and model parameters subject to non-uniqueness to perform well. Second, it was shown that model parameters cannot be individually linked to physical catchment characteristics but that the parameter sets as a whole, including interdependencies, can be related to a set of catchment descriptors. This is why the donor-based methods perform better than the regression based ones. Therefore it is recommended to use all the model parameters in regionalization, even when in presence of equifinality and strong parameter interactions. The results in this paper show that the uncertainty brought upon the regionalization results by the extra parameters are negligible as compared to the loss in performance by fixing them. Future work could make use of this information to either (1) add parameters to models to see if the opposite of this study would generate positive results, (2) attempt the same methodology using a more complex, more physical model and better catchment descriptors and (3) attempt to scale the improved Sobol' sensitivity analysis methods to the dimensions of hydrologic models to eliminate the parameter independence hypothesis of the Sobol' method.

9.8 Acknowledgments

The authors would like to thank Catherine Guay from IREQ for making available the CQ2 database, and Kenjy Demeester from École de technologie supérieure for providing the land cover data for the project.

9.9 References

Arsenault, R., Malo, J., Brissette, F., Minville, M. and Leconte, R. (2013) Structural and non-structural climate change adaptation strategies for the Péribonka water resource system. Water Resources Management, 27(7), 2075-2087. doi: 10.1007/s11269-013-0275-6.

Arsenault, R., and Brissette, F., 2014. Continuous streamflow prediction in ungauged basins: The effects of equifinality and parameter set selection on uncertainty in regionalization approaches. Water Resour. Res. 50 (7), 6135-6153. doi: 10.1002/2013WR014898.

Arsenault, R., Poulin, A., Côté, P. and Brissette, F., 2014. Comparison of stochastic optimization algorithms in hydrological model calibration. Journal of Hydrologic Engineering, 19(7), 1374-1384. doi: 10.1061/(ASCE)HE.1943-5584.0000938.

Arsenault, R. and Brissette, F., 2015. Multi-model averaging for continuous streamflow prediction in ungauged basins. Hydrolog. Sci. J., 38p. (Under re-review)

Bao, Z., J. Zhang, J. Liu, G. Fu, G. Wang, R. He, X. Yan, J. Jin, and H. Liu (2012), Comparison of regionalization approaches based on regression and similarity for predictions in ungauged catchments under multiple hydro-climatic conditions, J. Hydrol., 466–467, 37-46 doi:10.1016/j.jhydrol.2012.07.048.

Bardossy, A. (2007), Calibration of hydrological model parameters for ungauged catchments, Hydrol. Earth Syst. Sci., 11, 703–710.

Beven, K. (2006), A manifesto for the equifinality thesis, J. Hydrol, 320, 18–36.

Burn, D.H., and D.B. Boorman (1993), Estimation of hydrological parameters at ungauged catchments, J. Hydrol, 143(3–4), 429-454.

Chastaing, G., C. Prieur, and F. Gamboa (2014) Generalized Sobol sensitivity indices for dependent variables: numerical methods. J. Statist. Comput. Simulation, Taylor & Francis: STM, Behavioural Science and Public Health Titles, 1-28.

Chen, J., Arsenault, R. and Brissette, F. (2015) Reducing the parametric dimensionality for rainfall-runoff models: a benchmark for sensitivity analysis methods. Advances in Water Resources, 39p. (Manuscript under review).

Chen, J., F. P. Brissette, A. Poulin, and R. Leconte (2011), Overall uncertainty study of the hydrological impacts of climate change for a Canadian watershed, Water Resour. Res., 47, W12509, doi:10.1029/2011WR010602

Chen, J., F. P. Brissette, D. Chaumont, and M. Braun (2013), Finding appropriate bias correction methods in downscaling precipitation for hydrologic impact studies over North America, Water Resour. Res., 49, 4187-4205, doi:10.1002/wrcr.20331.

Duan, Q., Sorooshian, S. and Gupta, V. K., 1992. Effective and efficient global optimization for conceptual rainfall runoff models. Water Resources Research, 24(7), 1163-1173.

Duan, Q., Sorooshian, S. and Gupta, V. K., 1993. A shuffled complex evolution approach for effective and efficient optimization. Journal of Optimization Theory and Applications, 76(3), 501-521.

Duan, Q., Sorooshian, S. and Gupta, V. K., 1994. Optimal use of the SCE-UA global optimization method for calibrating watershed models. Journal of Hydrology, 158, 265-284.

Fortin, V. (2000), Le modèle météo-apport HSAMI: historique, théorie et application, Varennes: Institut de Recherche d'Hydro-Québec, 68 p.

Fortin, V. and Turcotte, R., 2007. Le modèle hydrologique MOHYSE (bases théoriques et manuel de l'usager). Note de cours pour SCA7420, Département des sciences de la terre et de l'atmosphère, Montréal : Université du Québec à Montréal, 1-17.

Hansen, N. and A. Ostermeier (1996), Adapting arbitrary normal mutation distributions in evolution strategies: The covariance matrix adaptation, In Proceedings of the 1996 IEEE International Conference on Evolutionary Computation, 312-317.

Hansen, N., and A. Ostermeier (2001), Completely derandomized self-adaptation in evolution strategies, Evolutionary Computation, 9(2), 159-195.

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

Hutchinson, M. F., D.W. McKenney, K. Lawrence, J. H. Pedlar, R. F. Hopkinson, E. Milewska, and P. Papadopol, (2009), Development and testing of Canada-wide interpolated spatial models of daily minimum-maximum temperature and precipitation for 1961-2003, J. Appl. Meteor. Climatol., 48, 725-741.

Lee H., N. R. McIntyre, H. S. Wheater, and A. R. Young (2006), Predicting runoff in ungauged UK catchments, Proceedings of the Institution of Civil Engineers. Water Management 159(2): 129–138.

Lee, H., McIntyre, N., Wheater, H., and Young, A., (2005). Selection of conceptual models for regionalisation of the rainfall-runoff relationship. Journal of Hydrology, 312(1–4), 125-147. http://dx.doi.org/10.1016/j.jhydrol.2005.02.016.

McIntyre, N., H. Lee, H. Wheater, A. Young, and T. Wagener (2005), Ensemble predictions of runoff in ungauged catchments, Water Resour. Res., 41, W12434, doi:10.1029/2005WR004289.

Merz, R., and G. Blöschl (2004), Regionalization of catchment model parameters, J. Hydrol, 287, 95–123.

Minville, M., F. Brissette, and R. Leconte (2008), Uncertainty of the impact of climate change on the hydrology of a Nordic watershed, J. Hydrol., 358(1-2), 70-83.

Minville, M., F. Brissette, S. Krau, and R. Leconte (2009), Adaptation to climate change in the management of a Canadian water resources system, Water Resour. Manage., 23(14), 2965-2986.

Minville, M., S. Krau, F. Brissette, and R. Leconte (2010), Behaviour and performance of a water resource system in Québec (Canada) under adapted operating policies in a climate change context, Water Resour. Manage., 24, 1333–1352.

Nash, J. E., and Sutcliffe, W. H., 1970. River flow forecasting through conceptual models: Part 1. A discussion of principles. Journal of Hydrology, 10(3), 282-290.

Nossent,J., P. Elsen, W. Bauwens (2011). Sobol' sensitivity analysis of a complex environmental model. Environmental Modelling and Software, 26 (12) , 1515–1525.

Nossent, J., and Bauwens, W. (2012), Optimising the convergence of a Sobol' sensitivity analysis for an environmental model: application of an appropriate estimate for the square of the expectation value and the total variance. In Proc. of the International Environmental Modelling and Software Society (iEMSs), 2012 International Congress on Environmental Modelling and Software, Managing Resources of a Limited Planet: Pathways and Visions under Uncertainty, Sixth Biennial Meeting, Leipzig, Germany, 1 - 5 July 2012.

Oudin, L., V. Andréassian, C. Perrin, C. Michel, and N. Le Moine (2008), Spatial proximity, physical similarity, regression and ungaged catchments: A comparison of regionalization approaches based on 913 French catchments. Water Resour. Res., 44, W03413. doi:10.1029/2007WR006240.

Oudin, L., A. Kay, V. Andréassian, and C. Perrin, (2010). Are seemingly physically similar catchments truly hydrologically similar? Water Resour. Res., 46, W11558, doi:10.1029/2009WR008887.

Parajka, J., A. Viglione, M. Rogger, J. L. Salinas, M. Sivapalan, and G. Blöschl (2013), Comparative assessment of predictions in ungauged basins – Part 1: Runoff hydrograph studies, Hydrol. Earth Syst. Sci. Discuss., 10, 375-409, doi:10.5194/hessd-10-375-2013.

Parajka, J., R. Merz, and G. Blöschl (2005), A comparison of regionalisation methods for catchment model parameters, Hydrol. Earth Syst. Sci., 9, 157–171.

Poulin, A., F. Brissette, R. Leconte, R. Arsenault, and J. Malo (2011), Uncertainty of hydrological modelling in climate change impact studies, J. Hydrol, 409(3–4), 626–636.

Razavi, T., and P. Coulibaly (2013), Streamflow prediction in ungauged basins: Review of regionalization methods, J. Hydrol. Eng., 18(8), 958–975.

Saltelli, A., and Tarantola, S. (2002). On the relative importance of input factors in mathematical models: safety assessment for nuclear waste disposal. J Am Stat Assoc, 97 (459), 702–709.

Samuel, J., P. Coulibaly, and R. Metcalfe (2011), Estimation of continuous streamflow in Ontario ungauged basins: Comparison of regionalization methods, J. Hydrol. Eng., 16(5), 447–459.

Sefton, C. E. M., and Howart, S. M. (1998). "Relationships between dynamic response characteristics and physical descriptors of catchments in England and Wales." J. Hydrol., 211(1–4), 1–16.

Seibert, J. (1999), Regionalisation of parameters for a conceptual rainfall runoff model, Agric. For. Meteorol., 98-99(31), 279–293.

Sivapalan, M., K. Takeuchi, S. W. Franks, V. K. Gupta, H. Karambiri, V. Lakshmi, X. Liang, J. J. McDonnell, E. M. Mendiondo, P. E. O'Connell, T. Oki, J. W. Pomeroy, D. Schertzer, S. Uhlenbrook, and E. Zehe (2003), IAHS Decade on Predictions in Ungauged Basins (PUB), 2003-2012: Shaping an exciting future for the hydrological sciences, Hydrolog. Sci. J., 48, 857-880.

Sobol', I. M. (1993), Sensitivity estimates for nonlinear mathematical models, Math. Model. Comput. Exp., 1, 407–17.

Tang, Y., Reed, P., Wagener, T., and van Werkhoven, K.: Comparing sensitivity analysis methods to advance lumped watershed model identification and evaluation, Hydrol. Earth Syst. Sci., 11, 793-817, doi:10.5194/hess-11-793-2007.

Vandewiele, G. L., and Elias, A. (1995). "Monthly water balance of ungauged catchments obtained by geographical regionalization." J. Hydrol., 170(1–4), 277–291.

van Werkhoven, K., T. Wagener, P. Reed, and Y. Tang (2008), Characterization of watershed model behavior across a hydroclimatic gradient, Water Resour. Res., 44, W01429, doi:10.1029/2007WR006271.

van Werkhoven, K., T. Wagener, P. Reed, and Y. Tang (2009), Sensitivity-guided reduction of parametric dimensionality for multi-objective calibration of watershed models, Adv. Water. Resou., 32, 1154–1169.

Velazquez, J.A., Anctil, F., Ramos, M. H., and Perrin, C., 2011. Can a multi-model approach improve hydrological ensemble forecasting? A study on 29 French catchments using 16 hydrological model structures. Adv. Geosci., 29, 33–42, doi:10.5194/adgeo-29-33-2011.

Viney, N. R., Vaze, J., Chiew, F. H. S., Perraud, J.-M., Post, D. A., and Teng, J., 2009. Comparison of multi-model and multi-donor ensembles for regionalisation of runoff generation using five lumped rainfall-runoff models, 18th World IMACS Congress and MODSIM09 International Congress on Modelling and Simulation, Cairns, Australia.

Wagener, T., and H. S. Wheater (2006), Parameter estimation and regionalisation for continuous rainfall-runoff models including uncertainty, J. Hydrol., 320, 132–154.

Zhang, Y., and F. H. S. Chiew (2009), Relative merits of different methods for runoff predictions in ungauged catchments, Water Resour. Res., 45, W07412, doi:10.1029/2008WR007504.

Zhang, C., J. G. Chu, G.T. Fu (2013), Sobol's sensitivity analysis for a distributed hydrological model of Yichun River Basin, China, J. Hydrol., 480, 58-68.

CHAPITRE 10

DISCUSSION GÉNÉRALE

10.1 Analyse de l’équifinalité en régionalisation

Un des aspects principaux des travaux de cette thèse est l’incertitude paramétrique en régionalisation. Le problème d’équifinalité, par lequel plusieurs jeux de paramètres différents produisent des hydrogrammes équivalents, est central à l’incertitude paramétrique. Plus un modèle contient de paramètres à ajuster, plus il risque d’y avoir d’équifinalité en raison du nombre de degrés de liberté supplémentaires. C’est pourquoi la communauté scientifique privilégie le concept de parcimonie afin de conserver le strict minimum de paramètres tout en maintenant un niveau acceptable de performance du modèle. De nombreuses études sont consacrées à ce sujet et des méthodes sophistiquées sont employées pour réduire le nombre de paramètres adéquatement, tel que vu au chapitre 9.

Cependant, dans le contexte de régionalisation, aucun essai réalisé dans cette thèse n’a montré que l’équifinalité était une variable importante. Bien au contraire, les travaux présentés ici ont clairement démontré que la réduction du nombre de paramètres ne permet pas d’améliorer la performance en régionalisation. De plus, les essais en présence d’équifinalité (Chapitres 6 et 9) ont montré des résultats clairs à cet égard. Les jeux de paramètres différents mais produisant des hydrogrammes équivalents en calage retournent des hydrogrammes similaires en régionalisation. Le choix de la méthode de régionalisation joue un rôle beaucoup plus important que celui de l’équifinalité. Les travaux du chapitre 3 (comparaison des méthodes de calage) n’auront donc pas contribué à améliorer la performance des méthodes de régionalisation. Cependant, à la lumière des résultats obtenus, il aurait pu être possible de trouver un algorithme qui permettrait d’élargir l’incertitude paramétrique. Ceci aurait pu contribuer à ajouter de la diversité lors de la régionalisation et ainsi permettre d’atteindre un plus haut niveau de robustesse.

Toutefois, les essais sur la réduction paramétrique (chapitre 9) ont permis de constater que les deux modèles mis à l'essai (HSAMI et MOHYSE) ne peuvent pas se permettre de fixer des paramètres importants. En effet, lorsque la variance d'ordre total cumulative des paramètres fixés atteint plus de 2.5% (par les méthodes d'analyse de sensibilité globale), la performance en régionalisation chute de façon marquée. Ceci contraste avec la littérature qui suggère de fixer les paramètres qui expliquent moins de 20% de la variance, tel qu'illustré au chapitre 9.

10.2 Caractéristiques physiques des bassins versants et paramètres des modèles

Une autre grande source d'incertitude en régionalisation est la sélection et la mesure des caractéristiques physiques des bassins versants. Certaines d'entre elles sont simples à mesurer, telles que la superficie, la pente moyenne, la latitude et la longitude du centroïde et l'élévation moyenne. D'autres sont mesurables avec un certain degré d'erreur, telles que la précipitation annuelle moyenne et la couverture du sol. Ces descripteurs introduisent nécessairement un biais dans l'identification des bassins similaires dû à la technique de mesure. Finalement, plusieurs descripteurs sont difficilement mesurables et sont donc ignorés ou grossièrement estimés. Par exemple, les types et profondeurs des couches de sol, la conductivité hydraulique du sol et les propriétés racinaires de la végétation sont des caractéristiques qui, croit-on, devraient influencer les caractéristiques des écoulements mais qui ne sont à toutes fins pratiques jamais mesurées à grande échelle. Il est donc difficile de déterminer quels descripteurs sont les plus utiles pour catégoriser les bassins versants en l'absence des plus influents.

Les travaux présentés aux chapitres 6, 8 et 9 ont montré qu'il était possible, en l'absence de descripteurs physiques adéquats, de sélectionner les bassins versants les plus rapporchés à titre de donneurs. Ceci suppose que les caractéristiques physiques sont plus similaires pour des bassins adjacents. Par contre, les résultats du chapitre 9 ont montré que les bassins versants où la méthode de proximité spatiale était performante étaient généralement très similaires également. Ce résultat montre que l'hypothèse de la similitude doit être valide pour qu'elle fonctionne correctement. Dans ce sens, il est clair que de sélectionner des bassins versants plus similaires est la méthode à privilégier tant que les descripteurs permettent de déterminer quels bassins sont effectivement les plus similaires. La qualité des descripteurs utilisés est donc primordiale.

En comparant les résultats du chapitre 6 et du chapitre 9, il est possible de constater que l'utilisation de descripteurs différents affecte les résultats de manière importante. Au chapitre 6, la régionalisation avait été effectuée avec des descripteurs sélectionnés parmi une liste afin de maximiser la performance de la régionalisation. Seuls quatre descripteurs avaient été nécessaires. Par ailleurs, dans le chapitre 9, tous les descripteurs avaient été utilisés. La différence est importante, de l'ordre de 0.02 sur l'échelle de Nash-Sutcliffe. Les 17 descripteurs utilisés dans le monde virtuel du chapitre 8 auraient pu être optimisés également, mais le but était de mieux comprendre les interactions entre les méthodes de régionalisation et les descripteurs alors l'ensemble a été conservé. Il est à noter que parmi les quatre descripteurs optimaux sélectionnés au chapitre 6 figurent la latitude et la longitude. La distance géographique semble donc être un indicateur adéquat dans le monde réel où la qualité des descripteurs hydrologiquement importants est faible.

10.3 Comparaisons entre le monde réel et le monde virtuel

Les travaux de régionalisation ont été effectués dans le monde réel (chapitre 6) et ses incertitudes ainsi que dans le monde virtuel (chapitre 9) libéré de la plupart de ces contraintes. De prime abord, le monde virtuel et le monde réel permettent aux méthodes de régionalisation de se comporter de manière très similaire. Ceci peut être considéré comme une démonstration du fait que le monde virtuel permette de faire de l'hydrologie numérique expérimentale.

Les avantages du monde virtuel sont nombreux par rapport au monde réel. En plus de n'avoir aucune donnée manquante et d'offrir un dense réseau de pseudo-observations, la qualité et la représentativité des données est la meilleure qui soit puisque connues sur l'ensemble du domaine de simulation. La qualité des descripteurs physiques a également profité aux méthodes de régionalisation. Ces travaux ont montré que les modèles hydrologiques et leurs paramètres sont aptes à simuler les processus physiques puisque les meilleurs bassins versants donneurs sont généralement ceux qui sont les plus similaires. Toutefois, l'approche de régression linéaire multiple ne permet pas d'estimer adéquatement la valeur des paramètres du modèle hydrologique à partir des descripteurs connus. L'hypothèse la plus probable est que les liens et les interactions entre les paramètres et les descripteurs physiques sont non-linéaires, ce qui expliquerait pourquoi les modèles, hautement non-linéaires, ont une meilleure habileté à utiliser l'information disponible. À partir de ces travaux, il resterait à déterminer quels paramètres physiques sont les plus importants dans le monde réel et de voir à quel point les modèles hydrologiques peuvent assimiler cette information. De plus, il serait intéressant de pouvoir prédire la valeur d'un paramètre uniquement à partir de caractéristiques physiques. Par contre, tel qu'il a été démontré, les liens complexes rendent cette tâche ardue.

10.4 Analyse des appoches multi-modèle

Les travaux en modélisation multi-modèle aux chapitres 4 et 5 ont permis de démontrer qu'il s'agit d'un outil puissant dans l'arsenal de l'hydrologue. En effet, le multi-modèle permet d'éliminer (ou réduire significativement) les erreurs structurelles des modèles afin de maximiser les forces de ceux-ci. De plus, la modélisation avec jeux de données météo multiples (« multi-input »), telle que proposée au chapitre 5, permet d'obtenir des résultats tout aussi convaincants que le multi-modèle mais à l'aide d'un seul modèle hydrologique. Les erreurs sur les jeux de données sont donc corrigées. La combinaison des deux méthodes (multi-modèle et multi-input) a permis des gains substantiels. Cette approche novatrice devrait être utilisée dès maintenant par tous les gestionnaires de systèmes hydriques puisque l'ajout d'un modèle très simple (MOHYSE par exemple) permet de faire des gains significatifs à très bas coût. Il s'agit également d'une source potentielle de réduction d'incertitude pour les études d'impacts en changements climatiques. En effet, les modèles climatiques ont des biais systémiques qu'il serait envisageable de réduire lors d'impacts de changements climatiques sur les systèmes hydriques.

10.4.1 Approches multi-modèle en régionalisation

Malgré les avantages de l'approche multi-modèle, certaines hypothèses ne sont pas respectées lorsqu'elle est appliquée en régionalisation. Premièrement, certaines méthodes intègrent la correction de biais lors de l'estimation des poids des membres. En régionalisation, le biais ne peut être considéré comme constant puisque la simulation est effectuée sur un site non-jaugé. Deuxièmement, puisque les pondérations sont obtenues en période de calage sur un site jaugé, elles peuvent difficilement être conservées lorsqu'elles sont transférées d'un site à un autre alors que la série hydrométrique est nécessairement différente.

Les travaux ont démontré que les modèles hydrologiques utilisés ne permettaient pas d'améliorer la qualité des simulations lorsque combinés avec des approches de pondération multi-modèle. La faible robustesse de deux des modèles et l'hétérogénéité de l'ensemble ont été identifiés comme des facteurs ayant contribué aux difficultés observées. Cependant, il serait envisageable d'analyser les méthodes de régionalisation avec le concept de multi-input. Ceci aurait pour effet de conserver la robustesse et l'homogénéité des modèles de l'ensemble puisqu'il s'agirait d'un seul modèle avec diverses sources de données météorologiques.

CONCLUSION ET CONTRIBUTIONS

L'objectif de ce projet de recherche était d'analyser et d'améliorer les méthodes de régionalisation paramétrique en prévision hydrologique aux sites non-jaugés. À travers des articles présentés aux chapitres 3 à 9, il a été possible de mieux comprendre l'effet de l'équifinalité et de l'identifiabilité des paramètres en régionalisation. L'étendue des travaux entrepris dans ce projet de recherche aura permis de proposer des contributions originales dans plusieurs domaines d'application, tels que la régionalisation paramétrique, la modélisation hydrologique, l'étude de l'incertitude paramétrique et le calage des modèles hydrologiques.

Pour le calage des modèles hydrologiques, il a été démontré que la forme de la surface de réponse avantageait certains algorithmes dans des cas particuliers. Suite aux travaux en ce sens présentés au chapitre 3, Rio Tinto Alcan a modifié ses pratiques de calage de modèles hydrologiques en remplaçant l'algorithme « SCE-UA » par un autre plus performant pour le calage du modèle CEQUEAU. Ce travail a permis d'ouvrir la porte à un autre projet, en cours de réalisation, qui analysera davantage la forme de la surface de réponse afin de prédire à priori quel algorithme est le plus adéquat à partir d'indicateurs de concavité, modalité et de niveau de bruit.

Dans l'aspect multi-modèle du projet, il a été montré que certaines méthodes de pondération permettaient d'améliorer avec constance la qualité des simulations par rapport aux membres individuels de l'ensemble. L'application à grande échelle réalisée dans ce travail aura permis de tirer des conclusions plus solides que ce qui avait été montré au préalable dans la littérature. De plus, une nouvelle approche a été proposée (l'approche « SCA », dans l'article 2 du chapitre 4) et elle a été en mesure de se comparer favorablement aux meilleures méthodes classiques utilisées dans l'étude. Cette contribution a été faite par un co-auteur de l'étude. Enfin, il a été démontré dans l'article 3 (chapitre 5) que l'utilisation de sources variées de données météorologiques sur grille permettait d'améliorer significativement la qualité des simulations une fois moyennées intelligemment. Cet aspect du concept multi-modèle n'avait jamais été exploré auparavant.

En prévision aux sites non-jaugés, plusieurs contributions ont été apportées. Seul un résumé est présenté ici en raison de la complexité du sujet et de la subtilité des implications. D'abord, une nouvelle métrique a été établie (Success Rate, chapitre 6) permettant de mesurer l'impact de l'équifinalité paramétrique en régionalisation. L'équifinalité a ainsi été évaluée dans ces circonstances et il a été montré qu'elle est négligeable comparativement à plusieurs autres facteurs. Ensuite, une nouvelle approche hybride de régionalisation a été proposée et a performé mieux que les méthodes traditionnelles (chapitre 6). Les méthodes de pondération multi-modèle ont par la suite été appliquées en régionalisation, avec des résultats peu encourageants (chapitre 7). Cependant, cette partie du projet a permis de mieux comprendre l'importance de la robustesse des modèles lors de la prévision aux sites non-jaugés.

Des contributions méthodologiques ont aussi été apportées. L'utilisation du monde virtuel issu d'un modèle régional de climat a permis de montrer que l'hydrologie des mondes réel et virtuel est similaire sur certains points, pavant la voie vers de nouvelles expériences. Ces travaux ont notamment montré que les paramètres du modèle hydrologique utilisé sont liés aux processus hydrologiques physiques d'une manière limitée. Il a également été démontré qu'il est possible d'estimer à l'avance si un bassin est un bon candidat à la régionalisation, mais que cette prévision est probabiliste (chapitre 8).

La dernière contribution importante de cette thèse concerne la réduction paramétrique des modèles hydrologiques. Il a été montré qu'il est possible de réduire le nombre de paramètres du modèle tout en préservant son niveau de performance en validation. Par contre, en régionalisation, la réduction des corrélations paramétriques et de l'équifinalité ne se traduit pas par des prévisions plus robustes aux sites non-jaugés. Il n'y a donc aucun gain à fixer des paramètres pour ce champ d'application (chapitre 9).

RECOMMANDATIONS

Les travaux de cette thèse auront permis de définir de nouvelles approches de modélisation et de régionalisation hydrologique. Ils ont également démontré des contraintes quant à l'utilisation des approches classiques de régionalisation paramétrique. Malgré les avancées proposées, il reste beaucoup d'améliorations possibles. Nous proposons ici quelques avenues de recherche pour des travaux futurs.

- la première recommandation serait de reproduire les travaux sur une plus grande échelle. Certaines incertitudes sont liées au faible échantillon de modèles hydrologiques (entre 1 et 3), et du nombre de bassins versants. Bien que plus de 250 bassins aient été utilisés pour les aspects de régionalisation, ceux-ci sont relativement homogènes en termes de climat et de physiographie. Intégrer des bassins versants du Canada, du Nord des États-Unis, et possiblement même d'Europe permettrait de construire une base de données plus hétérogène où les méthodes basées sur les paramètres physiques pourraient peut-être être mises à meilleure contribution. Notons ici que 94 bassins ont été utilisés en Ontario (Samuel et al. 2011), 267 au Québec (Arsenault et Brissette 2014), 913 dans une étude en France (Oudin et al. 2008), 320 en Autriche (Parajka et al. 2005) et plus de 30 au Royame Uni (Yadav et al. 2007) pour un total de plus de 1600 bassins versants. La richesse d'information pourrait permettre de mieux comprendre les limitations des méthodes tout en offrant des choix de jeux de paramètres plus judicieux en régionalisation. Dans la même lignée, dans l'éventualité où d'autres sources d'information quant aux descripteurs physiques des bassins seraient disponibles, il serait intéressant de les intégrer aux simulations.

- l'ajout de modèles hydrologiques serait un atout majeur pour les travaux portant sur la modélisation multi-modèle. Plus particulièrement, des modèles plus physiques et distribués pourraient être ajoutés en raison de leur structure différente de celle des modèles conceptuels et globaux utilisés dans cette thèse. Ceci permettrait de vérifier l’impact de la performance du modèle individuel sur la performance en simulation multi-modèle et multi-input.

- l’équifinalité paramétrique a été montrée comme étant un élément peu important en régionalisation. Il a également été montré que les donneurs multiples réduisaient l’impact de jeux de paramètres équifinaux. Le même constat a été fait lors du fixage de paramètres peu influents alors que la performance en régionalisation diminuait avec le nombre de paramètres fixés. Dans cette optique, il serait intéressant de voir si l’élargissement de bornes des paramètres ou l’ajout de paramètres permettrait d’obtenir de meilleures performances globales.

- le critère de Nash-Sutcliffe a été utilisé dans chacun des articles en raison de son fort potentiel de comparaison avec les autres résultats dans la littérature. Cependant, il est connu que ce critère pondère plus fortement les crues. Il serait intéressant de visualiser comment le choix de ce critère influence les résultats. Par exemple, un critère basé sur le volume de crue ou sur les étiages aurait certainement des caractéristiques différentes. Dans le contexte de simulation continue tel que dans cette thèse, d’autres critères tels que le biais, le critère de Kling-Gupta (Kling et Gupta, 2009) et le R^2 pourraient être testés.

- un des articles a proposé une méthode de pondération multi-input, où le modèle hydrologique lancé avec des données météorologiques de sources différentes était considéré comme plusieurs modèles indépendants dans l’ensemble multi-modèle. Un autre article traitait des méthodes de régionalisation basé sur les approches multi-modèle classiques, avec trois modèles différents. La conclusion de ce dernier article référait à la trop grande dissimilitude entre les modèles hydrologiques. Une prochaine étape serait d’utiliser le concept de multi-input en régionalisation avec un seul modèle hydrologique et plusieurs séries météorologiques. À première vue, la robustesse du modèle sera suffisante pour tirer profit de la pondération des modèles considérés comme indépendants. Ceci devrait se matérialiser peu importe le modèle choisi, mais l’hypothèse reste à confirmer.

- basé sur les travaux dans le monde virtuel issu du Modèle Régional de Climat (MRC), il serait possible de mettre à l'épreuve des méthodes de régionalisation non-stationnaires. Par exemple, l'effet des changements climatiques sur la météorologie ou les changements au niveau des caractéristiques physiques pourraient être simulés dans le temps, faisant en sorte de pouvoir évaluer des méthodes de régionalisation en non-stationnarité (Peel et Blöschl, 2011).

ANNEXE I

ARTICLE EN COLLABORATION 1 : POTENTIAL OF GRIDDED DATA AS INPUTS TO HYDROLOGICAL MODELING

Gilles R.C. Essou[1], Richard Arsenault[1] et François Brissette[1]

[1] Département de Génie de la Construction, École de technologie supérieure,

1100 Notre-Dame Ouest, Montréal, Québec, Canada H3C 1K3.

Article soumis à la revue « Journal of Hydrometeorology » en juillet 2014.

Abstract

Climate data measured by weather stations are crucially important and regularly used in hydrologic modelling. However, they are not always available due to the low spatial density and short record history of most station networks. On the other hand, gridded and interpolated datasets offer excellent network densities, but are seldom used in hydrologic applications mainly due to a combination of potential biases and smoothing of extremes. This study aims to evaluate the potential of various gridded datasets for hydrological modeling. In particular, it will focus on the quantification of biases and whether or not such biases can be compensated and filtered by the hydrological models. Three daily interpolated and gridded datasets covering the United-States were used in this study: Santa Clara, Daymet and CPC. They were compared to the MOdel Parameter Estimation eXperiment (MOPEX) dataset, used as a reference for comparative purposes. Hydrological simulations were performed on 424 basins in the United-States. A comparison between the various datasets shows that there are biases between the gridded and reference climate data. These biases result in a decreased hydrological modeling performance for all tested gridded datasets, when using the hydrological model calibrated on station data. However, when the hydrological model is calibrated using all specific gridded datasets, the calibration and validation Nash-Sutcliffe Efficiency values are not statistically different from one another. This leads to the conclusion that the use of gridded data allows equal performance levels to that of the observation-driven hydrology model, as long as proper model calibration is first performed.

Keywords: Gridded data, hydrological modeling, calibration, MOPEX, performance comparison.

I.1 Introduction

Weather station climate records are the main source of meteorological forcing in hydrological modelling. Their quality and availability is crucial to properly calibrate the models. However, because of their often poor spatial density and short historic records, measured datasets can be insufficient to adequately simulate the observed hydrograph. Also, climate data records regularly contain missing data and can contain biases due to the instrumentation (Arsenault and Brissette 2014a; Goodison et al. 1981, 1998). This necessarily affects the hydrological model's ability to produce quality hydrographs that are representative of the observed streamflow values, as was shown by Wilson et al. (1979), Krajewski and al. (1991), Obled and al. (1994) and Lopes (1996). It is consequently important to find alternatives to observed climate data in low weather station network density areas.

One of the alternatives to weather station data is remote sensing (Tang and al. 2009; Bastola and François 2012; Murray and al. 2013). However, this approach produces climate data for time periods shorter than required for hydrological modelling. Furthermore, with radar sensing, rainfall depth estimates are error-prone and biased and must be corrected in part by raingauge measurements at ground level (Steiner and al. 1999; Fulton and al. 1998; Seo 1998, Bastola and François 2012). Radar sensing can therefore only be robust in areas where there exists a dense raingauge network (Turk and al. 2008). Another challenge regarding radar remote sensing is that it is very sensitive to topographic and vegetation obstacles which block the electromagnetic waves, thus significantly degrading their precision and range (Warner and al. 2000; Westrick and al. 1999). Moreover, the precipitation estimation contains other uncertainties in cold climates where snowfall occurs. The unknown snowflake shape, density and vertical velocity parameters during the snowfall event reduce the estimation precision even further (Rasmussen and al. 2003). Remote sensing does show promise but evidently many problems remain.

Another solution to the low density climate data records is to use gridded datasets based on statistical interpolation between weather stations (Ruelland and al. 2008; Skaugen and Andersen 2010; Thornton and al. 1997; Daly and al. 1997; Taylor and al. 1997). As such, over the past decades several gridded datasets have been proposed, each using a unique interpolating algorithm and various spatial and temporal resolutions (Hulme 1992; Huffman and al. 2001; Adam and Lettenmaier 2003; Mitchell and Jones 2005; Yatagai and al. 2009; Hutchinson and al 2009). On the other hand, the introduction of an interpolation algorithm necessarily also introduces biases in the gridded datasets (Tozer et al., 2012). For this reason, gridded datasets have been seldom used in hydrological modelling applications (Muñoz and al. 2011; Mizukami and al. 2012). It is unknown if gridded datasets can be used in rainfall-runoff modelling nor are the impacts of doing so on the model performance.

This study aims to analyze the ability of a hydrological model to adequately simulate observed hydrographs using three important US gridded datasets (Santa Clara, Daymet and CPC) as meteorological input forcing. As a means of comparison, the reference precipitation and temperature data were provided by the MOPEX database, which is based on an average of observed climate data. The study revolves around two main points: (1) comparison of gridded data to MOPEX reference data to determine bias amplitude and (2) comparison of hydrological modelling performance using reference and gridded datasets as meteorological forcing.

I.2 Study area and datasets

I.2.1Study area

The study area is a group of 424 catchments in the continental United-States, within boundaries reaching from 67°W to 124.8°W longitude and 25°N to 49.4°N, as shown in figure-A I-1.

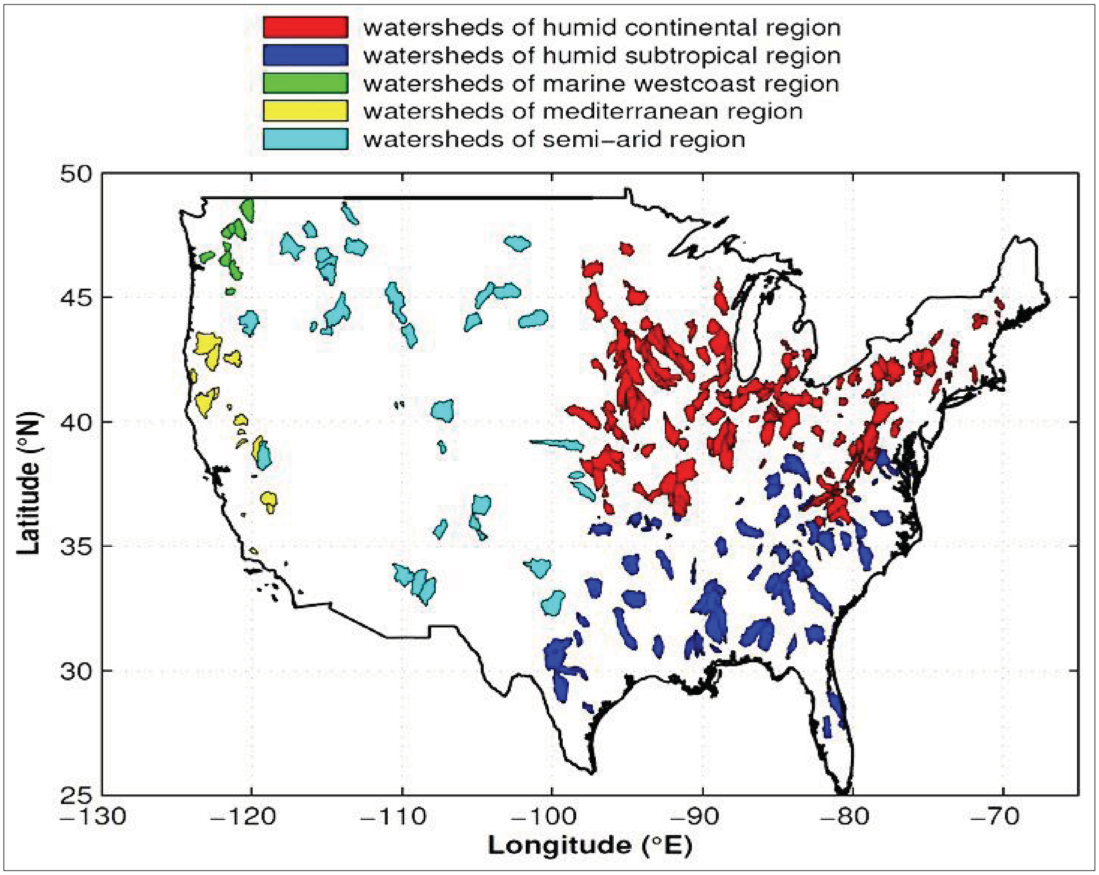
watersheds of humid continental region
watersheds of humid subtropical region
watersheds of marine westcoast region
watersheds of mediterranean region
watersheds of semi–arid region
50
45
40
35
30
25
Latitude (°N)
−130
−120
−110
−100
−90
−80
−70
Longitude (°E)

Figure-A I-1 Location and climate classification of the 424 catchments used in this study

The catchments are dispersed in 5 climatic zones according to the Köppen-Geiger classification system (Kottek and al. 2006). There are 236 basins classified as humid continental, 107 as humid subtropical, 13 in the marine west-coast region, 24 as Mediterranean and 44 as semi-arid. The catchments range between 66 km^2 and 10325 km^2 in size.

I.2.2 Datasets

All the comparisons and simulations were performed with daily climate data as well as daily discharge time series. Three gridded datasets were used and compared to the reference climate data.

I.2.2.1 Reference data

The reference climate data come from the MOPEX (***M**odel **P**arameter **E**stimation e**X**periment*). The MOPEX database contains precipitation, temperature (minimum and maximum) and streamflows on a daily time step. The database covers the years 1949-2003. Its conception stems from the National Climatic Data Center (NCDC) weather station observations (Duan and al. 2006). In fact, the MOPEX climate data are averaged observation values on the different catchments. An inverse distance weighting method was implemented to estimate the final MOPEX climate data. A detailed description of this data source is available in Schaake et al. (2006). It is important to note that each catchment in the database requires a minimal density of weather stations, which is determined by the size of the catchment as explained in Schaake and al. (2000). Furthermore, only time series of length greater than 10 years were admitted in the database. The reference streamflow data is also taken from this database. The MOPEX dataset is available online :
ftp://hydrology.nws.noaa.gov/pub/gcip/mopex/US_Data

I.2.2.2 Santa Clara gridded data

The University of Santa Clara gridded dataset were initially developed in Washington, but they were formatted into their current form at the University of Santa Clara. The daily precipitation and temperatures (minimum and maximum) are available for the years 1949-2003. They were interpolated on a 0.125° x 0.125° grid using the weather measurement data provided by the *National Oceanic and Atmospheric Administration* (NOAA) cooperative network, averaging 1 station per 700 km^2 (Maurer et al. 2002). The interpolation algorithm is based on the *Synergraphic Mapping System* (SYMAP) by Shepard (1984) and implemented as proposed by Widmann and Bretherton (2000). Particularly, the precipitations were downscaled to correspond to the long-term means of the precipitations from the *Parameter-elevation Regressions on Independent Slopes Model* (PRISM) (Daly and al. 1994, 1997). More precisely, it relies on 12 monthly means for the 1961-1990 period, which are statistically adjusted to capture the local variations on complex terrain. The Santa Clara dataset is available online: http://hydro.engr.scu.edu/files/gridded_obs/daily/ncfiles_2010

I.2.2.3 Climate Prediction Center gridded data

The *Climate Prediction Center (CPC)* data contains precipitation data only for the years 1949-2003 with a spatial resolution of 0.25° x 0.25°. The interpolation uses three main sources of observation data (Higgins and al. 2000). The first is the CPC cooperative network stations for the 1996-1999 period (15622 stations). The second is daily observations from the NCDC for the years 1948-1998 (approximately 16139 stations). The third is from the Hourly Precipitation Dataset (HPD) (approximately 5933 stations) (Higgins and al. 1996). The interpolation uses the Cressman method (Cressman 1959). The CPC dataset is available online: http://www.esrl.noaa.gov/psd/data/gridded/data.unified.daily.conus.html

I.2.2.4 Daymet gridded data

The Daymet dataset includes maximum and minimum temperatures and precipitation on a daily scale for the period 1980-2003. They were produced using the Daymet suite, an ensemble of algorithms and software designed to interpolate (and extrapolate) values at grid points with a 1km x 1km resolution (Thornton et al. 2012). Daymet uses observation network data to perform the interpolation with a Gaussian weighting scheme. A detailed description of Daymet is available in Thornton et al. (1997). The Daymet dataset is available online: http://daymet.ornl.gov/.

A summary of the dataset characteristics is presented in Table-A I-1.

Table-A I-1 Characteristics of datasets used in this study

Dataset	Spatial resolution	Source	Reference
MOPEX (observations)	---	ftp://hydrology.nws.noaa.gov/pub/gcip/mopex/US_Data	Duan et al., 2006
Santa Clara	0.125° x 0.125°	http://hydro.engr.scu.edu/files/gridded_obs/daily/ncfiles_2010	Maurer et al., 2002
CPC	0.25° x 0.25°	http://www.esrl.noaa.gov/psd/data/gridded/data.unified.daily.conus.html	Higgins et al. 2000
Daymet	1x1 km	http://daymet.ornl.gov/	Thornton et al., 2012

I.3 Methodology

The gridded datasets were first compared to the MOPEX reference time series to determine their relative differences. Second, their performance in hydrological modelling were analyzed and compared.

I.3.1 Dataset comparison

The interpolated data grid points inside each of the catchments were averaged using the inverse distance weighting method calculated with respect to the catchment centroid (Dirks and al. 1998). This method was shown to be amongst the best interpolation methods for such uses (Ruelland and al. 2008, Baillargeon and al. 2004). The comparison was performed on the daily, seasonal and extreme data. Moreover, the daily data was compared by climatic zone. Once again, the reference values were the ones taken from the MOPEX database. The first comparison criterion used in this study is the well-known Root Mean Squared Error (*RMSE),* which is defined as:

$$RMSE = \sqrt{\frac{1}{N}\sum_{i=1}^{N}\left(X_i - X_{MOPEX,i}\right)^2} \tag{A I-1}$$

Where X_i represents the gridded data value for day i, $X_{MOPEX,i}$ represents the MOPEX data value for day i and N is the length of the time series. The RMSE gives an indication on the difference amplitude between two series. An RMSE value of 0 is a perfect fit, and larger values indicate larger errors.

The second comparison criterion is the bias (***B***), defined as:

$$B = \frac{1}{N}\sum_{i=1}^{N}\left(X_i - X_{MOPEX,i}\right) \tag{A I-2}$$

The bias allows estimating how much one series underestimates or overestimates a second series. A bias of 0 indicates a perfect fit. A positive bias indicates an overestimation of the observations, while the opposite is true for negative biases.

The last criteria for the comparative analyses were chosen from the STARDEX project (Anagnostopoulou and al 2003; Hundecha and Bárdossy 2005, Schmidli and Frei 2005) and are intended to gain insight in comparing extreme values. They are the 90th percentile of daily precipitation (mm/day), the 90th percentile of daily maximum temperature (°C) and 10th percentile of daily minimum temperature (°C).

I.3.2 Hydrological model

The hydrological model used in this study is the HSAMI model (Fortin 2000; Minville and al. 2008). It is a lumped conceptual rainfall-runoff model developed and used operationally by Hydro-Québec for over 30 years. It is used to predict streamflow values on over 100 catchments in the province of Québec on an hourly and daily time scale. The HSAMI model has also been used extensively in streamflow prediction applications, climate change impact studies and rainfall-runoff modelling research projects (Minville and al. 2008, 2009; Chen and al. 2011a, 2011b, 2012; Poulin and al. 2011; Arsenault and al. 2013). It simulates the main hydrological cycle processes such as vertical and horizontal water transfer, evapotranspiration, snowmelt and soil freezing. It has up to 23 parameters that must be calibrated: 10 for the various production function processes, 5 for the horizontal transfer through reservoir-type soil layers, 2 for evapotranspiration and 6 for snow-related processes. There are four interconnected reservoirs that contribute to the vertical water transfer balance: Snow on ground, surface runoff, saturated soil layer and unsaturated soil layer. The horizontal water transfer is based on two unit-hydrographs (one for surface runoff and one for underground runoff) and a linear reservoir. HSAMI requires spatially averaged minimum and maximum temperatures as well as rainfall and snowfall depths. The cloud cover fraction and snow on ground may also be used if they are available.

Because of the large number of catchments, an automatic optimization algorithm was chosen to perform the model calibrations. Arsenault and al. (2014b) showed that the CMAES (Covariance Matrix Adaptation Evolution Strategy) (Hansen and Ostermeier 1996, 2001) algorithm was the optimal choice for calibrating the HSAMI model on 10 catchments, 8 of which were from the MOPEX database. Thus the CMAES optimization algorithm was used to perform the many calibrations in this project.

The calibration metric was computed on the odd years and cross-validated on the even years, and vice-versa. This allowed taking into account any climatic trends (such as decadal or multi-decadal natural variability) or modifications in underlying data from the addition or removal of weather stations. However there is a drawback to this method: the model must be run for the entire period in order to select the odd years for calibration, thus doubling the computation requirements compared to traditional block-type calibration. Also, 10 calibrations were performed in the odd/even approach, as well as 10 other calibrations in the even/odd approach, for a total of 20 calibrations. Only the best parameter set was taken for each case. This reduces the likelihood of having the calibration algorithm not converge during the optimization process.

The Nash-Sutcliffe Efficiency (NSE) metric (Nash and Sutcliffe 1970) was used to compare hydrologic simulation performance levels between groups. It is computed as follows:

$$NSE = 1 - \frac{\sum_{i=1}^{T}\left(Q_{obs,i} - Q_{sim,i}\right)^2}{\sum_{i=1}^{T}\left(Q_{obs,i} - \overline{Q_{obs}}\right)^2} \tag{A I-3}$$

Where NSE is the Nash-Sutcliffe Efficiency metric, Qsim,i is the simulated discharge for day i, Qobs,i is the observed discharge for day i and Qobs is the average observed discharge. Other metrics could have been used, but the NSE is the most widely used metric and was the obvious choice for this study.

Two calibration strategies were tested for comparison purposes. The first consisted in calibrating the model with the “observed” data from the MOPEX database. Then, the model was run in validation mode with the three gridded datasets as well as the MOPEX observed data. The NSE values were then compared between groups. This will give a general overview of the HSAMI models’ ability to adapt to different inputs than it was calibrated with. It will also answer the question as to whether or not gridded datasets can be directly inserted as substitutes to traditional station data and to quantify potential performance gains or losses. The second strategy consists in doing a specific calibration of the HSAMI model using each of the gridded datasets prior to calculating the corresponding validation performance. This is likely a more reasonable approach since the dataset that was used to calibrate the model is often the same dataset that serves in validation and in prediction.

The NSE values were compared between the gridded dataset groups as well as with the MOPEX-driven NSE scores. The Wilcoxon non-parametric test was used to identify statistically significant differences between the groups and the MOPEX reference group (Rakotomalala 2008).

Furthermore, the precipitation and temperature datasets were then mixed and recombined to produce a total of 12 distinct datasets, and the calibration, validation and comparison aspects were also performed on the newly created datasets. Table-A I-2 shows all of the resulting datasets used in this study.

From Table-A I-2, it is clear that the common period to all groups is 1980-2003. For this reason the entire study will be performed with these years to avoid any biases that could be caused by using different periods between the datasets.

Table-A I-2 List of datasets used in this study and coverage periods

Components		Period
Temperatures	Precipitation	
MOPEX	MOPEX	1949 – 2003
Santa Clara	Santa Clara	1949 – 2003
MOPEX	Santa Clara	1949 – 2003
Santa Clara	MOPEX	1949 – 2003
MOPEX	CPC	1949 – 2003
Santa Clara	CPC	1949 – 2003
Daymet	Daymet	1980 - 2003
Daymet	MOPEX	1980 - 2003
Daymet	Santa Clara	1980 - 2003
Daymet	CPC	1980 - 2003
MOPEX	Daymet	1980 - 2003
Santa Clara	Daymet	1980 - 2003

I.4 Results

I.4.1 Temperature comparison

I.4.1.1 Mean daily temperature

The results of the RMSE and bias between the mean daily temperature values of the Daymet and MOPEX datasets as well as between the Santa Clara and MOPEX datasets are presented in figure-A I-2.

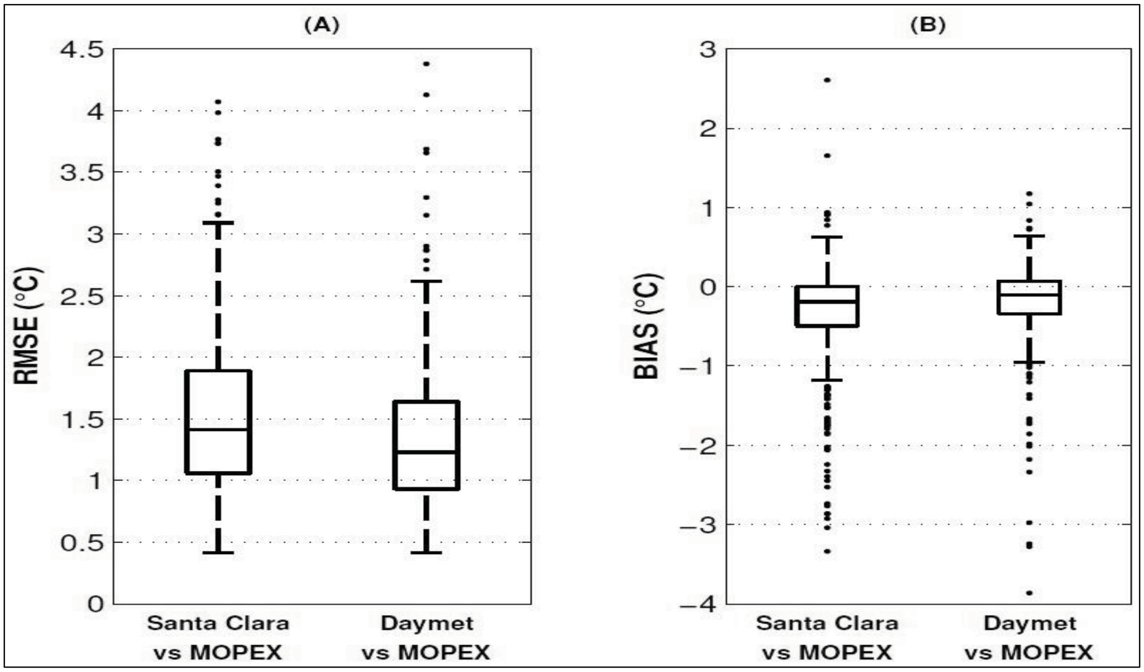

Figure-A I-2 RMSE (A) and bias (B) of the mean daily temperatures of the Santa Clara and Daymet datasets

The results show that the RMSE involving the Santa Clara and MOPEX temperatures range from 0.4°C to 4.1°C with a median RMSE of 1.4°C (Figure-A I-2A). The Daymet RMSE ranges from 0.4°C to 4.4°C with a median of 1.2°C. Globally, the Santa Clara mean daily temperatures deviate more than those of Daymet from the MOPEX reference as approximately 71% of the catchments reflect a higher RMSE for the Santa Clara dataset.

The Santa Clara mean daily temperatures show bias values ranging from -3.3°C to 2.6°C with a median of -0.2°C (Figure-A I-2B). As for Daymet, the biases range from -3.9°C to 1.2°C with a median of 0.1°C. However, for both datasets, there is a cold bias on the majority of catchments (75% and 65% of catchments respectively). Furthermore, on 64% of the catchments, the Santa Clara bias is lower than that of Daymet for the mean daily temperature. Therefore, both datasets are colder than the reference MOPEX dataset, but Santa Clara temperatures are colder on average.

I.4.1.2 Mean daily temperature by climatic zone

The results of the RMSE between the mean daily temperature values of the Daymet and MOPEX datasets as well as between the Santa Clara and MOPEX datasets for each of the climatic zones are presented in figures-A I-3A to I-3E.

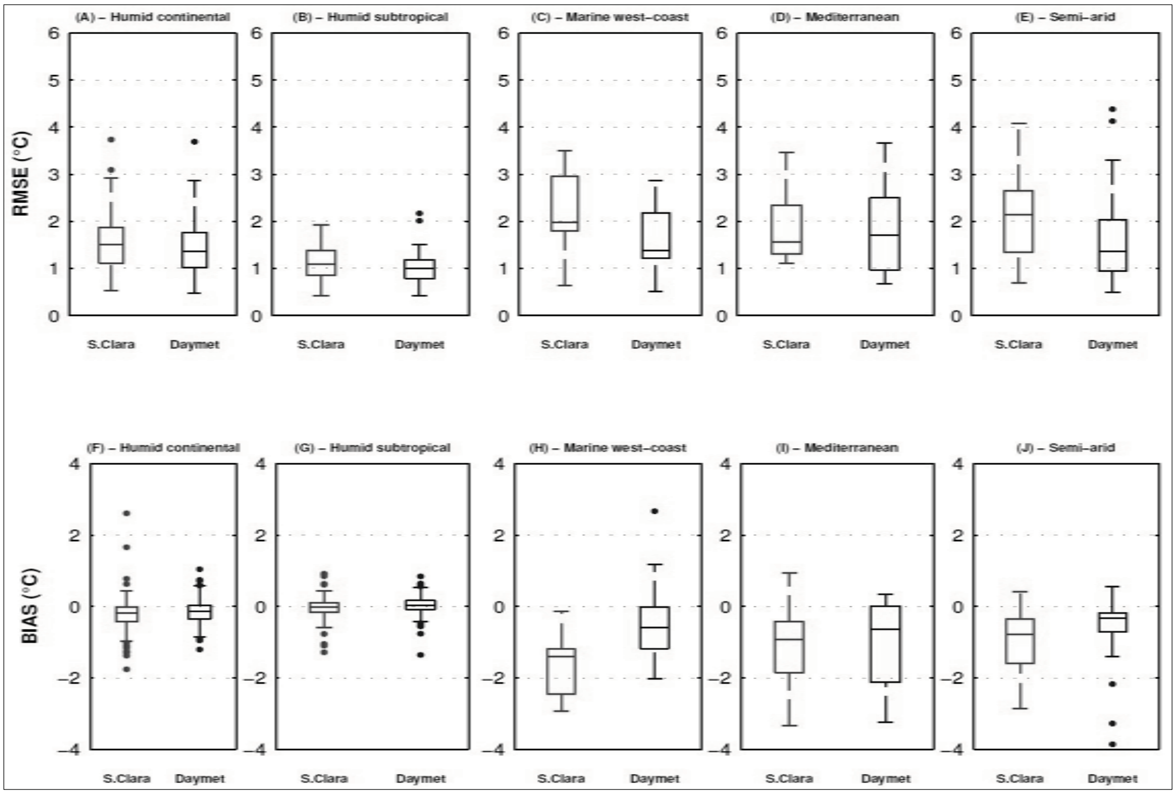

Figure-A I-3 Mean daily temperature RMSE and bias for the Santa Clara and Daymet datasets for the 5 climate zones

The results clearly demonstrate that the temperature RMSE values are higher for Santa Clara than for Daymet in all climatic zones except of the Mediterranean region. Moreover, for the Santa Clara dataset, the RMSE in semi-arid climate is relatively higher (median = 2.2°C) but the RMSE in humid subtropical climate are lower (median = 1.1°C). As for Daymet, the largest RMSE values were found in the Mediterranean region (median = 1.7°C), and the lowest, in the humid subtropical climate (median = 1.0°C).

The results for the bias for both datasets are presented in figures-A I-3F to I-3J. The results show that in all the climatic zones, the Santa Clara temperature biases are mainly cold (median bias <0°C). However, these biases are colder in the marine/west-coast climate region (median = -1.4°C) and relatively less so in the subtropical humid climate (median = -0.01°C). For the Daymet dataset, the results are mainly cold as well in all climate zones except for the humid subtropical climate where the median bias is slightly above zero (median = 0.02°C). The Mediterranean climate is relatively colder with a median bias of -0.6°C. In all the climate zones, the Santa Clara dataset is generally colder than the Daymet dataset.

I.4.1.3 Mean seasonal temperatures

Results are similar for seasonal temperatures and are not shown. The Santa Clara and Daymet mean seasonal temperature RMSE values are relatively low. However, for all seasons, the Santa Clara temperature RMSE values are larger than for Daymet. Also, for both datasets, the temperature RMSE values are generally higher in winter (median RMSE = 0.3°C) and lower in summer (median RMSE = 0.1°C). The temperature biases are cold for all seasons and for both datasets, although they are colder in winter and less so in summer. In all cases, the seasonal temperature biases are colder for the Santa Clara dataset than for Daymet.

I.4.1.4 Extreme temperatures : 90th percentile of maximum annual temperatures and 10th percentile of minimum annual temperatures

The extreme temperature biases are low for both the Santa Clara and the Daymet datasets (results not shown). The biases for the 90th percentile of the maximum annual temperatures range from -5.3°C to 2.2°C for both datasets. However, the median bias is small for both datasets. The Santa Clara extreme temperatures show a warm bias on 58% of the catchments (median = 0.1°C) and the Daymet extreme temperatures exhibit a cold bias on 64% of the catchments (median = 0.2°C). The same range also applies for the 10th percentile of minimum annual temperatures for both datasets.

I.4.2 Precipitation comparison

I.4.2.1 Daily precipitation

The results of the daily precipitation RMSE and bias for the Santa Clara, Daymet and CPC datasets are presented in Figure-A I-4.

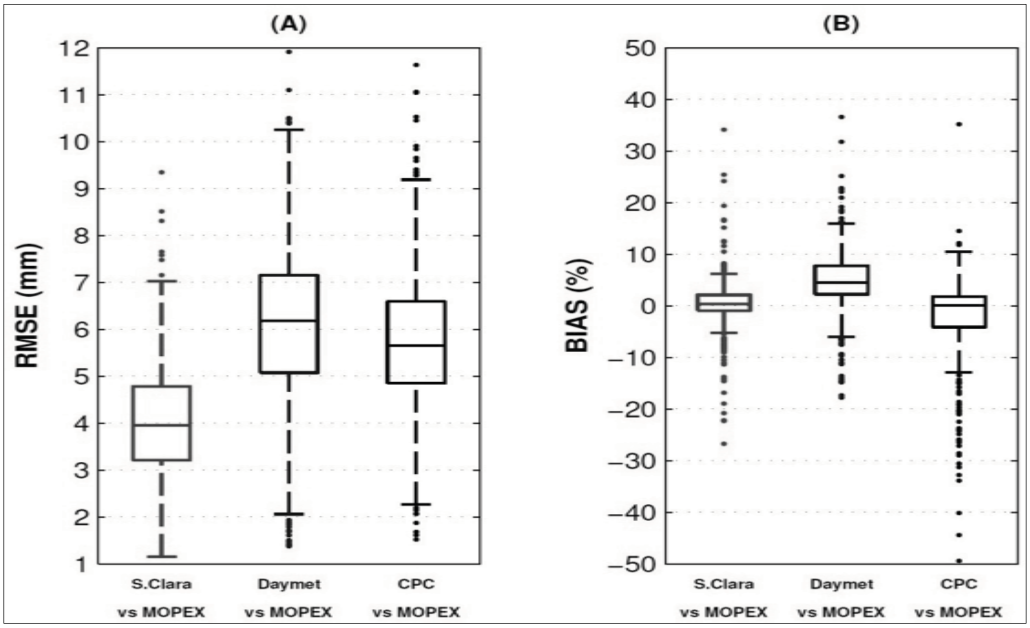

Figure-A I-4 RMSE (A) and bias (B) of the daily precipitation of the Santa Clara, Daymet and CPC datasets.

The results show that the precipitation RMSE range from 1.2mm to 9.3mm for Santa Clara, from 1.4mm to 11.9mm for Daymet and from 1.5mm to 11.6mm for CPC. In 97% of the catchments, the daily precipitation RMSE is lower than that of the CPC dataset. In turn, the CPC daily precipitation RMSE is lower than for Daymet in 75% of the catchments. Therefore the Santa Clara precipitation deviates the least from the MOPEX, and Daymet deviates the most.

The results show that the Santa Clara biases are humid for 57% of the catchments, whereas the Daymet and CPC biases are humid for 83% and 51% of catchments. These results indicate that the three gridded dataset precipitation series overestimate the MOPEX reference precipitation. However, this trend is stronger with Daymet. The bias comparison by catchment shows that Daymet precipitations are larger than Santa Claras' on 85% of catchments. In turn, the Santa Clara precipitation is larger than the CPC precipitation on 60% of the catchments. This means that the Daymet has the most precipitation and CPC has the least.

I.4.2.2 Daily precipitation by climatic zone

The results of the RMSE for the daily precipitation for Santa Clara, Daymet and CPC datasets for each of the climatic zones are presented in figures-A I-5A to I-5E.

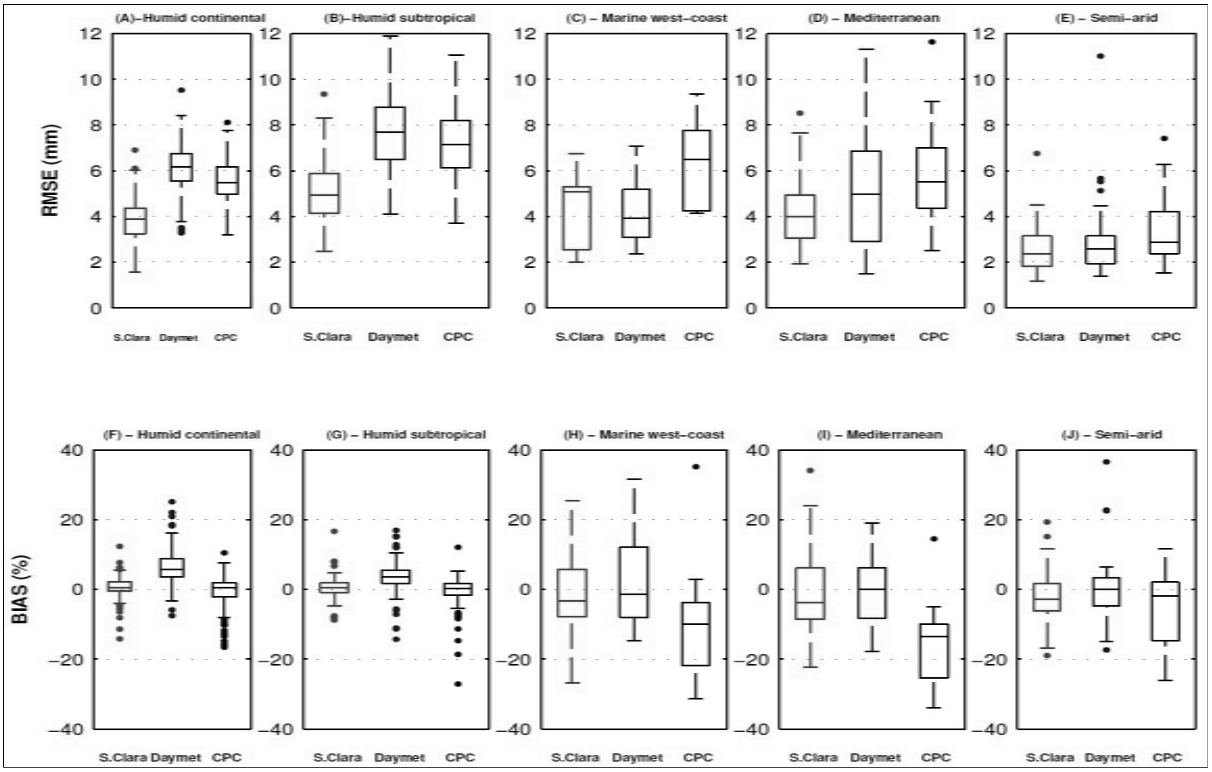

Figure-A I-5 Daily precipitation RMSE and bias for the Santa Clara, Daymet and CPC datasets for the 5 climate zones

The results show that the precipitation RMSE values are lower for Santa Clara than for Daymet and CPC in all climatic zones except for the Marine/West-Coast region, in which case Daymet has the lowest RMSE. For the three datasets, the largest RMSE values were found in the humid subtropical climate (Median RMSE= 4.9mm for Santa Clara, 7.2mm for Daymet and CPC). However, the lowest RMSE values were found in the semi-arid region (Median RMSE= 2.4mm for Santa Clara and Daymet, 2.9m for CPC).

The results for the biases are presented in figures-A I-5F to I-5J. They indicate that the Santa Clara precipitation biases are humid on over 57% of the catchments in the humid continental and humid subtropical climate zones. However, the biases are dry on over 58% of catchments in the other climatic zones. For Daymet, the biases are dry on approximately 54% of marine climate catchments, but they are humid in 84% of the catchments in the humid continental and humid subtropical zones. In the Mediterranean and semi-arid regions, no particular trend was detected. Finally, for CPC, the biases are humid on 56% of the humid continental and humid subtropical catchments, but are dry in at least 61% of the basins in the other climatic regions.

I.4.2.3 Total seasonal precipitation

Trends for seasonal precipitation are similar to annual ones (results not shown). The CPC RMSE values are generally lower than those of Daymet, but higher than those of Santa Clara. Furthermore, for the three gridded datasets, the largest RMSE values were found in summer (Median RMSE=21mm for Santa Clara, 29mm for Daymet and 25mm for CPC). The smallest RMSE values were found in winter (Median RMSE= 14mm for Santa Clara, 24mm for Daymet and 19mm for CPC).

For Santa Clara, the biases are mainly humid in all seasons except winter (median bias=-0.5%) The most humid biases were found in spring (median bias = 1.1%). However, the Daymet precipitations have wet biases for all the seasons, especially for spring (Median bias = 6.4%). Finally, for CPC, the biases are dry in all seasons except summer, where the bias is mainly humid (median bias = 1.5%). Therefore the Daymet seasonal precipitations are more abundant than for Santa Clara, which in turn is more humid than the CPC dataset.

I.4.2.4 Extreme precipitations: 90th percentiles of the maximum annual precipitation

The distributions of extreme precipitation biases for Santa Clara, Daymet and CPC are presented in Figure-A I-6.

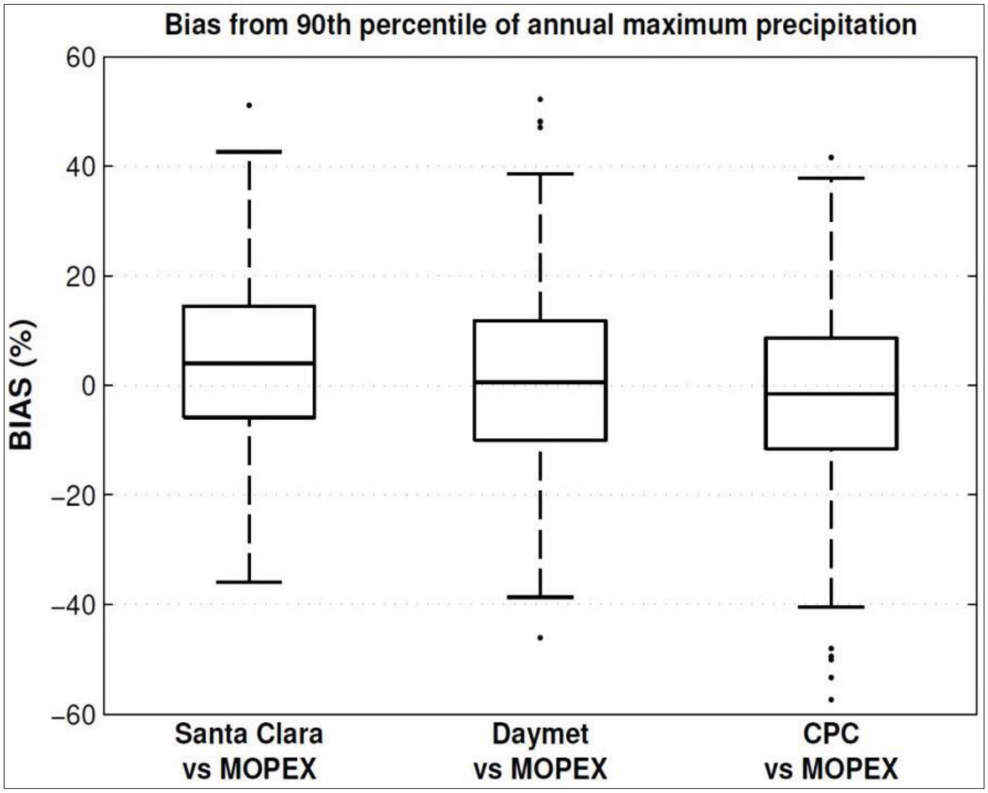

Figure-A I-6 Extreme precipitation biases for Santa Clara, Daymet and CPC

The results show that the extreme precipitation biases for Santa Clara spread from -36% to 51% with a median bias of 4.1%. These biases are humid on 61% of the catchments, which implies that the extreme precipitations are larger in the Santa Clara dataset than in the MOPEX reference dataset. For Daymet, the biases lie between -48% and 52%, with a median of 0.2%). On 51% of the basins, Daymets' extreme precipitations are larger than those of the MOPEX database. However, the CPC extreme precipitation biases range from -60% to 42% with a median of -1.5% and are dry on 55% of catchments. Therefore, the CPC extreme precipitations are mainly smaller than the MOPEX extremes.

I.4.3 Hydrological performance The performance of the HSAMI hydrological model is first assessed using the MOPEX database. Results are shown in Figure-A I-7 and indicate that the hydrology model perform very well, with a NSE median value of 0.783.

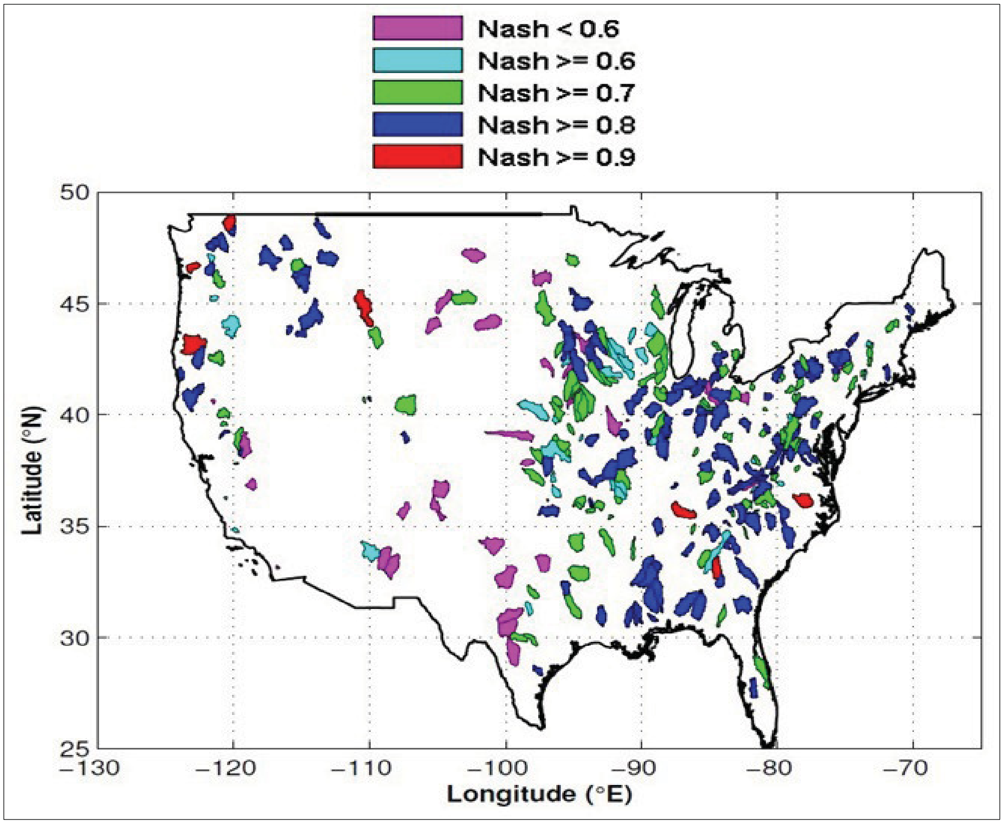

Figure-A I-7 Validation results (NSE) of the HSAMI hydrological model using the MOPEX database (Flow discharge, precipitation and temperature)

The model performs well over most of the United States with the exception of the semi-arid climate (see Figure-A I-1) where several catchments have a NSE value smaller than 0.6. This is not surprising considering that the hydrology model used in this study was developed for temperate climates and is not well adapted to the specific conditions of more arid landscapes. However, since the goal of this study is an inter-comparison of datasets, this relative lack of performance in semi-arid regions is of minimal concern.

The distribution of hydrological model performances using the various datasets is presented in Figure-A I-8.

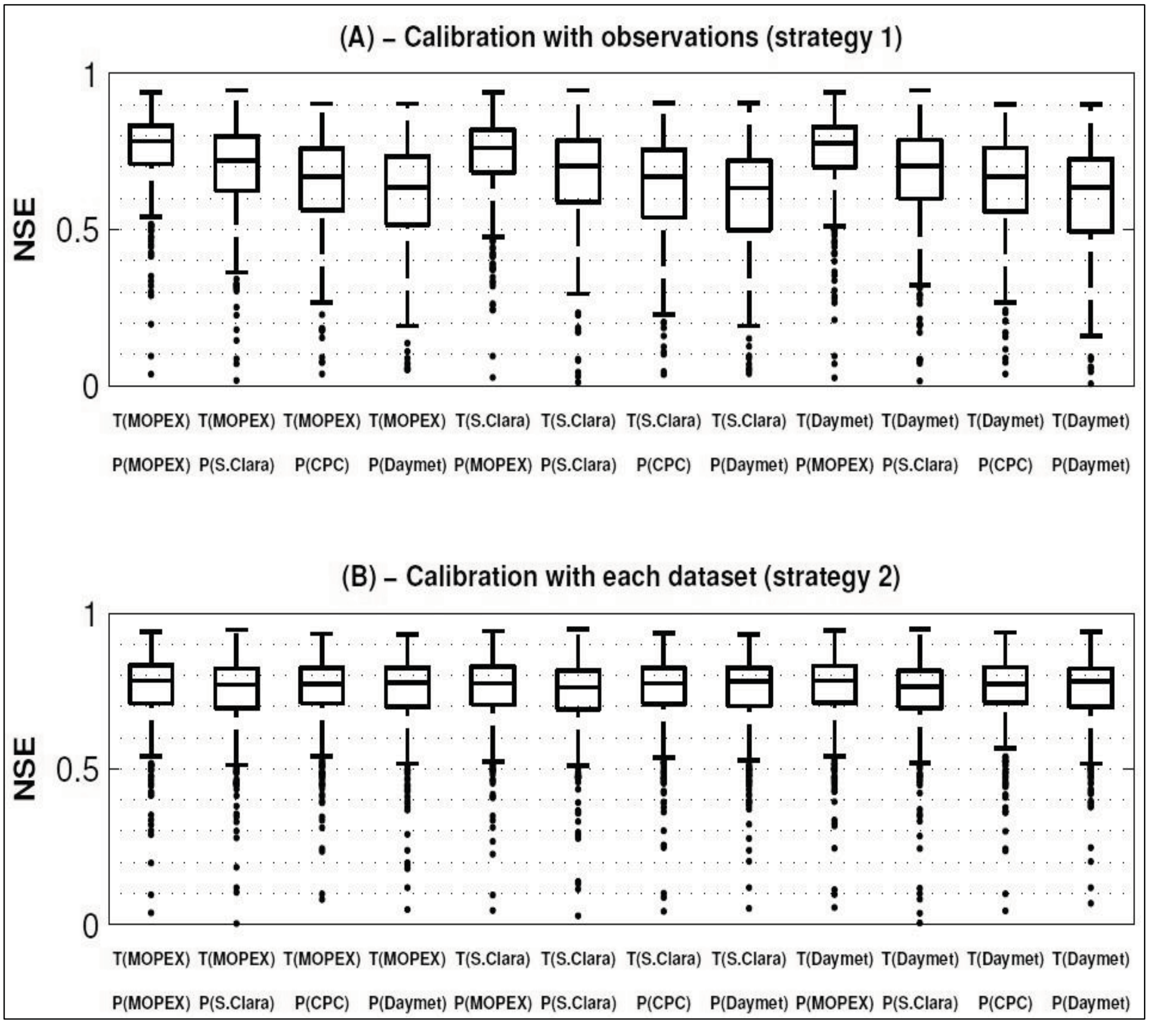

Figure-A I-8 Validation NSE distributions for the 12 climate datasets

With the first calibration strategy (calibration on MOPEX climate data and validation using the alternative datasets), shown in Figure-A I-8A, the NSE values in validation for the MOPEX datasets are all better than with any of the gridded climate data sources. The median NSE is 0.783 with the MOPEX data, which is the highest median score. The results demonstrate that when the MOPEX precipitation data is substituted by the gridded data precipitation while keeping the MOPEX temperature, there is a loss of performance that is dependent on the dataset. The Santa Clara precipitation was shown to be more similar to MOPEX, therefore it should be no surprise that the T-MOPEX/P-Santa Clara combination would fare better than the others (median NSE = 0.722). On the other hand, the Daymet precipitation lowers the overall skill to a median NSE of 0.634, which can be attributed to its relative poor similarity to MOPEX precipitation.

A similar observation was made when replacing the MOPEX temperatures by Santa Clara and Daymet temperatures. When replacing the temperatures (but maintaining the MOPEX precipitation), the loss in performance is much less steep. The Santa Clara hybrid drops to 0.761 while the Daymet hybrid drops to 0.776. This can be explained by the fact that the Daymet temperatures deviate less than the Santa Clara temperatures from the MOPEX database. Thus the HSAMI model calibrated with MOPEX performs better with the most similar gridded data inputs, but is more sensitive to precipitation than temperature.

With the second calibration strategy (independent calibration and validation for each of the datasets) shown in Figure-A I-8B, the NSE improved considerably. The results indicate that the lowest median value under this framework lies at 0.762, as compared to the 0.634 NSE obtained in the first calibration strategy. There was no performance loss when using Santa Clara temperatures and Daymet precipitation (same median NSE), but it is clear from Figure-A I-7B that the performance level is similar overall. A comparison was made catchment-by-catchment to determine the frequency with which each climate combination shows superior performance. The results are shown in Table-A I-3.

Table-A I-3 Frequency with which each climate combination shows superior performance.

	Precipitation (P)				
Temperature (T)	**MOPEX (%)**	**Santa Clara (%)**	**CPC (%)**	**Daymet (%)**	**Total (%)**
MOPEX	14.07	4.77	6.03	10.30	**35.17**
Santa Clara	7.79	7.79	8.04	11.81	**35.43**
Daymet	5.28	8.04	6.53	9.55	**29.40**
Total (%)	**27.14**	**20.60**	**20.60**	**31.66**	**100**

Table-A I-3 indicates that after a specific calibration strategy, all datasets perform at a very similar level. Still, Table-A I-3 indicates that the T_{mopex}-P_{mopex} dataset performs better on average, followed by $T_{santa\text{-}clara}$-P_{daymet}.

A Wilcoxon test was performed between each of the groups and the MOPEX reference group in Figure-A I-8B to determine which ones were statistically different from the reference values. This test determined that only three datasets performed differently from the MOPEX group: T_{mopex}-$P_{s.clara}$, $T_{s.clara}$-$P_{s.clara}$ and T_{daymet}-$P_{s.clara}$. These test results indicate that $P_{s.clara}$ appears slightly inferior to the CPC and Daymet precipitation datasets with respect to hydrological modeling. It also once again indicates that precipitation datasets are more critical than temperature datasets for hydrological modeling.

Further analyses based on catchment size and climate zone classifications were also performed. Following these tests, it was shown that basin size had no impact on the relative performances of the groups, while the climate type played a role only on the Mediterranean climate basins. The NSE distribution for the 12 climate datasets on the Mediterranean climate catchments are presented in Figure-A I-9.

It can be seen that for the 24 Mediterranean climate catchments, using Daymet precipitation results in much better simulations, independently of the temperature datasets used. The spread is also much smaller. The MOPEX precipitation is the least adequate for this climate zone resulting in a lower median performance value and a larger spread. It is not clear as to why this is the case. These catchments are located in mountainous regions, but so are the catchments from the west coast climate zone who do not exhibit a similar pattern.

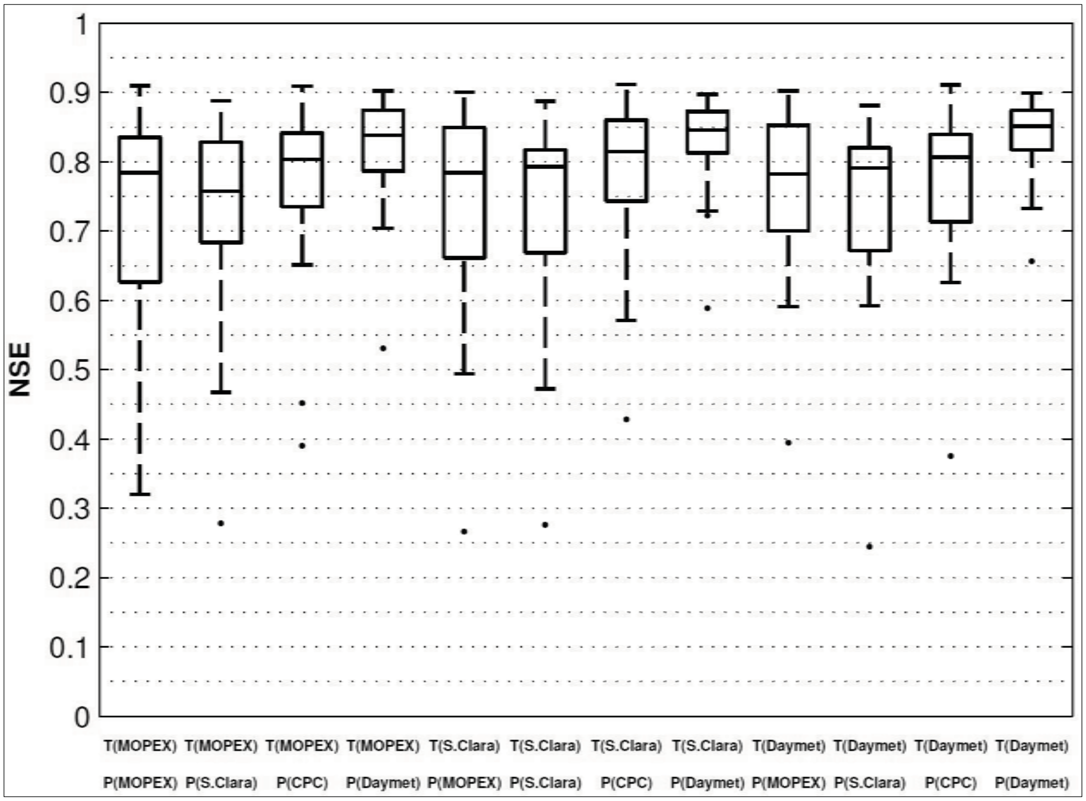

Figure-A I-9 Validation NSE distribution on the Mediterranean catchments for the 12 climate datasets. There are 24 catchments under a Mediterranean climate

I.5 Discussion

While weather station networks remain the most important source of information for hydrological modelling, their often low spatial resolution can sometimes lead to unrepresentative and poor model performance. The need to improve this resolution has been the driving force behind gridded and interpolated climate datasets. However, their potential use in hydrological modelling has been rather limited (Muñoz and al. 2011; Mizukami and al. 2012).

Gridded datasets also have the important advantage of having no-missing data and the potential ability to generate valuable information in areas not densely covered by weather stations, especially when taking into account external variables such as elevation (Tapsoba and al. 2005). On the other hand, interpolating algorithms are also limited in this potential ability, and 'spreading' very sparse station data onto a fine grid may results in artifacts not anchored in any real physics.

To shed light on these questions, this potential of three high-resolution datasets was investigated in this study, with an emphasis on hydrological modeling. The MOPEX dataset (precipitation and temperature) was used as the reference dataset. By mixing the 4 precipitation and 3 temperature datasets, flow discharge was simulated on the 424 catchments of the MOPEX database using the HSAMI hydrology model, resulting in 12 flow discharge time series for each catchment. A common 24-year period (1980-2003) was used for all datasets.

The results clearly indicate that all gridded datasets are biased when compared against the MOPEX reference dataset, and amongst themselves. In particular Daymet precipitation was the most biased when compared against the reference dataset. However, care must be used in the interpretation of the so-called biases. While it is widely agreed upon that station data is the closest approximation of the truth, they do nevertheless suffer from biases. Time series of observed relevant hydrometeorological variables are plagued with problems such as short temporal horizons, missing data, errors, instrument's biases and biases introduced through equipment change and modification of the environment of weather stations including station displacement. Additionally, observation stations are often located in convenient areas (electrical supply, easy access or maintenance) instead of in most relevant areas. As such, the position of the stations within a sparse network is likely to result in biased observations at the catchment scale, such as would result when higher-altitude elevations or remote areas are underrepresented. So while using the MOPEX database as a reference dataset makes sense, no judgment should be made on the suitability of each gridded datasets on the sole basis of the observed biases. Especially since overall, biases remain relatively small for most catchments and for all metrics considered.

Using all gridded datasets as inputs to an already calibrated hydrological model resulted in a decreased performance. This decrease in performance was predictable but not to the extent that was observed for some datasets. For example, the 0.1°C temperature bias and 4.5% precipitation bias of the Daymet dataset translated to a very large 20% decrease in the median NSE criteria (0.783 to 0.634) for the 424 catchments. This indicates that gridded datasets cannot be directly substituted to traditional observation datasets. However, when a specific hydrological model calibration is performed for each dataset, they all perform very similarly. In other words, within the limits of this study, for hydrological modeling purposes, all datasets appear to be equivalent as long as proper calibration is being done. This is not a problem for lumped models but may be a burden for more complex distributed models.

The resolution of the gridded dataset and the complexity of the interpolation scheme do not appear to have any effect in the results. This is likely partly due to the fact that a lumped model was used in the assessment and that all grid points were averaged at the catchment scale, perhaps hiding some potential advantages of the higher-resolution dataset. It is possible that advantages of higher resolution grids could be uncovered using distributed models on the larger catchments. But this would be a time-consuming and computationally-intensive task to set-up and calibrate distributed hydrological models on a large number of catchments. In this study 5088 (424x12) individual model calibrations were performed. This would be a daunting task for a complex distributed hydrological model, even on a subset of the catchments used in this study.

In this work, precipitation and temperature datasets were mixed and matched to form 12 different combinations. No ill-effects were observed in doing so, presumably because precipitation and temperature datasets are usually interpolated independently. As such, there is likely little physical coherence between values of precipitation and temperature in interpolated datasets. This is an aspect that could be better investigated through a comparison against high-resolution climate model or reanalysis of data, where physical consistency between datasets should arguably be much better preserved.

Using statistics averaged over the 424 catchments, this study showed that all gridded datasets behaved similarly for hydrological modeling. However, this study could not evaluate the impact of network density even though it is one of the most interesting scientific problems. The MOPEX database contains catchment-averaged temperature and precipitation data. Information about the number of stations used to generate the catchment-averaged data (which would be needed to estimate network density for each catchment) is not present in the database. Network density could be estimated using the existing NCDC stations. However, since watersheds in the MOPEX database were contributed by many different parties, such an estimation would be error-prone since stations from the CPC cooperative network could also have been used in some catchment and not in some others. Questions related to network density, such as whether or not gridded datasets offer benefits in areas with poor station coverage (as opposed to densely-covered regions where all datasets are expected to converge) would be better tackled using a small subset of carefully chosen watersheds for which precipitation and temperature data would be recalculated using NCDC stations for example.

Also worth noting is that the results are mostly similar from one climate region to the next, except in the Mediterranean climate zone where some differences are visible. However we must take into account the number of catchments in each zone. There are 24 Mediterranean and 13 Marine/West-coast catchments, whereas there are 343 catchments in the humid regions. The comparison between these groups is illustrative at best since there are an insufficient number of catchments for proper statistical significance testing in the small groups.

An advantage of using gridded datasets is that they are much easier to use than station data. They have uniform coverage and no missing data. Catchment-averaging can be done using a simple arithmetic mean, instead of using weight-based averaging as is commonly done, with weights constantly changing depending on which stations are reporting data on any given day. However, gridded datasets are not available in real-time, or near real-time like station data. As such they cannot be used in forecasting mode unless the interpolation is also done in near real-time. This is a process that is now done in-house by many water resources managers, but not yet available to the general public. It is however foreseeable that such data will be available in the near future. For example, such a product is currently in development by Environment Canada (Choi and al. 2013).

Finally this study opens the door to a more in-depth investigation of other gridded datasets. For example, more complex datasets such as PRISM (Daly and al. 1994, 1997) and even reanalysis datasets could be included in such a study. Reanalysis datasets offer the advantage of a much larger set of variables that could be useful for hydrological modeling.

I.6 Conclusion

This work compared 3 gridded climate datasets (Santa Clara, Daymet and CPC) to the MOPEX observation database and analysed their performance in hydrological modeling over 424 catchments in the continental US. The spatial heterogeneity of the catchments allowed comparing the HSAMI model performance relative to catchment size and climate attributes. The comparison was two-fold. First, the gridded climate characteristics were compared to the MOPEX observations with various metrics, and the RMSE and bias were compared between the groups. It was shown that there are non-negligible biases between the gridded datasets and the observations. Second, gridded datasets were used as direct inputs to a hydrological model calibrated on station data. In this case, the biases present in the precipitation and temperature datasets translated to a diminished performance in terms of the NSE criteria for most of the catchments. However, when the hydrological model was recalibrated on each specific gridded dataset (and combinations of precipitation and temperature from different datasets), the performance in validation was similar for most, with a few exceptions. Although some were statistically worse than the reference MOPEX set, none were significantly better. This leads to the conclusion that gridded datasets, while not perfect seem perfectly able to replace observation data where weather station networks are sparse for hydrological modelling, as long as a specific hydrological model calibration is performed using the chosen gridded dataset.

I.7 Acknowledgements

This work was funded through a Natural Science and Engineering Research Council collaborative research grant (NSERC-CRD) with Hydro-Québec, Rio-Tinto-Alcan, Ontario Power Generation and the Ouranos consortium on regional climatology and adaptation to climate change as industrial partners. These industrial research partners are greatly acknowledged for their direct and indirect contributions to this work. We also sincerely thank all the individuals and institutions who developed the datasets used in this work, and made them available to the scientific community. We hope that this work is a small contribution to their important effort.

I.8 References

Adam, J. C., and D. P. Lettenmaier, 2003: Adjustment of global gridded precipitation for systematic bias. J. Geophys. Res., 108, 4257, doi:10.1029/2002JD002499, D9.

Anagnostopoulou, Chr., P. Maheras, T. Karacostas, and M. Vafiadis, 2003: Spatial and temporal analysis of dry spells in Greece. Theoretical and Applied Climatology 74, 77-91.

Arsenault, R., J. S. Malo, F. Brissette, M. Minville, and R. Leconte, 2013: Structural and non-structural climate change adaptation strategies for the Péribonka water resource system. Water Resour. Manag., doi:10.1007/s11269-013-0275-6.

Arsenault, R., and F. Brissette, 2014a: Determining the Optimal Spatial Distribution of Weather Station Networks for Hydrological Modeling Purposes Using RCM Datasets: An Experimental Approach. J. Hydrometeor., 15, 517–526.

Arsenault, R., A. Poulin, P. Côté, and F. Brissette, 2014b: Comparison of Stochastic Optimization Algorithms in Hydrological Model Calibration. J. Hydrol., Eng., 19(7), 1374–1384.

Baillargeon, S., J. Pouliot, L-P. Rivest, V. Fortin, and J. Fitzback, 2004: Interpolation statistique multivariable de données de précipitations dans un cadre de modélisation hydrologique. Colloque national Géomatique de l'Association canadienne des sciences géomatiques (ACSG-CIG), Montréal, 27-28 octobre.

Bastola, S., and D. François, 2012: Temporal extension of meteorological records for hydrological modelling of Lake Chad Basin (Africa) using satellite rainfall data and reanalysis datasets. Meteorol. Appl. 19, 54-70.

Chen J., F. Brissette, and R. Leconte, 2011a: Uncertainty of downscaling method in quantifying the impact of climate change on hydrology. Journal of Hydrology. J. Hydrol., 401, 190–202.

Chen, J., F. P. Brissette, A. Poulin, and R. Leconte, 2011b: Overall uncertainty study of the hydrological impacts of climate change for a Canadian watershed, Water Resour. Res., doi:10.1029/2011WR010602.

Chen, J., F. P. Brissette, and R. Leconte, 2012: Downscaling of weather generator parameters to quantify the hydrological impacts of climate change, Climate Research, doi: 10.3354/cr01062.

Choi, H., P. F. Rasmussen, and V. Fortin, 2013: Evaluation and Comparison of Historical Gridded Data Sets of Precipitation for Canada. AGU Fall Meeting Abstracts. Vol. 1.

Cressman, G. P., 1959: An operational objective analysis system. Monthly Wea. Review, 87(10), 367-374.

Daly, C., G. H. Taylor, and W. P. Gibson, 1997: The PRISM approach to mapping precipitation and temperature. 10th Conference on Applied Climatology, Reno, NV, Amer. Meteor. Soc., 10-12.

Daly, C., R. P. Neilson, and D. L. Phillips, 1994: A statistical–topographic model for mapping climatological precipitation over mountainous terrain, J. Appl. Meteor. 33, 140–158.

Dirks, K. N., J. E. Hay, C. D. Stow, and D. Harris, 1998: High-resolution studies of rainfall on norfolk island. part ii : The interpolation of high-spatial-resolution rainfall data on norfolk island. J. Hydrol. 208,187-193.

Duan, Q., and Coathors, 2006: Model Parameter Estimation Experiment (MOPEX): An overview of science strategy and major results from the second and third workshops, J. Hydrol. 320, 3–17, doi:10.1016/j.jhydrol.2005.07.031.

Fortin, V., 2000: Le modèle météo-apport HSAMI : historique, théorie et application. Rapport de recherche, Institut de Recherche d'Hydro-Québec. Varennes, 68p.

Fulton, R. A., J. P. Breidenbach, D.-J. Seo, D. A. Miller, and T. O'Bannon, 1998: The WSR-88D rainfall algorithm. Wea. Forecasting 13 (2), 377–395.

Goodison, B. E., H. L. Ferguson, and G. A. McKay, 1981: Measurement and data analysis. In: Gray, D.M., and Male, D.H., eds. Handbook of snow: Principles, Processes, Management and Use, 191–274.

Goodison, B. E., P. Y. T. Louie, and D. Yang, 1998: WMO solid precipitation measurement intercomparison. Instruments and Observing Methods Rep. 67 (WMO/TD 872), World Meteorological Organization, Geneva, Switzerland, 212 pp.

Hansen, N., and A. Ostermeier, 2001: Completely Derandomized Self-Adaptation in Evolution Strategies, Evolutionary Computation, 9(2), 159-195.

Hansen, N., and A. Ostermeier, 1996: Adapting arbitrary normal mutation distributions in evolution strategies : The covariance matrix adaptation, In Proceedings of the 1996 IEEE International Conference on Evolutionary Computation, 312-317.

Higgins, R. W., W. Shi, E. Yarosh, and R. Joyce, 2000: Improved United States precipitation quality control system and analysis. NCEP/Climate Prediction Center ATLAS N°6. Available: www.cpc.ncep.noaa.gov/research_papers/ncep_cpc_atlas/7/index.html

Higgins, R. W., J. E. Janowiak, and Y. Yao, 1996: A gridded hourly precipitation data base for the United States (1963-1993). NCEP/Climate Prediction Center ATLAS No. 1., 46 p.

Huffman, G. J., R. F. Adler, M. Morrissey, D. Bolvin, S. Curtis, R. Joyce, B. McGavock, and J. Susskind, 2001: Global precipitation at one-degree daily resolution from multisatellite observations, J. Hydrometeor., 2, 36-50.

Hulme, M., 1992: A 1951-80 global land precipitation climatology for the evaluation of General Circulation Model. Climate Dynamics 7, 57-72.

Hundecha, Y., and A. Bárdossy, 2005: Trends in daily precipitation and temperature extremes across western Germany in the second half of the 20th century. International Journal of Climatology 25, 1189-1202.

Hutchinson, M. F., D. McKenney, K. Lawrence, J. Pedlar, R. Hopkinson, E. Milewska, and P. Papadopol, 2009: Development and Testing of Canada-Wide Interpolated Spatial Models of Daily Minimum-Maximum Temperature and Precipitation for 1961-2003. Amer. Meteor. Soc., Bulletin, 48, 725-741.

Jones, D. A., and B. Trewin, 2000: The spatial structure of monthly temperature anomalies over Australia. Aust. Meteorol. Mag., 49, 261–276.

Kottek, M., J. Grieser, C. Beck, B. Rudolf, and F. Rubel, 2006: World Map of the Köppen-Geiger climate classification updated. Meteorol. Z., 15, 259-263. DOI: 10.1127/0941-2948/2006/0130.

Krajewski, F.W., V. Lakshmi, K. P. Georgakakos, C. J. Subhash, 1991: A Monte Carlo study of rainfall sampling effect on a distributed catchment model. Wat. Resour. Res. 27, 119-128.

Lopes, V.L., 1996: On the effect of uncertainty in spatial distribution of rainfall on catchment modelling. Catena 28, 107-119

Maurer, E. P., A. W. Wood, J. C. Adam, D. P. Lettenmaier, and B. Nijssen, 2002: A Long-Term Hydrologically-Based Data Set of Land Surface Fluxes and States for the Conterminous United States. J. Climate 15, 3237-3251.

Minville, M., F. Brissette, and R. Leconte, 2008: Uncertainty of the impact of climate change on the hydrology of a Nordic watershed. J. Hydrol., 358(1-2), 70-83.

Minville, M., F. Brissette, S. Krau, and R. Leconte. 2009: Adaptation to Climate Change in the Management of a Canadian Water-Resources System Exploited for Hydropower. Water Resour. Manag., 23 (14), 2965-2986, DOI: 10.1007/s11269-009-9418-1.

Mitchell T. D., and P. D. Jones, 2005: An improved method of constructing a database of monthly climate observations and associated high-resolution grids. International Journal of Climatology 25(6): 693–712.

Mizukami, N., and M. B. Smith, 2012: Analysis of inconsistencies in multi-year gridded quantitative precipitation estimate over complex terrain and its impact on hydrologic modeling. J. Hydrol. 428, 129–141.

Muñoz, E., C. Álvarez, M. Billib, J. L. Arumí, and D. Rivera, 2011: Comparison of gridded and measured rainfall data for basin-scale hydrological studies. Chilean Journal of Agricultural Research 71(3), 459-468.

Murray S. J., I. M. Watson, and I. C. Prentice, 2013: The use of dynamic global vegetation models for simulating hydrology and the potential integration of satellite observations. Progress in Physical Geography, 37(1) 63–97.

Nash, J. E., and J. V. Sutcliffe, 1970: Rivver flow forecasting through conceptual models, Part 1. A discussion of principle. J. Hydrol., 10, 282-290.

National Climatic Data Center. 1994. Surface Land daily Cooperative. Summary of the Day (TD-3200). National Environmental Satellite and Data Information Service, NOAA, U.S. Department of Commerce

Obled, C., J. Wendling, and K. Beven, 1994: The sensitivity of hydrological models to spatial rainfall patterns : an evaluation using observed data. J. Hydrol., 159, 305-333.

Poulin, A., F. Brissette, R. Leconte, R. Arsenault, and J. S. Malo, 2011: Uncertainty of hydrological modelling in climate change impact studies in a Canadian, snow-dominated river basin, J. Hydrol., 409 (3-4), 626-636.

Rakotomalala, R. 2008: Comparaison de populations-Tests non paramétriques. Université Lumière Lyon 2. 201pp.

Rasmussen, R., M. Dixon, S. Vasiloff, F. Hage, S. Knight, J. Vivekanandan, and M. Xu, 2003: Snow nowcasting using a real-time correlation of radar reflectivity with snow gauge accumulation. J. Appl. Meteor. 42 (1), 20–36.

Ruelland, D., S. Ardoin-Bardin, G. Billen, and E. Servat, 2008: Sensitivity of a lumped and semi-distributed hydrological model to several methods of rainfall interpolation on a large basin in West Africa. J. Hydrol., 361, 96– 117, doi:10.1016/j.jhydrol.2008.07.049

Schaake, J. C., Q. Duan, M. Smith, and V. Koren, 2000: Criteria to Select Basins for Hydrologic Model Development and Testing. Preprints, 15th Conference on Hydrology, 10–14 January 2000, Amer. Meteor. Soc., Long Beach, CA, Paper P1.8.

Schaake, J., S. Cong, and Q. Duan, 2006: The U. S. MOPEX Data Set, IAHS Publication 307 (2006), 9-28.

Schmidli, J., and C. Frei, 2005: Trends of Heavy Precipitation and Wet and Dry Spells in Switzerland during the 20th Century International J. Climatol., 25, 753-771

Seo, D.-J., 1998: Real-time estimation of rainfall fields using radar rainfall and rain gage data. J. Hydrol., 208 (1–2), 37–52.

Shepard, D.S., 1984: Computer mapping: The SYMAP interpolation algorithm. In : Spatial Statistics and Models. Springer Netherlands, 133-145.

Skaugen, T., and J. Andersen, 2010: Simulated precipitation fields with variance-consistent interpolation. Hydrol. Sci. J., 55(5), 676-686.

Steiner, M., J. A. Smith, S. J. Burges, C. V. Alonso, and R. W. Darden, 1999: Effect of bias adjustment and rain gauge data quality control on radar rainfall estimation. Water Res. Research 35 (8), 2487–2503.

Tapsoba, D., V. Fortin, and F. Anctil, 2005: Apport de la technique du krigeage avec dérive externe pour une cartographie raisonnée de l'équivalent en eau de la neige: Application aux bassins de la rivière Gatineau. Canadian Journal of Civil Engineering 32.1 289-297.

Tang Q., H. Gao, H. Lu, and D. P Lettenmaier, 2009: Remote sensing: hydrology. Progress in Physical Geography, 33 (3), 490-509.

Taylor, G., C. Daly, W. Gibson, and J. Sibul-Weisburg, 1997: Digital and Map Products Produced Using PRISM, 10th Conf. on Applied Climatology, Reno, NV, Amer. Meteor. Soc., 217-218.

Thornton, P. E., S. W. Running, and M. A., White, 1997: Generating surfaces of daily meteorological variables over large regions of complex terrain. J. Hydrol., 190:214-251. [Available online at http://dx.doi.org/10.1016/S0022-1694(96)03128-9]

Thornton, P. E., M. M. Thornton, B. W. Mayer, N. Wilhelmi, Y. Wei, and R. B. Cook, 2012: Daymet: Daily surface weather on a 1 km grid for North America, 1980-2012. Oak Ridge National Laboratory Distributed Active Archive Center, Oak Ridge, T, N. Doi, 10. [Available online at http://daymet.ornl.gov/]

Tozer, C. R., A. S. Kiem, and D. C. Verdon-Kidd, 2012: On the uncertainties associated with using gridded rainfall data as a proxy for observed. Hydrol. Earth Sys. Sci., 16, 1481–1499.

Turk, F. J., P. Arkin, E. Ebert, and M. Sapiano, 2008: Evaluating high resolution precipitation products. Bull. Amer. Meteor. Soc., 89, 1911–1916.

Yatagai A., O. Arakawa, K. Kamiguchi, H. Kawamoto, M. I. Nodzu, and A. Hamada, 2009: A 44-year daily precipitation dataset for Asia based on dense network of rain gauges. SOLA, 5, 137–140. Doi:10.2151/sola.2009-035.

Warner, T. T., E. A. Brandes, J. Sun, D. N. Yates, and C. K. Mueller, 2000: Prediction of a flash flood in complex terrain. Part I: A comparison of rainfall estimates from radar, and very short range rainfall simulations from a dynamic model and an automated algorithmic system. J. Appl. Meteor., 39 (6), 797–814.

Westrick, K. J., C. F. Mass, and B. A. Colle, 1999: The limitations of the WSR-88D radar network for quantitative precipitation measurement over the Coastal Western United States. Bull. Amer. Meteor. Soc., 80 (11), 2289–2298.

Widmann, M., C. S. Bretherton, 2000: Validation of mesoscale precipitation in the NCEP reanalysis using a new gridcell dataset for the northwestern United States. J. Climate., 13, 1936–1950.

ANNEXE II

ARTICLE EN COLLABORATION 2 : REDUCING THE PARAMETRIC DIMENSIONALITY FOR RAINFALL-RUNOFF MODELS : A BENCHMARK FOR SENSITIVITY ANALYSIS METHODS.

Jie Chen[1], Richard Arsenault[1] et François Brissette[1]

[1] Département de Génie de la Construction, École de technologie supérieure,

1100 Notre-Dame Ouest, Montréal, Québec, Canada H3C 1K3.

Article soumis à la revue « Advances in Water Resources » en janvier 2014.

Abstract: The appropriate reduction of the number of model parameters can be an important tool in reducing the effect of parameter uncertainty on hydrological modeling. Sobol' sensitivity analysis has been used successfully in the past to identify the relative importance of hydrological model parameters and fixing them accordingly. However, the Sobol' method assumes an independence of parameters which is known to be incorrect. The effects of its limitations on reducing hydrological model parameters need to be investigated. This study proposes an experimental approach to assess the commonly used Sobol' analysis for reducing the parameter dimensionality of hydrological models. Specifically, a new approach based on a multi-objective genetic algorithm (MOGA) is proposed. In this approach, the number of model parameters is directly pitted against an efficiency criterion within the MOGA, thus allowing both the identification of key model parameters and the optimal number of parameters to be used within the same analysis. The proposed approach was tested over two different Canadian Nordic watersheds using a conceptual lumped hydrological model (HSAMI) with 23 free parameters. Its performance was then compared with the Sobol' method for one watershed. The results show that both methods performed very similarly and allowed 11 out of 23 HSAMI parameters to be reduced with little loss in model performance, although the relative importance of some parameters was different. Based on this comparison, Sobol' appears to be an effective and robust method despite its limitations. On the other hand, the MOGA algorithm outperformed Sobol' analysis for further reduction of the parametric space and found optimal solutions with as little as 8 parameters with minimal performance loss in validation. However, this gain was achieved at the expense of a much larger computational burden.

Keywords: hydrological model; parameter reduction; Sobol' sensitivity analysis; multi-objective genetic algorithm

II.1 Introduction

Hydrological models have been used in a wide range of water resources management activities, such as watershed streamflow quantification, reservoir system operations, groundwater protection, water distribution systems, water use [Wurbs, 1998; Pechlivanidis et al., 2011] and climate change impacts assessment [e.g. Wilby and Harris, 2006; Kay et al., 2009; Chen et al., 2011a, b, and 2012]. The successful use of hydrological models largely depends on how well they are calibrated and the complexity of model calibration depends, amongst others, on the number of parameters that must be optimized [van Werkhoven et al., 2009]. Model performance over the calibration period is generally improved when introducing additional parameters due to the added degrees of freedom; however, this may lead to over-fitting and poorer performance when using the model over different time period [Myung and Pitt, 2002]. The number of "free parameters" is generally too large for most commonly used hydrological models and the problem of equifinality is ubiquitous [Beven and Binley, 1992; Wagener et al., 2001; Tonkin and Doherty, 2005]. Because of the equifinality problem, the uncertainty associated with the choice of an optimal parameter set can be large, since a single optimal parameter set for a hydrologic model may not be found during the calibration process [Klepper et al., 1991; van Straten and Keesman, 1991; Beven and Binley, 1992; Yapo et al., 1996]. This is particularly true for hydrological models with a large parametric dimensionality, since both over-parameterization and parameter interactions can cause model parameters to be not uniquely identifiable [Gan et al., 2014]. Additionally, the use of different efficiency metrics for model optimization may result in different optimal parameter sets [van Werkhoven et al., 2009].

Several studies [e.g. Bastidas et al., 1999; Huang and Liang, 2006; Cox et al., 2006; Hogue et al., 2006; Wagener and Kollat, 2007; Pechlivanidis et al., 2010] suggest that the most appropriate approach in dealing with equifinality consists in reducing the number of free parameters by fixing the less important parameters to constant values. The reduction in model parameters is also an efficient way to reduce the impacts of parameter uncertainty on hydrological simulations. It also has the advantage of simplifying the optimization problem linked to model calibration. Additionally, if a reliable parameter identification technique is developed over regions where data is available, the reduced parameter set can be much more easily regionalized for predicting flows in ungauged regions [Lee et al., 2005; Pechlivanidis et al., 2010; Arsenault and Brissette, 2014].

To reduce the number of parameters during the model calibration process, it is necessary to quantitatively evaluate the influence of each parameter on model performance. This has been done in several studies. For example, Huang and Liang [2006] introduced an alternative subsurface flow parameterization into a hydrological model (Three-Layer Variable Infiltration Capacity). Two out of three parameters were reduced in the calibration process and the results showed that the performance of the hydrological model with the one-parameter subsurface flow formulation was comparable to the model with the three-parameter version. This study further indicated that the reduction of model parameters is an effective way to reduce the parameter uncertainty for hydrological simulations. More recently, Pechlivanidis et al. [2010] overcame problems in distributed modelling associated with the lack of parameter identifiability through reduction of parameter dimensionality. The semi-distributed Probability Distributed Moisture model was calibrated based on parameter regionalisation and a good balance in the model complexity was achieved assuming that some parameters directly correspond to their physical characteristics while other insensitive parameters were held constant. Cox et al. [2006] also showed that parametrically reduced models have lower prediction residual sums of squares (the sum of squared differences between the observed and predicted values) than the original model, strongly suggesting that the original model was over-fitted.

Sensitivity analysis (local and global sensitivity analysis) is one of the efficient ways to identify the influence of each parameter on the model performance. Thus, it can be used to reduce the parameter dimensionality for hydrological models by fixing and ignoring insensitive parameters during the calibration process [Wilson et al., 1987a, b; Pitman, 1994; Gao et al., 1996; Bastidas et al., 1999]. Among all sensitivity analysis methods [Gan et al., 2014], the Sobol' method [Sobol', 1993] is one of the most widely used global sensitivity analysis methods [e.g., Pappenberger et al., 2008; van Werkhoven et al., 2008; Nossent et al., 2011; Zhang et al., 2013]. Current Sobol' analysis methods are capable of assessing the effect of each parameter and its interactions with other parameters on the model output and have been used successfully in several studies. For example, Tang et al. [2007] compared the Sobol' method with three other sensitivity analysis tools (Parameter Estimation Software, Regional Sensitivity Analysis, and Analysis of Variance) and found that the Sobol' method yielded more robust sensitivity rankings than the other methods. Van Werkhoven et al. [2009] used Sobol' analysis as a screening tool to reduce the parametric dimensionality of hydrological models and found that parameters explaining at least 20% of the model output variance should be included in the calibration process. This threshold generally reduced the number of model parameters by at least 30% for the Sacramento Soil Moisture Accounting model, while maintaining good performance of predictions. This study further indicated that the reduced parameter sets changed across different hydroclimatic gradients and that multiple metrics may be necessary for the selection of optimized parameters. More recently, Nossent et al. [2011] used the Sobol' method to analyze the parameter sensitivity of Soil & Water Assessment Tool for flow simulations and found that no more than 9 parameters out of 26 are needed to adequately represent the model output variability. However, the Sobol' method assumes no correlation between parameters, whereas strong inter-dependence of parameters is usually found in hydrological models [Pechlivanidis et al., 2010]. It is thus necessary to investigate the effects of this independence hypothesis on Sobol' performance in the reduction of hydrological model parameters by comparing it to other methods that do not depend on this hypothesis.

Accordingly, this work proposes a multi-objective genetic algorithm (MOGA)-based approach to assess the most commonly used Sobol' sensitivity analysis in the reduction of the hydrological model parametric dimensionality. The proposed approach directly uses the number of model parameters as one of the objectives, thus directly quantifying the value of using additional free parameters, as well as identifying the best combination of model parameters. The proposed method was tested using a lumped conceptual rainfall-runoff model over two Canadian watersheds in the Province of Quebec, and then compared with the Sobol' method for one watershed.

II.2 Studied Watersheds

Two Canadian watersheds (Peribonka and Yamaska, Figure-A II-1) located in the Province of Quebec were selected to test the proposed method and evaluate its applicability under different watershed characteristics.

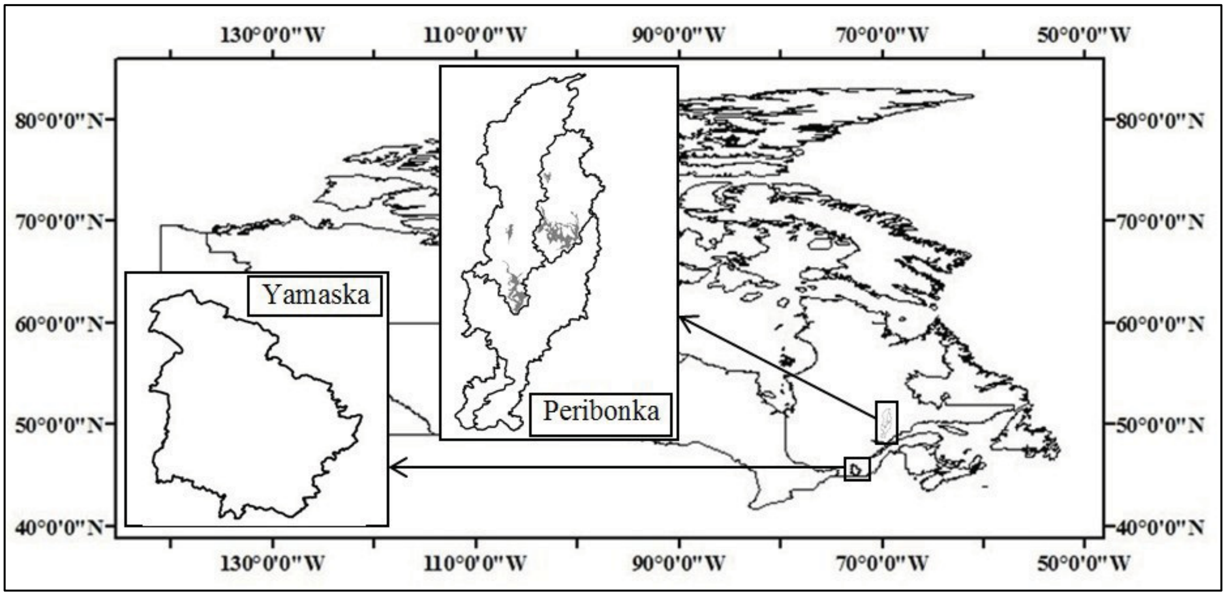

Figure-A II-1 Location map of the two watersheds

Both Peribonka and Yamaska watersheds are composed of several tributaries draining basins of approximately 27000 km^2 and 4843 km^2 in southeastern and southern Quebec, respectively. The southern parts of Peribonka and Yamaska watersheds, respectively named as Chute-du-Diable and Cowansville watersheds, are used in this study. The details on both watersheds are presented below.

II.2.1 Chute-du-Diable

The Chute-du-Diable watershed is located in the central part of the province of Quebec, and covers 9700 km^2 of mostly forested areas with sparse population. The basin is part of the northern Quebec subarctic region, characterized by wide daily and annual temperature ranges, heavy wintertime snowfall, and pronounced rainfall and/or snowmelt peaks in the spring. The average annual precipitation in the area is 962 mm, of which about 36% is snowfall. The average annual maximum and minimum temperatures between 1979 and 2003 were 5.49°C and -5.85°C, respectively. The Chute-du-Diable watershed contains a large hydropower reservoir managed by Rio Tinto Alcan for hydroelectric power generation. River flows are regulated by two upstream reservoirs. Snow plays a crucial role in the watershed management, with 35% of the total yearly discharge occurring during the spring flood. The mean annual discharge of the Chute-du-Diable watershed is 211.4m^3/s. Snowmelt peak discharge usually occurs in May and averages about 1220 m^3/s. Natural inflows series have been reconstructed by Rio Tinto Alcan using operation data from the upstream reservoir. Hydrologic model performance (to be detailed later) testifies to the quality of the reconstructed inflow data.

II.2.2 Cowansville

Cowansville is a non-regulated, 210 km^2 watershed located in southern Quebec. The average annual rainfall in the Cowansville watershed is 1267 mm with about 22% of snow. The average annual maximum and minimum temperatures were 11.59 °C and 1.14 °C, respectively over the last 30 years. As opposed to Chute-du-Diable watershed, the Cowansville watersheds' southern location and much smaller size results in hydrographs that are less dominated by snowmelt. In fact, annual maximum peak discharge often occurs during the summer season, a feature never seen in the Chute-du-Diable watershed. The mean annual discharge of the Cowansville watershed is 4.5m^3/s. Snowmelt peak discharge usually occurs in April and averages about 70 m^3/s.

II.3 Methodology

II.3.1 Hydrological modeling

HSAMI is a 23-parameter, lumped, conceptual, rainfall-runoff model developed by Hydro-Québec, and which has been used to forecast natural inflows for over 20 years [Fortin, 2000]. HSAMI is used by Hydro-Québec for daily forecasting of natural inflows on nearly 100 watersheds with drainage areas ranging from 160 km^2 to 69,195 km^2. HSAMI was also used in several flow forecasting and climate change impact studies [e.g. Minville et al., 2008, 2009; Chen et al., 2011a, b, 2012; Poulin et al., 2011, Arsenault et al., 2013]. Of HSAMIs' 23 parameters, two account for evapotranspiration, 6 for snow accumulation/melting, 10 for vertical water movement, and 5 for horizontal water movement (see Table-A II-1). Vertical flows are simulated with 4 interconnected linear reservoirs (snow on the ground, surface water, unsaturated and saturated zones). Horizontal flows are routed through 2 unit hydrographs and one linear reservoir. In addition, the model takes into account snow accumulation, snowmelt, soil freezing/thawing and evapotranspiration. Model calibration was done automatically using the Covariance Matrix Adaptation Evolution Strategy (CMAES) [Hansen and Ostermeier, 1996, 2001], following the conclusions of Arsenault et al. [2014].

The basin-averaged daily input data required for HSAMI are liquid and solid precipitation, as well as maximum and minimum temperatures. Cloud cover fraction and snow water equivalent can also be used as input, if available. A natural inflow or discharge time series is also needed for calibration/validation.

Two different metrics were used to evaluate model's adequacy on representing the high and low flows. The first efficiency metric is the commonly used Nash-Sutcliffe Efficiency (NSE) [Nash and Sutcliffe, 1970], which is a normalized statistic that determines the relative magnitude of the residual variance compared to the observed data variance. The second efficiency metric is a Box-Cox transformed version of the root mean square error (TRMSE) as used in the study of [van Werkhoven et al., 2009]. To ensure that conflicting objectives exist, in the case of the NSE, the objective to minimize will be (1-NSE). No such problem exists with TRMSE whose value naturally decreases with a better fit.

Table-A II-1 HSAMI 23 free parameters

Sub-model	**ID**	**Physical meaning**	**Unit**	**Parameter range**
Evapo-transpiration	P1	Factor multiplying potential evapotranspiration (PET) for the estimation of summer real evapotranspiration (RET)	--	[0.6 3]
	P2	Factor multiplying PET for estimating the RET in winter	--	[0 0.3]
Snowmelt	P3	Snow melting rate during daytime. ΔT in Celsius is calculated as the difference between Tmax and parameter of Tmax threshold for snowmelt (P5).	cm/Δ° C/day	[0.05 0.4]
	P4	Snow melting rate during nighttime. ΔT in Celsius is calculated as the difference between P5 and Tmin.	cm/Δ° C/day	[0.05 0.5]
	P5	Tmax threshold for snowmelt	°C	[-6 7]
	P6	Tmin threshold for accelerated snowmelt	°C	[-6 6]
	P7	Reference temperature for calculating the heat supplied by the rain to the snow cover	°C	[-6 4]
	P8	Empirical parameter used to connect the state variables describing snow cover and cumulated snowmelt to the proportion of the basin covered by snow	--	[0.8 5]
Surface runoff	P9	Empirical parameter used to connect the state variables describing soil freezing and thawing to the proportion of snowmelt water flowing on the surface	--	[0.8 15]
	P10	24-hour rainfall amount needed to generate 50% runoff with completely dry soil.	cm	[10 45]
	P11	24-hour rainfall amount needed to generate 50% runoff with completely saturated soil.	cm	[1 8]
Vertical water movement	P12	Water amount in the unsaturated zone that cannot drain by gravity	cm	[0 7]
	P13	Maximum water amount that can be contained in the unsaturated soil zone	cm	[3 25]
	P14	Maximum water amount that can be contained in the aquifer before generating surface runoff	cm	[4 30]
	P15	Proportion of surface water flowing through the intermediate hydrograph instead of moving through the soil column	--	[0.15 0.7]
	P16	Proportion of soil water that is directed to the intermediate hydrograph when the unsaturated zone overflows	--	[0.3 1]
	P17	Emptying rate of the unsaturated zone to the groundwater reservoir	$24h^{-1}$	[0.09 0.07]
	P18	Emptying rate of the groundwater reservoir (base flow)	$24h^{-1}$	[0.006 0.018]
Horizontal water movement	P19	Emptying rate of the intermediate reservoir, through the intermediate hydrograph	$24h^{-1}$	[0.6 1.2]
	P20	Time to peak for the surface unit hydrograph	day	[0.3 5]
	P21	Shape parameter of the surface hydrograph (using a gamma distribution function)	--	[0.4 5]
	P22	Time to peak for the intermediate unit hydrograph	day	[1.5 13]
	P23	Shape parameter of the intermediate hydrograph (using a gamma distribution function)	--	[0.15 1.5]

Note: Tmax=maximum temperature and Tmin=minimum temperature

In this study, ten years of data were used for model calibration, and another ten years of data were used for model validation as presented in Table-A II-2. HSAMI with all 23 free parameters was independently calibrated and validated 1000 times to quantify the parameter equifinality. The median of NSE and TRMSE for these 1000 calibrations showed good performances of HSAMI for both watersheds. The results for the Chute-du-Diable watershed were consistently better than those of the Cowansville watershed. This is because the input data quality of the former is considered better than that of the latter, due to more weather stations, and because the daily time step used in this study was less suited to the smaller watershed. In addition, snowmelt dominated basins are also usually easier to model, because the winter streamflow is not sensitive to the precipitation input. Since TRMSE is watershed size-dependent, its values are smaller for the Cowansville watershed, even though HSAMI was better optimized for the Chute-du-Diable watershed.

II.3.2 Multi-objective genetic algorithm (MOGA)

Multi-objective optimization has been quite commonly used in hydrology over the past decade [Sawaragi et al., 1985; Steuer, 1986; Yapo et al., 1998; Vrugt et al., 2003; Shafii and Smedt, 2009; Reed et al., 2013; Efstratiadis and Koutsoyiannis, 2010] . The central goal of using a multi-objective approach in model calibration is to increase model robustness over different efficiency metrics (best-compromise solution) as well as to reduce parameter uncertainty. However, most hydrology studies have focused on finding an optimal parameter set (usually using an automated procedure), and not on the more fundamental aspect of reducing the parameter uncertainty. Since multi-objective optimization is the process of simultaneously optimizing two or more conflicting objectives it allows evaluating the correlation and trade-offs between different efficiency metrics.

In this work, the relationship between the number of parameters that must be optimized during the calibration process and the model performance is viewed as a multi-objective problem. The multi-objective optimization problem will be solved using MOGA which is a popular meta-heuristic multi-objective optimizer [Deb, 2001; Konak et al., 2006]. Two multi-objective optimization problems will be solved independently. With MOGA, the population is randomly initialized and further evolves until it eventually converges to a non-dominated solution. A Pareto front is generated along which all solutions can be considered optimal. In other words, the Pareto front should indicate the best model performance in calibration when using an increasing number of model parameters. Conversely, minimization of the number of model parameters will only be achieved through a reduction in model performance due to the reduction in degrees of freedom. The Pareto front, and its translation into the validation dataset, should allow for a quantitative evaluation of the true benefits of increasing the number of model parameters, as well as the identification of the most important model parameters, all in the same step.

In this study, a controlled elitist genetic algorithm was used. It consists of a variant of the Non-dominated Sorted Genetic Algorithm (NSGA-II) [Deb, 2001; Deb et al., 2002]. The algorithms' hyperparameters were set in such a way to emphasize exploration of the search space because of the large amount of possible combinations. As such, a population of 200 individuals was selected. To compensate for the long processing time of generating 200 candidates per iteration (where each candidate requires a complete calibration of HSAMI), only 20 generations were performed. Furthermore, since one of the objectives was to minimize the number of model parameters, the initial populations were forced using all 23 parameters. This ensured that the search would sample along the entire Pareto front rather than start at a low point and climb the Pareto front. The latter could be biased if no optimal candidates are found for a few iterations past the current highest number of parameters value.

When the algorithm chooses less than 23 parameters, an important issue arises with the value that must be assigned to all non-chosen parameters. If one decides not to calibrate all of HSAMI's 23 parameters, forced values have to be set for the remaining parameters. In this case, fixed parameter values were randomly assigned within a reduced search space defined for each parameter. The reduced search space is a subset of the total parameter space (Table-A II-1) used by the optimization process. The total parameter space is chosen, while maintaining a physical sense to minimize the chances that parameters are constrained by the search boundaries. The reduced search space was evaluated using the range of plausible values for each parameter following 1000 different calibrations using all 23 parameters. This procedure eliminated the possibility of choosing an impossible fixed value for several parameters. Multi-objective optimization was performed 8 times for each efficiency metric, with each iteration using different parameter sets for fixed parameters. The 8 different parameter sets were selected randomly from the reduced search space. The random assignments partly eliminated biases linked to a deliberate choice in fixed parameter values. Overall, thirty-two multi-objective optimization runs were conducted since each of the 8 random assignments was run for the 2 objective functions and both basins.

II.3.3 Sobol' sensitivity analysis

The Sobol' method [Sobol', 1993] is also used to assess parameter sensitivity and reduce the parameter dimensionality. The Sobol' method assesses the sensitivity of each parameter and parameter interactions based on their contributions to the total model variance. These contributions are called Sobol' sensitivity indices (SI), which are expressed by the ratio of the partial variance to the total variance. The first order SI measures the variance contribution of the individual parameter to the total model variance. The second order SI measures the variance contribution of the two-parameter interaction to the total model variance, and so on. The total SI measures the sensitivity due to the combined effect of one parameter and its interactions with all other parameters. When using the Sobol' method to reduce the parametric dimensionality for a hydrological model, parameters that do not make important contributions to the total variance can be fixed with constant values and ignored during the optimization process [Sobol', 1993; Pappenberger et al., 2008; van Werkhoven et al., 2008; Nossent et al., 2011; Zhang et al., 2013].

In this study, the Sobol' implementation was as follows. First, 100000 parameter sets were sampled following a Sobol' sequence in the available parameter space. The Sobol' Indices were then computed for the first and total orders from the selected parameter set using the Saltelli algorithm [Saltelli, 2002]. For this work, only the total order parameters are of importance since the idea is to fix the least important parameters to the total variance, thus their interactions must be considered.

The parameters were then sorted by order of importance and the HSAMI model was calibrated with the least important parameter fixed to a random value within the predetermined boundaries. This operation was repeated 100 times to eliminate (or at least greatly reduce) the stochastic effects on the results. The results were noted in calibration and in validation, then the next least-important parameter was fixed and the process was launched again. However, the fixed parameters are kept at the same values throughout the study, i.e. the first parameter to be fixed keeps its value for the duration of the study, and thus the parameter space is kept intact for each calibration. This means that for each supplementary fixed parameter, the parameter space shrinks by 1 dimension but the parameter space is a subset of the previous parameter space. This methodology is continued until 22 parameters are fixed and a single one is left to calibrate. The idea to keep the previously fixed parameters to a constant value allows comparing the effect of fixing the next parameter in a stable environment, instead of a constantly changing one (due to random selection of the parameters at each step).

II.4 Results

This section consists of four sub-sections. The first sub-section involves analyzing the parametric equifinality for the hydrological model HSAMI, as a first step prior to the parameter reduction analysis. The second sub-section includes two steps to reduce HSAMI parameters using the multi-objective approach. The multi-objective optimization is initially conducted to obtain a Pareto front. Through the Pareto front, a judgment can be made about the number of parameters that can be fixed, while keeping an adequate model performance. The third step involves ordering parameters based on their importance with an experimental approach. Finally, the proposed method is compared to the Sobol' sensitivity analysis method in terms of their performances in reducing HSAMI parameters.

II.4.1 Parametric equifinality analysis

Figure-A II-2 presents the cumulative distribution function (CDF) of all parameter values obtained after performing 1000 automatic calibrations with the CMAES algorithm.

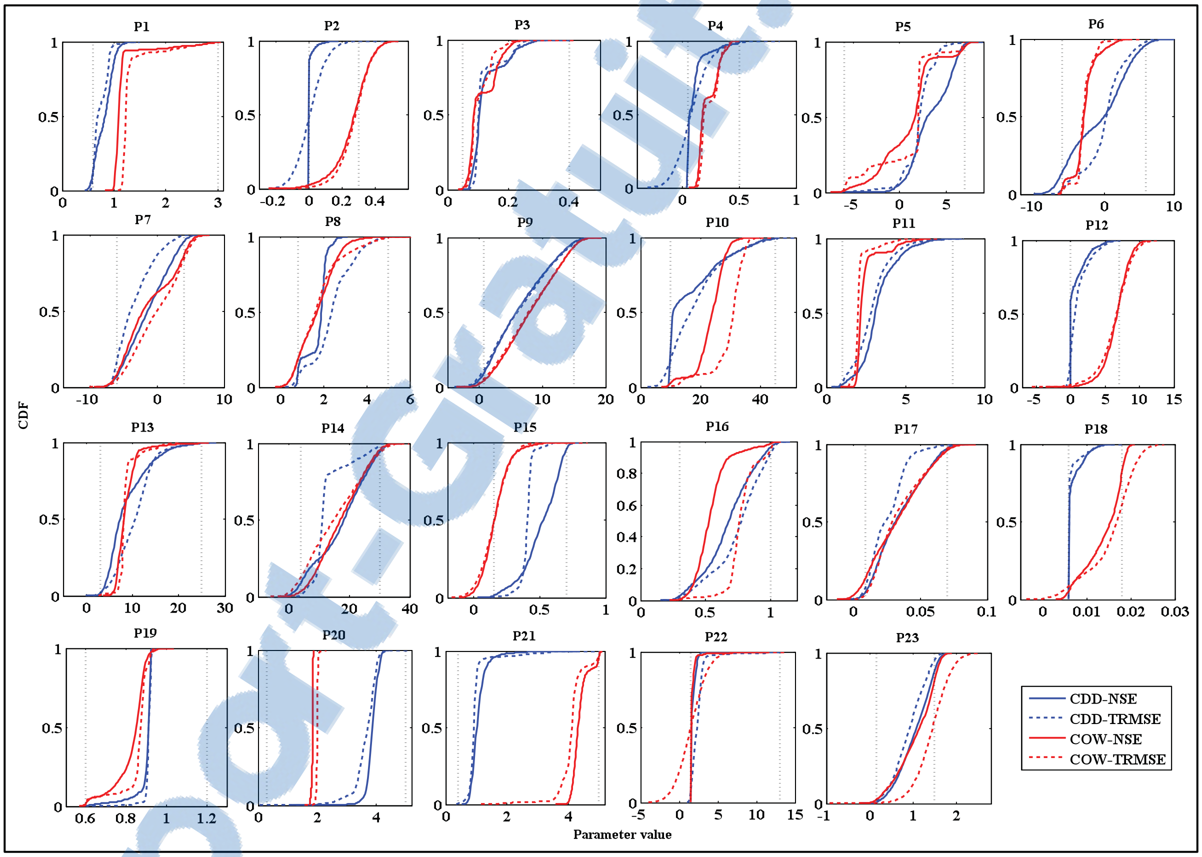

Figure-A II-2 Cumulative distribution function (CDF) of parameter values over 1000 calibrations (23 optimized parameters) using Nash Sutcliffe efficiency (NSE) and transformed root mean square error (TRMSE) as efficiency metrics for both Chute-du-Diable (CDD) and Cowansville (COW) watersheds. The x-axis covers the full boundary range while the dash lines show the upper and lower boundaries of the reduced search space for each parameter

Model performance was very good for both watersheds, as indicated by the NSE and TRMSE values in Table-A II-2.

Table-A II-2 Median, maximum (Max) and minimum (Min) values of the Nash-Sutcliffe efficiency (NSE) and transformed root mean squared error (TRMSE) for one thousand calibrations and validations of HSAMI with all 23 free parameters for two watersheds. Optimum NSE and TRMSE values are 1 and 0, respectively

Watershed	Source	Period	NSE			TRMSE		
			Median	Min	Max	Median	Min	Max
Chute-du-Diable	Calibration	1979-1988	0.89	0.87	0.90	1.52	1.49	2.42
	Validation	1989-1998	0.88	0.79	0.90	1.78	1.70	2.75
Cowansville	Calibration	1990-1999	0.75	0.63	0.76	0.65	0.64	0.98
	Validation	2000-2009	0.72	0.59	0.74	0.78	0.75	1.11

However, the problem associated with the determination of an optimal set of parameters is quite clear from results presented in Figure-A II-2. Most of the parameters had calibrated values covering the entire parameter space. In fact, the boundary values delimiting parameter space were made larger than the operational range used by Hydro-Québec to make sure that parameters were not constrained over the possible range of physical parameter values. For parameters having a physical sense, boundaries were set as large as possible to encompass the full range of realistic possible values. For empirical parameters, boundaries were fixed using good judgment and user experience with the model. Figure-A II-2 shows that despite being liberal in the boundary delineation process, most of the parameters would at one point very likely migrate outside the physically reasonable range of values if given the chance. This reflects on the problems of parameter identifiability and uniqueness that are the root cause of equifinality [Ebel and Loague, 2006]. In other words, Figure-A II-2 shows that the optimization process found many local minima over 1000 calibrations, whereas differences between largest and smallest NSE were 0.03 for the Chute-du-Diable watershed and 0.13 for the Cowansville watershed. Differences increased to 0.1 and 0.15 respectively for the validation period (Table-A II-2).

II.4.2 Multi-objective approach

II.4.2.1 Multi-objective optimization

Figure-A II-3 presents the mean of all 8 Pareto fronts for both watersheds and using both efficiency metrics. The circle data points in this graph define the Pareto front and each circle corresponds to the best obtained calibration result using 'n' model parameters.

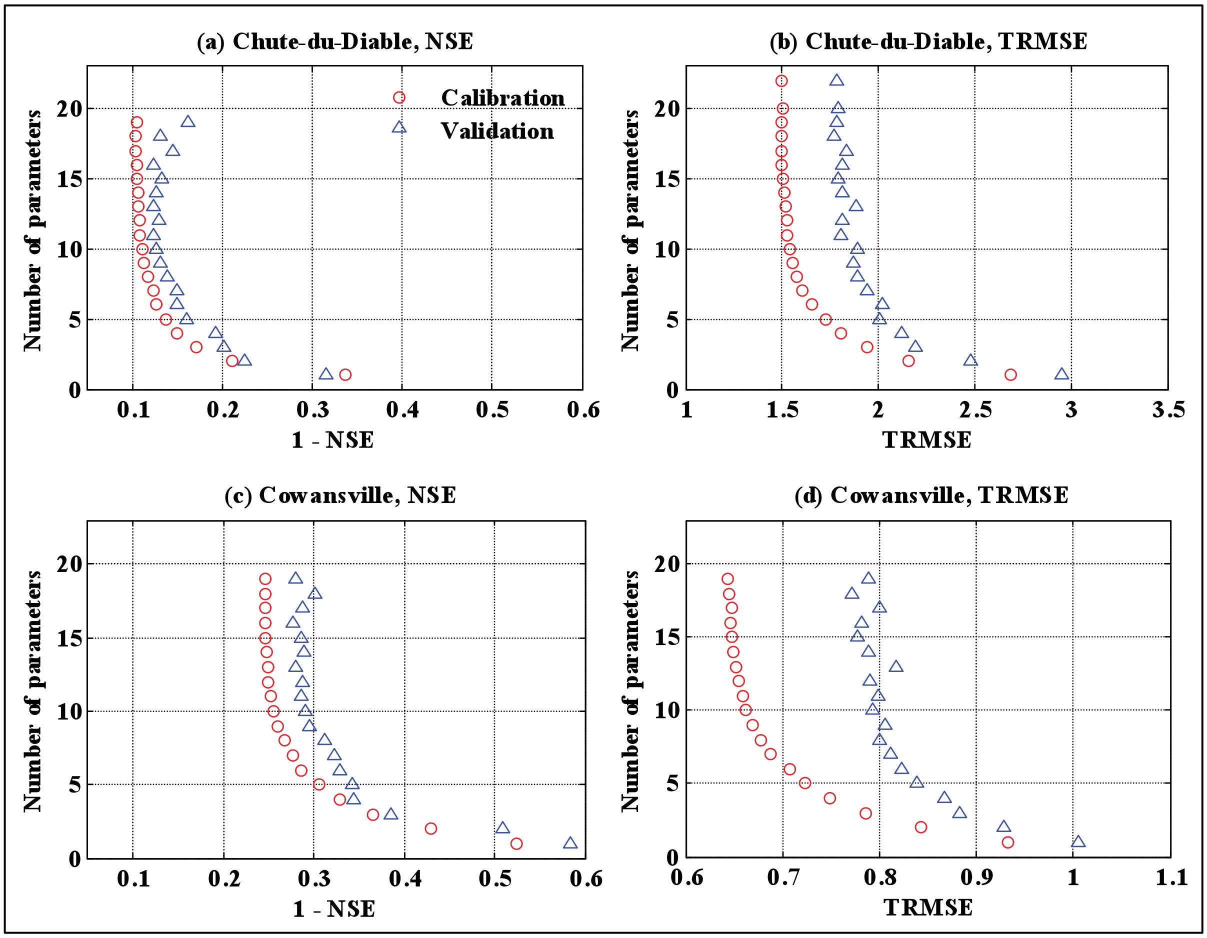

Figure-A II-3 Mean of all 8 Pareto fronts deriving from the 8 multi-objective optimizations using the NSE and TRMSE as efficiency metrics for calibration and validation periods over both Chute-du-Diable and Cowansville watersheds

The shape of the Pareto front indicates that adding additional free parameters to the model will necessarily result in an improvement in model performance. If adding a given parameter was to result in the same model performance, it would not appear on the Pareto front. This could explain why there is no point on the Pareto front with more than 19 parameters for the Chute-du-Diable watershed using NSE as the efficiency metric and for the Cowansville watershed using both efficiency metrics. However, in this case, the most likely reason for the lack of points using more than 19 parameters is the relatively low number of model iterations during the multi-objective optimization coupled with uncertainty associated with each calibration. Displaying more points (up to 23) on the Pareto front can be achieved, as increasing the iteration number. However, this may not be necessary, as the 'return on investment' of adding additional parameters rapidly diminishes (red circles in Figure-A II-3). In this case, the use of more than 10 parameters results in only very modest performance improvements. The key question is whether or not these improvements are real and not simply the result of over-fitting. This can be very easily verified by using each combination of parameters along the Pareto front over the validation period. These results are shown with the triangles in Figure-A II-3. A similar pattern is observed for the validation, with performance improvements up to 10 parameters. However, above 10 parameters, model performance is not improved anymore. This demonstrates quite clearly that any improvement observed during calibration with more than 10 parameters is due to over-fitting and not to a better representation of physical processes within the hydrological model. Any additional parameter will only increase parameter uncertainty with no additional benefit, and, in some cases, will be detrimental to model performance (see upper-left panel in Figure-A II-3).

II.4.2.2 Parameters' importance ordering

Examination of Figure-A II-3 indicates that at least 11 parameters out of 23 do not contribute much to model performance. The next step involves ordering parameters based on their importance. However, one problem arises in which the results of a choice of best parameters are not perfectly consistent from one multi-objective optimization to the other (results not shown). In other words, the best 10 parameters resulting from one multi-objective optimization will not necessarily be the same 10 parameters resulting from another. As discussed earlier, this is the consequence of the random assignation of fixed parameter values for each multi-objective optimization. To circumvent this problem, results from all multi-objective optimizations were summed up to identify the leading parameters. Table-A II-3 presents the number of times that each parameter was optimized for the Chute-du-Diable watershed using the NSE, when moving from 1 to 19 parameters for the 8 multi-objective optimizations.

Table-A II-3 Number of parameters selected as being important for the 8 MOGA optimizations using the NSE for the Chute-du-Diable watershed

PN		**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**11**	**12**	**13**	**14**	**15**	**16**	**17**	**18**	**19**	**Sum**	**Ratio**
P1	Number of times that each parameter is selected during optimization for all 8 multi-objective optimization runs	0	0	2	3	3	4	5	7	8	7	8	8	8	6	7	4	4	5	1	90	0.69
P2		0	0	0	0	0	0	0	0	0	3	1	4	5	5	6	2	2	4	1	33	0.25
P3		0	1	1	3	4	5	5	4	5	5	7	7	8	6	7	4	4	5	1	82	0.63
P4		0	0	0	0	2	4	5	5	6	6	6	6	7	5	6	3	3	4	1	69	0.53
P5		2	2	3	2	3	3	2	5	4	4	4	4	6	4	5	2	3	4	1	63	0.48
P6		0	3	2	4	4	4	5	5	6	6	6	7	6	5	6	3	4	3	1	80	0.61
P7		0	0	0	0	0	0	0	0	0	1	1	1	1	1	2	2	1	2	0	12	0.09
P8		1	1	2	3	3	2	2	5	4	6	7	7	7	5	7	4	4	5	1	76	0.58
P9		0	0	0	0	0	0	0	0	0	0	0	0	0	0	2	1	1	2	0	6	0.05
P10		0	0	0	0	1	1	1	0	0	0	0	0	0	2	3	3	1	5	1	18	0.14
P11		0	0	0	0	0	0	1	1	3	3	5	5	5	4	4	4	4	5	1	45	0.34
P12		0	0	0	0	0	0	1	1	2	3	4	2	3	3	3	3	3	5	1	34	0.26
P13		0	0	1	2	2	4	3	6	5	5	6	6	6	5	6	3	3	3	1	67	0.51
P14		2	3	1	2	3	3	3	2	2	2	1	1	2	1	1	1	2	3	0	35	0.27
P15		0	1	4	3	3	2	3	1	1	1	1	3	4	3	4	3	3	3	1	44	0.34
P16		0	0	0	0	0	1	1	1	1	2	2	2	2	1	2	2	2	4	1	24	0.18
P17		0	0	0	0	0	0	0	1	0	1	1	2	1	1	1	0	2	3	1	14	0.11
P18		0	0	1	0	1	1	1	1	2	1	3	3	6	3	3	3	3	4	1	37	0.28
P19		1	3	3	3	5	4	5	5	6	7	7	7	8	6	7	4	4	5	1	91	0.7
P20		0	0	0	0	1	5	5	6	8	8	8	8	8	6	7	4	4	5	1	84	0.64
P21		0	0	0	2	2	1	2	2	2	2	3	5	3	5	5	2	3	3	1	43	0.33
P22		2	2	4	4	3	4	6	6	6	6	6	6	6	5	6	4	4	5	1	86	0.66
P23		0	0	0	1	0	0	0	0	1	1	1	2	2	2	5	3	4	3	0	25	0.19

* Note: Green, yellow, no colour and pink denote the most important, important, average and least important, respectively; PN =the number of free parameters; RN = Number of runs using PN; Sum = summation of the number of times the parameter was present in a Pareto-optimal set.

The horizontal axis presents the 19 combination of parameters from the Pareto front (as discussed earlier, the combination of 20, 21, 22 and 23 parameters were not optimal as these combinations were no better than the one with 19-parameter model). The vertical axis represents the number of times each of the 23 parameter was selected for each combination defining the Pareto front. For example, the first column indicates that when one-parameter was selected to define the Pareto front, parameters 5, 14 and 22 were chosen twice while parameters 8 and 19 were selected once for a total of 8 corresponding to the 8 performed optimizations. The last row of each column is the summation of all chosen parameters divided by the number of parameters and should be equal to 8. If the number is less than 8, it simply indicates that one of the 8 multi-objective optimizations did not retain this number of parameter as optimal, thus indicating better performance with a lesser number of parameters. The construction of Table-A II-3 followed these 4 steps:

1- The total number of selections for each parameter was first summed in the horizontal direction. For example, parameter 1 was selected 90 times (never when the total number of parameters was 1 or 2, twice when 3 parameters were retained, etc);
2- The total number of points selected on the 8 Pareto fronts was summed up to 131 (sum of last row in Table-A II-3). If one parameter was always selected for all 8 multi-objective optimization at every combination of parameters, its summation in step 1 should be equal to 131;
3- The selection ratio was calculated for each parameter as the sum obtained in step 1 divided by 131;
4- Three thresholds were used to qualitatively classify the parameters into 4 classes of importance according to the ratios calculated in step 3. The classes were defined as *most important* (greater than or equal to 0.4), *important* (less than 0.4, while greater than or equal to 0.25), *average* (less than 0.25, while greater than or equal to 0.15), and *least important* (less than 0.15).

It is important to restate that the uncertainty in parameter identification presented in Table-A II-3 is the result of the random assignation of fixed parameter value and of equifinality itself.

For example, if two parameters are strongly correlated, a better assigned fixed value may result in the other one being selected. Technically speaking, if the most significant parameters in the hydrological model are randomly given optimal fixed values, they will not be retained by the algorithm. This is why multiple trials have to be done and also why no parameter was always selected across the board. This approach did not maximize the differences between parameters because clearly, a parameter being selected in the lower part of the Pareto front (few parameters) should have more weight than a selection in the steepest part of the front. Weighting schemes were tested but they had little effect on the identification of dominant parameters.

Table-A II-3 only presented the parameter importance for the Chute-du-Diable watershed using the NSE. However, all above procedures were applied to both watersheds and for both efficiency metrics. Table-A II-4 presents the selection ratios and classes of importance for the four considered combinations.

A first examination of Table-A II-4 reveals that 5 parameters (P1, P3, P5, P13 and P22) were consistently identified as the *most important* in all four cases. At the other end of the spectrum, parameters P7 and P9 were the *least important* for all cases. The importance of each parameter can be ordered according to the summation of selection ratios for both watersheds and efficiency metrics (the last column in Table-A II-4).

Table-A II-4 Selection ratio for parameters being important for 8 MOGA optimizations using the NSE and TRMSE for both watersheds

Watershed	**Chute-du-Diable**		**Cowansville**		**Sum**	**Order**
Metric	**NSE**	**TRMSE**	**NSE**	**TRMSE**		
P1	0.69	0.71	0.72	0.57	2.68	1
P2	0.25	0.34	0.19	0.31	1.09	16
P3	0.63	0.73	0.55	0.63	2.53	2
P4	0.53	0.67	0.7	0.32	2.22	5
P5	0.48	0.54	0.73	0.51	2.26	3
P6	0.61	0.31	0.31	0.4	1.64	12
P7	0.09	0.13	0.08	0.04	0.34	22
P8	0.58	0.39	0.52	0.38	1.87	9
P9	0.05	0.03	0.06	0.05	0.18	23
P10	0.14	0.19	0.41	0.29	1.03	17
P11	0.34	0.19	0.69	0.59	1.81	10
P12	0.26	0.35	0.52	0.67	1.8	11
P13	0.51	0.41	0.6	0.44	1.97	7
P14	0.27	0.51	0.03	0.06	0.87	18
P15	0.34	0.38	0.59	0.79	2.1	6
P16	0.18	0.12	0.3	0.06	0.67	20
P17	0.11	0.48	0.1	0.04	0.73	19
P18	0.28	0.49	0.18	0.42	1.37	14
P19	0.69	0.59	0.31	0.37	1.96	8
P20	0.64	0.26	0.19	0.44	1.53	13
P21	0.33	0.5	0.09	0.33	1.25	15
P22	0.66	0.62	0.53	0.44	2.25	4
P23	0.19	0.09	0.07	0.26	0.61	21

* Note: Green, yellow, no colour and pink denote the most important, important, average and least important, respectively.

II.4.3 Sobol' sensitivity analysis

The Sobol' sensitivity analysis method was also used to reduce the parametric dimensionality of the HSAMI model. A case study was conducted for the Chute-du-Diable watershed and using the NSE as the efficiency metric. The total-order Sobol' Indices (SI) reflect the full impact of each parameter on the model output and therefore are most relevant in model calibration [van Werkhoven et al., 2009]. Thus, the total-order SI is used to determine the parameter importance.

Table-A II-5 Total order sensitivity index (SI) and its cumulative sum of Sobol' sensitivity analysis for 23 parameters of HSAMI using the NSE for the Chute-du-Diable watershed

Parameter	Total order SI	Cumulative total order SI	Importance Order
P9	0	0	23
P7	0	0	22
P18	0.001	0.001	21
P23	0.001	0.002	20
P16	0.004	0.005	19
P14	0.005	0.01	18
P10	0.006	0.016	17
P22	0.007	0.022	16
P2	0.009	0.031	15
P17	0.013	0.044	14
P21	0.013	0.057	13
P1	0.014	0.071	12
P15	0.02	0.091	11
P20	0.022	0.113	10
P12	0.032	0.145	9
P19	0.04	0.184	8
P8	0.057	0.242	7
P4	0.078	0.32	6
P11	0.087	0.408	5
P5	0.097	0.505	4
P3	0.15	0.655	3
P6	0.16	0.815	2
P13	0.185	1	1

Table-A II-5 presents the total SI and its cumulative sum for all 23 parameters. If contributing 90% of the total output variance is used as a threshold to identify the important parameters, 10 parameters have to be kept for model optimization. Comparing these 10 parameters (the last column of Table-A II-5) from the Sobol' method with the 10 important parameters from the MOGA (the last column of Table-A II-4), it can be found that 7 parameters are identified to be commonly important by both methods, even though the importance order is not exactly the same. Especially, 8 parameters are identified to be commonly important for the same watershed and metric (column #2 of Table-A II-4 versus the last column of Table-A II-5).

To further understand the ability of Sobol' sensitivity analysis with respect to reducing the parametric dimensionality, the trade-off between the number of parameters and the NSE is presented for the Chute-du-Diable watershed. Figure-A II-4 presents the mean value of 1-NSE over 100 calibrations with the number of calibrated parameters ranging from 22 to 1. Values of 1-NSE for the validation period are also presented. Similarly to the Pareto front presented in Figure-A II-3, Figure-A II-4 shows that the increase in the number of fixed parameters generally results in reduction in model performance for the calibration period.

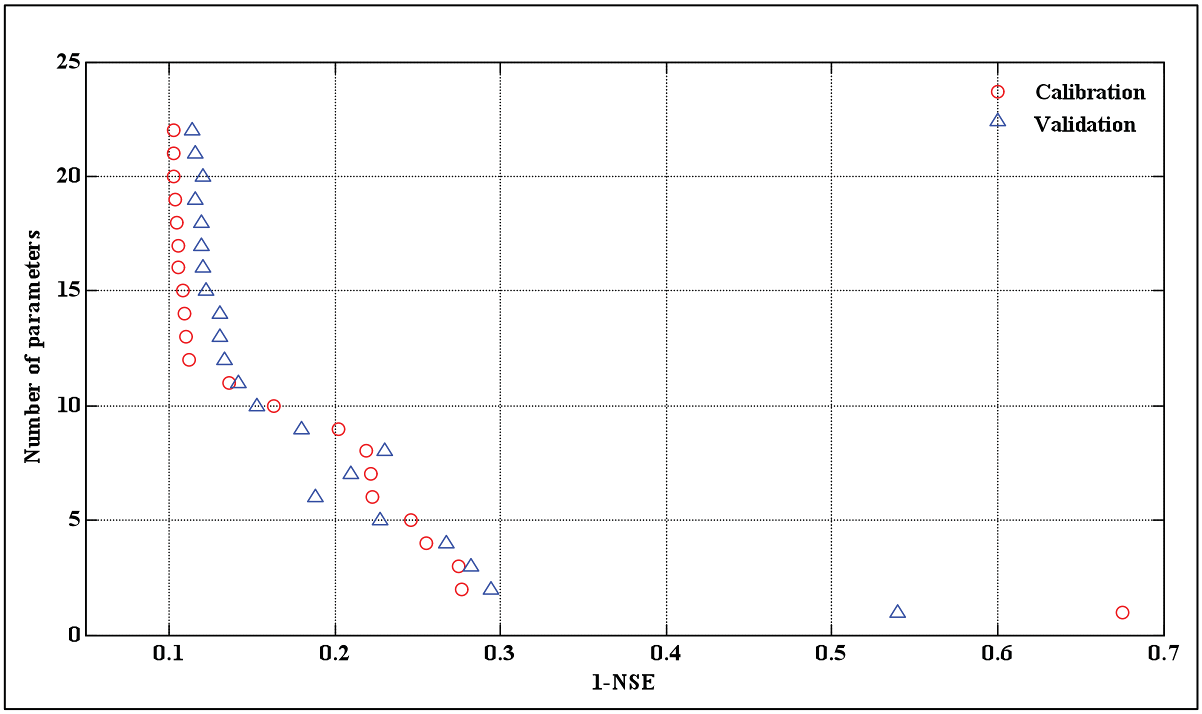

Figure-A II.4 Mean value of 1-NSE for 100 calibrations and validations with free parameters decreasing from 22 to 1. The non-calibrated parameters were fixed with random numbers. The fixed parameters were determined based on their contributions to total output variance (from the least to the most) according to Sobol' sensitivity analysis

In particular, the model performance degenerates considerably for both calibration and validation periods when the number of free parameters is less than 12. In other words, 11 parameters can be reduced with little loss in model performance. When keeping 12 parameters, MOGA and Sobol' methods show even more similarity with respect to identifying the commonly important parameters. Specifically, 11 out of 12 parameters are identified to be commonly important for the same watershed and metrics, as well as for different watersheds and metrics.

II.5 Discussion and conclusion

This study presented a benchmark for assessing the performance of sensitivity analysis in reducing the parametric dimensionality during the hydrological model calibration process. A new method was proposed based on MOGA and tested using the conceptual lumped rainfall-runoff model HSAMI with 23 free parameters over two Canadian watersheds in the Province of Quebec. The model was calibrated using two efficiency metrics: NSE and TRMSE. Results indicated that at least 11 parameters can be reduced with little degeneration in model performance for both watersheds.

The proposed method was used to validate the Sobol' sensitivity analysis method in reducing the HSAMI parameter dimensionality. Both methods indicated that a nearly equivalent model performance could be preserved when the number of HSAMI parameters was reduced from 23 to 12. In particular, 11 out of 12 parameters are identified to be commonly important, even though the relative importance of some parameters differ between methods. However, the proposed method allowed to even further reduce the number of parameters with minimal performance loss. In particular, there was little degradation in model performance for the validation period when reducing the number of parameters to 8. Sobol' method was significantly less successful at finding appropriate combinations of smaller numbers of parameters as indicated by a sharp decrease in model performance below 12 parameters. For small parameter sets, the MOGA algorithm has the advantage of being able to find independent optimal combinations at each step in the parameter reduction process, whereas

Sobol' analysis has to keep the parameters fixed in previous steps. In other words, the MOGA method is free to drop parameters fixed at an earlier step if advantageous. For example, a given parameter may be considered important for the 12-parameter version of the model, only to be dropped and replaced by two other parameters in the 11-parameter version.

Although the proposed benchmark method displayed significant benefits for a large reduction of the parameter-space, the advantages of Sobol' are numerous, such as execution time, ability to understand interactions and quantification of the variance explanation for each parameter. Of course there are some drawbacks with the Sobol' analysis, such as the definition of the boundary space. For example, if a parameter has an a priori unknown value, setting the boundaries too narrow or too wide will result in an under (over)-estimation of the parameter's importance. Therefore care must be taken in setting the parameter boundaries. It is important to note that these drawbacks are effectively present in all parameter reduction methods. Also, in a case where two parameters are strongly or perfectly correlated, the analysis method will set both parameters to equal levels of importance. However, in reality one of the parameters could be fixed without harming the model's performance. Fundamental research is ongoing to address these issues [Chastaing et al. 2014], but this work has shown that the Sobol' method's limitations should not be problematic for hydrological model parameter dimensionality reduction in the interim.

It should be pointed out that the computational cost is the major disadvantage of the benchmark method. The lumped hydrological model used in this study was highly optimized and one run of the model for a 10-year period took about one tenth of a second on a single core of a modest processor. The multi-objective optimization algorithm demands thousands of calibrations to be performed, each demanding in turn several thousand model evaluations. Almost one week of computational time (one-core equivalent on a modest processor) was required for one multi-objective optimization. Consequently, the computational burden would be very high for models that take a longer time to run. A model that takes one second per evaluation over a 10-year period would require about 70 computational days on a single core. Since complex distributed models may take a few minutes to run, the proposed approach would not be efficient without access to parallel computing power. However, access to 8 or 12 cores is relatively simple nowadays on a single desktop machine, and the proposed approach remains feasible with many hydrological models. Moreover, multi-objective optimization algorithms are very efficient at using parallel computing and access to high-performance computing is getting relatively easier, especially with the advent of Graphics Processing Unit (GPU) computing and its accompanying breakthroughs. As such, this should not be viewed as an insurmountable problem.

The need to run several multi-objective optimizations to overcome the problem of choosing values for the parameters when not used in the calibration also adds to the computational burden. A possible option to circumvent this problem is to deliberately use “bad” fixed values for all parameters. By using known “bad” values for each parameter, it is possible to avoid the pitfall of having one parameter deemed unimportant (i.e. its fixed a priori value was too good to begin with). The choice is obvious for several parameters but not as clear for parameters whose uncertainty covers the entire search space. While this approach could have allowed only one multi-objective calibration to be done (per basin and per optimization metric), the underlying need to run hundreds of calibration to determine the “bad” parameters zones within the search space partly negates the advantage of this option. It was felt more appropriate to use a random assignation to remove the possible bias in parameter value selection.

A few important considerations must be taken with respect to the methodology used in this paper. First, the MOGA optimization algorithm was setup using hyperparameters defined by users’ experiences with the algorithm and the model. Other hydrological models (depending on their complexity) could require different population/generation ratios. In this case, the 10:1 ratio was considered adequate to obtain a good exploration of the space while permitting some exploitation for refining the results. The number of generations and population size were computed according to the maximum allotted time for performing the computations. If more computing power was available, more model evaluations would have been performed. The final point to consider is the stochastic nature of model calibration. Uncertainty is also present due to initial conditions the calibration algorithm uses. The initial seed value was left variable to ensure this uncertainty is present when calibrating the models. This gives more confidence since there was no bias in “forcing” the model into a specific region due to a particularly advantageous fixed seed.

Though the approach presented herein is based on two-objective metrics, other similar yet more targeted approaches could be introduced (i.e. exploring a three-objective space). Both the NSE and TRMSE could be minimized as well as the number of parameters. This would be a more objective way of defining the importance of each parameter [Rosolem et al., 2012, 2013]. This would require more computing power, but the uncertainty would also be reduced when using the parameters in the validation space, as is the case when working on a single objective. Other approaches could use constraints to bind certain parameters together if they are known to be independent. This would reduce the search space while maintaining the same objectives.

II.6 Acknowledgments

This work was partially supported by the Natural Science and Engineering Research Council of Canada (NSERC), Hydro-Québec and the Ouranos Consortium on Regional Climatology and Adaptation to Climate Change.

II.7 References

Arsenault, R., A. Poulin, P. Côté, and F. Brissette (2014), A comparison of stochastic optimization algorithms in hydrological model calibration. J. Hydrol. Eng., 19(7), 1374-1384. doi: 10.1061/(ASCE)HE.1943-5584.0000938.

Arsenault, R., and F. P. Brissette (2014), Continuous streamflow prediction in ungauged basins: The effects of equifinality and parameter set selection on uncertainty in regionalization approaches, Water Resour. Res., 50(7), 6135–6153, doi:10.1002/2013WR014898.

Arsenault, R., J.S. Malo, F. Brissette, M. Minville, and R. Leconte (2013), Structural and non-structural climate change adaptation strategies for the Péribonka water resource system, Water Resour. Manag., 27(7), 2075-2087. doi:10.1007/s11269-013-0275-6.

Bastidas, L. A., H. Gupta, S. Sorooshian, W. J. Suttleworth, and Z. L. Yang (1999), Sensitivity analysis of a land surface scheme using multicriteria methods, J. Geophys. Res., 104 (D16), 19481–90.

Beven, K., and A. Binley (1992), The future of distributed models – model calibration and uncertainty prediction, Hydrol. Process., 6(3), 279–98.

Chastaing, G., C. Prieur, and F. Gamboa (2014) Generalized Sobol sensitivity indices for dependent variables: numerical methods. J. Statist. Comput. Simulation, Taylor & Francis: STM, Behavioural Science and Public Health Titles, 1-28.

Chen, J., F. P. Brissette, , and R. Leconte (2011b), Uncertainty of downscaling method in quantifying the impact of climate change on hydrology, J. Hydrol., 401, 190-202.

Chen, J., F. P. Brissette, A. Poulin, and R. Leconte (2011a), Overall uncertainty study of the hydrological impacts of climate change for a Canadian watershed, Water Resour. Res., doi:10.1029/2011WR010602.

Chen, J., F. P. Brissette, and R. Leconte (2012), Downscaling of weather generator parameters to quantify the hydrological impacts of climate change, Climate Research, doi: 10.3354/cr01062.

Cox, G. M., J. M. Gibbons, and A. T. A. Wood (2006), Craigon J, Ramsden SJ, Crout NMJ. Towards the systematic simplification of mechanistic models, Ecol Model., 198 (1–2), 240–6.

Deb, K. (2001). Multi-Objective Optimization using Evolutionary Algorithms, 1st edition. Coll. « Wiley-Interscience series in systems and optimization ». Chichester, England : John Wiley & Sons, Ltd, 497 p.

Deb, K., A. Pratap, S. Agarwal, and T. Meyarivan (2002), A fast and elitist multiobjective genetic algorithm: NSGA-II, IEEE T. Evolut. Comput., 6, 2, 182-197.

Ebel, B., and K. Loague (2006), Physics-based hydrologic-response simulation: Seeing through the fog of equifinality, Hydrol. Process., 20, 2887–2900.

Efstratiadis, A., and D. Koutsoyiannis (2010), One decade of multi-objective calibration approaches in hydrological modelling: a review, Hydrol. Sci. J., 55, 1, 58-78.

Fortin, V. (2000), Le modèle météo-apport HSAMI: historique, théorie et application. Institut de Recherche d'Hydro-Québec, Varennes, p. 68.

Gan Y., Q. Duan, W. Gong, C. Tong, Y. Sun, W. Chu, A. Ye, C. Miao, and Z. Di (2014), A comprehensive evaluation of various sensitivity analysis methods: A case study with a hydrological model, Environ. Model. Softw., 51, 269–285.

Gao, X., S. Sorooshian, and H. V. Gupta (1996), Sensitivity analysis of the biosphere-atmosphere transfer scheme, J. Geophys. Res., 101, 7279–7289.

Hansen, N., and A. Ostermeier (1996), Adapting arbitrary normal mutation distributions in evolution strategies: The covariance matrix adaptation, In Proceedings of the 1996 IEEE International Conference on Evolutionary Computation, 312-317.

Hansen, N., and A. Ostermeier (2001), Completely Derandomized Self-Adaptation in Evolution Strategies, Evolutionary Computation, 9(2), 159-195.

Huang, M.Y., and X. Liang (2006), On the assessment of the impact of reducing parameters and identification of parameter uncertainties for a hydrologic model with applications to ungauged basins, J. Hydrol., 320(1–2), 37–61.

Kay, A. L., H. N. Davies, V. A. Bell, and R. G. Jones (2009), Comparison of uncertainty sources for climate change impacts: flood frequency in England, Climatic Change, 92, 41–63.

Klepper, O., H. Scholten, and J. P. G. van de Kamer (1991), Prediction Uncertainty in an Ecological Model of the Oosterschelde Estuary, J. Forecasting, 10, 191–209.

Konak, A., D. W. Coit, and A. E. Smith (2006), Multi-objective optimization using genetic algorithms: A tutorial, Reliability Engineering and System Safety, 91, 992–1007.

Lee H., N. McIntyre, H. Wheater, and A. Young (2005), Selection of conceptual models for regionalisation of the rainfall-runoff relationship, J. Hydrol., 312(1-4), 125-147.

Minville, M., F. Brissette, S. Krau, and R. Leconte(2009), Adaptation to Climate Change in the Management of a Canadian Water-Resources System Exploited for Hydropower, Water Resour. Manag., 23 (14), 2965-2986, DOI: 10.1007/s11269-009-9418-1.

Minville, M., F. Brissette, and R. Leconte (2008), Uncertainty of the impact of climate change on the hydrology of a nordic watershed, J. Hydrol., 358: 70-83.

Myung, J., and M. A. Pitt (2002), When a good fit can be bad, Trends Cogn. Sci., 6, 421–425.

Nash, J. E., and W. H. Sutcliffe (1970), River flow forecasting through conceptual models: Part 1. A discussion of principles, J. Hydrol., 10, 282-290.

Nossent, J., P. Elsen, and W. Bauwens (2011), Sobol' sensitivity analysis of a complex environmental model, Environ. Model. Softw., 26 (12), 1515-1525.

Pappenberger, F., K.J. Beven, M. Ratto, and P. Matgen (2008), Multi-method global sensitivity analysis of flood inundation models, Adv. Water Resour., 31 (1), 1-14.

Pechlivanidis, I. G., B. Jackson, N. McIntyre, and H. S. Wheater (2011), Catchment scale hydrological modelling: A review of model types, calibration approaches and uncertainty analysis methods in the context of recent developments in technology and applications, Global NEST Journal, 13 (3), 193-214.

Pechlivanidis, I. G., N. McIntyre, and H. S. Wheater (2010), Calibration of the semi-distributed PDM rainfall-runoff model in the Upper Lee catchment, UK, J. Hydrol., 386 (1-4), 198-209.

Pitman, A. J. (1994), Assessing the sensitivity of a land-surface scheme to the parameter values using a single column model, J. Clim., 7(12), 1856– 1869.

Poulin, A., F. Brissette, R. Leconte, R. Arsenault, and J. S. Malo (2011), Uncertainty of hydrological modellling in climate change impact studies in a Canadian, snow-dominated river basin, J. Hydrol., 409 (3-4), 626-636.

Reed, P.M., D. Hadka, J. D. Heman, J. R. Kasprzyk, J. B. Kollat (2013), Evolutionary multiobjective optimization in water resources: The past, present, and future, Adv. Water. Resou., 51, 438-456.

Rosolem, R., H. V. Gupta, W. J. Shuttleworth, L. G. G. Gonçalves, and X. Zeng (2013), Towards a comprehensive approach to parameter estimation in land surface parameterization schemes, Hydrol. Process., doi:10.1002/hyp.9362.

Rosolem, R., H. V. Gupta, W. J. Shuttleworth, X. Zeng, and L. G. G. de Gonçalves (2012), A fully multiple-criteria implementation of the Sobol' method for parameter sensitivity analysis, J. Geophys. Res., 117(D7), D07103, doi:10.1029/2011JD016355.

Saltelli, A. (2002), Sensitivity Analysis for Importance Assessment. Risk Analysis, 22: 579–590. doi: 10.1111/0272-4332.00040

Sawaragi, Y., H. Nakayama, T. Tanino (1985), Theory of Multiobjective Optimization (vol. 176 of Mathematics in Science and Engineering). Orlando, FL: Academic Press Inc. ISBN 0126203709.

Shafii, M., and F. De Smedt (2009), Multi-objective calibration of a distributed hydrological model (WetSpa) using a genetic algorithm, Hydrol. Earth Syst. Sci., 13, 2137–2149.

Sobol’, I. M. (1993), Sensitivity estimates for nonlinear mathematical models, Math. Model. Comput. Exp., 1, 407–17.

Steuer, R. E. (1986), Multiple Criteria Optimization: Theory, Computations, and Application, New York: John Wiley & Sons, Inc. ISBN 047188846X.

Tang, Y., P. Reed, T. Wagener, and K. van Werkhoven (2007), Comparing sensitivity analysis methods to advance lumped watershed model identification and evaluation, Hydrol. Earth Syst. Sci., 11 (2), 793-817.

Tonkin, M. J., and J. Doherty (2005), A hybrid regularized inversion methodology for highly parameterized environmental models, Water Resour Res., 41:W10412. doi:10.1029/2005WR003995.

van Straten, G., and K. J. Keesman (1991), Uncertainty propagation and speculation in projective forecasts of environmental change: A lake-eutrophication example, J. Forecasting, 10, 163–190.

van Werkhoven, K., T. Wagener, P. Reed, and Y. Tang (2009), Sensitivity-guided reduction of parametric dimensionality for multi-objective calibration of watershed models, Adv. Water. Resou., 32, 1154–1169.

van Werkhoven, K., T. Wagener, P. Reed, and Y. Tang (2008), Characterization of watershed model behavior across a hydroclimatic gradient, Water Resour. Res., 44, W01429, doi:10.1029/2007WR006271.

Vrugt, J. A., H. V. Gupta, L. A. Bastidas, W. Bouten, and S. Sorooshian (2003), Effective and efficient algorithm for multiobjective optimization of hydrologic models, Water Resour. Res., 39(8), 1214, doi:10.1029/2002WR001746.

Wagener, T, and J. Kollat (2007), Numerical and visual evaluation of hydrologic and environmental models using the Monte Carlo Analysis Toolbox (MCAT), Environ. Model. Softw., 22, 1021–33.

Wagener, T., D. P. Boyle, M. J. Lees, H. S. Wheater, H. V. Gupta, and S. Sorooshian (2001), A framework for development and application of hydrological models, Hydrol. Earth Syst. Sci., 5(1), 13–26.

Wilby, R. L., and I. Harris (2006), A framework for assessing uncertainties 1118 in climate change impacts: Low-flow scenarios for the River Thames, 1119 UK, Water Resour. Res., 42, W02419, doi:10.1029/2005WR004065.

Wilson, M. F., A. Henderson-Sellers, R. E. Dickinson, and P. J. Kennedy (1987a), Sensitivity of the Biosphere-Atmosphere Transfer Scheme (BATS) to the inclusion of variable soil characteristics, J. Clim. Appl. Meteorol., 26, 341–362.

Wilson, M. F., A. Henderson-Sellers, R. E. Dickinson, and P. J.Kennedy (1987b), Investigation of the sensitivity of the land surface parameterization of the NCAR Community Climate Model in regions of tundra vegetation, Int. J. Climatol., 7, 319–343.

Wurbs, R. A. (1998), Dissemination of generalized water resources models in the United States, Water Int., 23, 190–198.

Yapo, P. O., H. V. Gupta, and S. Sorooshian (1996), Calibration of conceptual rainfall-runoff models: Sensitivity to calibration data, J. Hydrol., 181, 23–48.

Yapo, P. O., H. V. Gupta, and S. Sorooshian (1998), Multi-objective global optimization for hydrologic models, J. Hydrol., 204, 83-97.

Zhang, C., J. G. Chu, G.T. Fu (2013), Sobol's sensitivity analysis for a distributed hydrological model of Yichun River Basin, China, J. Hydrol., 480, 58-68.

ANNEXE III

LISTE DES CONTRIBUTIONS SCIENTIFIQUES

La liste complète des publications et des présentations effectuées suite à la réalisation des travaux dans cette thèse est présentée ici.

1- Articles scientifiques - journaux avec jury (publiés/acceptés)

Arsenault, R. et Brissette, F. (2014). Multi-model averaging for continuous streamflow prediction in ungauged basins. Hydrolog. Sci. J., 38p. (Soumis juillet 2014, en révision finale depuis janvier 2015)

Arsenault, R. et Brissette, F. (2014). Continuous streamflow prediction in ungauged basins: The effects of equifinality and parameter set selection on uncertainty in regionalization approaches. Water Resour. Res. (accepté le 3 juillet 2014, sous presse).

Arsenault, R., Poulin, A., Côté, P. et Brissette, F. (2014) A comparison of stochastic optimization algorithms in hydrological model calibration. J. Hydrol. Eng. 19(7), 1374-1384. Doi: 10.1061/(ASCE)HE.1943-5584.0000938.

2- Articles scientifiques - journaux avec jury (soumis)

Arsenault, R., Gatien, P., Renaud, B., Brissette, F. et Martel, JL. (2014). A comparative analysis of 9 multi-model averaging approaches in hydrological continuous streamflow prediction, J. Hydrol. 41p. (Soumis juin 2014, 2e révision depuis février 2015)

Arsenault, R. et Brissette, F. (2015). Analysis of continuous streamflow regionalization methods using a Regional Climate Model environment framework. 41p, (soumis à Hydrological Sciences Journal en février 2015)

Arsenault, R., Poissant, D. et Brissette, F. (2015). Parameter dimensionality reduction of a conceptual model for streamflow prediction in ungauged basins. 47p., (Soumis à Advances in Water Resources en février 2015)

Arsenault, R., Essou, G. et Brissette, F. (2014). Improving hydrological model simulations using four gridded climate datasets in a multi-model averaging framework. 36p. (Soumis à Journal of Hydrology en janvier 2015)

Essou, G., **Arsenault, R**. et Brissette, F. (2014) Potential use of gridded data as inputs to hydrological modeling. J. Hydrometeor., 34p. (Soumis juillet 2014)

Chen, J., **Arsenault, R**. et Brissette, F. (2014). A framework for the reduction of parametric dimensionality in hydrology models. Advances in Water Resources., 55p. (Re-soumis janvier 2015)

3- Conférences internationales/nationales/provinciales - avec jury

Arsenault, R. et Brissette, F. (2014). Étude des méthodes de prévision hydrologique aux sites non-jaugés dans le monde virtuel du MRCC 4.2.4 (15 km). 6e Symposium sur les changements climatiques d'Ouranos, Québec, Canada, (4 décembre 2014).

Arsenault, R., *Martel, JL., Brissette, F., Poulin, A. et Côté, P. (2014). Stochastic optimization algorithm selection in hydrological model calibration based on fitness landscape characterization. European Geophysical Union (EGU), Vienne, Autriche, (29 avril 2014).

4- Conférences internationales/nationales/provinciales - sans jury

Arsenault, R. et Brissette, F. (2015). Régionalisation dans le monde réel et virtuel. Deuxième rencontre annuelle du projet de recherche sur l'utilisation des données alternatives en hydrologie. Institut National en Recherche Scientifique – Centre Eau, Terre et Atmosphère, Québec, 3 février 2015.

Arsenault, R., Brissette, F. et Essou, G. (2014). Régionalisation dans le monde réel et virtuel et modélisation multi-modèle/multi-input: L'utilité des données alternatives. Première rencontre annuelle CRSNG-RDC du projet de recherche sur l'utilisation des données alternatives en hydrologie. Rio tinto Alcan, Chicoutimi, 3-5 juin 2014.

Arsenault, R., Poulin, A., Côté, P. et Brissette, F. (2013). Comparaison de la performance d'algorithmes d'optimisation stochastique en calage de modèles hydrologiques. Colloque « La recherche hydrologique au Québec dans un contexte de changements climatiques : État des lieux et perspectives », 25-26 avril 2013.

LISTE DE RÉFÉRENCES BIBLIOGRAPHIQUES

Abrahart, R. J. et L. See. 2002. « Multi-model data fusion for river flow forecasting: an evaluation of six alternative methods based on two contrasting catchments ». *Hydrology and Earth System Sciences,* 6(4), 655-670.

Abramson, M. A., C. Audet et J.E. Dennis Jr. 2004. « Generalized pattern searches with derivative information ». *Mathematical Programming*, Series B, 100, 3–25.

Adam, J. C. et D. P. Lettenmaier. 2003. « Adjustment of global gridded precipitation for systematic bias ». *Journal of Geophysical Research*, 108, 4257, doi:10.1029/2002JD002499, D9.

Ajami, N.K., Q. Duan, X. Gao et S. Sorooshian. 2006. « Multimodel combination techniques for analysis of hydrological simulations: application to Distributed Model Intercomparison Project results ». *Journal of Hydrometeorology*, 7, 755–768.

Akaike, H. 1974. « A new look at the statistical model identification ». *IEEE Transactions on Automation and Control*, 19(6), 716-723.

Anagnostopoulou, C., P. Maheras, T. Karacostas et M. Vafiadis. 2003. « Spatial and temporal analysis of dry spells in Greece ». *Theoretical and Applied Climatology* 74, 77-91.

Arsenault, R., J-S. Malo, F. Brissette, M. Minville et R. Leconte. 2013. « Structural and non-structural climate change adaptation strategies for the Péribonka water resource system ». *Water Resources Management*. 13p. DOI: 10.1007/s11269-013-0275-6.

Arsenault, R. et F. Brissette. 2014a. « Determining the Optimal Spatial Distribution of Weather Station Networks for Hydrological Modeling Purposes Using RCM Datasets: An Experimental Approach ». *Journal of Hydrometeorology*, 15, 517-526. doi: /10.1175/JHM-D-13-088.1

Arsenault, R. et F. Brissette. 2014b. « Multi-model averaging for continuous streamflow prediction in ungauged basins ». *Hydrological Sciences Journal*, Under re-review, 35p.

Arsenault, R. et F. Brissette. 2014c. « Continuous streamflow prediction in ungauged basins: The effects of equifinality and parameter set selection on uncertainty in regionalization approaches ». *Water Resources Research*, 50(7), 6135–6153, doi:10.1002/2013WR014898.

Arsenault, R., A. Poulin, P. Côté, et F. Brissette. 2014a. « Comparison of stochastic optimization algorithms in hydrological model calibration ». *Journal of Hydrologic Engineering,* 19(7), 1374-1384. Doi: 10.1061/(ASCE)HE.1943-5584.0000938.

Arsenault, R., P. Gatien, B. Renaud, F. Brissette et J-L. Martel. 2014b. « A comparative analysis of 9 multi-model averaging approaches in hydrological continuous streamflow simulation ». *Journal of Hydrology*, (Under re-review).

Audet, C. et J.E. Dennis Jr. 2006. « Mesh adaptive direct search algorithms for constrained optimization ». *SIAM Journal on Optimization*, 17 (2), 188-217.

Baillargeon, S., J. Pouliot, L-P. Rivest, V. Fortin et J. Fitzback. 2004. « Interpolation statistique multivariable de données de précipitations dans un cadre de modélisation hydrologique ». Colloque national Géomatique de l'Association canadienne des sciences géomatiques (ACSG-CIG), Montréal, 27-28 octobre.

Bao, Z., J. Zhang, J. Liu, G. Fu, G. Wang, R. He, X. Yan, J. Jin, et H. Liu. 2012. « Comparison of regionalization approaches based on regression and similarity for predictions in ungauged catchments under multiple hydro-climatic conditions ». *Journal of Hydrology*, 466–467, 37-46 doi:10.1016/j.jhydrol.2012.07.048.

Bardossy, A. 2007. « Calibration of hydrological model parameters for ungauged catchments ». *Hydrology and Earth System Sciences*, 11, 703–710.

Bastidas, L. A., H. Gupta, S. Sorooshian, W. J. Suttleworth et Z. L. Yang. 1999. « Sensitivity analysis of a land surface scheme using multicriteria methods ». *Journal of Geophysical Research*, 104 (D16), 19481–90.

Bastola, S. et D. François. 2012. « Temporal extension of meteorological records for hydrological modelling of Lake Chad Basin (Africa) using satellite rainfall data and reanalysis datasets ». Meteorological Applications, 19, 54-70.

Barrette, M., T. Wong, B. de Kelper et P. Côté. 2008. « Statistical multi-comparison of evolutionary algorithms ». *Bioinspired Optimization Methods and Applications*, October 13-14, Ljubljana, Slovenia, pp. 71-80.

Bates, J.M. et C.W.J. Granger. 1969. « The Combination of Forecasts ». *Operational Reseach Quarterly*. 20(4), 451-468.

Beauchamp, J., R. Leconte, M. Trudel et F. Brissette. 2013. « Estimation of the summer-fall PMP and PMF of a northern watershed under a changed climate ». *Water Resources Research*, 49(6), 3852-3862 DOI: 10.1002/wrcr.20336

Beven, K. et A. Binley. 1992. « The future of distributed models – model calibration and uncertainty prediction ». *Hydrological Processes*, 6(3), 279–98.

Beven, K. 2001. « Rainfall-runoff modelling – The primer ». John Wiley and Sons, Chichester, UK, 372p.

Beven, K. 2006a. « A manifesto for the equifinality thesis ». *Journal of Hydrology*, 320, 18–36.

Beven, K. 2006b. « Searching for the Holy Grail of scientific hydrology: Qt=(S, R, Δt)A as closure ». *Hydrology and Earth System Sciences*, 10, 609-618, doi:10.5194/hess-10-609-2006

Blasone, R.S., H. Madsen et D. Rosbjerg. 2007. « Parameter estimation in distributed hydrological modelling: comparison of global and local optimisation techniques ». *Nordic hydrology*, 38 (4-5), 451-476.

Bougeault, P. et Coauteurs. 2010. « The THORPEX Interactive Grand Global Ensemble ». *Bulletin of the American Meteorological Society*, 91, 1059–1072. doi:10.1175/2010BAMS2853.1

Bowler, N.E., A. Arribas et K. R. Mylne. 2008. « The Benefits of Multianalysis and Poor Man's Ensembles ». *Monthly Weather Review*, 136, 4113–4129. doi:10.1175/2008MWR2381.1

Buckland, S.T., K.P. Burnham et N.H. Augustin. 1997. « Model Selection: An Integral Part of Inference ». *Biometrics*. 53(2), 603-618.

Burn, D.H. et D.B. Boorman. 1993. « Estimation of hydrological parameters at ungauged catchments ». *Journal of Hydrology*, 143(3–4), 429-454.

Burnham, K.P. et D.R. Anderson. 2002. « Model Selection and Multi Model Inference: A Practical Information-Theoretic Approach ». Second Edition. United-States: Springer-Verlag, New-York. 487p.

Cavadias, G. et G. Morin. 1986. « The Combination of Simulated Discharges of Hydrological Models ». *Nordic Hydrology*, 17(1), 21-32.

Caya, D. et R. Laprise. 1999. « A semi-implicit semi-Lagrangian regional climate model: The Canadian RCM ». *Monthly Weather Review*, 127 (3), 341-362.

Charbonneau, R., J.-P. Fortinet G. Morin. 1977. « The CEQUEAU model: description and examples of its use in problems related to water resource management / Le modèle CEQUEAU: description et exemples d'utilisation dans le cadre de problèmes reliés à l'aménagement ». *Hydrological Sciences Bulletin*, 22(1), 193-202.

Chastaing, G., C. Prieur et F. Gamboa. 2014. « Generalized Sobol sensitivity indices for dependent variables: numerical methods ». *Journal of Statistical Computation and Simulation*, Taylor & Francis: STM, Behavioural Science and Public Health Titles, 1-28.

Chen, M. et Q. Lu. 2005. « A hybrid model based on genetic algorithm and ant colony algorithm ». *Journal of Information & Computational Science*, 2, 647-653.

Chen, J., F.P. Brissette, A. Poulin et R. Leconte. 2011a. » Uncertainty of downscaling method in quantifying the impact of climate change on hydrology ». *Journal of Hydrology*, 401(3-4), 190-202.

Chen, J., F.P. Brissette, A. Poulin et R. Leconte. 2011b. « Global uncertainty study of the hydrological impacts of climate change for a Canadian watershed ». *Water Resources Research*, 47, W12509.

Chen, J., F. P. Brissette et R. Leconte. 2012. « Downscaling of weather generator parameters to quantify the hydrological impacts of climate change ». *Climate Research*, doi: 10.3354/cr01062.

Chen, J., F. P. Brissette, D. Chaumont et M. Braun. 2013. « Finding appropriate bias correction methods in downscaling precipitation for hydrologic impact studies over North America ». *Water Resources Research*, 49, 4187-4205,
doi:10.1002/wrcr.20331.

Chen, J., R. Arsenault et F. Brissette. 2015. « Reducing the parametric dimensionality for rainfall-runoff models: a benchmark for sensitivity analysis methods ». *Advances in Water Resources*, 39p. (Manuscript under review).

Choi, H., P. F. Rasmussen et V. Fortin. 2013. « Evaluation and Comparison of Historical Gridded Data Sets of Precipitation for Canada ». *AGU Fall Meeting Abstracts*. Vol. 1.

Cox, G. M., J. M. Gibbons, A. T. A. Wood, J. Craigon, S.J. Ramsden et N.M.J. Crout. 2006. « Towards the systematic simplification of mechanistic models ». *Ecological Modelling*, 198 (1–2), 240–6.

Cressman, G. P. 1959. « An operational objective analysis system ». *Monthly Weather Review*, 87(10), 367-374.

Daly, C., R. P. Neilson et D. L. Phillips. 1994. « A statistical–topographic model for mapping climatological precipitation over mountainous terrain ». *Journal of Applied Meteorology and Climatology,* 33, 140–158.

Daly, C., G. H. Taylor et W. P. Gibson. 1997. « The PRISM approach to mapping precipitation and temperature ». *10th Conference on Applied Climatology*, Reno, NV, American Meteorological Society, 10-12.

Davolio, S., M.M. Miglietta, T. Diomede, C. Marsigli, A. Morgillo et A. Moscatello. 2008. « A meteo-hydrological prediction system based on a multi-model approach for precipitation forecasting ». *Natural Hazards and Earth System Sciences*, 8, 143–159, doi:10.5194/nhess-8-143-2008.

Deb, K. 2001. « Multi-Objective Optimization using Evolutionary Algorithms : 1st edition ». Wiley-Interscience series in systems and optimization. Chichester, United-Kingdom: John Wiley & Sons, Ltd, 497p.

Deb, K., A. Pratap, S. Agarwal et T. Meyarivan. 2002. « A fast and elitist multiobjective genetic algorithm: NSGA-II ». *IEEE Transactions on Evolutionary Computation*, 6(2), 182-197.

DiCiccio, T.J. et J.P. Romano. 1995. « On bootstrap procedures for second-order accurate confidence limits in parametric models ». *Statistica Sinica*, 5, 141-160.

Diks, C.G.H. et J.A. Vrugt. 2010. « Comparison of point forecast accuracy of model averaging methods in hydrologic applications ». *Stochastic Environmental Research and Risk Assessment*, 24(6), 809-820.

Dirks, K. N., J. E. Hay, C. D. Stow et D. Harris. 1998. « High-resolution studies of rainfall on norfolk island. part ii : The interpolation of high-spatial-resolution rainfall data on norfolk island ». *Journal of Hydrology*, 208,187-193.

Duan, Q., S. Sorooshian et V.K. Gupta. 1992. « Effective and efficient global optimization for conceptual rainfall runoff models ». *Water Resources Research*, 24(7), 1163-1173.

Duan, Q., S. Sorooshian et V.K. Gupta. 1993. « A shuffled complex evolution approach for effective and efficient optimization ». *Journal of Optimization Theory and Applications*, 76(3), 501-521.

Duan, Q., S. Sorooshian et V.K. Gupta. 1994. « Optimal use of the SCE-UA global optimization method for calibrating watershed models ». *Journal of Hydrology,* 158, 265-284.

Duan, Q., J. Schaake, V. Andreassian, S. Franks, H.V. Gupta, Y.M. Gusev, F. Habets, A. Hall, L. Hay, T.S. Hogue, M. Huang, G. Leavesley, X. Liang, O.N. Nasonova, J. Noilhan, L. Oudin, S. Sorooshian, T. Wagener et E.F. Wood. 2006. « Model parameter estimation experiment (MOPEX): Overview and summary of the second and third workshop results ». *Journal of Hydrology*, 320, 3-17.

Ebel, B. et K. Loague. 2006. « Physics-based hydrologic-response simulation: Seeing through the fog of equifinality ». *Hydrological Processes*, 20, 2887–2900.

Ebtehaj, M., H. Moradkhani et H.V. Gupta. 2010. « Improving robustness of hydrologic parameter estimation by the use of moving block bootstrap resampling ». *Water Resources Research*, 46, W07515, doi:10.1029/2009WR007981.

Efron, B. 1979. « Bootstrap methods: Another look at jackknife ». *Annals of Mathematical Statistics*, 7, 1–26.

Efron, B. 1987. « Better bootstrap confidence intervals ». *Journal of the American Statistical Association*, 82(397), 171–200.

Efstratiadis, A. et D. Koutsoyiannis. 2010. « One decade of multi-objective calibration approaches in hydrological modelling: a review ». *Hydrological Sciences Journal*, 55, 1, 58-78.

Essou, G.R.C., R. Arsenault et F. Brissette. 2014. « Potential of gridded data as inputs to hydrological modeling ». *Journal of Hydrometeorology,* (Under Review)

Fang, L., P. Chen et L. Shihua. 2007. « Particle swarm optimization with simulated annealing for TSP ». *Proceedings of the 6th Conference on 6th WSEAS Int. Conf. on Artificial Intelligence, Knowledge Engineering and Data Bases* - Volume 6. Corfu Island, Greece, World Scientific and Engineering Academy and Society (WSEAS), 206-210.

Fortin, V. 2000. « Le modèle météo-apport HSAMI: historique, théorie et application ». Varennes: Institut de Recherche d'Hydro-Québec, 68p.

Fortin, V. et R. Turcotte. 2007. « Le modèle hydrologique MOHYSE (bases théoriques et manuel de l'usager) ». Notes de cours pour SCA7420, Département des sciences de la terre et de l'atmosphère, Université du Québec à Montréal, Montréal, Canada, 17p.

Fortin, J.-P., R. Turcotte, S. Massicotte, R. Moussa, J. Fitzback et J.-P. Villeneuve. 2001. « Distributed watershed model compatible with remote sensing and GIS data. 1: Description of the model ». *Journal of Hydraulic Engineering*, 6(2), 91-99.

Fortnow, L. 2009. « The status of the P versus NP problem ». *Communications of the ACM*, 52(9), 78–86. doi:10.1145/1562164.1562186

Franchini, M., G. Galeati et S. Berra. 1998. « Global optimization techniques for the calibration of conceptual rainfall-runoff models ». *Hydrological Sciences Journal*, 43(3), 443-458.

Friedman, M. 1937. « The use of ranks to avoid the assumption of normality implicit in the analysis of variance ». *Journal of the American Statistical Association*, 32, 675–701.

Friedman, M. 1940. « A comparison of alternative tests of significance for the problem of m rankings ». *Annals of Mathematical Statistics*, 11, 86–92.

Fulton, R. A., J. P. Breidenbach, D.-J. Seo, D. A. Miller et T. O'Bannon. 1998. « The WSR-88D rainfall algorithm ». *Weather Forecasting*, 13 (2), 377–395.

Gan Y., Q. Duan, W. Gong, C. Tong, Y. Sun, W. Chu, A. Ye, C. Miao et Z. Di. 2014. « A comprehensive evaluation of various sensitivity analysis methods: A case study with a hydrological model ». *Environmental Modelling and Software*, 51, 269–285.

Gao, X., S. Sorooshian et H. V. Gupta. 1996. « Sensitivity analysis of the biosphere-atmosphere transfer scheme ». *Journal of Geophysical Research*, 101, 7279–7289.

Geem, Z.W., J.H. Kim et G.V. Loganathan. 2001. « A new heuristic optimization algorithm: Harmony search ». *Simulation*, 76(2), 60-68.

Gneiting, T., A.E. Raftery et A.H. Westveld. 2005. « Calibrated Probabilistic Forecasting Using Ensemble Model Output Statistic and Minimum CRPS Estimation ». *Monthly Weather Review,* 133, 1098-1118.

Goldberg, D.E. 1989. « Genetic algorithms in search, optimization & machine learning ». Addison-Wesley, Boston, 432p.

Goodison, B. E., H. L. Ferguson et G. A. McKay. 1981. « Measurement and data analysis ». In: Gray, D.M., and Male, D.H., eds. Handbook of snow : Principles, Processes, Management and Use, 191–274.

Goodison, B. E., P. Y. T. Louie et D. Yang. 1998. « WMO solid precipitation measurement intercomparison ». *Instruments and Observing Methods* Rep. 67 (WMO/TD 872), World Meteorological Organization, Geneva, Switzerland, 212 pp.

Goswami, M., K.M. O'Connor et K.P. Bhattarai. 2007. « Development of regionalisation procedures using a multi-model approach for flow simulation in an ungauged catchment ». *Journal of Hydrology*, 333 (2–4), 517-531.

Granger, C.W. et P. Newbold. 1977. « Forecasting economic time series ». First Edition. New-York, United-States : Academic Press, 333p.

Granger, C.W.J. et R. Ramanathan. 1984. « Improved methods of combining forecasts ». *Journal of Forecasting,* 3(2), 197-204.

Hansen, N. et A. Ostermeier. 1996. « Adapting arbitrary normal mutation distributions in evolution strategies: The covariance matrix adaptation ». In *Proceedings of the 1996 IEEE International Conference on Evolutionary Computation*, 312-317.

Hansen, N. et A. Ostermeier. 2001. « Completely derandomized self-adaptation in evolution strategies ». *Evolutionary Computation*, 9(2), 159-195.

Hansen, B.E. 2008. « Least-squares forecast averaging ». *Journal of Econometrics*, 146(2), 342–350.

He, Y., A. Bárdossy et E. Zehe. 2011. « A review of regionalisation for continuous streamflow simulation ». *Hydrology and Earth System Sciences*, 15, 3539-3553, doi:10.5194/hess-15-3539-2011

Higgins, R. W., J. E. Janowiak et Y. Yao. 1996. « A gridded hourly precipitation data base for the United States (1963-1993) ». NCEP/Climate Prediction Center ATLAS No. 1., 46 p.

Higgins, R. W., W. Shi, E. Yarosh et R. Joyce. 2000. « Improved United States precipitation quality control system and analysis ». NCEP/Climate Prediction Center ATLAS N°6.

Hochberg, Y. 1988. « A sharper Bonferroni procedure for multiple tests of significance ». *Biometrika*, 75(4), 800-802.

Hoeting, J.A., D. Madigan et A.E. Raftery. 1999. « Bayesian Model Averaging: A Tutorial ». *Statistical Science*, 14(4), 382-401.

Holland, J. 1975. « Adaptation in natural and artificial systems ». University of Michigan Press, Oxford, 183p.

Hrachowitz, M., H.H.G. Savenije, G. Blöschl, J.J. McDonnell, M. Sivapalan, J.W. Pomeroy, B. Arheimer, T. Blume, M.P. Clark, U. Ehret, F. Fenicia, J.E. Freer, A. Gelfan, H.V. Gupta, D.A. Hughes, R.W. Hut, A. Montanari, S. Pande, D. Tetzlaff, P.A. Troch, S. Uhlenbrook, T. Wagener, H.C. Winsemius, R.A. Woods, E. Zehe et C. Cudennec. 2013. « A decade of Predictions in Ungauged Basins (PUB) – a review ». *Hydrological Sciences Journal*, 58 (6), 1198–1255.

Hu, T.S., K.C. Lam et S.T. Ng. 2001. « River flow time series prediction with a range-dependent neural network ». *Hydrological Sciences Journal*, 46, 729–745.

Huang, M.Y. et X. Liang. 2006. « On the assessment of the impact of reducing parameters and identification of parameter uncertainties for a hydrologic model with applications to ungauged basins ». *Journal of Hydrology*, 320(1–2), 37–61.

Huffman, G. J., R. F. Adler, M. Morrissey, D. Bolvin, S. Curtis, R. Joyce, B. McGavock et J. Susskind. 2001. « Global precipitation at one-degree daily resolution from multisatellite observations ». *Journal of Hydrometeorology*, 2, 36-50.

Hulme, M. 1992. « A 1951-80 global land precipitation climatology for the evaluation of General Circulation Model ». *Climate Dynamics,* 7, 57-72.

Hundecha, Y. et A. Bárdossy. 2005. « Trends in daily precipitation and temperature extremes across western Germany in the second half of the 20th century ». *International Journal of Climatology,* 25, 1189-1202.

Hutchinson, M. F., D.W. McKenney, K. Lawrence, J. H. Pedlar, R. F. Hopkinson, E. Milewska, et P. Papadopol. 2009. « Development and testing of Canada-wide interpolated spatial models of daily minimum-maximum temperature and precipitation for 1961-2003 ». *Journal of Applied Meteorology and Climatology*, 48, 725-741.

Ingber, L. 1989. « Very fast simulated re-annealing. » *Mathematical Computer Modelling*, 12, 967-973.

Ingber, L. 1993. « Adaptive simulated annealing (ASA) ». McLean, VA, Lester Ingber Research.

Ingber, L. 1996. « Adaptive simulated annealing (ASA): Lessons learned ». *Control and Cybernetics*, 25, 33-54.

Jefferys, W.H. et J.O. Berger. 1992. « Ockham's Razor and Bayesian Analysis ». *American Scientist*, 80(1), 64-72.

Jones, D. A. et B. Trewin. 2000. « The spatial structure of monthly temperature anomalies over Australia ». *Australian Meteorological Magazine*, 49, 261–276.

Kay, A.L., H.N. Davies, V.A. Bell et R.G. Jones. 2009. « Comparison of uncertainty sources for climate change impacts: flood frequency in England ». *Climatic Change*, 92, 41–63

Kennedy, J. et R. Eberhart. 1995. « Particle swarm optimization ». *Proceedings of IEEE International Conference on Neural Networks*, 4. 1942–1948. doi:10.1109/ICNN.1995.488968

Kirkpatrick, S., C.D. Gelatt et M.P. Vecchi. 1983. « Optimization by simulated annealing ». *Science*, 220(4598), 671–680. doi:10.1126/science.220.4598.671

Klepper, O., H. Scholten et J. P. G. van de Kamer. 1991. « Prediction Uncertainty in an Ecological Model of the Oosterschelde Estuary ». *Journal of Forecasting*, 10, 191–209.

Kling, H. et H.V. Gupta. 2009. « On the development of regionalization relationships for lumped watershed models: the impact of ignoring sub-basin scale variability ». *Journal of Hydrology*, 373, 337–351.

Konak, A., D. W. Coit et A. E. Smith. 2006. « Multi-objective optimization using genetic algorithms: A tutorial ». *Reliability Engineering and System Safety*, 91, 992–1007.

Kottek, M., J. Grieser, C. Beck, B. Rudolf et F. Rubel. 2006. « World Map of the Köppen-Geiger climate classification updated ». *Meteorologische Zeitschrift* , 15, 259-263. DOI: 10.1127/0941-2948/2006/0130.

Krajewski, F.W., V. Lakshmi, K. P. Georgakakos et C.J. Subhash. 1991. « A Monte Carlo study of rainfall sampling effect on a distributed catchment model ». *Water Resources Research*, 27, 119-128.

Krasnopolsky, V. M. et Lin, Y. 2012. « A Neural Network Nonlinear Multimodel Ensemble to Improve Precipitation Forecasts over Continental US ». *Advances in Meteorology*, Article ID 649450, 11p. doi:10.1155/2012/649450

Kruskal, W. H. et W.A. Wallis. 1952. « Use of ranks in one-criterion variance analysis ». *Journal of the American Statistical Association,* 47, 583-621.

Lee, H., N.R. McIntyre, H. Wheater et A. Young. 2005. « Selection of conceptual models for regionalisation of the rainfall-runoff relationship ». *Journal of Hydrology*, 312(1–4), 125-147. http://dx.doi.org/10.1016/j.jhydrol.2005.02.016.

Lee, H., N.R. McIntyre, H. Wheater et A. Young. 2006. « Predicting runoff in ungauged UK catchments ». Proceedings of the Institution of Civil Engineers. *Water Management*, 159(2), 129–138.

Li, H., Y. Zhang, F.H.S. Chiew et X. Shiguo. 2009. « Predicting runoff in ungauged catchments by using Xinanjiang model with MODIS leaf area index ». Journal of Hydrology, 370(1–4), 155–162.

Li, X., D.E. Weller, et T.E. Jordan. 2010. « Watershed model calibration using multi-objective optimization and multi-site averaging ». Journal of Hydrology, 380(3–4), 277–288.

Liu, Y. et H. V. Gupta. 2007. « Uncertainty in hydrologic modeling: Toward an integrated data assimilation framework ». *Water Resources Research*, 43, W07401, doi:10.1029/2006WR005756.

Lopes, V.L. 1996. « On the effect of uncertainty in spatial distribution of rainfall on catchment modelling ». *Catena,* 28, 107-119.

Lucas-Picher, P., P. Riboust, S. Somot et R. Laprise. 2015. « Reconstruction of the Spring 2011 Richelieu River Flood by Two Regional Climate Models and a Hydrological Model ». *Journal of Hydrometeorology,* 16, 36–54. doi: http://dx.doi.org/10.1175/JHM-D-14-0116.1

Lunacek, M. et D. Whitley. 2006. « The dispersion metric and the CMA evolution strategy ». *Proceedings of the 8th annual conference on Genetic and evolutionary computation*, July 08-12, 2006, Seattle, USA. doi:10.1145/1143997.1144085

Maraun, D. 2012. « Nonstationarities of regional climate model biases in European seasonal mean temperature and precipitation sums ». *Geophysical Research Letters*, 39, L06706, doi:10.1029/2012GL051210.

Mariani, V.C., L.G.J. Luvizotto, F.A. Guerraet D.S.C. Leandro. 2011. « A hybrid shuffled complex evolution approach based on differential evolution for unconstrained optimization ». *Applied Mathematics and Computation*, 217(12), 5822-5829.

Maurer, E. P., A. W. Wood, J. C. Adam, D. P. Lettenmaier et B. Nijssen. 2002. « A Long-Term Hydrologically-Based Data Set of Land Surface Fluxes and States for the Conterminous United States ». *Journal of Climate,* 15, 3237-3251.

McCuen, R. H. et B.S. Levy. 2000. « Evaluation of peak discharge transposition ». *Journal of Hydrologic Engineering*, 5 (3), 278–289.

McIntyre, N., H. Lee, H. Wheater, A. Young et T. Wagener. 2005. « Ensemble predictions of runoff in ungauged catchments ». *Water Resources Research*, 41, W12434, doi:10.1029/2005WR004289.

McLachlan, G.J. 1992. « Discriminant analysis and statistical pattern recognition ».Wiley Series in Probability and Mathematical Statistics: Applied Probability and Statistics. Wiley-Interscience. John Wiley & Sons, Inc., New York. 526 pp. ISBN: 0-471-61531-5

Merz, R. et G. Blöschl. 2004. « Regionalization of catchment model parameters ». *Journal of Hydrology*, 287, 95–123.

Minville, M., F. Brissette et Leconte, R. 2008. « Uncertainty of the impact of climate change on the hydrology of a Nordic watershed ». *Journal of Hydrology*, 358(1-2): 70-83

Minville, M., F. Brissette, S. Krau, et R. Leconte. 2009. « Adaptation to climate change in the management of a Canadian water-resources system ». *Water Resources Management*, 23(14), 2965-2986.

Minville, M., S. Krau, F. Brissette et R. Leconte. 2010. « Behaviour and performance of a water resource system in Québec (Canada) under adapted operating policies in a climate change context ». *Water Resources Management*, 24, 1333–1352.

Minville, M., D. Cartier, C. Guay, L.-A. Leclaire, C. Audet, S. Le Digabel et J. Merleau. 2014. « Improving process representation in conceptual hydrological model calibration using climate simulations ». *Water Resources Research*, 50, 5044–5073, doi:10.1002/2013WR013857.

Mitchell T. D. et P. D. Jones. 2005. « An improved method of constructing a database of monthly climate observations and associated high-resolution grids ». *International Journal of Climatology,* 25(6): 693–712.

Mizukami, N. et M. B. Smith. 2012. « Analysis of inconsistencies in multi-year gridded quantitative precipitation estimate over complex terrain and its impact on hydrologic modeling ». *Journal of Hydrology,* 428, 129–141.

Mohamoud, Y. M. 2008. « Prediction of daily flow duration curves and streamflow for ungauged catchments using regional flow duration curves ». *Hydrological Sciences Journal*, 53(4), 706–724.

Moradkhani, H. et S. Sorooshian. 2009. « General review of rainfall-runoff modeling: model calibration, data assimilation and uncertainty analysis ». In Hydrological Modelling and the Water Cycle: Coupling the Atmospheric and Hydrological Models, Sorooshian, S., Hsu, K.-l., Coppola, E., Tomassetti, B., Verdecchia, M., Visconti, G. (Eds.), Springer, 1-24.

Moriasi D.N., J.G. Arnold, M.W. Van Liew, R.L. Bingner, R.D. Harmel et T.L. Veith. 2007. « Model evaluation guidelines for systematic quantification of accuracy in watershed simulations ». Transactions of the American Society of Agricultural Engineers, 50(3), 885-900.

Muñoz, E., C. Álvarez, M. Billib, J. L. Arumí et D. Rivera. 2011. « Comparison of gridded and measured rainfall data for basin-scale hydrological studies ». *Chilean Journal of Agricultural Research,* 71(3), 459-468.

Murray S. J., I. M. Watson et I. C. Prentice. 2013. « The use of dynamic global vegetation models for simulating hydrology and the potential integration of satellite observations ». *Progress in Physical Geography*, 37(1) 63–97.

Music, B. et D. Caya. 2007. « Evaluation of the hydrological cycle over the Mississippi River basin as simulated by the Canadian RCM (CRCM) ». *Journal of Hydrometeorology*, 8, 969–988.

Music, B. et D. Caya. 2009. « Investigation of the sensitivity of the water cycle components simulated by the Canadian Regional Climate Model (CRCM) to the land surface parameterization, the lateral boundary data and the internal variability ». *Journal of Hydrometeorology*, 10, 3–21.

Music, B., A. Frigon, M. Slivitzky, A. Musy, D. Caya et R. Roy. 2009. « Runoff modelling within the Canadian Regional Climate Model (CRCM): analysis over the Quebec/Labrador watersheds ». In: New Approaches to Hydrological Prediction in Data Sparse Regions (Proc. of Symposium HS.2 at the Joint IAHS & IAH Convention, Hyderabad, India, September 2009). *International Association of Hydrological Sciences (IAHS) Red Book Series Publ*. 333, 183-194.

Mylne, K.R., R.E. Evans et R.T. Clark. 2002. « Multi-model multi-analysis ensembles in quasi-operational medium-range forecasting ». *Quarterly Journal of the Royal Meteorological Society*, 128, 361–384. doi: 10.1256/00359000260498923

Myung, J. et M. A. Pitt. 2002. « When a good fit can be bad ». *Trends in Cognitive Sciences*, 6, 421–425.

Nash, J. E. et J.V. Sutcliffe. 1970. « River flow forecasting through conceptual models part I — A discussion of principles ». *Journal of Hydrology*, 10 (3), 282–290.

National Climatic Data Center. 1994. « Surface Land daily Cooperative. Summary of the Day (TD-3200) ». *National Environmental Satellite and Data Information Service*, NOAA, U.S. Department of Commerce.

Neitsch, S.L., J.G. Arnold, J.R. Kiniry, J.R. Williams et K.W. King. 2002. « Soil water assessment tool theoretical documentation ». Grassland, Soil and Water Research Laboratory, Temple, Texas. GSWRL Report 02-01.

Neuman, S.P. 2003. « Maximum likelihood Bayesian averaging of uncertain model predictions ». *Stochastic Environmental Research and Risk Assessment*, 17(5), 291-305.

Nossent, J., P. Elsen et W. Bauwens. 2011. « Sobol' sensitivity analysis of a complex environmental model ». Environmental Modelling and Software, 26 (12) , 1515–1525.

Nossent, J. et W. Bauwens. 2012. « Optimising the convergence of a Sobol' sensitivity analysis for an environmental model: application of an appropriate estimate for the square of the expectation value and the total variance ». In *Proc. of the International Environmental Modelling and Software Society (iEMSs)*, 2012 International Congress on Environmental Modelling and Software, Managing Resources of a Limited Planet: Pathways and Visions under Uncertainty, Sixth Biennial Meeting, Leipzig, Germany, 1 - 5 July 2012.

Obled, C., J. Wendling et K. Beven. 1994. « The sensitivity of hydrological models to spatial rainfall patterns : an evaluation using observed data ». *Journal of Hydrology*, 159, 305-333.

Oudin, L., F. Hervieu, C. Michel, C. Perrin, V. Andréassian, F. Anctil et C. Loumagne. 2005. « Which potential evapotranspiration input for a rainfall-runoff model? Part 2 – Towards a simple and efficient PE model for rainfall-runoff modelling ». *Journal of Hydrology*, 303(1-4), 290-306, DOI: 10.1016/j.jhydrol.2004.08.026.

Oudin, L., V. Andréassian, C. Perrin, C. Michel et N. Le Moine. 2008. « Spatial proximity, physical similarity, regression and ungaged catchments: A comparison of regionalization approaches based on 913 French catchments ». *Water Resources Research*, 44, W03413. doi:10.1029/2007WR006240.

Oudin, L., A. Kay, V. Andréassian et C. Perrin. 2010. « Are seemingly physically similar catchments truly hydrologically similar? ». *Water Resources Research*, 46, W11558, doi:10.1029/2009WR008887.

Pappenberger, F., K.J. Beven, M. Ratto et P. Matgen. 2008. « Multi-method global sensitivity analysis of flood inundation models ». *Advances in Water Resources*, 31 (1), 1-14.

Parajka, J., R. Merz et G. Blöschl. 2005. « A comparison of regionalisation methods for catchment model parameters ». *Hydrology and Earth System Sciences*, 9, 157–171.

Parajka, J., G. Blöschl et R. Merz. 2007. « Regional calibration of catchment models: Potential for ungauged catchments ». *Water Resources Research*, 43, W06406, doi:10.1029/2006WR005271.

Parajka, J., A. Viglione, M. Rogger, J. L. Salinas, M. Sivapalan et G. Blöschl. 2013. « Comparative assessment of predictions in ungauged basins – Part 1: Runoff hydrograph studies ». *Hydrology and Earth System Sciences*, 10, 375-409, doi:10.5194/hessd-10-375-2013.

Pechlivanidis, I. G., N. McIntyre et H.S. Wheater. 2010. « Calibration of the semi-distributed PDM rainfall-runoff model in the Upper Lee catchment, UK ». *Journal of Hydrology*, 386 (1-4), 198-209.

Pechlivanidis, I. G., B. Jackson, N. McIntyre et H.S. Wheater. 2011. « Catchment scale hydrological modelling: A review of model types, calibration approaches and uncertainty analysis methods in the context of recent developments in technology and applications ». *Global NEST Journal*, 13 (3), 193-214.

Pedersen, M.E.H. 2010. « Good parameters for differential evolution ». Technical Report HL1002, Hvass Laboratories.

Peel, M.C. et G. Blöschl. 2011. « Hydrological modelling in a changing world ». *Progress in Physical Geography*, 35(2), 249-261.

Perrin, C., C. Michel et V. Andréassian. 2003. « Improvement of a parsimounious model for streamflow simulation ». *Journal of Hydrology*, 279(1-4), 275-289.

Pitman, A. J. 1994. « Assessing the sensitivity of a land-surface scheme to the parameter values using a single column model ». *Journal of Climate*, 7(12), 1856– 1869.

Poulin, A., F. Brissette, R. Leconte, R. Arsenault et J.S. Malo. 2011. « Uncertainty of hydrological modelling in climate change impact studies in a Canadian, snow-dominated river basin ». *Journal of Hydrology*, 409(3-4), 626-636.

Raftery, A.E. 1993. « Change point and change curve modeling in stochastic processes and spatial statistics ». *Journal of Applied Statistical Science*, vol.1, no°4, p. 403-424.

Raftery, A.E. et Y. Zheng. 2003. « Discussion: Performance of Bayesian Model Averaging ». *Journal of the American Statistical Association,* 98(464), 931-938.

Raftery A.E., T. Gneiting et F. Bakabdaoui. 2005. « Using Bayesian Model Averaging to Calibrate Forecast Ensembles ». *Monthly Weather Review,* 133(5), 1155-1174.

Rakotomalala, R. 2008. « Comparaison de populations-Tests non paramétriques ». Université Lumière Lyon 2. 201pp.

Rasmussen, R., M. Dixon, S. Vasiloff, F. Hage, S. Knight, J. Vivekanandan et M. Xu. 2003. « Snow nowcasting using a real-time correlation of radar reflectivity with snow gauge accumulation ». *Journal of Applied Meteorology and Climatology,* 42 (1), 20–36.

Razavi, T. et P. Coulibaly. 2013. « Streamflow prediction in ungauged basins: Review of regionalization methods ». *Journal of Hydrologic Engineering*, 18(8), 958–975.

Reed, P.M., D. Hadka, J.D. Heman, J.R. Kasprzyk et J.B. Kollat. 2013. « Evolutionary multiobjective optimization in water resources: The past, present, and future ». *Advances in Water Resources*, 51, 438-456.

Reichl, J. P. C., A. W. Western, N. R. McIntyre et F. H. S. Chiew. 2009. « Optimization of a similarity measure for estimating ungauged streamflow ». *Water Resources Research*, 45, W10423, doi:10.1029/2008WR007248.

Ricard, S., R. Bourdillon, D. Roussel et R. Turcotte. 2013. « Global Calibration of Distributed Hydrological Models for Large-Scale Applications ». *Journal of Hydrologic Engineering*, 18(6), 719–721.

Ronkkonen, J., S. Kukkonen et K.V. Price. 2005. « Real-parameter optimization with differential evolution ». *Proceedings of the IEEE Congress in Evolutionary Computation*, 506 -513.

Rosolem, R., H. V. Gupta, W. J. Shuttleworth, X. Zeng et L. G. G. de Gonçalves. 2012. « A fully multiple-criteria implementation of the Sobol' method for parameter sensitivity analysis ». *Journal of Geophysical Research*, 117(D7), D07103, doi:10.1029/2011JD016355.

Rosolem, R., H. V. Gupta, W. J. Shuttleworth, L. G. G. Gonçalves et X. Zeng. 2013. « Towards a comprehensive approach to parameter estimation in land surface parameterization schemes ». *Hydrological Processes*, doi:10.1002/hyp.9362.

Ruelland, D., S. Ardoin-Bardin, G. Billen et E. Servat. 2008. « Sensitivity of a lumped and semi-distributed hydrological model to several methods of rainfall interpolation on a large basin in West Africa ». *Journal of Hydrology*, 361, 96–117, doi:10.1016/j.jhydrol.2008.07.049

Salomon, R. 1996. « Reevaluating genetic algorithm performance under coordinate rotation of benchmark functions ». *BioSystems*, 39, 263-278.

Saltelli, A. 2002. « Sensitivity Analysis for Importance Assessment ». *Risk Analysis*, 22: 579–590. doi: 10.1111/0272-4332.00040

Saltelli, A. et S. Tarantola. 2002. « On the relative importance of input factors in mathematical models: safety assessment for nuclear waste disposal ». *Journal of the American Statistical Association*, 97 (459), 702–709.

Samuel, J., P. Coulibaly et R. Metcalfe. 2011. « Estimation of continuous streamflow in Ontario ungauged basins: Comparison of regionalization methods ». *Journal of Hydrologic Engineering*, 16(5), 447–459.

Sawaragi, Y., H. Nakayama et T. Tanino. 1985. « Theory of Multiobjective Optimization (vol. 176 of Mathematics in Science and Engineering) ». Orlando, FL: Academic Press Inc. ISBN 0126203709.

Schaake, J. C., Q. Duan, M. Smith et V. Koren. 2000. « Criteria to Select Basins for Hydrologic Model Development and Testing ». *15th Conference on Hydrology*, 10–14 January 2000, *American Meteorological Society*, Long Beach, CA, Paper P1.8.

Schaake, J., S. Cong et Q. Duan. 2006. « The U. S. MOPEX Data Set ». *IAHS Publication*, 307, 9-28.

Shafii, M. et F. De Smedt. 2009. « Multi-objective calibration of a distributed hydrological model (WetSpa) using a genetic algorithm ». *Hydrology and Earth System Sciences*, 13, 2137–2149.

Schmidli, J. et C. Frei. 2005. « Trends of Heavy Precipitation and Wet and Dry Spells in Switzerland during the 20th Century ». *International Journal of Climatology*, 25, 753-771

Schmitt, L. M. 2001. « Theory of genetic algorithms ». *Theoretical Computer Science*, 259, 1–61.

Schulla, J. et K. Jasper. 2000. « Model description WASIM-ETH (Water balance simulation model ETH) ». ETH-Zurich, Zurich, Switzerland.

Schwarz, G.E. 1978. « Estimating the dimension of a model ». *Annals of Statistics*, 6 (2), 461–464.

See, L. et S. Openshaw. 2000. « A hybrid multi-model approach to river level forecasting ». *Hydrological Sciences Journal*, 45, 523–536.

Seibert, J. 1999. « Regionalisation of parameters for a conceptual rainfall runoff model ». *Agricultural and Forest Meteorology*, 98-99(31), 279–293.

Sefton, C. E. M. et S.M. Howart. 1998. « Relationships between dynamic response characteristics and physical descriptors of catchments in England and Wales ». *Journal of Hydrology*, 211(1–4), 1–16.

Sellami, H., I. La Jeunesse, S. Benabdallah, N. Baghdadi et M. Vanclooster. 2014. « Uncertainty analysis in model parameters regionalization: a case study involving the SWAT model in Mediterranean catchments (Southern France) ». *Hydrology and Earth System Sciences*, 2393-2413.

Seo, D.-J. 1998. « Real-time estimation of rainfall fields using radar rainfall and rain gage data ». *Journal of Hydrology*, 208 (1–2), 37–52.

Shamseldin, A., K.M. O’Connor et G. Liang. 1997. « Methods for combining the output of different rainfall-runoff models ». *Journal of Hydrology*, 197, 203-229.

Shamseldin, A.Y., K.M. O'Connor et A.E. Nasr. 2007. « A comparative study of three neural network forecast combination methods for simulated river flows of different rainfall-runoff models ». *Hydrological Sciences Journal,* 52(5), 896-916.

Shepard, D.S. 1984. « Computer mapping : The SYMAP interpolation algorithm, Spatial Statistics and Models », G, L, Gaile and C, J, Willmott, Eds,, D, Reidel, 133–145.

Shu, C. et D.H. Burn. 2003. « Spatial patterns of homogeneous pooling groups for flood frequency analysis ». *Hydrological Sciences Journal*, 48(4), 601-618, DOI:10.1623/hysj.48.4.601.51417.

Singh, V. P. et D.K. Frevert. 2001. « Mathematical models of large watershed hydrology ». *Water Resources Publications*. Chapter 13.

Singh, V.P. et D.A. Woolhiser. 2002. « Mathematical modeling of watershed hydrology ». *Journal of Hydrologic Engineering*, 7, 270–292.

Sivapalan, M., K. Takeuchi, S. W. Franks, V. K. Gupta, H. Karambiri, V. Lakshmi, X. Liang, J. J. McDonnell, E. M. Mendiondo, P. E. O'Connell, T. Oki, J. W. Pomeroy, D. Schertzer, S. Uhlenbrook et E. Zehe. 2003. « IAHS Decade on Predictions in Ungauged Basins (PUB), 2003-2012: Shaping an exciting future for the hydrological sciences ». *Hydrological Sciences Journal*, 48, 857-880.

Skaugen, T. et J. Andersen. 2010. « Simulated precipitation fields with variance-consistent interpolation ». *Hydrological Sciences Journal*, 55(5), 676-686.

Smith, M. B., D.-J. Seo, V.I. Koren, S. Reed, Z. Zhang, Q. Duan, F. Moreda et S. Cong. 2004. « The distributed model intercomparison project (DMIP): Motivation and experiment design ». *Journal of Hydrology*, 298, 4–26.

Sobol', I. M. 1993. « Sensitivity estimates for nonlinear mathematical models ». *Mathematical and Computer Modelling*, 1, 407–17.

Spears, W.M., S.T. Green et D.F. Spears. 2010. « Biases in particle swarm optimization ». *International Journal of Swarm Intelligence Research*, 1(2), 34-57.

Steiner, M., J. A. Smith, S. J. Burges, C. V. Alonso et R. W. Darden. 1999. « Effect of bias adjustment and rain gauge data quality control on radar rainfall estimation ». *Water Resources Research*, 35 (8), 2487–2503.

Steuer, R. E. 1986. « Multiple Criteria Optimization: Theory, Computations, and Application ». New York: John Wiley & Sons, Inc. ISBN 047188846X.

Storn, R. et K. Price. 1997. « Differential evolution - a simple and efficient heuristic for global optimization over continuous spaces ». *Journal of Global Optimization*, 11, 341–359.

Tang, Y., P. Reed, T. Wagener et K. van Werkhoven. 2007. « Comparing sensitivity analysis methods to advance lumped watershed model identification and evaluation ». *Hydrology and Earth System Sciences*, 11, 793-817, doi:10.5194/hess-11-793-2007.

Tapsoba, D., V. Fortin et F. Anctil. 2005. « Apport de la technique du krigeage avec dérive externe pour une cartographie raisonnée de l'équivalent en eau de la neige: Application aux bassins de la rivière Gatineau ». *Canadian Journal of Civil Engineering,* 32.1 289-297.

Taylor, G., C. Daly, W. Gibson et J. Sibul-Weisburg. 1997. « Digital and Map Products Produced Using PRISM ». *10th Conf. on Applied Climatology*, Reno, NV, American Meteorological Society, 217-218.

Thornton, P.E., S.W. Running et M.A. White. 1997. « Generating surfaces of daily meteorological variables over large regions of complex terrain ». *Journal of Hydrology,* 190, 214-251. doi:10.1016/S0022-1694(96)03128-9

Thornton, P.E., M.M. Thornton, B.W. Mayer, N. Wilhelmi, Y. Wei et R.B. Cook. 2012. « Daymet: Daily surface weather on a 1 km grid for North America, 1980 – 2012 ». Acquired online (http://daymet.ornl.gov/)

Tolson, B.A. et C.A. Shoemaker. 2007. « Dynamically dimensioned search algorithm for computationally efficient watershed model calibration ». *Water Resources Research*, 43, W01413.

Tonkin, M. J. et J. Doherty. 2005. « A hybrid regularized inversion methodology for highly parameterized environmental models ». *Water Resources Research*, 41, W10412. doi:10.1029/2005WR003995.

Tozer, C. R., A. S. Kiem et D. C. Verdon-Kidd. 2012. « On the uncertainties associated with using gridded rainfall data as a proxy for observed ». *Hydrology and Earth System Sciences*, 16, 1481–1499.

Trelea, I.C. 2003. « The particle swarm optimization algorithm: convergence analysis and parameter selection ». *Information Processing Letters*, 85(6), 317–325.

Tuba, M., M. Subotic et N. Stanarevic. 2011. « Modified cuckoo search algorithm for unconstrained optimization problems ». In Proceedings of the 5th European conference on European computing conference (ECC'11), Remi Leandre, Metin Demiralp, Milan Tuba, Luige Vladareanu, and Olga Martin (Eds.). World Scientific and Engineering Academy and Society (WSEAS), Stevens Point, Wisconsin, USA, 263-268.

Turk, F. J., P. Arkin, E. Ebert et M. Sapiano. 2008. « Evaluating high resolution precipitation products ». *Bulletin of the American Meteorological Society*, 89, 1911–1916.

Ülker, E.D. et A. Haydar. 2012. « Comparison of the performances of differential evolution, particle swarm optimization and harmony search algorithms on benchmark functions ». *Academic Research International*, 3(2), 85-92.

Valéry, A. 2010. « Modélisation precipitations – debit sous influence nivale. Élaboration d'un module neige et évaluation sur 380 bassins versants ». *Agro Paris Tech*., 417p.

Valéry, A., V. Andréassian et C. Perrin. 2014. « 'As simple as possible but not simpler': What is useful in a temperature-based snow-accounting routine? Part 2 - Sensitivity analysis of the Cemaneige snow accounting routine on 380 catchments ». *Journal of Hydrology*, 517(0), 1176-1187.

Vandewiele, G. L. et A. Elias. 1995. « Monthly water balance of ungauged catchments obtained by geographical regionalization ». *Journal of Hydrology*, 170(1-4), 277–291.

van Straten, G. et K. J. Keesman. 1991. « Uncertainty propagation and speculation in projective forecasts of environmental change: A lake-eutrophication example ». *Journal of Forecasting,* 10, 163–190.

van Werkhoven, K., T. Wagener, P. Reed et Y. Tang. 2008. « Characterization of watershed model behavior across a hydroclimatic gradient ». *Water Resources Research*, 44, W01429, doi:10.1029/2007WR006271.

van Werkhoven, K., T. Wagener, P. Reed et Y. Tang. 2009. « Sensitivity-guided reduction of parametric dimensionality for multi-objective calibration of watershed models ». *Advances in Water Resources*, 32, 1154–1169.

Velazquez, J.A., F. Anctil et C. Perrin. 2010. « Performance and reliability of multimodel hydrological ensemble simulations based on seventeen lumped models and a thousand catchments ». *Hydrology and Earth System Sciences*, 14, 2303-2317.

Velazquez, J.A., F. Anctil, M.H. Ramos et C. Perrin. 2011. « Can a multi-model approach improve hydrological ensemble forecasting? A study on 29 French catchments using 16 hydrological model structures ». *Advances in Geosciences*, 29, 33–42, doi:10.5194/adgeo-29-33-2011

Verseghy, D.L., N.A. McFarlane et M. Lazare. 1993. « Class—A Canadian land surface scheme for GCMS, II. Vegetation model and coupled runs ». *International Journal of Climatology*, 13, 347–370. doi: 10.1002/joc.3370130402.

Viney, N.R., J. Vaze, F.H.S. Chiew, J.-M. Perraud, D.A. Post et J. Teng. 2009. « Comparison of multi-model and multi-donor ensembles for regionalisation of runoff generation using five lumped rainfall-runoff models ». 18th World IMACS Congress and MODSIM09 International Congress on Modelling and Simulation, Cairns, Australia.

Vrugt, J.A., H.V. Gupta, W. Bouten et S. Sorooshian. 2003a. « A Shuffled Complex Evolution Metropolis algorithm for optimization and uncertainty assessment of hydrologic model parameters ». *Water Resources Research*, 39, 1201, doi:10.1029/2002WR001642, 8.

Vrugt, J. A., H. V. Gupta, L. A. Bastidas, W. Bouten et S. Sorooshian. 2003b. « Effective and efficient algorithm for multiobjective optimization of hydrologic models ». *Water Resources Research*, 39(8), 1214, doi:10.1029/2002WR001746.

Vrugt, J. A. et B.A. Robinson. 2007a. « Improved evolutionary optimization from genetically adaptive multimethod search ». *Proceedings of the National Academy of Sciences*, 104, 708– 711.

Vrugt, J. A. et B.A Robinson. 2007b. « Treatment of uncertainty using ensemble methods: Comparison of sequential data assimilation and Bayesian model averaging ». *Water Resources Research,* 43, W01411, doi:10.1029/2005WR004838.

Wagener, T. et H. S. Wheater. 2006. « Parameter estimation and regionalisation for continuous rainfall-runoff models including uncertainty ». *Journal of Hydrology*, 320, 132–154.

Wagener, T. et J. Kollat. 2007. « Numerical and visual evaluation of hydrologic and environmental models using the Monte Carlo Analysis Toolbox (MCAT) ». *Environmental Modelling and Software*, 22, 1021–33.

Warner, T. T., E. A. Brandes, J. Sun, D. N. Yates et C. K. Mueller. 2000. « Prediction of a flash flood in complex terrain. Part I: A comparison of rainfall estimates from radar, and very short range rainfall simulations from a dynamic model and an automated algorithmic system ». *Journal of Applied Meteorology and Climatology*, 39 (6), 797–814.

Westrick, K. J., C. F. Mass et B. A. Colle. 1999. « The limitations of the WSR-88D radar network for quantitative precipitation measurement over the Coastal Western United States ». *Bulletin of the American Meteorological Society*, 80 (11), 2289–2298.

Whitley, D., M. Lunacek et A. Sokolov. 2006. « Comparing the niches of CMA-ES, CHC and pattern search using diverse benchmarks ». *In Parallel problem solving from nature* (PPSN IX), LNCS, 4193, 988–997.

Widmann, M. et C.S. Bretherton. 2000. « Validation of mesoscale precipitation in the NCEP reanalysis using a new gridcell dataset for the northwestern United States ». *Journal of Climate*, 13, 1936–1950.

Wilby, R. L. et I. Harris. 2006. « A framework for assessing uncertainties in climate change impacts: Low-flow scenarios for the River Thames, UK ». *Water Resources Research*, 42, W02419, doi:10.1029/2005WR004065.

Wilcoxon, F. 1945. « Individual comparisons by ranking methods ». *Biometrics*, 1(6), 80–83.

Wilson, M. F., A. Henderson-Sellers, R. E. Dickinson et P. J. Kennedy. 1987a. « Sensitivity of the Biosphere-Atmosphere Transfer Scheme (BATS) to the inclusion of variable soil characteristics ». *Journal of Applied Meteorology and Climatology*, 26, 341–362.

Wilson, M. F., A. Henderson-Sellers, R. E. Dickinson et P. J.Kennedy. 1987b. « Investigation of the sensitivity of the land surface parameterization of the NCAR Community Climate Model in regions of tundra vegetation ». *International Journal of Climatology*, 7, 319–343.

Wurbs, R. A. 1998. « Dissemination of generalized water resources models in the United States ». *Water International*, 23, 190–198.

Yadav, M., T. Wagener et H. Gupta. 2007. « Regionalization of constraints on expected watershed response behavior for improved predictions in ungauged basins ». *Advances in Water Resources*, 30, 1756–1774.

Yang, X.-S. et S. Deb. 2009. « Cuckoo search via Levy flights ». *In Proc. of World Congress on Nature & Biologically Inspired Computing, India*. IEEE Publications, USA, 210-214.

Yapo, P. O., H. V. Gupta et S. Sorooshian. 1996. « Calibration of conceptual rainfall-runoff models: Sensitivity to calibration data ». *Journal of hydrology*, 181, 23–48.

Yapo, P. O., H. V. Gupta et S. Sorooshian. 1998. « Multi-objective global optimization for hydrologic models ». *Journal of Hydrology*, 204, 83-97.

Yatagai A., O. Arakawa, K. Kamiguchi, H. Kawamoto, M. I. Nodzu et A. Hamada. 2009. « A 44-year daily precipitation dataset for Asia based on dense network of rain gauges ». *Scientific Online Letters on the Atmosphere*, 5, 137–140. Doi:10.2151/sola.2009-035.

Ye, M., S.P. Neuman et P.D. Meyer. 2004. « Maximum likelihood Bayesian averaging of spatial variability models in unsaturated fractured tuff ». *Water Resources Research*, 40, W05113, doi:10.1029/2003WR002557.

Yen, J., J.C. Liao, B. Lee et D. Randolph. 1998. « A Hybrid approach to modeling metabolic systems using genetic algorithms and simplex method ». *IEEE Transactions on Systems, Man, and Cybernetics*, 28, 173-191.

Zelelew, M.B. et K. Alfredsen. 2014. « Transferability of hydrological model parameter spaces in the estimation of runoff in ungauged catchments ». *Hydrological Sciences Journal*, 59 (8), 1470–1490. http://dx.doi.org/10.1080/02626667.2013.838003

Zhang, Y. et F. H. S. Chiew. 2009. « Relative merits of different methods for runoff predictions in ungauged catchments ». *Water Resources Research*, 45, W07412, doi:10.1029/2008WR007504.

Zhang, C., J. G. Chu et G.T. Fu. 2013. « Sobol's sensitivity analysis for a distributed hydrological model of Yichun River Basin, China ». *Journal of Hydrology*, 480, 58-68.