7-27-2015 12:00 AM

Assessment of the Regionalization of Precipitation in Two Canadian Climate Regions: A Fuzzy Clustering Approach

Sarah E. Irwin
The University of Western Ontario

Supervisor
Professor Slobodan P. Simonovic
The University of Western Ontario

Graduate Program in Civil and Environmental Engineering
A thesis submitted in partial fulfillment of the requirements for the degree in Master of Engineering Science

Recommended Citation

Irwin, Sarah E., "Assessment of the Regionalization of Precipitation in Two Canadian Climate Regions: A Fuzzy Clustering Approach" (2015). *Electronic Thesis and Dissertation Repository*. 3052.
https://ir.lib.uwo.ca/etd/3052

ASSESSMENT OF THE REGIONALIZATION OF PRECIPITATION IN TWO CANADIAN CLIMATE REGIONS: A FUZZY CLUSTERING APPROACH

(Thesis format: Monograph)

by

Sarah E. Irwin

Graduate Program in Civil and Environmental Engineering

A thesis submitted in partial fulfillment
of the requirements for the degree of
Master of Engineering Science

The School of Graduate and Postdoctoral Studies
The University of Western Ontario
London, Ontario, Canada

Abstract

Regional frequency analysis (RFA) is used to obtain reliable estimates of local precipitation events for a variety of applications in water resources engineering. The focus of the presented research is on an initial step of the RFA process; that is the formation of precipitation regions (also referred to as regionalization). The aim of this study is to dissect the regionalization procedure into its individual components that require subjective user input, and to evaluate their respective influences on the results. All assessments are conducted in two of Canada's climate regions; namely the Prairie and Great Lakes-St. Lawrence lowlands. Additionally, a new fuzzy clustering approach to regionalization that uses optimization is proposed. It is evident that the outcomes are sensitive to the choice of the regionalization method, the number of regions into which the sites of the study area are partitioned, the climate site attributes and the temporal resolution of the precipitation data. Recommendations for the selection of such factors are provided based on their application.

Keywords

Regional frequency analysis, regionalization, climate, precipitation, ANUSPLIN, atmospheric variables, fuzzy *c*-means, *L*-moments, optimization, differential evolution, fuzzy Compromise programming

Acknowledgements

First and foremost I would like to thank my Supervisor, Professor Slobodan P. Simonovic for providing me with this unforgettable learning experience. His guidance and expertise in the field of water resources engineering and systems engineering have been invaluable. It has been a privilege and an honour to work with such a passionate, inspiring professor throughout my undergraduate and graduate education.

I would like to extend a special thank you to Professor Donald H. Burn of the University of Waterloo for offering his expert knowledge in the field of water resources engineering and particularly on the topic of regional frequency analysis. I would also like to thank Dr. Roshan Srivastav for his mentorship and generously spending time to teach me statistics and computer programming skills. Thank you to all of my mentors for broadening my horizons and instilling confidence in myself and our work.

To the students of the Facility for Intelligent Decision Support - Abhishek, Amin, Andre, Angela, Arunkumar, Benyou, Bogdan, Elaine, Leanna, Nick, Nikhil, Patrick, Roshan, Sohom, Stuart, Vladimir - thank you for your support. I am lucky to have gained valuable and lifelong friendships during my time spent in FIDS.

And finally, I would like to thank my parents and family for providing me with utmost support while I pursue my undergraduate and graduate education at the University of Western Ontario.

Table of Contents

List of Tables

List of Figures

Nomenclature

List of symbols

c – The number of clusters into which the objects (climate sites of the study area) are partitioned using the fuzzy *c*-means algorithm.

k - The number of clusters into which the objects (climate sites of the study area) are partitioned using the k-means algorithm.

$\boldsymbol{\tilde{p}}$ – The fuzzy parameter that represents the significance of the maximum deviation from the ideal point (deviations are assigned equal weight for values of *p* equal to 1; otherwise deviations are weighted according to the magnitude of their deviation). It is the exponent of the fuzzy distance metric for the fuzzy Compromise programming procedure.

List of acronyms

ANUSPLIN - A package developed by the Australian National University that interpolates noisy, multi-variate data using smoothing splines.

AS-1 - Attribute Set 1: Climate site attributes that include latitude, longitude and distance to major water bodies.

AS-2 - Attribute Set 2: Climate site attributes that include latitude, longitude, distance to major water bodies and elevation.

AS-3 - Attribute Set 3: Climate site attributes that include atmospheric variables that exhibit a strong linear correlation with the local precipitation.

AS-4 - Attribute Set 4: Climate site attributes that include all atmospheric variables considered in the analysis.

CanESM2 - Canadian Earth Systems Model (second generation). Earth's systems models are advanced global climate models (see **GCM** below).

CCA - Canonical correlation analysis: A statistical approach that solves for the linear combinations of variables in two multi-variate vectors that together, form a maximum correlation.

CVI - Cluster validity index: Provide a measurement of the "goodness" of a clustering of data.

DE - Differential evolution: A type of evolutionary algorithm (see **EA**).

DEM - Digital elevation model: A three dimensional representation of the Earth's surface that is derived from surface elevation data.

DJF - December, January, February season.

EA - Evolutionary algorithm: An optimization technique that searches within a population of solutions for the solution set that optimizes the problem criteria/objectives. It is a general term that encompasses several different search algorithms.

GCM - Global climate model: Computer models that simulate the physical processes of the atmosphere, oceans, cryosphere and land surface.

JJA - June, July, August season.

***L*-moment** – Linear moments: Linear combinations of ordered observations that are used to describe the shapes of probability distributions.

MAM - March, April, May season.

PCA - Principle component analysis: A statistical technique that is used to transform a large set of possibly correlated observations into a lower dimensional set of uncorrelated (orthogonal) vectors that are referred to as principle components.

RFA – Regional frequency analysis: A stochastic method for estimating at-site precipitation magnitudes.

ROI - Region of Influence: A site focused regionalization technique.

SON - September, October, November season.

Chapter 1

1. Introduction

1.1 General

Water resources engineers are responsible for the development of plans, designs and operational procedures for systems that control the distribution of water within a region of interest (Chin, 2006). Their ultimate challenge is to manage uncertain hydrologic processes while balancing the needs of the natural and socio-economic environments. Hydrologic processes (overland flow, groundwater flow, evaporation, infiltration, etc.) are primarily driven by precipitation events that are characterized by complex spatial and temporal distributions. Precipitation is considered to be a random event and therefore its occurrence is estimated through probabilistic (stochastic) methods. The accepted stochastic approach for estimating at-site precipitation magnitudes (of certain duration) is known as frequency analysis. The general approach to frequency analysis involves fitting a statistical probability distribution to the historical precipitation record measured at the site of interest (Chin, 2006). Estimates of precipitation magnitudes are extracted from the frequency distribution and used in a variety of water resources applications including water infrastructure design, forecasting and downscaling functions, drought frequency analysis, insurance risk calculations and hydrologic modeling for reservoir operations, among others (Hosking and Wallis, 1997).

A major challenge facing water resources engineers is the absence of complete, sufficiently long precipitation records (Burn, 1990). For example, some engineering applications (water infrastructure design) must accommodate rare precipitation events that correspond to a large return period (50, 100 years) that is defined as the average number of years between events of a certain magnitude. Complete records that are at least 50 or 100 years in length are difficult to obtain, and frequency distributions that are developed from datasets that are shorter than the return period of interest are unable to capture the true variability of the local precipitation. This is a major limitation of the traditional frequency analysis. To overcome this problem, data from several neighboring sites that exhibit similar statistical properties of precipitation can be combined into a single probability distribution from which precipitation occurrence is more reliably estimated. The modified approach is referred to as regional frequency analysis (RFA) (Chin, 2006; Chebana and Ouarda, 2008). An important assumption of the RFA procedure is that the at-site frequency distributions of the observed precipitation records (that are to be combined into a single distribution) are identically distributed; that is also known as the rule of homogeneity (Hosking and Wallis, 1997). This criterion is approximately achieved by assigning sites to homogeneous precipitation regions in a process called regionalization; thus introducing the fundamental topic of the presented research.

The general regionalization procedure involves the employment of a tool to partition climate sites of the study area into contiguous or non-contiguous regions based on the similarity of their attributes (that are typically site characteristics that drive the local precipitation) (Ouarda *et al*., 2001). Depending on the chosen method, similarity can be measured using correlation coefficients or distance metrics such as the Euclidean or Mahalanobis distance (Rao and Srinivas, 2005). The precipitation regions are subsequently validated for uniform precipitation test statistics. The most common method for validation is the *L*-moment regional heterogeneity test developed by Hosking and Wallis (1997).

Regionalization methods aim to achieve two objectives/criteria in their output: (i) maximization of the number of station-years in the regional precipitation records; and (ii) maximization of the number of homogeneous precipitation regions. The criteria are in conflict with one another because in order to maximize the number of station-years in the regional precipitation record, the climate sites should be assigned to a small number of large-sized regions; however, a large number of regions that consist of fewer statistically similar sites are more likely to be classified as homogeneous. It is unlikely for climate sites to be partitioned into a set of regions that are entirely homogeneous without compromising the number of station-years of the regional precipitation record and manual adjustments to the sites' memberships are often required to improve the results.

The compositions of the precipitation regions depend upon several user preferences including the regionalization method, the number of regions to which the climate sites are assigned, the climate site attributes and the temporal resolution of the precipitation data. The regions derived for one application (that require a specific set of user preferences) may not be well suited to other uses (Srinivas, 2013) and therefore, the aim of this study is to dissect the regionalization procedure into its individual components that require subjective user input, and to evaluate their respective influences on the results. The fuzzy *c*-means clustering algorithm is employed to delineate precipitation regions in two climatically diverse study areas; namely, the Prairie and Great Lakes-St. Lawrence lowlands climate regions of Canada. Homogeneity of the precipitation regions is validated using the *L*-moment regional heterogeneity test developed by Hosking and Wallis (1997). The final step of performing manual adjustments to site membership is omitted from the analysis.

Key research objectives are as follows:

- Justify the application of the fuzzy *c*-means algorithm to partition climate sites into precipitation regions in the subsequent tests;
- Measure the ability of different climate site attributes to achieve the objective criteria in the regionalization outcomes. Several combinations of location parameters (latitude, longitude, elevation and distance to water bodies) and atmospheric variables recorded at different pressure levels are considered as potential attributes because of their strong influence on the variability of precipitation;
- Study the effect of the temporal resolution on the formation of precipitation regions. Monthly, seasonal and annual time scales are considered as well as the maximum annual series of the precipitation data;
- Assess the available methods for determining the preferred number of clusters for the climate sites to be assigned to (the *c*-value parameter of the fuzzy *c*-means clustering algorithm);
- Introduce a new method (based on optimization and fuzzy Compromise programming) for the selection of input parameters to the fuzzy *c*-means algorithm; namely the fuzzifier (that controls the degree of fuzziness of a partitioning) and the *c*-value (that is the number of regions to which the climate sites are assigned).

1.2 Organization of the Thesis

Chapter 2 contains a literature review of research pertaining to the regionalization of precipitation. First an evaluation of regionalization methods is presented, followed by a review of the approaches used to select the optimal number of regions to which the climate sites are assigned; typically referred to as cluster validity indices. Subsequently a comparison of regionalization outcomes for the employment of different site attributes is presented, followed by a review of the related studies that have applied different temporal resolutions of precipitation data. The proposed model for selecting input parameters to the fuzzy *c*-means clustering algorithm is introduced in Chapter 3. Application of the tests including the study areas and data used are described in Chapter 4. Chapter 5 explains the methodology used in four main steps: (i) formation of the attribute sets; (ii) selection of the number of regions; (iii) delineation of precipitation regions; and (iv) validation of regional homogeneity. Afterward the method for the proposed model for regionalization is described. Chapter 6 presents the results of the following investigations: (i) justification for the use of the fuzzy *c*-means algorithm; (ii) influence of the choice of site attributes on the regionalization outcomes; (iii) effect of the temporal resolution of the precipitation data on the regional compositions; (iv) analysis of the output from the proposed regionalization model. Finally, a summary of the results and concluding remarks are provided in Chapter 7.

Chapter 2

2. Literature Review

In this Chapter a review of the literature pertaining to various aspects of the regionalization procedure is presented including: (i) regionalization methods; (ii) techniques for determining the preferred number of clusters into which the climate sites are partitioned; (iii) selection of climate site attributes that drive the local precipitation; and (iv) choice of the temporal resolution of the precipitation datasets.

2.1 Regionalization Methods

Several regionalization methods are available including: (i) correlation analysis; (ii) principle component analysis; (iii) region of influence; and (iv) cluster analysis, among others. The following sections analyze some of the most commonly used regionalization techniques (and their variations) for the delineation of coherent hydrologic and pluviometric (precipitation) regions. A substantial amount of literature is available on the formation of hydrologic regions for flood frequency analysis and therefore, major findings from the aforementioned works are included where they are also applicable to the regionalization of precipitation.

2.1.1 Correlation Analysis

Correlation analysis is one approach used to delineate precipitation regions. The general procedure involves partitioning climate sites that exhibit strong, positive correlations between their precipitation records into uniform pluviometric regions. Certain extensions of the correlation analysis technique are explained below.

Elementary linkage analysis is a simple variation of correlation analysis and its procedure is as follows: Two climate sites that form the strongest positive correlation with one another are identified and combined into a provisional region. Additional climate sites that demonstrate a significant correlation with those belonging to the provisional region are also assigned as members; (a correlation coefficient that exceeds a predefined threshold value is considered to be significant). After all significantly correlated sites are assigned to the current group the process is repeated until the entire study area is partitioned into coherent precipitation regions. Adelekan (1998) used elementary linkage analysis to better understand the spatial and temporal variations of thunderstorm patterns in Nigeria and to study the effects of high intensity rainfall on flooding and soil erosion in the resultant regions.

Saikranthi *et al.* (2013) and Gadgil *et al.* (1993) used another variation of correlation analysis to identify homogeneous rainfall regions in India. The technique differs from elementary linkage analysis in that the initial step is to identify the sites that form the weakest correlation between their precipitation records. These sites become seed points for the development of two distinct regions. Additional seed points are established based on the following criteria: (i) the average correlation between the site (potential seed point) and all other seed points is a minimum value; and (ii) the correlation between the site (potential seed point) and all other seed points is insignificant (less than a predefined threshold value). After all seed points have been established, the precipitation regions are formed around them. Sites are assigned to the region corresponding to the seed point that they are most strongly correlated with.

Canonical correlation analysis (CCA) is another variation of the correlation analysis approach that has been used in regional flood frequency analyses. Cavadias (1990) and Ouarda *et al.* (2001) employed this technique in the Canadian provinces of Newfoundland and Ontario, respectively. Using CCA the correlation structure between sets of basin characteristics and hydrologic parameters is assessed. The linear combination of the basin characteristics that forms the strongest correlation with the hydrologic parameters is identified and employed in the regional estimation procedure.

CCA is advantageous as it can be used to form homogeneous regions for gauged and ungauged basins. Furthermore, it is superior to the Pearson correlation approaches (elementary linkage analysis and its variations) as it provides additional information on the correlation structure between basin characteristics and hydrologic parameters (Ouarda, 2001). Major limitations of all correlation analysis methods include: (i) the subjectivity involved in selecting a threshold value for site membership; and (ii) their inability to detect differences in the magnitudes of the correlated datasets (Unal, 2003).

2.1.2 Principal Component Analysis

Principal component analysis (PCA) is another common approach to regionalization. The procedure first requires computation of the principal components of the sites' precipitation records; i.e. the transformation of the sites' precipitation records into a set of linearly uncorrelated (orthogonal) vectors. The regions are subsequently formed using one of the following techniques: (i) mapping; or (ii) maximum loading.

Using the former approach the sites' component loadings are plotted on a map of the study area. (Note that the loadings indicate the degree of variance of the sites' precipitation records that are described by each component; higher loadings are representative of greater variances). The loadings are contoured and the contoured areas that exceed a predefined threshold value are classified as coherent precipitation regions. The process is conducted such that there is one map per component. Typically each component map forms one unique region; however it is possible for certain sites to have significant loadings under multiple components resulting in overlapping regions (Serra *et al.*, 1996). Maximum loading is another approach to regionalization that uses PCA. For this method, each component is representative of a region and sites are assigned to the region (component) in which they have the highest loading (Chen, 2009).

PCA is not an effective means of regionalization when a large number of components account for the total variance of the precipitation records because the process of assigning sites to regions becomes more subjective and difficult to manage (Srivinas, 2013). Serra *et al.* (1996) employed PCA to delineate homogeneous rainfall regions in Spain. They found that the local orography and land-sea interactions contributed to significant spatial variations in rainfall patterns across the country. As a result the leading principal components were unable to capture much of the total variance of the study area's precipitation records. The sites had significant loadings under many components resulting in a highly subjective decision-making process when determining their regional memberships. Evidently, PCA may not be an effective regionalization method for areas with complex topography or local influences such as significant water bodies.

2.1.3 Region of Influence

The concept of the site-focused pooling approach to regionalization was originally conceived by Acreman and Wiltshire (1987) and was later extended by Burn (1990) into what is known as the region of influence (ROI) method. Using this method, contiguous regions are formed around individual target sites such that they each have their own, unique region. Sites of the study area are assigned to a target site's region of influence if the similarity between their attributes is greater than a predefined threshold value and similarity is measured using a distance metric.

The procedure requires the subjective selection of a threshold value. The subjectivity, however, can be eliminated by assigning the minimum number of sites needed to obtain a desirable precipitation record length to the region, where the desired record length depends upon the application. The ROI approach is advantageous as it is flexible, meaning that the regions shift according to the site that is under investigation and as such, each region is comprised of the sites that are most similar to the target. Ga’al *et al* (2008) found that the flexible regions formed by the ROI approach provide superior results for regional frequency analysis compared to the fixed regions that may be formed using correlation analysis, principal component analysis, cluster analysis, among other methods. Overall ROI is a choice approach to regionalization and it is particularly useful for site specific applications.

2.1.4 Cluster Analysis

Clustering algorithms are perhaps the most popular regionalization methods in climate literature (Rao and Srinivas, 2005; Satyanarayana and Srinivas, 2008; 2011; Srinivas, 2013; Asong, 2015). Many clustering algorithms are available for delineating precipitation regions. Their general procedure involves partitioning climate sites into regions according to the similarity of their attributes where attributes are drivers of the local precipitation and similarity is measured using a distance metric such as the Euclidean or Mahalanobis distance (Rao and Srinivas, 2005).

Clustering algorithms are categorized as hierarchical and partitional; and hierarchical algorithms are further divided into agglomerative and divisive classifications. Agglomerative algorithms merge individual sites into larger clusters and conversely, divisive algorithms divide one large cluster (that is composed of all sites in the study area) into smaller regions. Divisive algorithms are uncommon in regionalization literature; however several types of agglomerative algorithms have been used including single linkage, complete linkage, average linkage and Ward's algorithm. The underlying difference between most agglomerative algorithms is the means by which similarity is measured between site attributes (Kalkstein, 1978; Rao and Srinivas, 2005).

Partitional clustering algorithms partition/divide sites of a study area into regions. They work to minimize the value of an objective function that measures the sum of the distances between climate sites belonging to the same region in the attribute space; as such, the within cluster similarity and likewise, the between cluster separation are maximized (Zalik and Zalik, 2011). The k-means clustering algorithm is a very common partitional algorithm, developed by MacQueen (1967), that measures similarity as the distance between the climate sites and the cluster centroids (the average attribute value of the member sites of the cluster) in the attribute space (Burn and Goel, 2000; Pelchzer, 2008; Satyanarayana and Srinivas, 2008; Dikbas, 2013). At each step of the iterative process a climate site is assigned to the region to which it is most similar and the value of the cluster centroid is updated to incorporate its new member. Following recalculation of the cluster centroid, it is possible for its member sites to exhibit stronger similarities to other clusters and therefore, site memberships may be reassigned. Other partitional algorithms include: (i) the k-medoids approach where similarity is measured between a climate site and the median value of the member site attributes (Kaufman, 1987); and (ii) the k-modes algorithm where similarity is measured between the climate site and mode of the member site attributes (Huang, 1998).

The ability of the algorithms to update site membership values at each iteration is considered to be a major advantage of all partitional algorithms over the hierarchical approaches. A disadvantage of partitional algorithms is their sensitivity to the initialization of the cluster centres; however, to address this limitation the algorithm is evaluated several times until the objective function yields a global minimum value. Gong and Richman (1995) compared hierarchical (single linkage, complete linkage, average linkage, Ward's method) and non-hierarchical (k-means, principle component analysis) clustering methods and determined that the non-hierarchical algorithms provided more accurate results.

Rao and Srinivas (2005) recognized the merits and limitations of both hierarchical and partitional algorithms and therefore, combined them into a hybrid clustering algorithm that uses a hierarchical algorithm to initialize the cluster centres and the k-means technique to partition the climate sites into regions. They evaluated the outcomes of the single linkage, complete linkage, Ward's and k-means algorithms, and combinations of single linkage - k-means, complete linkage - k-means and Ward's - k-means algorithms and found that the latter produced the best results in terms of the maximization of within cluster similarity and between cluster separation.

2.1.5 Fuzzy clustering algorithm

All algorithms described in the previous section form hard clusters such that each site belongs to only one region (Zalik and Zalik, 2011); therefore, implying that member sites of the same region fully resemble one another, which is not a valid assumption (Srinivas, 2013). To address this limitation the fuzzy *c*-means algorithm was developed (Bezdek, 1981). The fuzzy *c*-means algorithm is very similar to the k-means technique except that it computes the degree to which a site belongs to each cluster on a scale of 0 to 1 where a value of 1 represents full membership. As such, each climate station can partially belong to several clusters theoretically providing a more accurate partitioning of the sites. The outcomes of the regionalization procedure often require subjective and manual adjustments to site membership in order to improve the regional homogeneity of precipitation variability to an acceptable level. Its membership function provides useful information for removing or relocating discordant stations; thereby offering another advantage of the fuzzy *c*-means algorithm over traditional hard clustering techniques (Srinivas, 2013). Rao and Srinivas (2006) and Goyal and Gupta (2014) have conducted comparative analyses between the fuzzy *c*-means and k-means algorithms for regional flood frequency analysis. They concluded that the former technique outperformed the latter by achieving a greater number of homogeneous hydrologic regions. Although the analyses were conducted for the regionalization of flood quantiles, their findings also apply to the formation of precipitation regions.

Evidently the fuzzy *c*-means algorithm provides several advantages over the traditional hard clustering algorithms; however it does have certain limitations. A major drawback is its requirement for subjective input parameters; namely, the fuzzifier (the parameter that controls the fuzziness of the membership function) and the *c*-value (the number of regions to which the climate sites are assigned). The magnitudes of the input parameters significantly impact the regionalization outcomes. The requirement for the number of clusters to which the climate sites are assigned to be solved for *a priori* is a disadvantage of all clustering algorithms; they are all incapable of establishing the number of clusters that provides for the best partitioning of the sites.

2.1.6 Summary

The available approaches to regionalization include but are not limited to the methods that have been introduced above. Other techniques include spectral analysis (Azad *et al*., 2010), common factor analysis (Carter and Elsner, 1997) and artificial neural networks (Michaelidies *et al.,* 2001). In addition, new methods and variations of existing techniques are continuing to emerge in climate literature. Of the regionalization methods that are available at this time, clustering algorithms are preferred for their innate ability to recognize underlying patterns in complex datasets.

Traditionally, hard clustering methods have been used to partition climate sites such that they each belong to exactly one region. The k-means clustering algorithm is favoured for its ability to update site membership at each iteration in order to achieve the global minimum value of its objective function. More recently, however, the fuzzy *c*-means clustering algorithm has been gaining popularity in climate literature for the inclusion of its membership function that assigns partial membership values to the sites for each cluster. This feature provides a more accurate partitioning of climate sites into coherent precipitation regions.

Based on the information provided in literature, the fuzzy *c*-means algorithm has been elected as the choice regionalization method to be employed in the presented research. Chapter 6, section 1 presents a comparison between the performances of the k-means and fuzzy *c*-means clustering algorithms. The purpose of this investigation is to further justify the application of the fuzzy *c*-means technique for the subsequent analyses.

2.2 Number of Precipitation Regions

A review of the literature pertaining to the delineation of precipitation regions recognizes clustering algorithms as superior regionalization approaches. They are limited, however, by their inability to identify the number of clusters that provide for the natural partitioning of the climate sites into precipitation regions. Consequently, the number of clusters must be solved for prior to the employment of the algorithm (Gurrutxaga, 2013). The fuzzy *c*-means algorithm (that is elected as the choice regionalization method for this research) requires two subjective input parameters to be solved for *a priori*; namely, the *c*-value and the fuzzifier. These parameters are most reliably solved for using the trial and error method where their values are varied for a range of magnitudes and used as input to the algorithm. The resultant partitionings are validated for regional homogeneity of the statistics of the precipitation records. The parameters that result in the partitioning with the highest proportion of homogeneous regions are retained and used for regional frequency analysis applications. Although the trial and error method provides accurate results it is very computationally demanding because the algorithm must be evaluated for all combinations of the *c* and fuzzifier parameters.

To address this issue cluster validity indices (CVIs) are employed to solve for the optimal values of the input parameters, particularly for the *c*-value. CVIs are used to evaluate a partitioning of objects (climate sites) for their compactness (similarity between climate site attributes belonging to the same region) and separation (distinction between clusters that is measured as the distance between cluster centres in the attribute space) for a range of different input parameters (Kim and Ramakrishna, 2005). Many different CVIs are available. Together they produce a variety of outcomes and currently, a single, universally accepted measure has not been identified. Satyanarayana and Srinivas (2011) employed five CVIs to determine the magnitudes of the *c* and fuzzifier input parameters to the fuzzy *c*-means algorithm for a regional frequency analysis application. The following CVIs were considered: (i) the fuzzy partition coefficient; (ii) the fuzzy partition entropy; (iii) the fuzziness performance index; (iv) the normalized classification entropy; and (v) the extended Xie-Beni index. Of these indices, the first four demonstrated trends that increased or decreased monotonically and as such, they were deemed unsuitable for the solving for the magnitudes of the input parameters. They found that the extended Xie-Beni index performed relatively well.

Besides the lack of consistency between their outputs, a major drawback of the CVIs for their applications in the regionalization of precipitation is the disconnect between the natural grouping of climate sites in the attribute space (that is solved for using CVIs) and the inherent partitioning of the precipitation data (that is desired for regional frequency analysis). Since climate site attributes (that are drivers of the local precipitation) are used as input to the clustering algorithm and precipitation is reserved as an independent dataset for validation, the natural grouping of the site attributes is unlikely to directly correspond to that of the precipitation data.

Note that in Chapter 6, for the assessments of the impacts of the chosen regionalization method, climate site attributes, and temporal resolution of the precipitation data on the regionalization outcomes, the value of the fuzzifier is set equal to 2 and the *c*-value is solved for using a trial and error method. The fuzzifier can take on a value in the range of 1 to infinity; however the range is more commonly reduced to 1 to 10. Although it is more accurate to solve for the value of the fuzzifier parameter that provides for the optimal partitioning of the data, it is not important to do so for the aforementioned assessments as long as a consistent value is used. Fixing the magnitude of the fuzzifier significantly reduces the computational time required to run the tests. As such, the fuzzifier is set equal to 2 that is the most common value used in other research initiatives (Pal and Bezdek, 1995).

2.3 Climate Site Attributes

Clustering algorithms are used to partition climate sites into regions based on the similarity of their attributes and therefore, attribute selection has a significant influence on the composition of the precipitation regions. In the past, precipitation statistics were assigned to sites as attributes (Easterling, 1989; Kulkarni and Kripalani, 1998; Dikbas *et al.,* 2012). There are two disadvantages associated with this: (i) it eliminates the availability of an independent dataset to validate regional homogeneity; and (ii) it requires a large number of sites with long records to accurately represent the precipitation statistics (Burn, 1997; Satyanarayana and Srinivas 2008; 2011). Complete and sufficiently long observed climate records are frequently unavailable, and in certain applications it is the lack of complete data that necessitates the implementation of the regionalization procedure. This applies to the development of climatic design values such as rainfall magnitudes of a specific return period. Observed precipitation records must contain a sufficiently large number of station-years to provide for reliable estimates of rainfall for the return period of interest.

Evidently an alternative set of attributes is needed to form precipitation regions. Burn (1997) used seasonality measures to regionalize catchments in Western Canada. Seasonality is defined as the timing of flooding (precipitation) events within the year. If the timing of precipitation events varies across the study area, the use of seasonality may be extended to regionalize precipitation; however this may not be effective for smaller areas. Comrie and Glenn (1998) successfully adopted the seasonal timing of monthly precipitation to delineate precipitation regions in the Southwestern United States and Northern Mexico. They attributed the variation in seasonality to the different atmospheric drivers and changes in elevation across the large study area.

Satyanarayana and Srinivas (2008) proposed an alternative set of attributes including large scale atmospheric variables that influence precipitation processes. In their study the k-means clustering algorithm was employed to partition climate grid points into summer monsoon rainfall (SMR) regions in India using location parameters and a suite of atmospheric variables that drive SMR. Satyanarayana and Srinivas (2011) later extended their work to a fuzzy environment, where climate grid points are assigned to more than one region. Seasonality measures were used as input to the fuzzy *c*-means algorithm (Bezdek, 1981) in addition to the atmospheric variables and location parameters. More recently, Asong *et al.* (2015) conducted a study in the Canadian Prairie provinces where the fuzzy *c*-means algorithm was employed to partition climate stations into precipitation regions based on the similarity of atmospheric variables, teleconnection indices and location parameters.

Any set of attributes can be used as input to the pooling or clustering algorithms; however in order to obtain meaningful results it is important to select attributes that are relevant to the problem under consideration. It is therefore recommended to use variables that influence local precipitation processes. Statistical analyses are used to assess the significance of the relationship between precipitation and the potential attributes. Asong *et al.* (2015) performed principle component analysis and canonical correlation analysis to determine the inter-relationships between precipitation and a set of potential attributes including geographical site parameters and 21 atmospheric variables. The attributes that formed statistically significant relationships with the local precipitation were retained and used in the regionalization procedure. Selecting attributes that are physically meaningful to the problem under investigation can have several merits including reduced computational time and improved regional homogeneity (Wagener, 2004; Jafaar, 2011).

In Chapter 6, section 2 of this thesis the performances of four attribute sets are evaluated for their ability to form large regions with uniform precipitation statistics in a timely fashion. Topography, geographic location and atmospheric circulation patterns are considered to be the most influential factors affecting precipitation variability; therefore, location parameters and a suite of atmospheric variables are employed as potential attributes (Johnson and Hanson, 1995). The main objective of this investigation is to evaluate the need for a screening procedure and the selection of statistically relevant site attributes.

2.4 Temporal Resolution of Precipitation

The composition of precipitation regions depends upon several factors including the implemented time scale. Certain applications require precipitation to be obtained for a specific temporal resolution. Several uses of precipitation regions and their associated temporal resolutions are described below.

Homogeneous precipitation regions are often used in climate forecasting applications where relationships between precipitation and atmospheric variables are formed and used to project weather patterns. Projections are found to be more accurate over uniform regions than large heterogeneous areas. Long-term precipitation projections (annual, seasonal and monthly temporal scales) are used in applications involving water availability including the development of a water budget (Johnson and Hanson, 1995; Saikranith *et al.*, 2012). Short-term precipitation projections (hourly and sub-hourly temporal scales) are used in hydrologic model calibration and the estimation of soil erosion and infiltration rates for agricultural purposes (Jebari, 2007). Extreme hydrologic events such as flooding and drought are related to precipitation variability at different time scales; thus requiring a multi-temporal scale analysis to enhance the predictability of hydrologic models (Jiang *et al*, 2013). Regions of extreme precipitation are used to derive climatic design values such as the magnitude of the rainfall for a specified return period, as is done with Intensity-Frequency-Duration (IDF) curves. Precipitation data measured within the same pluviometric area is combined into a regional probability distribution from which the design value is attained (Burn, 2014).

Research has been conducted to study the effect of the temporal scale on the formation of precipitation regions. Saikranthi *et al.*, (2012) studied the spatial and temporal variability of monsoon rainfall over India using annual and seasonal resolutions. The spatial distribution of rainfall varied for the employment of the two time scales. Johnson and Hanson (1995) studied the relationships of daily, monthly and seasonal precipitation with topography and atmospheric variables over a mountainous area in Idaho, USA. Again, unique relationships were observed for the implementation of different time scales. Conversely, Jebari *et al.* (2007) found similarities between the distribution of sub-hourly and daily rainfall in Tunisia, thus permitting the disaggregation of relatively coarse to fine resolutions of precipitation within the study area.

An assessment of the effect of the temporal scale on the composition of the precipitation regions is presented in Chapter 6, section 3 of this thesis. Annual, seasonal and monthly resolutions are considered in addition to the maximum annual series that is used in design applications.

Chapter 3

3. Proposed regionalization model using a fuzzy clustering technique

In this section a new approach to the regionalization of precipitation using a fuzzy clustering technique is presented. First the need for a new method is explained; then the proposed model is introduced; and finally, a description of the model components is provided.

3.1 Problem Description

As discussed in Chapter 2, section 2 many researchers employ cluster validity indices (CVIs) to select the input parameters (the *c*-value and fuzzifier) to the fuzzy *c*-means clustering algorithm (Satyanarayana and Srinivas, 2011; Goyal and Gupta, 2014). CVIs are used to determine the values of the input parameters that provide a natural partitioning of the climate sites in the attribute space, and they are a substitute for the trial and error method that is computationally demanding. Several different validity indices are available and together they produce a wide range of results (Bhatia *et al.*, under review) that are highly uncertain. In addition they do not take into account the inherent grouping of the sites in terms of the variability of their precipitation records. As such, there is a great need for an alternative method to select input parameters to the fuzzy *c*-means algorithm that accounts for the quality of the regionalization outcomes (a high percentage of homogeneous precipitation regions and a large number of station-years per region) in a timely fashion.

3.2 Problem Objective

A model is presented for the regionalization of precipitation using a fuzzy clustering approach. It uses optimization to select input parameters (the *c*-value and fuzzifier) to the fuzzy *c*-means clustering algorithm such that two objectives are achieved in the results: (i) maximization of the numbers of station-years of the regional precipitation records; and (ii) maximization of the number of homogeneous precipitation regions. The model employs a programming technique that is comprised of two main steps: (i) differential evolution optimization (Storn and Price, 1995; Price and Storn, 1997) to generate a set of optimal solutions; and (ii) fuzzy Compromise programming (Bender and Simonovic, 2000) to assist in the selection of a single preferred solution that is the pair of input parameters that best satisfy the objectives of the regionalization output.

In Chapter 6, section 4, the model is demonstrated in the Great Lakes-St Lawrence lowlands climate region of Canada using monthly precipitation data that is organized into four seasons. Climate sites are partitioned into regions according to the similarity of their geographic locations including latitude, longitude and distance to major water bodies. The optimal solutions are ranked based on the preferences of three different decision makers: (i) Decision Maker 1 assigns equal criteria weights to the objective functions; (ii) Decision Maker 2 assigns a criteria weight of 1 to the percentage of regions with homogeneous precipitation statistics and a criteria weight of 2 to the minimum number of climate sites belonging to a region; and (iii) Decision Maker 3 allocates criteria weights that are opposite to Decision

Maker 2. Model output is presented and assessed for its ability to satisfy the objective functions and reflect the priorities of three unique decision makers.

3.3 Model Description and Theoretical Background

This section explains how systems analysis methods are applied to solve the presented regionalization problem. Justification for selection of the differential evolution and fuzzy Compromise programming methods for inclusion in the model is provided.

The systems approach to solving engineering problems utilizes mathematical models to represent a physical system. Three fundamental engineering systems techniques exist: (i) optimization; (ii) simulation; and (iii) multi-objective analysis. Optimization is the procedure for determining the set of decision variables (subject to constraints) that optimize an objective function. Simulation is used to assess a system's response to various input parameters. It is useful when the system is too complicated to obtain an optimal solution. Multi-objective analysis is similar to optimization; however there are several objective functions that are in conflict with one another. As such, a single optimal solution does not exist and a set of trade-off (non-dominated) solutions are generated instead (Simonovic, 2009). The models can be solved for using deterministic, probabilistic or fuzzy techniques. Deterministic models employ fixed, known decision variables and parameters while models that contain random or uncertain variables and parameters are classified as probabilistic or stochastic. Probabilistic methods require knowledge of historical events in order to develop a frequency distribution from which the probability of occurrence of an event is estimated. The amount of available historical data, however, is often insufficient to capture the true statistics of the event occurrence and as such, fuzzy set theory can be applied to address the issues of human input, subjectivity and deficient historical records. Fuzzy set theory was ultimately developed to incorporate the uncertainty that is caused by a lack of knowledge into a model and it is used to measure the degree to which an event occurs (Simonovic, 2009). The proposed systems tool uses a combination of optimization and fuzzy set theory to derive solutions to two objective functions simultaneously and to select a preferred outcome.

Several methods are available for solving optimization problems. The proposed regionalization model employs a combination of differential evolution (DE) and fuzzy Compromise programming. DE generates a set of solutions (decision variables subject to a set of defined constraints) that best satisfy the objective functions. Fuzzy Compromise programming is subsequently used to evaluate and rank the alternative solutions to assist in the selection of a single preferred solution.

The DE belongs to a family of evolutionary algorithms (EA). They are simple search mechanisms that mimic evolutionary theory and they work by generating an initial population of solutions (decision variables) subject to a set of defined constraints that are referred to as the parent solutions. Evolutionary operators (mutation and crossover functions) are employed to perturb and recombine the parent solutions to form a population of provisional solutions. The fitness levels of the parent and provisional solutions are computed. The fitness values provide a measure of the solutions' abilities to satisfy the objective functions. The parent and provisional solutions are compared and those that achieve a higher level of fitness are retained and become the next generation of parent solutions. The process continues until a maximum number of generations (as specified by the user) is reached or the solutions converge. The procedure is described in more detail in section 5.5.1.

EAs are systems tools that are capable of solving optimization problems that exhibit complexities such as very large search spaces, uncertainty, noise and disjoint solution sets (Coello Coello, 2007). They are also capable of simultaneously generating multiple solutions that are close to the global optimum (Simonovic, 2009). EAs are elected as the most appropriate method for incorporation in the proposed regionalization model for their ability to generate solutions to multiple objective functions simultaneously and to solve for the percentage of homogeneous precipitation regions (that involves a very complex procedure) without any simplification (see section 5.3 and 5.4). In addition, EAs are capable of generating solutions while maintaining discrete values of *c* (that is, the number of regions to which the climate sites are assigned). The fuzzy *c*-means algorithm requires discrete, whole numbers of the *c*-value in order to operate. The DE algorithm is chosen over the traditional genetic algorithm for its ability to provide better coverage of the outcomes of the objective functions corresponding to the optimal solutions (Schardong and Simonovic, 2011).

The proposed optimization technique produces a set of solutions that best satisfy at least one of the objective functions. A single ideal solution that optimizes all objective functions does not exist since the ideal point (that provides optimal solutions for all objectives) is infeasible for a convex region. In order to select the preferred solution, the Compromise programming method is employed to evaluate and rank the optimal (alternative) solutions based on their closeness to the ideal solution in the objective space. Closeness is measured using a distance metric. The Compromise programming method also incorporates criteria weights (that provide an indication of the relative importance of the objective functions) into the evaluation of alternative solutions (Simonovic, 2009).

Fuzzy Compromise programming is used to evaluate and rank the set of alternative solutions for the proposed regionalization model. It is an extension of the traditional Compromise programming method to a fuzzy environment where the inputs including the objective functions, decision maker preferences (criteria weights) and parameter of the distance metric (*p* - exponent) are represented by fuzzy sets opposed to crisp values. Fuzzy sets are "characterized by a membership function mapping the elements of a domain, space or universe of discourse *X* to the unit interval [0,1]" (Simonovic, 2009). The domain signifies the range of plausible values of the inputs and the membership values indicate the degree to which the inputs belong to the domain. Fuzzy Compromise programming provides an advantage of incorporating the uncertainty of the objective function, criteria weights and the *p*-values to the distance metric. The ideal and alternative solutions are represented by small regions (membership functions) in the decision space as opposed to points. As such, the alternative solutions are evaluated using a fuzzy distance metric that computes the distance between the weighted centres of gravity of the ideal and alternative solutions. A detailed explanation of the procedure is provided in section 5.5.2.

Chapter 4

4. Application

4.1 Study Area

The case studies are applied in two Canadian climate regions including the Prairie and the Great Lakes-St. Lawrence lowlands. The diverse climates of the Prairie and Great Lakes regions are attributed to differences in latitude, topography and their proximity to major water bodies. Application of the methodology (presented in section 5) in two climatically different areas introduces additional variability to the regionalization procedure, and understanding the effect of the choice of the regionalization method, climate site attributes and the temporal scale of the precipitation data in different climates can provide valuable information to hydro-climatologists.

The study areas are selected such that the processes that drive precipitation and therefore, the structure of the attribute sets are consistent (although their magnitudes vary). In order to identify these areas, linear correlation coefficients are calculated between a suite of atmospheric variables and precipitation recorded at a 2 by 2-degree grid that extends across Canada. The precipitation values are extracted from a high resolution gridded dataset (that is the ANUSPLIN dataset introduced in Chapter 4.2). A coarse resolution is used to improve the computational efficiency of the analysis. The correlation coefficients are assigned as attributes to the climate grid points and employed as input to the fuzzy *c*-means algorithm. Clustering algorithms partition objects that share similar characteristics into homogeneous regions and therefore, in this analysis the fuzzy *c*-means algorithm groups climate sites that exhibit comparable relationships between their precipitation records and atmospheric variables (that are represented by linear correlation coefficients) into coherent regions. The atmospheric variables employed in this analysis are listed in Chapter 4.2. Figure 1 presents the resultant regions of uniform precipitation drivers on a map of Canada. Colours are used to distinguish the regional boundaries and the colours to which the regions are assigned are not significant. It is evident that the outcomes resemble the Canadian climate regions shown in Figure 2. As such, two unique Canadian climate regions are selected as reasonable study areas for the present investigation in which the structure of the attribute sets are consistent. If the study areas were selected such that they crossed the regional boundaries presented in Figure 1, the site attributes would consist of different combinations of variables and therefore, the distance metric (that computes the similarity between sites of the study area) of the clustering algorithm could not be computed.

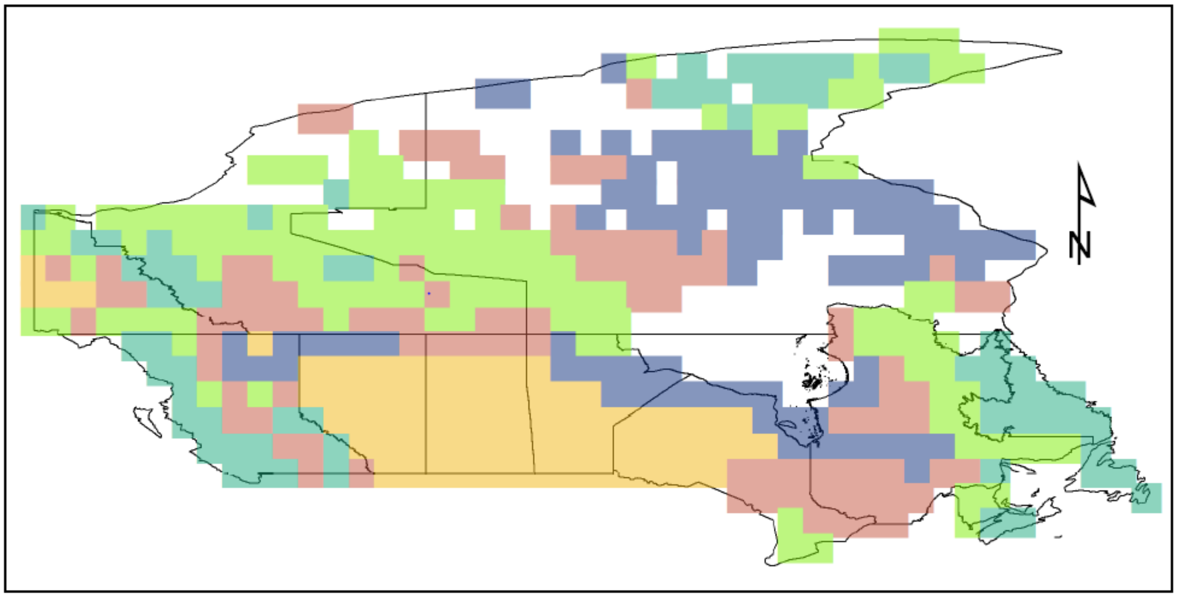
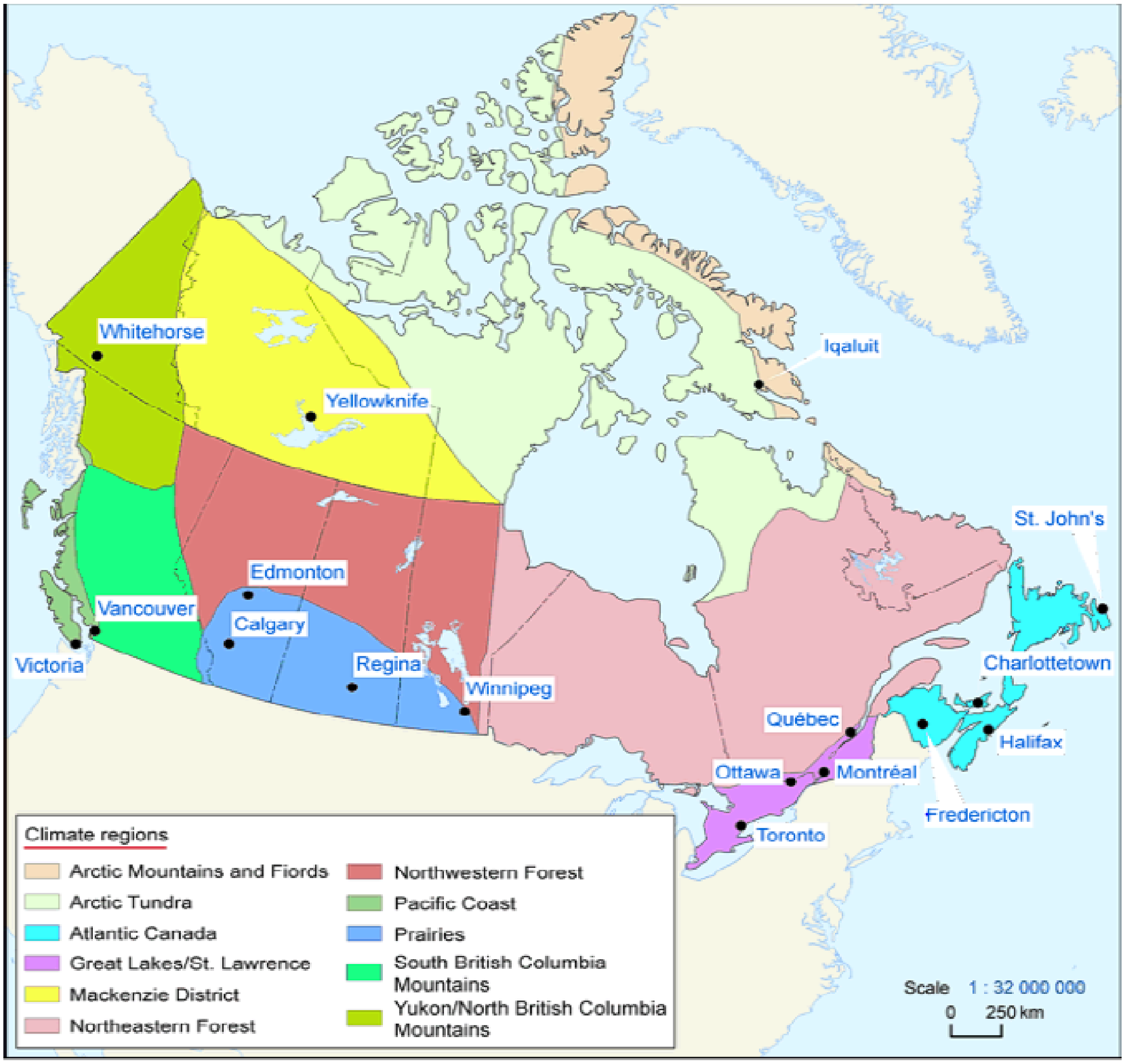


Figure 1: Results for regionalization of linear correlation coefficients calculated between large scale atmospheric variables and precipitation for Canada.

Descriptions of the study areas are provided herein: The Prairie region lies East of the Rocky Mountain range along the Southern provincial borders of Alberta, Saskatchewan and Manitoba. Its geographic position extends from approximately 49 to 54 degrees latitude and -117 to -95 degrees longitude, and it is shown as the blue area on the Canadian Climate regions map in Figure 2. The regional climate is classified as continental because of the absence of major water bodies (oceans, large lakes) that typically generate moderate temperatures and increased precipitation; as such the climate is dry and temperatures are more extreme. The climate is predominantly driven by large scale atmospheric circulation patterns and the area receives approximately 300 – 550 mm of annual precipitation (McGinn, 2010).



Source(s): Environment Canada, Atmospheric Environment Service, Climate Research Branch, 1998, Climate Trends and Variations Bulletin for Canada, Ottawa.

Figure 2: Map depicting the Canadian Climate Regions (Source: Statistics Canada, 2012).

The Great Lakes-St. Lawrence lowlands region is located along the Southern provincial borders of Ontario and Quebec. It is bounded by Lake Huron and Georgian Bay to the West, and Lake Erie and Lake Ontario to the South and East. Its geographic position is approximately 42 to 48 degrees latitude and -83 to -70 degrees longitude. The prevailing winds from the West, humid air from the Gulf of Mexico and cold, dry air from the North significantly influence the regional climate in addition to the presence of the Great Lakes and their interactions with the lower atmosphere (USEPA, 2012). Lake effect precipitation is common during the fall and winter seasons when the temperatures of the lake decrease at a slower rate than the surrounding air. This process occurs when a cold air mass passes over the relatively warm lakes and a significant amount of moisture is evaporated, held in the lower atmosphere and precipitated downwind of the lakeshore often in the form of snow (Sousounis, 2001; Lapen and Hayhoe, 2003). In the summer season convective rainfall and thunderstorms are typical in the Great Lakes region (Ashmore and Church, 2001). In the late spring and early summer season the relatively cool lake temperatures have a stabilizing effect on the lower atmosphere and reduce the magnitude of convective rainfall by approximately 10 – 20% over and downwind of the lakes (Scott and Huff, 1996).

4.2 Data

Four main datasets are used in this study: (i) atmospheric variables obtained from the Canadian CanESM2 Earth Systems model (http://www.cccma.ec.gc.ca/ last accessed Nov, 2014); (ii) ANUSPLIN precipitation data that has been interpolated to a high resolution grid (Hutchinson *et al.*, 2009; Hopkinson *et al.*, 2011); (iii) elevation data extracted from digital elevation models (DEMs) in ArcGIS 10.2 (http://resources.arcgis.com/ last accessed Nov, 2014) and (iv) a shape-file containing the geographical locations of major inland water bodies in Canada (http://geo2.scholarsportal.info/ last accessed Feb, 2015). The atmospheric variables considered as potential attributes include air temperature, geopotential height, specific humidity, relative humidity, and the Northward and Eastward wind components.

Most weather occurs in the troposphere that extends from the Earth's surface to an altitude of approximately 12 km. The air pressure ranges from approximately 1000 – 200 mb (100 – 20 kPa) and therefore, the atmospheric variables considered in the analysis are obtained for pressure levels of 20, 30, 40 50, 60, 70, 85, 92.5 and 100 kPa

The ANUSPLIN high resolution gridded precipitation dataset is used in this investigation. Hutchinson *et al.,* (2009) generated the dataset using a trivariate thin-plate smoothing spline technique to interpolate daily precipitation recorded at Canadian climate stations to a 10 by 10 km grid that covers the country. The dataset was generated for a time period of 1961 to 2003. The dataset was further improved by reducing the residuals between the observed and interpolated gridded values (Hopkinson *et al.*, 2011). This was achieved by correcting the alignment between the climatological days of the observed data recorded at different climate stations. The temporal window was also expanded to 1950 to 2011. Gridded datasets are preferred over other observed precipitation datasets because of their good spatial coverage and complete, consistently generated record lengths. For a comparative analysis between the ANUSPLIN dataset and other sources of gridded precipitation data for Canada such as the North American Regional Reanalysis (NARR) dataset and the Canadian Precipitation Analysis (CaPA) dataset refer to Eum *et al.* (2014).

Latitude and longitude are derived from the ANUSPLIN dataset while elevation values are extracted from DEM files that are obtained from the Canadian GeoBase website (http://www.geobase.ca/ last accessed Nov 2014). The DEM files are imported to an ArcGIS environment and mosaicked (merged) to form one continuous layer for each study area. The positions of the ANUSPLIN grid points are also imported to ArcGIS and the elevation data are assigned to the points according to their spatial relationship. The shape-file containing geographic information of Canadian water bodies is used to calculate the minimum distance between the grid points and major water bodies located upwind; that is the additional location attribute used for the Great Lakes study area. Distance to water bodies is included as a location parameter for the Great Lakes region because the presence of the lakes has a significant influence on the regional precipitation. Since the prevailing winds flow from the Northwest direction the significant water bodies located on the windward side of the study area are considered to be either Lake Huron or Georgian Bay.

For the assessment of the influence of the choice of regionalization method and the attribute sets on the regionalization outcomes, the analysis calls for all datasets to be converted to a monthly temporal resolution and analyzed for four separate seasons: (i) December, January, February (DJF); (ii) March, April, May (MAM); (iii) June, July, August (JJA); and (iv) September, October, November (SON). Atmospheric variables are extracted for a monthly temporal resolution and used directly. Precipitation data are available for a daily resolution and the values are added together to obtain monthly values. For the evaluation of the effect of the temporal resolution on the regionalization results monthly, seasonal and annual resolutions are considered in addition to the maximum annual series of the data. The daily ANUSPLIN precipitation data is added together to obtain monthly, seasonal and annual precipitation values. The maximum annual series record is computed as the maximum daily precipitation magnitude within each year. All data are obtained for a 55 year time period ranging from 1951 to 2005. The Great Lakes and Prairie study areas contain 959 and 2674 ANUSPLIN grid points, respectively.

Chapter 5

5. Methodology

In this section the methodology employed in the presented research is explained in four main steps: (i) formation of the attribute sets; (ii) determination of the preferred number of clusters to which the sites are assigned (selection of the *c*-value); (iii) regionalization of precipitation using the fuzzy *c*-means clustering algorithm; and (iv) validation of regional homogeneity using *L*-moment statistics.

5.1 Formation of Attribute Sets

The attribute sets are prepared using location parameter and atmospheric variable records.

1) In the Prairie region AS-1 and AS-2 are simply matrices containing the latitude, longitude and the latitude, longitude and elevation of the grid points, respectively.

2) In the Great Lakes region AS-1 contains latitude, longitude and distance to water bodies parameters; AS-2 includes the latitude, longitude, distance to water bodies and elevation parameters. The distance to water bodies parameter is calculated as the minimum distance between the grid point and the significant water body located upwind; see [Eq. 1]:

$$dist = \min_{1 \le i \le length(xv)} \{ \sqrt{(xv_i - xq)^2 + (yv_i - yq)^2} \} \qquad (1)$$

where, *xv* and *yv* are vectors containing the longitudinal and latitudinal coordinates of the perimeter of the water body, respectively, and i represents a single point on the perimeter; xq and yq are the longitudinal and latitudinal coordinates of the grid points.

3) AS-3 is composed of the atmospheric variables that form strong linear correlations with precipitation for the Prairie and Great Lakes regions. Atmospheric variables are interpolated from the global climate model (GCM) grid to the ANUSPLIN grid using the Inverse Distance Weighting method as shown in [Eq. 2] and [Eq. 3].

$$w_i = \frac{{}^{1}\!/_{d_i^2}}{\sum_{i=1}^{4}({}^{1}\!/_{d_i^2})} \tag{2}$$

$$V = \sum_{i=1}^{k} w_i v_i \tag{3}$$

where the distance between the i^{th} GCM grid point and the ANUSPLIN grid point of interest is represented by d_i; v_i and w_i are the magnitude of the atmospheric variable at the i^{th} GCM grid point and the weight assigned to it, respectively. The spatially interpolated variables are represented by V.

Linear correlation coefficients are calculated between precipitation and the atmospheric variables at each ANUSPLIN grid point (or climate site) using the Pearson correlation formula.

4) AS-4 is composed of the values of all atmospheric variables (air temperature, geopotential height, relative humidity, specific humidity and the Northward and Eastward wind components) modeled at nine pressure levels (20, 30, 40, 50, 60, 70, 85, 92.5 100 kPa) for all sites in the Prairie and Great Lakes regions.

5.2 Determination of the preferred number of clusters (*c*-value selection)

The number of regions for the ANUSPLIN grid points to be partitioned into, is an important input parameter to the fuzzy *c*-means clustering algorithm that must be determined prior to regionalization.

1) The number of regions (the *c*-value parameter) is varied between 5 to 50 and 5 to 100 in the Great Lakes and Prairie regions, respectively and used as input to the fuzzy *c*-means algorithm (see section 5.3). The resultant regions are validated using the *L*-moment regional heterogeneity test (see section 5.4).

2) The percentages of regions that are classified as homogeneous (for each partitioning) in the validation procedure are computed for a range of *c*-values. The lowest *c*-values to attain a minimum of 80% regional homogeneity are selected as the preferred number of regions for the sites to be partitioned into. The threshold value of 80% is selected as a suitable trade-off that provides for a small number of clusters (and therefore larger number of station-years of the precipitation record) and a high

percentage of regions with homogeneous precipitation statistics for a single partitioning. (Further justification for selection of the 80% threshold value is provided in Chapter 6.2.2).

5.3 Regionalization of Precipitation (Fitness Function Part I)

This section explains the fuzzy *c*-means clustering process. There are *N* sites that are to be assigned to *c* clusters. Each site has one feature vector that contains M attributes (of an attribute set) that are a combination of atmospheric variables and location parameters. The procedure is as follows:

1) Rescale the attributes of the feature vectors in order to standardize their variance and magnitude, otherwise variables that are larger in magnitude will have a greater influence on the resultant clusters (Satyanarayana and Srinivas, 2011).

$$z_{ji} = \frac{(y_{ji} - \overline{y_j})}{\sigma_j} \quad i \in \{1, \dots, N\}; j \in \{1, \dots, M\} \tag{4}$$

where z_{ji} is the rescaled value of y_{ji} for attribute j and site i; $\overline{y_j}$ and σ_j are the mean and standard deviation of attribute j for all sites, respectively.

2) Initialize the c cluster centroids and assign each site to the closest centre that is measured using the squared Euclidean distance metric. At each step the cluster

centroids are updated and the sites may be re-assigned in order to minimize the objective function presented in [Eq. 5], [Eq. 6] and [Eq. 7]:

$$J = \sum_{k=1}^{c} \sum_{i=1}^{M} u_{ik}^{m} || z_i - C_k ||^2 \tag{5}$$

$$u_{ik} = \frac{1}{\sum_{l=1}^{c} \frac{|z_i - c_k|}{|z_i - c_i|}^{2/(m-1)}} \tag{6}$$

$$C_k = \sum_{i=1}^{N} \frac{u_{ik}^{m} z_i}{\sum_{i=1}^{N} u_{ik}^{m}} \tag{7}$$

where *J* is the value of the objective function; z_i is the feature vector (attribute set) of site *i*; C_k is the centroid of cluster *k;* u_{ik} is the degree of membership of site z_i in cluster *k*; and *m* is a weight exponent of fuzzy membership (the fuzzifier) that is equal to 2 for the first three assessments and solved for using optimization in the proposed regionalization model.

Repeat the previous step for the same value of *c* until the objective function converges to a minimum value, known as the global minimum.

3) The fuzzy *c*-means algorithm produces a matrix that contains the climate site membership values; that is, the degree that the climate sites belong to each cluster. Climate sites are assigned to the cluster in which their membership value exceeds the defined threshold criteria, thereby hardening the fuzzy clusters; see [Eq. 8] (Satyanarayana and Srinivas 2011):

$$T_i = \max\{\frac{1}{c}, \frac{1}{2}[max_{1 \le k \le c}(u_{ik})]\} \quad (8)$$

where T_i is the defined threshold value.

5.4 Validation of Regional Homogeneity (Fitness Function Part II)

A test based on *L*-moment statistics is used to validate the regional homogeneity of precipitation. *L*-moments describe the probability distribution of the dataset from which they are calculated.

The site to site variability of the sample *L*-moment ratios (*L*-moment ratio of scale (*L*-Cv), *L*-Skewness, *L*-Kurtosis) that are calculated from the observed precipitation records provide three separate measures of regional heterogeneity. The metric utilizing the variability of *L*-Cv has proven to be the most useful indicator of heterogeneity. Its value is denoted by H_1 and the procedure for its computation is presented below. The methodology and equations are adopted from Hosking and Wallis (1997).

1) Rank the climate data for each member site in ascending order, then compute *L*-moment ratios for scale (t), skewness (t_3) and kurtosis (t_4) as follows:

$$t = \frac{l_2}{l_1} = \frac{(2b_1 - b_0)}{b_0} \quad (9)$$

$$t_3 \ = \frac{l_3}{l_2} \ = \frac{(6b_2 \ - 6b_1 + b_0)}{(2b_1 - b_0)} \qquad (10)$$

$$t_4 \ = \frac{l_4}{l_2} \ = \frac{(20b_3 \ - 30b_2 + 12b_1 - b_0)}{(2b_1 - b_0)} \qquad (11)$$

where,

$$b_0 \ = l = \ n^{-1}\sum_{j=1}^{n} x_j$$

$$b_1 \ = \ n^{-1}\sum_{j=2}^{n} x_j [\frac{(j-1)}{(n-1)}]$$

$$b_2 \ = \ n^{-1}\sum_{j=3}^{n} x_j [\frac{(j-1)(j-2)}{(n-1)(n-2)}]$$

$$b_3 \ = \ n^{-1}\sum_{j=4}^{n} x_j [\frac{(j-1)(j-2)(j-3)}{(n-1)(n-2)(n-3)}]$$

where *x* is precipitation measured at a single site and *n* is the record length.

2) To measure heterogeneity of a cluster, compare the observed between site dispersion to the between site dispersion that would be expected from a homogeneous cluster. Between site dispersion is measured as the standard deviation of the *L*-moment ratio measure of scale (*L*-Cv) for all sites in the cluster, which is represented by V_1.

$$V_1 \ = \ \{\frac{\sum_{i=1}^{N_C} n_i(t_i - t^R)^2}{\sum_{i=1}^{N_C} n_i}\}^{1/2} \qquad (12)$$

where N_c is the number of sites in a cluster; n is the site record length; t_i is the L-moment ratio measure of scale for site *i;* and t^R is the regionally averaged L-moment ratio of scale.

3) Establish a homogeneous region for comparison. Compute the regional average L-moment ratios for the cluster, and fit the average ratios to a kappa distribution. The regional L-moment ratios are weighted based on the sites' record lengths and are calculated as follows:

$$t^R = \frac{\sum_{i=1}^{Nc} n_i(t_i)}{\sum_{i=1}^{Nc} n_i} \qquad (13)$$

$$t_3^R = \frac{\sum_{i=1}^{Nc} n_i(t_{3i})}{\sum_{i=1}^{Nc} n_i} \qquad (14)$$

$$t_4^R = \frac{\sum_{i=1}^{Nc} n_i(t_{4i})}{\sum_{i=1}^{Nc} n_i} \qquad (15)$$

4) Simulate N_{sim} realizations of the observed region from the kappa distribution. N_{sim} is typically a large number; i.e. 500. Compute the between site dispersion (V_1) for each set of the simulated sites that together are considered to be homogeneous.

5) Evaluate the homogeneity of the cluster using the homogeneity measure (H_1) where μ_V and σ_v are the mean and standard deviation of the N_{sim}values of V_1:

$$H_1 = \frac{(V_1 - \mu_V)}{\sigma_v} \qquad (16)$$

6) Apply the corrective measure proposed by Castellarin *et al*., (2008) to account for the effect of inter-site cross-correlations on the outcomes of the *L*-moment regional heterogeneity test.

$$H_{1,adj} = H_1 + 0.122\ x\ \overline{p^2}(N_c - 1) \tag{17}$$

where, $\overline{p^2}$ is the mean of squares of the cross-correlations of the precipitation records that is computed for all N_c climate sites.

7) Accept the cluster as homogeneous if $H_{1,adj} < 1$; reject the cluster as heterogeneous if $H_{1,adj} \geq 2$. When $1 \leq H_{1,adj} < 2$ the cluster is considered to be possibly heterogeneous. For this analysis all clusters with corresponding $H_{1,adj}$values equal to or greater than 1 are considered to be heterogeneous.

5.5 Proposed Approach to Regionalization of Precipitation using Fuzzy Cluster Analysis

In this section the methodology of the proposed regionalization model is explained in two main steps: (i) differential evolution to generate a set of optimal, alternative solutions; and (ii) fuzzy Compromise programming to rank the alternative solutions.

5.5.1 Differential Evolution

The DE optimization process is described below. The goal is to solve for the decision variables (the value of c and the fuzzifier) that are subject to a set of constraints that optimize the objective functions. A mathematical representation of the optimization problem is presented below:

$$z_1 = \max\left(\min\left(n_{sites,i}\right)_g\right) \qquad i = 1, 2, \ldots, c; \quad g = 1, 2, \ldots, G \tag{18}$$

$$z_2 = \max\left(\frac{\sum_{i=1}^{c} n_{H,i}}{c}\right)_g \qquad i = 1, 2, \ldots, c; \quad g = 1, 2, \ldots, G \tag{19}$$

$$n_{H,i} = \begin{bmatrix} 1 & \quad if\ H_{1i,} \leq 1 \\ 0 & \quad otherwise \end{bmatrix} \qquad i = 1, 2, \ldots, c \tag{20}$$

where, z_1 and z_2 represent the objective functions; $n_{sites,i,g}$ represents the number of climate sites in a region; n_H represents the number of climate regions that are classified as homogeneous; and G is the number of generations for which the evolutionary algorithm is performed.

1) Randomly generate an initial population of solutions (vectors of decision variables) that are subject to the defined constraints.
2) Compute the fitness of the initial (parent) solutions; that is, the values of the objective functions for each set of solutions. Refer to section 5.3 and 5.4 for a detailed description of the fitness function.
3) Employ evolutionary operators (mutation and crossover) to generate a set of provisional solutions in an attempt to improve the overall level of fitness of the current population. Mutation is performed first. It involves the perturbation of a

solution by applying the weighted difference between two randomly selected parent solutions; the weight is referred to as the scaling factor shown in [Eq. 21]. The scaling factor is a dynamic parameter that linearly varies between 0.6 to 0.4 for each generation as recommended by Schardong and Simonovic (2011). The linear variation of the scaling factor helps to reduce the probability of stagnation that occurs when the solutions' fitness levels do not improve over several generations and the solutions have not yet converged.

$$SF_g = 0.6 - (0.6 - 0.4\ x\ \tfrac{g}{G}) \qquad g = 1, 2, \ldots, G \qquad (21)$$

where *SF* represents the scaling factor; *g* is the number of the current generation; and *G* is the number of maximum generations.

The mutation calculation is shown in [Eq. 22] where *Mu* is the mutant solution; *Pa* is the parent solution; *i* is the current solution vector; *D* is the total number of solutions in the current generation; and *r0, r1* and *r2* are indices that correspond to three randomly chosen solutions of the parent population. Ensure that the *c*-values are whole numbers after performing mutation.

$$Mu_{i,g} = Pa_{r0} + SF\ x\ (Pa_{r1} - Pa_{r2}) \qquad i = 1, 2, \ldots, D; \ \ g = 1, 2, \ldots, G \qquad (22)$$

4) Perform the crossover operator where randomly selected mutant solutions are exchanged with their parent counterpart to complete the formation of the provisional population. A randomly generated number between 0 and 1 is compared to a

crossover ratio that is presented in [Eq. 23]. If the randomly generated number is lower than the crossover ratio the parent solution is replaced by its corresponding mutant value. Otherwise the parent solution remains in the provisional population. See [Eq. 24] for the crossover formula. Similar to the scaling factor of the mutation equation the cross over ratio linearly varies between 1 and 0.8 for each generation (Schardong and Simonovic, 2011).

$$CR = 1 - (1 - 0.8\, x\, \tfrac{g}{G}) \qquad g = 1, 2, \ldots, G \tag{23}$$

where *CR* is the crossover ratio.

$$PS_{i,g} = \begin{bmatrix} Mu_{i,g} & if\ rand(0,1) \leq CR \\ Pa_{i,g} & otherwise \end{bmatrix} \qquad i = 1, 2, \ldots, D; \ \ g = 1, 2, \ldots, G \tag{24}$$

where *PS* represents the provisional solution.

5) Calculate the fitness of the solutions in the provisional population using the fitness function (see sections 5.3 and 5.4).
6) Introduce the selection criterion. The fitness levels of the provisional solutions are compared with their corresponding parent solution. Those that exhibit a higher level of fitness are retained and used to form the new generation of solutions as shown in [Eq. 25].

$$Pa_{i,g+1} = \begin{bmatrix} PS_{i,g} & if\ f\left(PS_{i,g}\right) \leq f(S_{i,g}) \\ S_{i,g} & otherwise \end{bmatrix} \qquad i = 1, 2, \ldots, D; \ \ g = 1, 2, \ldots, G \tag{25}$$

where *f(P)* and f(*PS)* represent the fitness of the parent and provisional solutions, respectively.

7) Steps 3 to 6 are repeated until the stopping criterion is reached; that is the maximum number of generations (*G*, specified by the decision maker) or solution convergence.

5.5.2 Fuzzy Compromise Programming

The DE process produces a set of alternative, optimal solutions that have acquired the highest levels of fitness. Fuzzy Compromise programming is subsequently employed to evaluate and rank the alternative solutions according to the preferences of the decision maker. The procedure is explained below.

8) Membership functions (MFs) are created to represent the objective functions. The MFs for the maximization of the number of regions with homogeneous precipitation statistics and the maximization of the minimum number of climate stations (station-years) per region are represented by [Eq. 26] and [Eq. 27], respectively. Compute the degrees of membership for the solutions in their respective MFs.

$$MF_{1,i} = \left[\begin{array}{ll} \frac{nd_{1,i}}{80} & if\ nd_1 < 80 \\ 1 & otherwise \end{array}\right. \qquad for\ i = 1, 2, \dots, NDS \qquad (26)$$

$$MF_{2,i} = \left[\begin{array}{ll} \frac{nd_{2,i}}{50} & if\ nd_2 < 50 \\ 1 & otherwise \end{array}\right. \qquad for\ i = 1, 2, \dots, NDS \qquad (27)$$

where MF_1 and MF_2 represent the membership functions of the first and second objective functions; nd_1 and nd_2 stand for the value of the alternative solution for their corresponding objective function; i signifies the current solution vector and *NDS* is the total number of solutions in the set.

9) The membership values of the alternative solutions for each criterion (objective function) are used as input to the fuzzy Compromise programming software in addition to the *p* exponent of the fuzzy distance metric and the two criteria weights. The value of *p* is provided as a triangular membership function with three defining parameters equal to 1, 2 and 10 respectively (Simonovic, 2009).

10) Run the fuzzy Compromise programming software to evaluate and rank the alternative solutions according to their proximity to the ideal solution (an infeasible solution that optimizes both objective functions). The fuzzy distance metric that is employed to compute the distance between the alternative and ideal solutions is reproduced from Simonovic (2009) and presented in [Eq. 28].

$$\widetilde{L_p} = \left[\sum \tilde{\alpha}_i^{\tilde{P}} \left(\frac{\widetilde{Z_\iota^*} - \tilde{Z}_\iota(S)}{\widetilde{Z_\iota^*} - \widetilde{Z_\iota^{**}}}\right)^{\tilde{P}}\right]^{1/\tilde{P}} \tag{28}$$

where $\widetilde{L_p}$ is the fuzzy distance metric; $\tilde{\alpha}$ is the fuzzy criteria weight; $\tilde{p}$ is a fuzzy parameter that represents the significance of the maximum deviation from the ideal point (deviations are assigned equal weight for values of p equal to 1; otherwise deviations are weighted according to the magnitude of their deviation); Z represents the ideals of the solutions (Simonovic, 2009).

Chapter 6

6. Results and Discussion

The results for the assessments of the impacts of the regionalization components (method, site attributes, temporal resolution) on the formation of precipitation regions are presented here. The regionalization outcomes are evaluated for their ability to achieve the following objective criteria: (i) maximization of the number of station-years in the regional precipitation record (minimization of the *c*-value); and (ii) maximization of the percentage of homogeneous precipitation regions. Subsequently the proposed regionalization model is evaluated for its ability to attain the aforementioned objective criteria for the regionalization output and to satisfy the decision maker preferences.

6.1 Regionalization Method

Justification for application of the fuzzy *c*-means clustering algorithm is provided in this section. The k-means and fuzzy *c*-means algorithms are compared for their ability to attain the objective criteria in the regionalization output. The algorithms are employed several times, using a range of parameter values to represent the number of clusters to which the climate sites are assigned. The lowest number of regions to achieve a minimum of 80% regional homogeneity is elected as the preferred result (refer to section 6.2.2 for a detailed explanation of the selection procedure).

Parameters *k* and *c* represent the number of clusters to which the climate sites are assigned for the k-means and fuzzy *c*-means algorithms, respectively. Parameter values that first reach the minimum requirement for regional homogeneity are summarized in Table 1 and Table 2 for the Great Lakes and Prairie regions, respectively. The performances are evaluated for three of the four attribute sets that are considered in the selection of climate site attributes assessment including: AS-1 (latitude, longitude, distance to water bodies); AS-3 (atmospheric variables that form strong linear correlations with the local precipitation); and AS-4 (a comprehensive set of atmospheric variables that are recorded for a range of pressure levels). The precipitation data is obtained for a monthly resolution and organized into four seasons including: DJF (December, January, February); MAM (March, April, May), JJA (June, July, August); and SON (September, October, November).

Table 1: Number of regions required to attain a minimum of 80% homogeneous precipitation regions in the Great Lakes region. A comparison of outcomes from the k-means and fuzzy *c*-means algorithms.

	AS-1		AS-3		AS-4	
	k	**c**	**K**	**c**	**k**	**c**
DJF	30	30	36	36	36	34
MAM	18	15	15	17	10	15
JJA	15	12	-	-	15	15
SON	17	15	-	-	19	19

Table 2: Number of regions required to attain a minimum of 80% homogeneous precipitation regions in the Prairie region. A comparison of outcomes from the k-means and fuzzy *c*-means algorithms.

	AS-1		AS-3		AS-4	
	k	**c**	**K**	**c**	**k**	**c**
DJF	76	70	-	-	100	90
MAM	60	55	65	62	65	55
JJA	69	65	-	-	83	80
SON	70	63	64	61	65	55

For instances where the *c*-value is less than the *k*-value the results are highlighted in green; equal results are coloured in yellow; and for the occurrences where *k* is less than the *c*-value the results are highlighted in red. Evidently the fuzzy *c*-means algorithm outperforms the k-means algorithm for the majority of cases. Table 1 reveals that the fuzzy *c*-means algorithm requires fewer regions to attain the minimum homogeneity requirement using AS-1 for the MAM, JJA and SON seasons; in addition to using AS-4 in the DJF season. The k-means algorithm only outperforms the fuzzy *c*-means algorithm for the MAM season using AS-3 and AS-4. The algorithms produce the same results for all other combinations of site attributes and seasons. Table 2 reveals that the fuzzy *c*-means algorithm is even more successful in the Prairie region as it outperforms the k-means algorithm for all scenarios. As such, the fuzzy *c*-means algorithm has proven to be the most suitable regionalization method for the presented research.

6.2 Selection of Site Attributes

In this section four different combinations of site attributes are assessed for their abilities to form high quality precipitation regions in a timely fashion. The attribute sets used are listed in Chapter 6.1 in addition to AS-2 that consists of the location parameters in AS-1 and site elevation. The main objective of this investigation is to evaluate the importance of the attribute selection process. The fuzzy *c*-means algorithm is employed to partition the climate sites into regions that are subsequently validated using observed precipitation and an *L*-moment regional heterogeneity test. All data is converted to a monthly temporal resolution and organized into four seasons for evaluation including the DJF, MAM, JJA and SON seasons.

A description of the correlation analysis results is presented first. The outcomes guide the formation of AS-3. Afterward, the results for the determination of the preferred number of regions to which the sites are assigned, and maps of the corresponding precipitation regions are revealed for all combinations of attribute sets, seasons and study areas. Finally, significant findings are highlighted and discussed.

6.2.1 Formation of Attribute Sets (Attribute Selection)

The composition of Attribute Set 3 (AS-3) is driven by the outcomes of linear correlation analysis that are presented in this section. More specifically, AS-3 contains the atmospheric variables that form strong linear correlations with the local precipitation. Correlation coefficients that are greater than 0.2 across the study area are considered to be strong. A coefficient of 0.2 is selected because very few values exceed 0.3 across the Great Lakes region. AS-3 is the only attribute set that has a temporally and spatially inconsistent structure because the magnitudes of the correlation coefficients change according to the season and study area.

Colour plots (Figures A32 - A39 - provided in Appendix A) are used to display the correlation coefficients for each climate region, season and atmospheric variable. The x and y axes represent the number of climate sites and pressure levels, respectively. The magnitudes of the correlation coefficients for all sites and pressure levels are represented by colours corresponding to a scale where red and blue represent the positive and negative limits, respectively. The results of the correlation analysis (lists of atmospheric variables that comprise of AS-3) are summarized in Table 3. In the Great Lakes study area and the DJF season air temperature and the Northward wind component modeled at 30 kPa and 20 kPa, respectively, form relatively strong linear correlations with precipitation over a significant portion of the climate region. In the MAM season geopotential height modeled at pressure levels that range from 20 to 70 kPa are well correlated with precipitation. Alternatively, the colour plots reveal that all atmospheric variables are poorly correlated with precipitation in the JJA and SON seasons. As a result AS-3 can only be derived for the DJF and MAM seasons in the Great Lakes region. This may be explained by the fact that atmospheric variables do not drive the local precipitation during these seasons (local influences such as topography may be more influential); alternatively, linear correlations may not be an adequate measure of the relationship formed between the variables for these periods.

Table 3: List of atmospheric variables that comprise of Attribute Set 3 (AS-3).

GL-DJF	va (20, 30 kPa), ta (30 kPa)
GL-MAM	zg (20, 30, 40, 50, 60, 70 kPa)
GL-JJA	no correlations
GL-SON	no correlations
Prairie-DJF	no correlations
Prairie-MAM	hus (all), ta (all), zg (20, 30, 40, 50, 60, 70, 85 kPa)
Prairie-JJA	no correlations
Prairie-SON	hus (all), ta (all), zg (20, 30, 40, 50, 60, 70, 85 kPa), ua (85, 92.5, 100 kPa)

Hur	- relative humidity
Hus	- specific humidity
Ta	- air temperature
Zg	- geopotential height
Va	- Northward wind component
Ua	- Eastward wind component

Similarly, in the Prairie region AS-3 is only derived for two seasons; MAM and SON. In the MAM season specific humidity and air temperature modeled at pressure levels extending from 20 to 100 kPa and geopotential height modeled at pressure levels of 20 to 85 kPa exhibit strong linear relationships with the local precipitation across the majority of the climate region. The same set of atmospheric variables is selected for the SON season in addition to the Eastward wind components modeled at pressure levels of 85, 92.5 and 100 kPa.

6.2.2 Determination of the Preferred Number of Regions (Attribute Selection)

The number of clusters (the *c*-value) for the climate sites to be partitioned into is an important input parameter to the fuzzy *c*-means algorithm. Its magnitude significantly impacts the composition and spatial distribution of the resultant precipitation regions. As such, sufficient reasoning should be applied to select an appropriate value.

In this analysis the *c*-values are selected as the minimum numbers to achieve 80% regional homogeneity. That is, at least 80% of the resultant precipitation regions are classified as homogeneous according to the *L*-moment heterogeneity test. The *c*-values are varied between 5 to 50 and 5 to 100 (at increments of 5) for the Great Lakes and Prairie study areas, respectively and used as input to the fuzzy *c*-means algorithm. The search window is significantly lower in the Great Lakes region to reduce the computational time. The value of *c* that achieves the minimum homogeneity requirement is less than 50 for all scenarios in the Great Lakes region, and greater than 50 for all scenarios in the Prairie region. This discrepancy is likely attributed to the number of sites per study area (there are twice as many sites in the Prairie region). Each partitioning is validated for regional homogeneity of the at-site frequency distributions (measured as the regional homogeneity of the at-site *L*-moment ratios of scale). The results are presented as four separate line plots (one plot per season) in Figures 3-4 for the Great Lakes and Prairie regions, respectively. The *c*-values and corresponding percentages of regional homogeneity are plotted on the x and y axes. The lines with circle symbols and upward, downward and left facing triangle symbols correspond to the attribute set that is employed for regionalization.

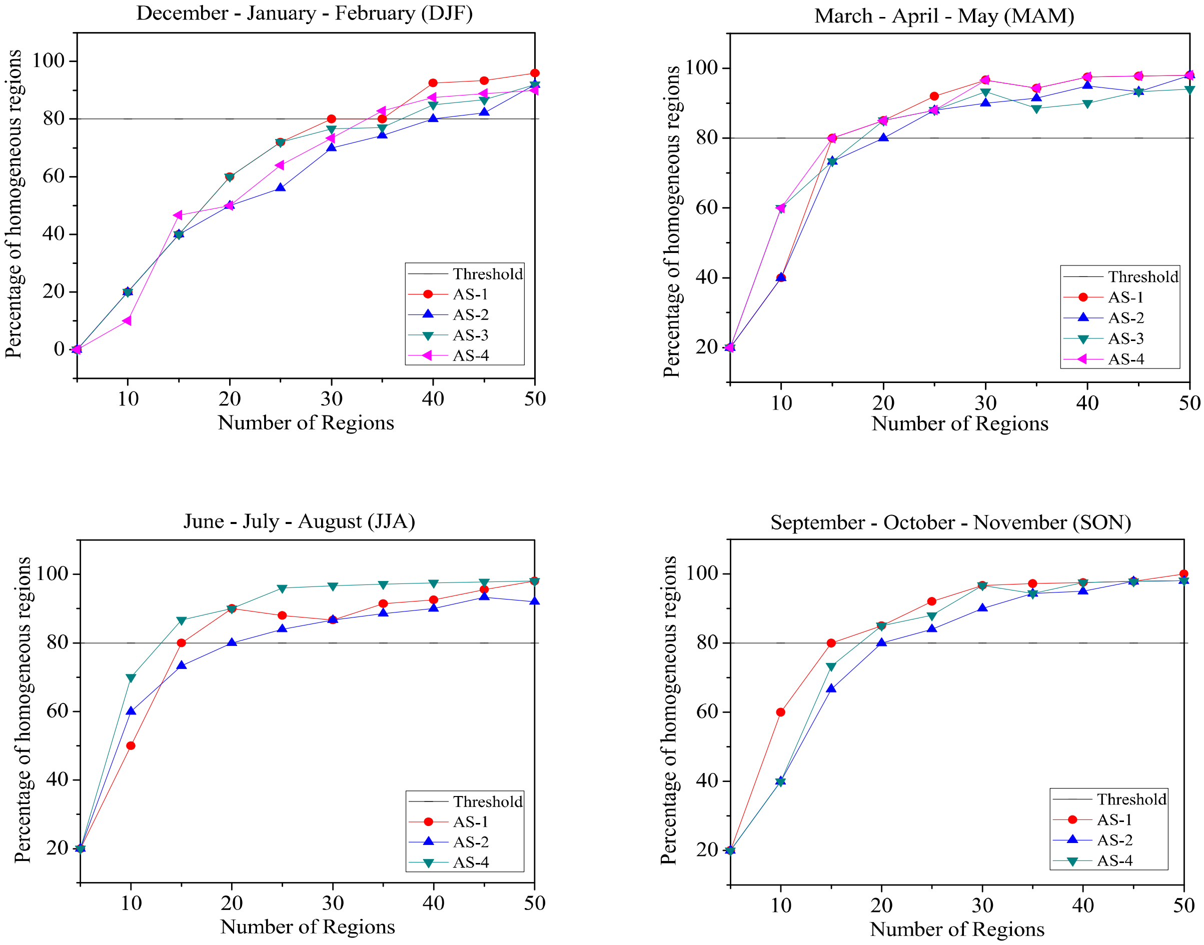
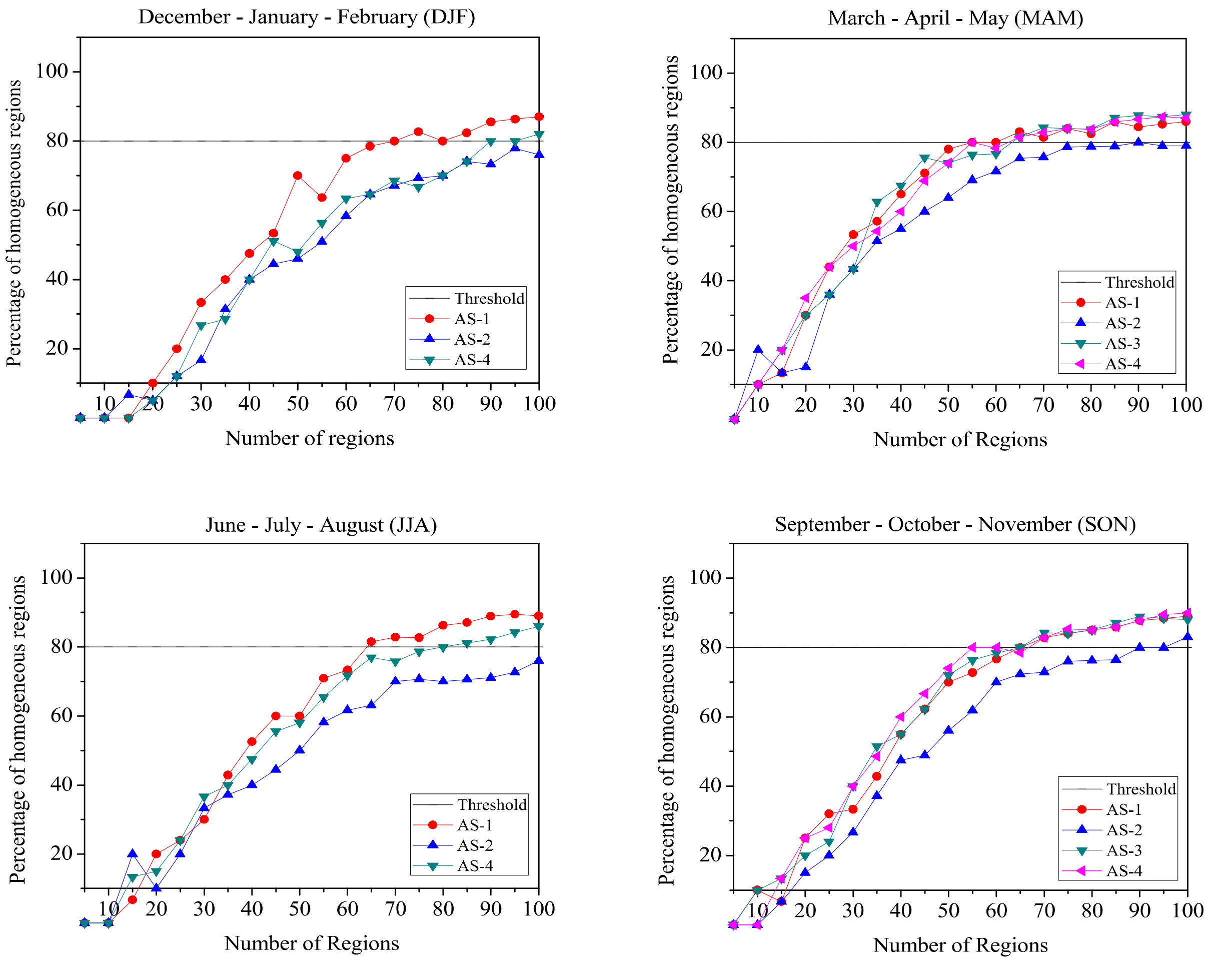


Figure 3: Outcomes of the *c*-value selection process in the Great Lakes region.


December - January - February (DJF)

March - April - May (MAM)

June - July - August (JJA)

September - October - November (SON)

Figure 4: Outcomes of the *c*-value selection process in the Prairie region.

The plots indicate the lowest value of *c* (at an increment of 5) that meets or exceeds the threshold value of 80% regional homogeneity (denoted by the black, horizontal line). A localized search is performed to find the exact whole number that meets the selection criteria. The incremental selection process is employed to reduce computational demands. Eighty percent is selected as the threshold value because it is located at the elbow of the plot for all scenarios. Figures 3-4 show that relatively large increases in the *c*-values (beyond the plot elbows) are required to improve the percentage of homogeneous regions beyond 80%; thereby compromising the lengths of the regional precipitation records.

The final values of *c* that correspond to each attribute set, season and study area are presented in Table 4. In the Great Lakes region, there are small variations in the *c*-values between attribute sets. The difference in outcomes is more pronounced between seasons. The *c*-values are larger in the DJF season (30 to 37) compared to all other seasons (15 to 20) for all four attribute sets. Similar trends are revealed for the Prairie region. Comparing the outcomes between climate regions it is evident that the *c*-values are much greater in the Prairie than in the Great Lakes regions. This is likely because the Prairie region is over two times larger than the Great Lakes region; there are 2674 and 959 grid points in the respective areas.

Table 4: The *c*-values for all combinations of attribute sets, seasons and study areas. (LSAV - large scale atmospheric variable).

	AS-1		AS-2		AS-3		AS-4	
	Prairie	**Great Lakes**	**Prairie**	**Great Lakes**	**Prairie**	**Great Lakes**	**Prairie**	**Great Lakes**
DJF	70	30	> 100	37	-	36	90	34
MAM	55	15	> 100	19	62	17	55	15
JJA	65	12	> 100	20	-	-	80	15
SON	63	15	90	19	61	-	55	19

The results of this analysis divulge more information than the preferred number of clusters for the climate sites to be partitioned into. They reveal valuable information regarding the performance of the potential attribute sets in terms of forming large, homogeneous precipitation regions for different seasons and climate regions. Further analysis of the results is presented in section 6.2.3 and 6.2.4.

6.2.3 Assessment of the Final Precipitation Regions (Attribute Selection)

After the preferred numbers of regions have been selected, the fuzzy *c*-means algorithm is employed to delineate precipitation regions for all combinations of attribute sets, seasons and study areas. The precipitation regions are plotted on maps such that each region is assigned a unique colour. The maps show the general spatial pattern of the precipitation regions and therefore, the fuzzy boundaries are not visible to allow for clearer depictions. Certain results are not shown to avoid redundancies.

Attribute Set 1 (AS-1) is composed of location parameters that are temporally fixed and therefore their spatial composition does not change between seasons. However, because the number of regions to which the climate sites are assigned (the *c*-value) is selected as the minimum number that achieves 80% regional homogeneity, the spatial distribution of the clusters is dependent upon the spatial variability of precipitation statistics. For example, in the Great Lakes region the *c*-values are 30, 15, 12 and 15 for the DJF, MAM, JJA and SON seasons, respectively. Figures 5-6 illustrate the comparison between the precipitation regions formed for the DJF and MAM seasons, respectively. The composition of precipitation regions are similar for the MAM, JJA and SON seasons (results for the JJA and SON seasons are not shown to avoid redundancy) and very different for the DJF season due to the values of *c*. Similar observations are made for the Prairie region. Figures 7-8 present the Prairie precipitation regions formed by AS-1 for the DJF and MAM seasons, respectively. The *c*-values of the respective seasons are 70, 55, 65 and 63 for DJF, MAM, JJA and SON.

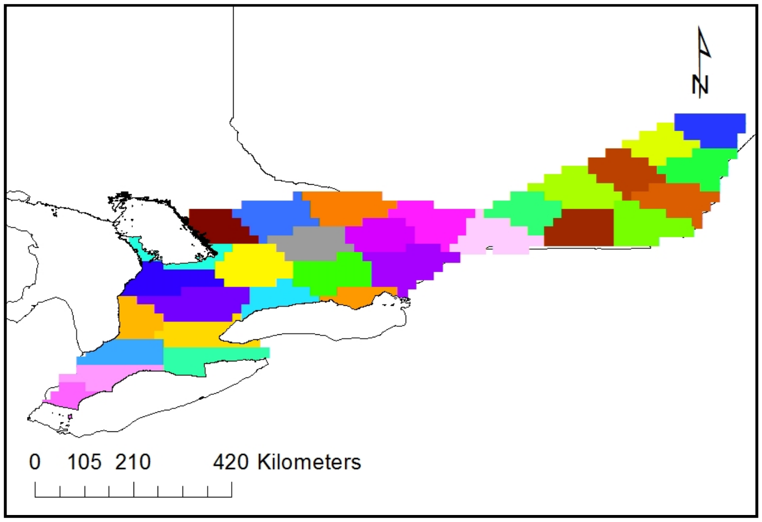


Figure 5: Precipitation regions formed by AS-1 for the DJF season in the Great Lakes study area.

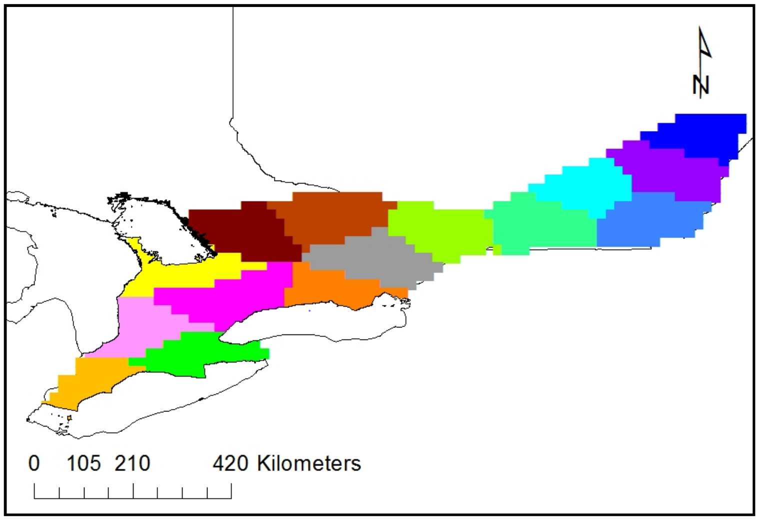


Figure 6: Precipitation regions formed by AS-1 for the MAM season in the Great Lakes region.

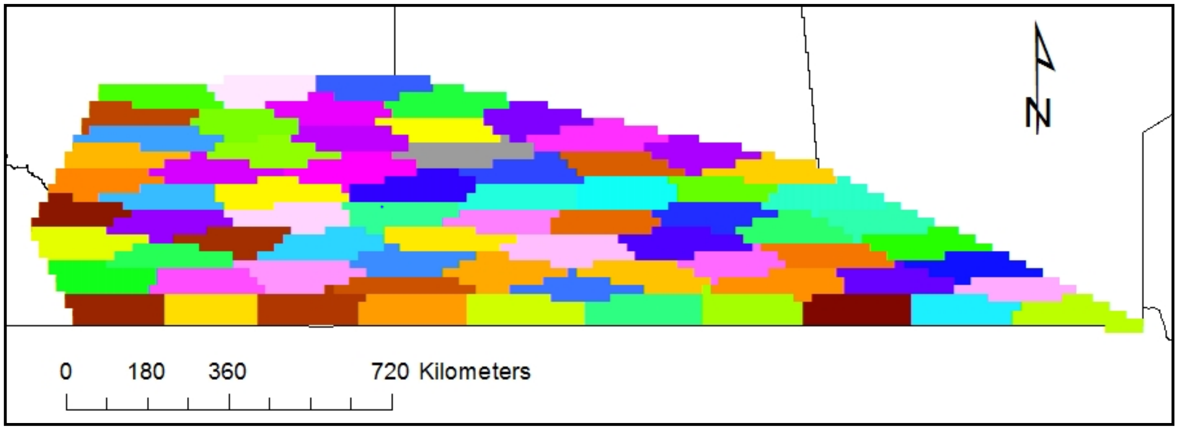


Figure 7: Precipitation regions formed by AS-1 for the DJF season in the Prairie region.

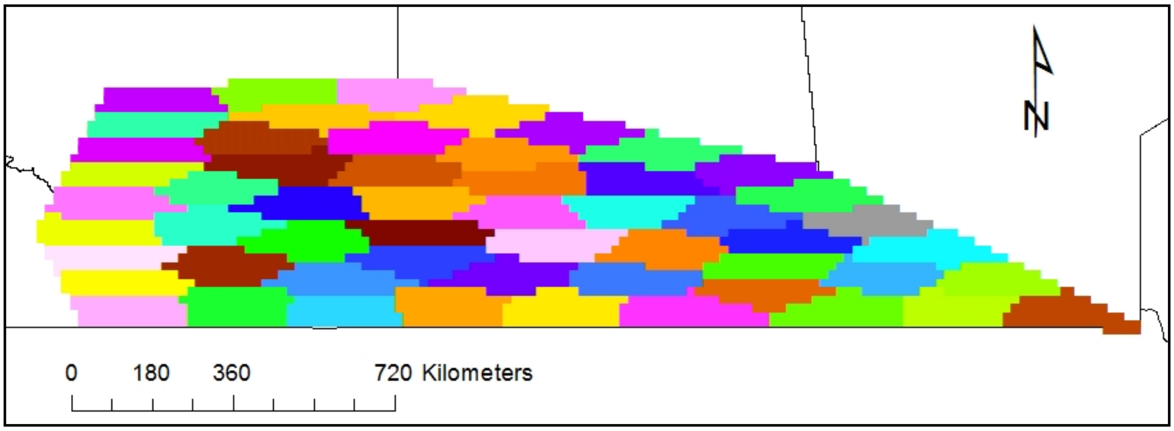


Figure 8: Precipitation regions formed by AS-1 for the MAM season in the Prairie region.

Comparable results are observed for AS-2 that consists of the location parameters of AS-1 in addition to the site elevations. In the Great Lakes region the climate sites are grouped into 37, 19, 20 and 19 regions for the DJF, MAM, JJA and SON seasons, respectively. Evidently the magnitudes of the *c* parameter are greater; although the overall trend is the same as for AS-1 where the largest *c*-value corresponds to the DJF season. In the Prairie region, however, climate sites must be partitioned into more than 100 regions for the DJF, MAM and JJA seasons and 90 regions for the SON season in order to achieve the minimum regional homogeneity requirement. Figure 9 presents the results for the SON season. The regional compositions are distinctly different from those produced by AS-1, particularly at the Western extent of the study area where the regions are exceptionally non-contiguous.

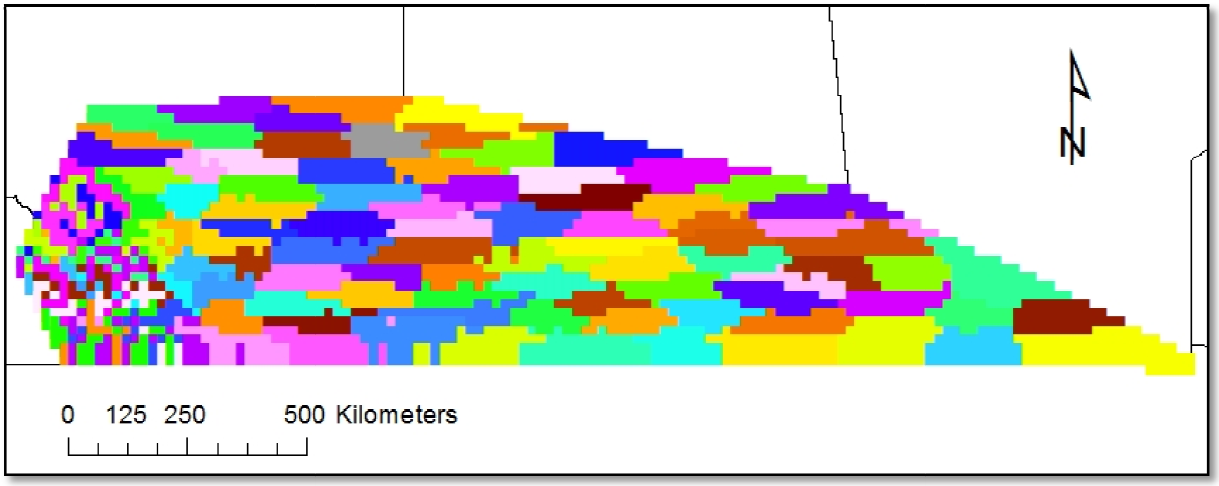


Figure 9: Precipitation regions formed by AS-2 for the SON season in the Prairie study area.

In the Great Lakes study area, regions delineated using AS-3 and AS-4 are comparable for the same season. The regional compositions are also alike for the MAM, JJA and SON seasons. Figures 10-11 present a comparison of the precipitation clusters formed by AS-4 for the MAM and SON seasons. Once more, the figures demonstrate that although there is a difference in the *c*-values (12 and 19 for the respective seasons) the same regional patterns are formed. Similar observations are made for the Prairie precipitation regions. Figures 12-13 illustrate a comparison of the precipitation regions formed by AS-3 and AS-4 for the MAM season. The climate sites are partitioned into 62 and 55 clusters by the respective attribute sets and their spatial patterns are very comparable. Figures 14-15 present the partitioning of climate sites by AS-4 for the DJF season in the Great Lakes and Prairie regions, respectively. The *c*-values corresponding to the DJF season are high relative to the other seasons.

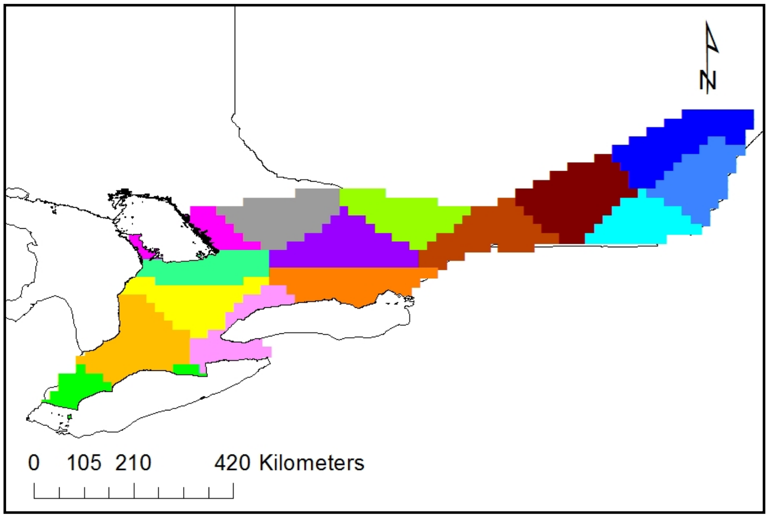


Figure 10: Precipitation regions formed by AS-4 for the MAM season in the Great Lakes study area.

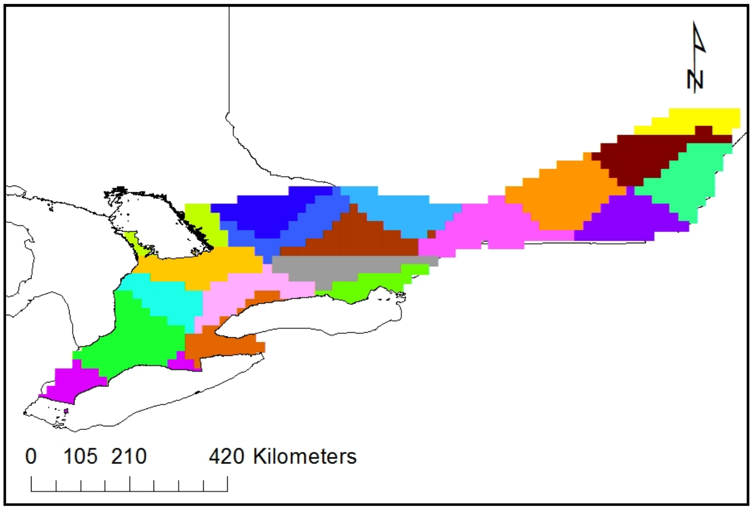


Figure 11: Precipitation regions formed by AS-4 for the SON season in the Great Lakes study area.

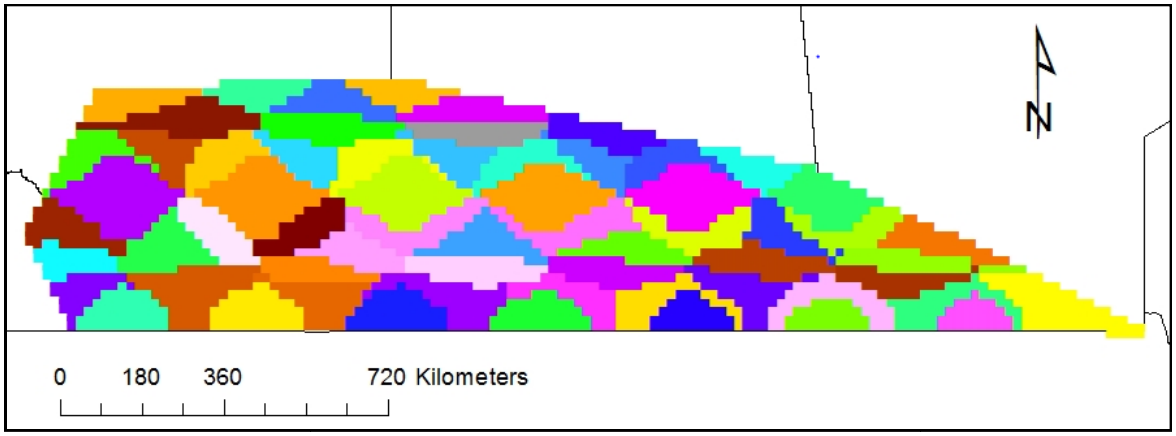


Figure 12: Precipitation regions formed by AS-3 for the MAM season in the Prairie study area.

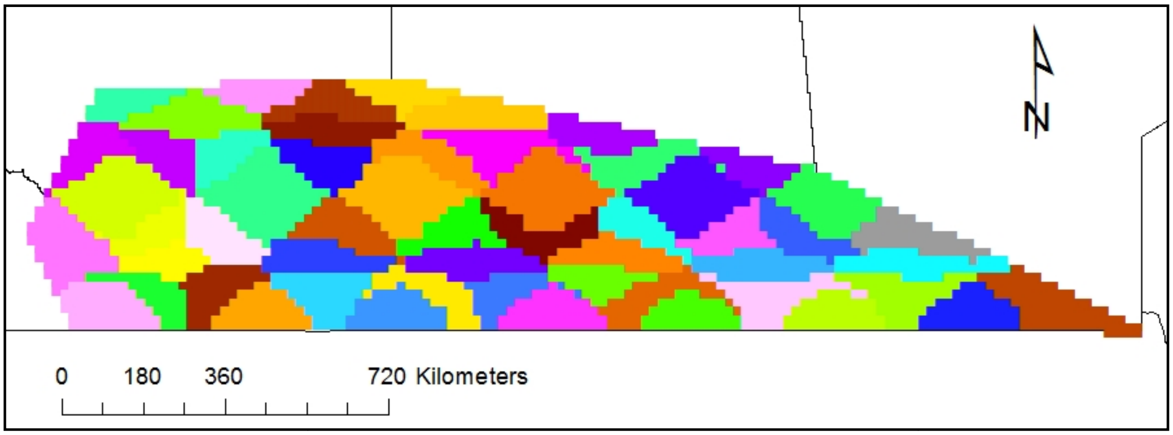


Figure 13: Precipitation regions formed by AS-4 for the MAM season in the Prairie region.

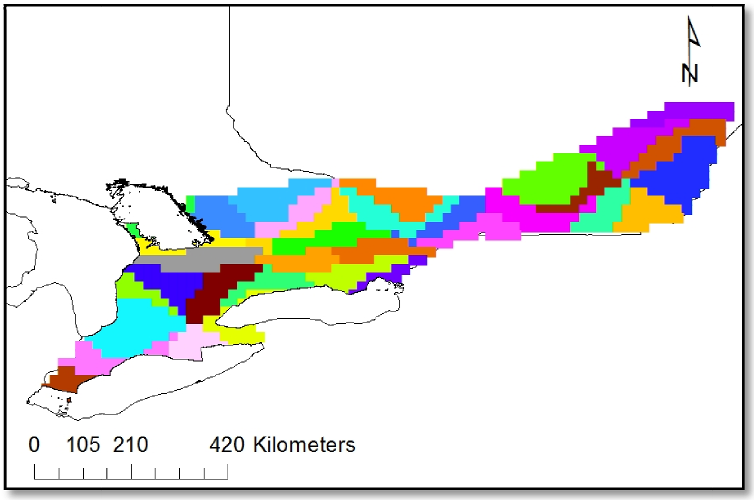


Figure 14: Precipitation regions formed by AS-4 for the DJF season in the Great Lakes study area.

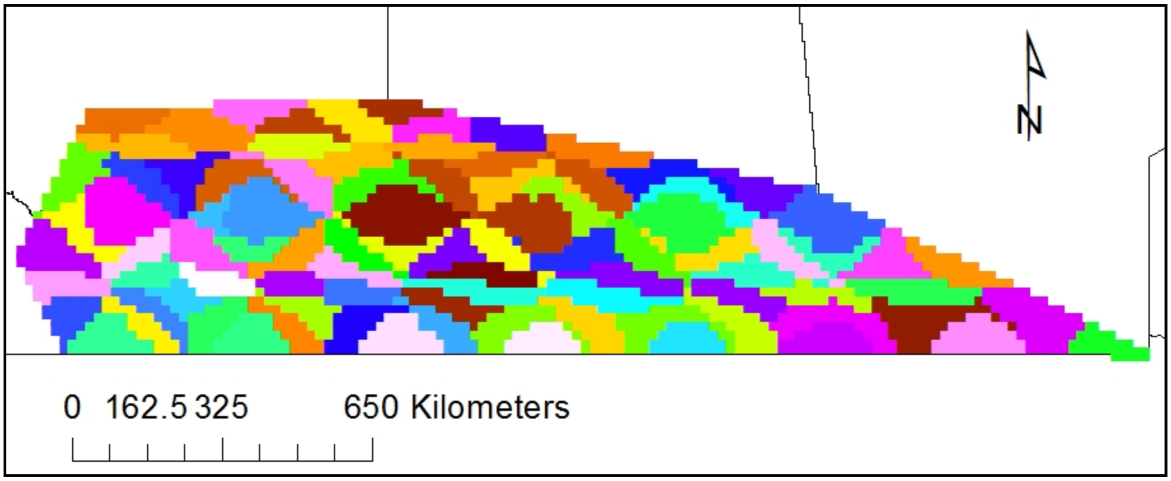


Figure 15: Precipitation regions formed by AS-4 for the DJF season in the Great Lakes study area.

6.2.4 Discussion (Attribute Selection)

The number of clusters for the climate sites to be partitioned into provides a good indication of the performance of the attribute set. A low *c*-value is indicative of a superior performance. The spatial compositions of the precipitation regions formed by AS-1 are unlike those formed by AS-2, AS-3 and AS-4. Table 4 reveals that AS-1 matches or marginally outperforms the other attribute sets in forming homogeneous precipitation regions for direct comparisons of almost all seasons for both climate regions (with the exception of the SON season in the Prairie region). That being said, for the same season, the *c*-values between all attribute sets are still very comparable; (see Figures 3-4). A disadvantage of using location parameters as attributes is that they are limited to applications where the assumption of a stationary climate is acceptable; in other words the statistical properties of the precipitation record (the mean, standard deviation, etc.) are constant overtime. This assumption may be valid over shorter time periods. The merit of using atmospheric variables as attributes is that they are capable of detecting potential shifts in precipitation regions with time that can provide valuable information to planning applications.

Overall the patterns of precipitation regions formed by AS-2 are comparable to those formed by the other attribute sets for application in the Great Lakes region; however, in the Prairie region the results are vastly different and inferior to all other outcomes. In order to achieve the minimum requirement for regional homogeneity, the climate sites must be partitioned into more than 100 regions in the DJF, MAM and JJA seasons and 90 regions in the SON season. This observation is likely attributed to the inclusion of the elevation parameter and the presence of the Canadian Rocky Mountain range located at the Western extent of the study area. It is the significant differences in site elevation in the concentrated area that result in the non-contiguous precipitation regions.

Regions formed by AS-3 and AS-4 are very similar for a direct comparison of the seasons. It is therefore evident that attribute screening through linear correlation analysis does not impact the formation of precipitation regions in a significant way. The only advantage of employing AS-3 over AS-4 is the reduced computational time to run the regionalization and validation codes because AS-3 is composed of fewer variables. However, AS-1 is more computationally efficient than AS-3 (and AS-4) for most seasons (with the exception of the DJF season for the Great Lakes region). A major limitation of employing AS-3 for regionalization is that it cannot be created for seasons or study areas in which the atmospheric variables and local precipitation form weak linear correlations. It is well known that large scale atmospheric variables influence precipitation processes (among other factors); however it is possible that the linear correlation analysis is not an appropriate measure of their relationship. This may be due to the presence of a non-linear relationship between the variables and local precipitation. To further support this argument, the linear correlation analysis results presented in Figures A32 - A39 show that the selected atmospheric variables form much stronger correlations with the local precipitation in the Prairie region. It may therefore be inferred that AS-3 should provide better results in the Prairie than in the Great Lakes region because there is a stronger relationship between the selected attributes and precipitation. Figures 3-4 show that the overall performance of AS-3 does not exceed AS-1 or AS-4 in terms of achieving a high percentage of regional homogeneity for a low number of regions in either study area.

The attribute sets have also been compared for their performance in different seasons. It is apparent that the regional compositions are fairly consistent for the MAM, JJA and SON seasons. For the DJF season, however, the spatial distributions of precipitation regions are very different as the climate sites are partitioned into many more clusters. This observation applies to both study areas. A large number of clusters may be indicative of a poor performance of the attribute set, potential difficulties with snow measurements or a large variation of the regional precipitation statistics. The latter theory is tested by analyzing the results produced by AS-1. Since the location parameters are temporally fixed, the only seasonal difference between the regions derived using AS-1 is the number of clusters to which the climate sites are assigned. This is because the criteria for selection of the number of regions (the *c*-values) is dependent upon the results of the *L*-moment regional heterogeneity test that provide a measure of the variation between the sample *L*-moment ratios of scale of the climate sites' precipitation records. A region is classified as homogeneous if it is characterized by a low variation of the *L*-Cv of the at-site precipitation records; hence, there is a direct relationship between the precipitation variability and the number of regions to which the sites are assigned. The large spatial variation in the DJF seasonal precipitation may be attributed to isolated snowfall events and lake effect precipitation in the Great Lakes region. Due to the continental climate of the Prairie region, lake effect precipitation does not occur here and the significant spatial variability of precipitation in the DJF season is likely attributed to isolated snowfall events and difficulties with snowfall measurements.

It is the assessment of the attribute set performance that distinguishes this work from similar regionalization studies. In addition, to the best of the author's knowledge this is the first assessment of the regionalization of precipitation that utilizes the ANUSPLIN 10 by 10 km gridded precipitation dataset in a Canadian setting. Asong *et al.* (2015) performed a similar analysis in the Canadian Prairie provinces of Alberta, Saskatchewan and Manitoba (covering the entire geopolitical regions). They also utilized statistical methods (principal component analysis and canonical correlation analysis) to screen and select relevant site attributes (geophysical site parameters, atmospheric covariates and teleconnection indices), and employed the fuzzy c-means algorithm to delineate precipitation regions. Their results, however, are drastically different than those presented in this document. Asong *et al*., (2015) derived five homogeneous precipitation regions while the results of this analysis recommend the climate sites to be partitioned into 55 to over 100 regions in order to attain 80% regional homogeneity in the Prairie climate region (that encompasses an even smaller area than the Canadian Praire provinces).

Plausible explanations for such a large discrepancy in the results are: (i) the difference in the temporal resolution of the precipitation data used to validate regional homogeneity; (ii) the inclusion/absense of the corrective measure to account for the effect of the inter-site cross-correlations on the *L*-moment regional heterogeneity test; and (iii) the difference in the number of climate sites within the study area. Asong *et al*., (2015) used attributes that formed statistically sigificant relationships with monthly precipitaiton to form the precipitation regions and subsequently tested for regional homogeneity using the *L*-moment regional heterogeneity test and total annual, total seasonal and extreme seasonal precipitation values. The current analysis tests for regional homogeneity using precipitaiton data at a monthly temporal resolution (organized into four separate seasons). Since variations in the precipitation data tend to become dampened at higher temporal resolutions it makes sense for annual and seasonal datasets to provide more statistically homogeneous regions as compared to the monthly precipitation data.

The present analysis utilizes a measure to correct for the impact of inter-site cross-correlations on outcomes of the *L*-moment regional heterogeneity test (Castellarin *et al*¸ 2008). The presence of spatial correlations between the sites' precipitation records can cause the precipitaiton regions to appear homogeneous when they are not. As such, implementation of the corrective measure improves the power of the validation test and provides for more realistic results. Employment of the corrective measure is very important to this study that utilizes the fine ANUSPLIN gridded precipitation dataset. The climate sites are in close proximity with one another and therefore, are very likely to exhibit strong spatial correlations. There is no indication that Asong *et al*., (2015) used a measure to correct for cross-correlations between the climate stations' preciptiation records and this may have had a positive effect on their results. However, the number of climate stations employed in their analysis are sparsely distributed and therefore, there is less of a need for such a corrective measure.

It is believed that the difference in the number of climate sites within the study areas has a significant effect on the disparity in the results of the two analyses. Asong *et al*. (2015) utilized 120 Environment Canada climate stations (http://www.ec.gc.ca/dccha-ahccd/, last accessed 2015 June), while this study employs 2674 ANUSPLIN grid points as climate sites (that is over 20 times more objects to be partitioned by the clustering algorithm). Baeriswyl and Rebetez (1997) performed the regionalization of precipitation in Switzerland using two datasets consisting of 47 and 101 climate stations. Regionalization of the datasets resulted in 7 and 13 precipitation regions, respectively. They attribtued the "finer and more detailed results" of the latter analysis to the larger number of stations within the region. This is likely to be the main contribution to the difference between the study by Asong *et al*., (2015) and the work presented here. The use of a fine gridded dataset is advantageous for gaining a complete understanding of the precipitation processes in a study area. Such fine datasets may not be available for certain temporal resolutions of precipitation (for example, sub daily) that are needed for many applications of regional frequency analysis.

6.3 Temporal Resolution of Precipitation Data

The effect of the temporal resolution of the precipitation data on the regionalization outcomes is assessed in this section. Monthly, seasonal and annual resolutions are considered as well as the maximum annual series of the data. The purpose of this investigation is to determine if there is a change in or link between the precipitation regions derived for different resolutions such that precipitation data can be accurately combined or disaggregated into different resolutions for applications in regional frequency analysis. This information is useful for studies that require precipitation to be recorded at a specific resolution that is limited or unavailable. Again, the fuzzy *c*-means algorithm is employed to partition the climate sites into regions based on the similarity of their location parameters and elevations (Attribute Set 2 of the assessment presented in section 6.2). The regions are subsequently validated using observed precipitation and an *L*-moment regional heterogeneity test.

The results for the determination of the preferred number of regions to which the sites are assigned, and the maps of the corresponding precipitation regions are revealed for all temporal resolutions in both study areas. Important observations are summarized in the discussion section.

6.3.1 Determination of the preferred number of regions (Temporal Resolution)

The same *c*-value selection method that is used for the assessment of the site attributes is employed for this analysis. The preferred number of regions to which the climate sites are assigned is chosen as the lowest value to achieve a minimum of 80% regional homogeneity; that is at least 80% of the precipitation regions are classified as homogeneous using the *L*-moment regional heterogeneity test.

Figures 16 – 17 present the results of the *c*-value selection process for the monthly and seasonal temporal resolutions in the Great Lakes and Prairie regions, respectively. Figure 18 shows the results of the *c*-value selection procedure for the maximum annual series and annual temporal resolutions in both study areas. The *c*-values and corresponding percentages of homogeneous regions are plotted on the x and y axes, respectively.

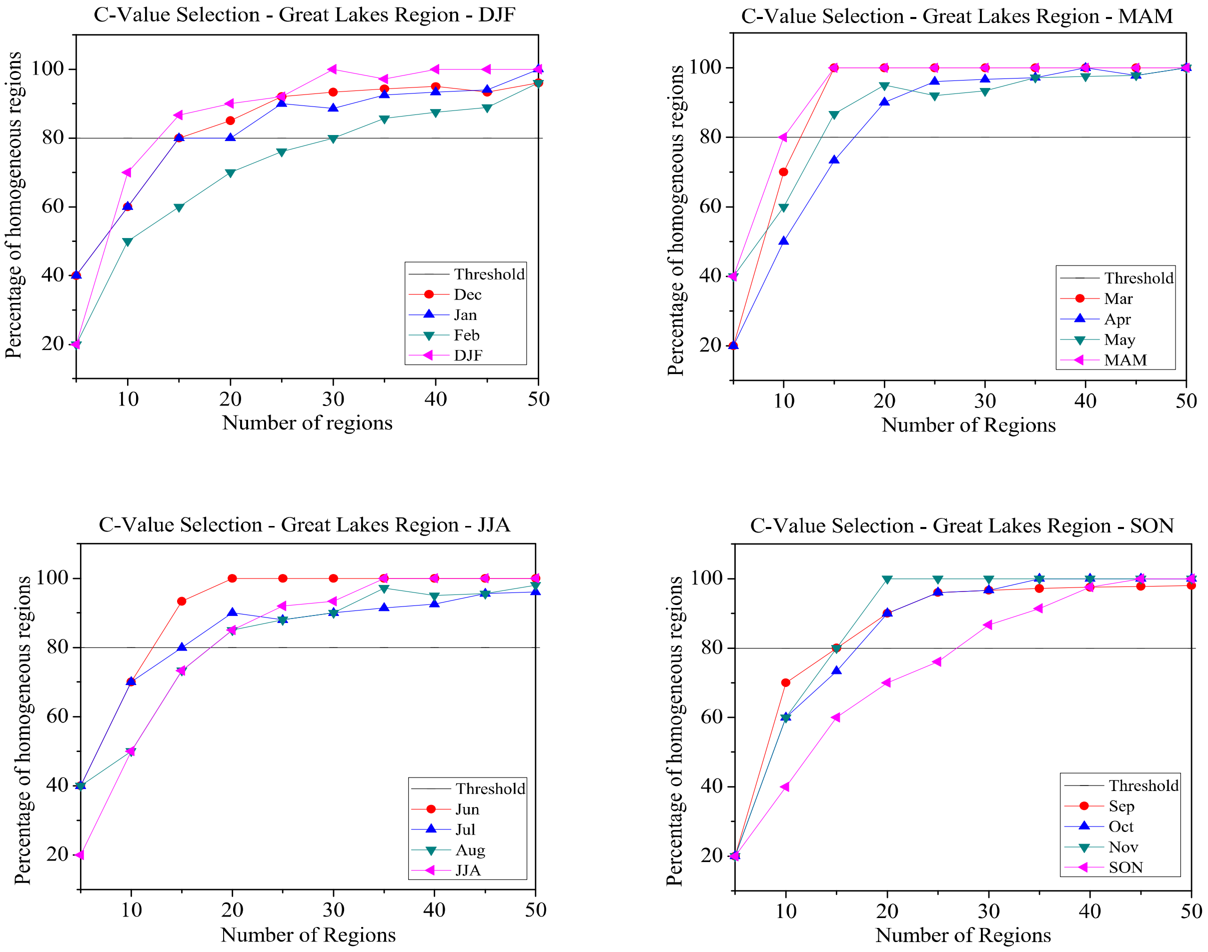


Figure 16: Outcomes of the *c*-value selection process for the monthly and seasonal temporal resolutions in the Great Lakes study area.

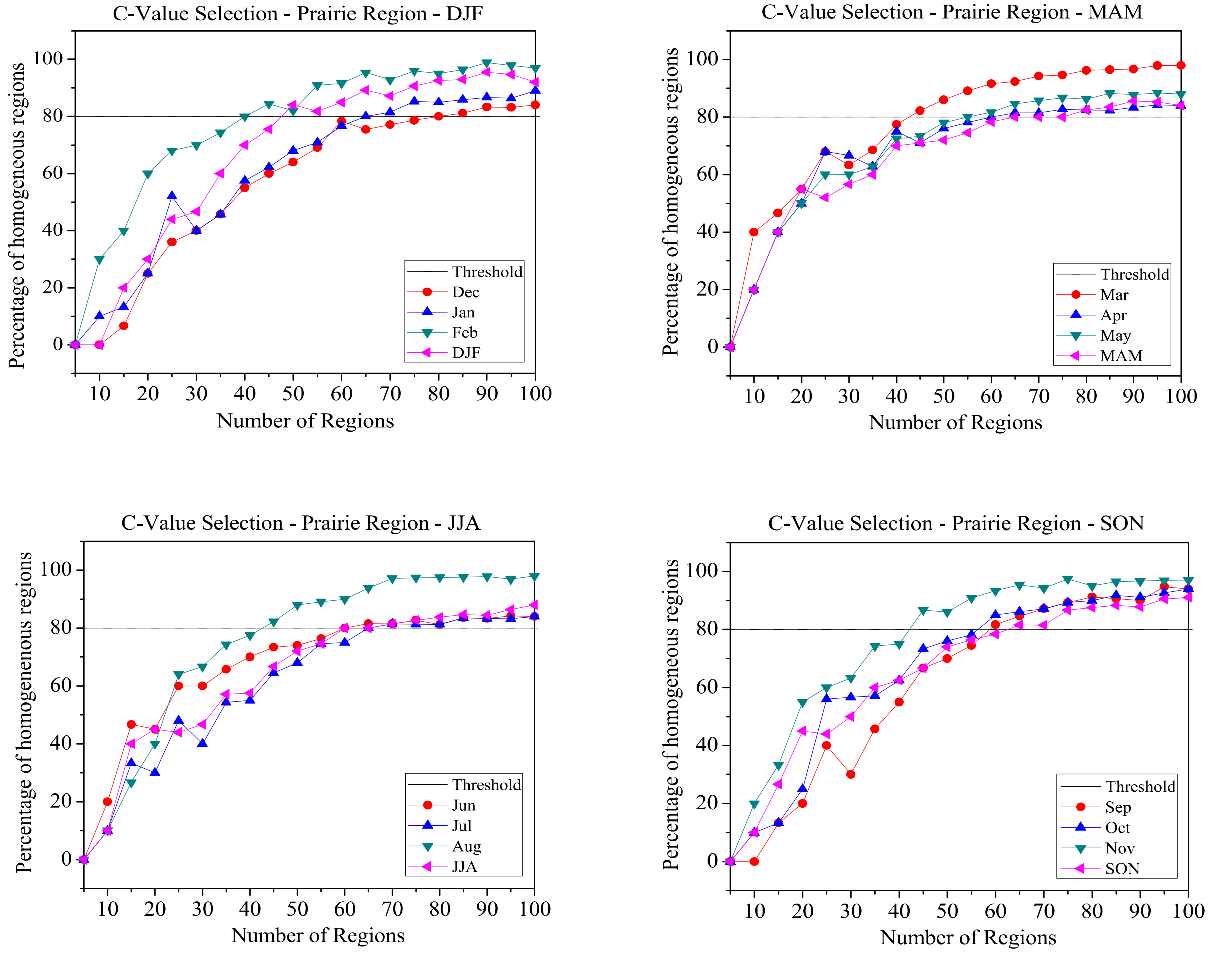


Figure 17: Outcomes of the *c*-value selection process for the monthly and seasonal temporal resolutions in the Prairie study area.

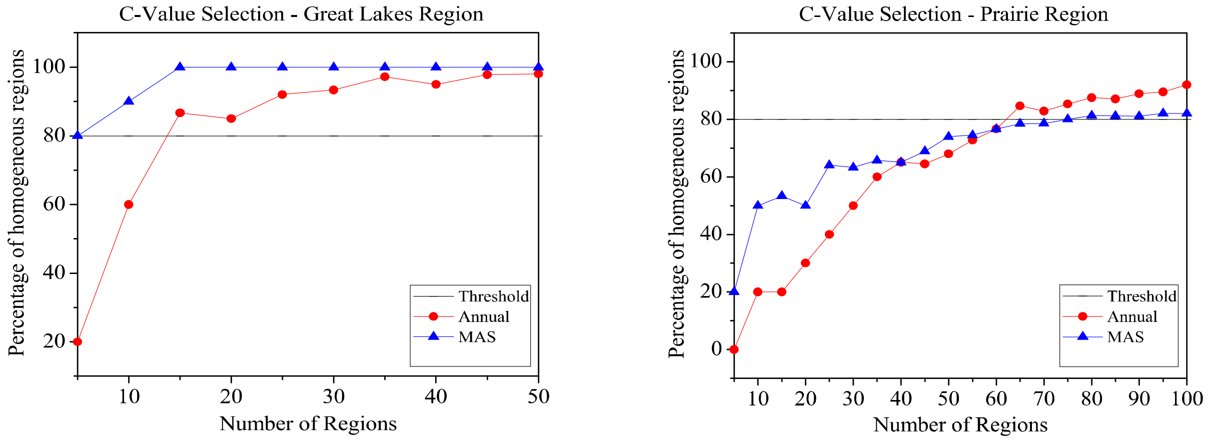


Figure 18: Outcomes of the *c*-value selection process for the annual maximum series and the annual temporal resolution in the Great Lakes and Prairie study areas on the left and right, respectively.

The plots identify the lowest *c*-value to cross the minimum threshold of 80% homogeneous regions that is denoted by the black, horizontal line. To determine the exact whole number that meets the selection criteria, a localized search is performed. Evidently the monthly resolution reduces the number of regions required to achieve the homogeneity criterion for some scenarios but not for others. There is significant variability in the *c*-values required for different months and this is observed in both study areas.

The final preferred *c*-values are summarized in Table 5 and 6 for the monthly and seasonal, annual and maximum annual series resolutions, respectively. In the Great Lakes region and for the monthly temporal resolution the magnitudes of the *c*-values range from 11 to 29. For January and February climate sites are partitioned into a greater number of clusters to achieve the selection criteria of 80% regional homogeneity; their respective *c*-values are 20 and 29. Alternatively, the *c*-values for March and June are 10 and 11, respectively. The magnitudes of the *c*-values for the seasonal temporal resolution exhibit a similar range to the monthly scale that is, 10 to 27; although the *c*-values are relatively high for the months of January and February (20 and 29), their magnitude for the DJF season is only 15. In addition, the preferred *c*-value that is derived for the SON season is 27 while the *c*-values for the individual months of September, October and November extend from 12 to 17. The value of *c* for the annual temporal resolution is 13 that is also within the range for the monthly and seasonal scales; however the *c*-value of the maximum annual series is 5 that is well outside of the typical range.

Table 5: Final *c*-values for the annual, maximum annual series and seasonal temporal resolutions in the Great Lakes and Prairie climate regions.

Region	Annual	MAS	DJF	MAM	JJA	SON

Great Lakes	13	5	15	10	19	27
Prairie	61	71	46	65	60	65

Table 6: Final *c*-values for the monthly temporal resolution in the Great Lakes and Prairie climate regions.

Region	Jan	Feb	Mar	Apr	May	Jun	Jul	Aug	Sep	Oct	Nov	Dec
Great Lakes	20	29	10	16	13	11	15	17	13	17	12	15
Prairie	65	36	41	56	52	56	61	42	60	56	43	76

For all temporal resolutions the *c*-values are much greater in the Prairie than in the Great Lakes region. The magnitudes of the *c*-values extend from 36 to 76 for the monthly temporal resolution. Contrary to the Great Lakes region, February calls for the lowest number of regions for the climate sites to be partitioned into (36) in order to achieve the minimum regional homogeneity requirement. The *c*-value for the month of December is 76 that is more than twice the magnitude of the value for February. The values for all other months are within the range of 41 to 65. The *c*-values for the seasonal, annual and maximum annual series resolutions fall within the range of the monthly values; they are 46 to 65, 61 and 71, respectively.

6.3.2 Regionalization of Precipitation (Temporal Resolution)

After the *c*-values have been determined they are used as input to the fuzzy *c*-means clustering algorithm to partition the climate sites into precipitation regions for all temporal resolutions and in both study areas. The resultant regions are plotted on maps such that they are distinguishable by colour. Climate sites are assigned a colour corresponding to the region in which they have maximum membership in order to show the general spatial pattern of the regions and as such, the fuzzy boundaries are not depicted. Certain results are not shown to avoid redundancies.

In the Great Lakes study area a comparison of the precipitation regions formed for the monthly temporal resolution is conducted. Figures 19, 20 and 21 present three different outcomes for the months of February, August and September that are partitioned into 29, 17 and 13 regions, respectively.

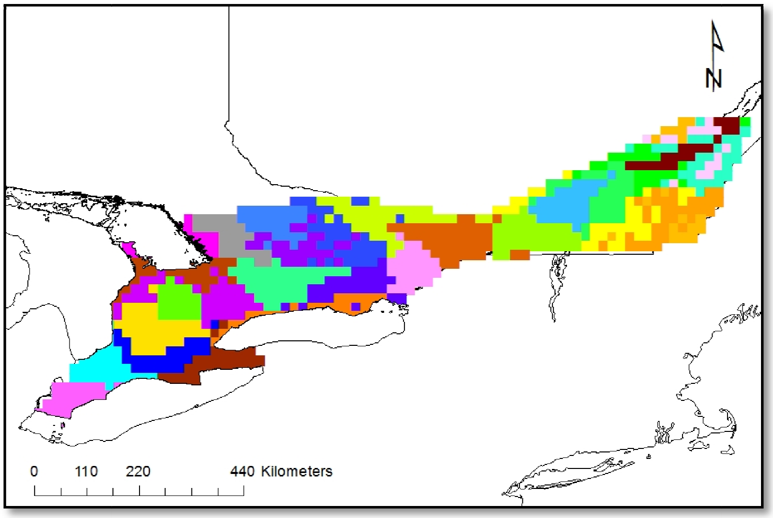


Figure 19: Precipitation regions delineated for the month of February in the Great Lakes study area.

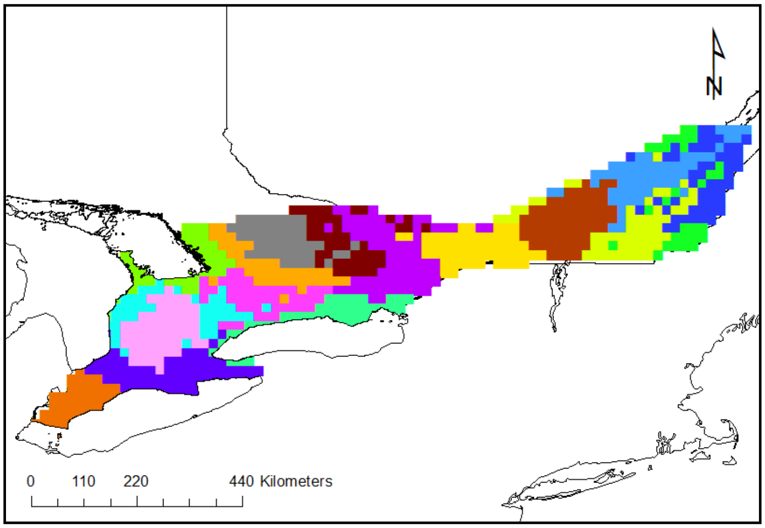


Figure 20: Precipitation regions delineated for the month of August in the Great Lakes study area.

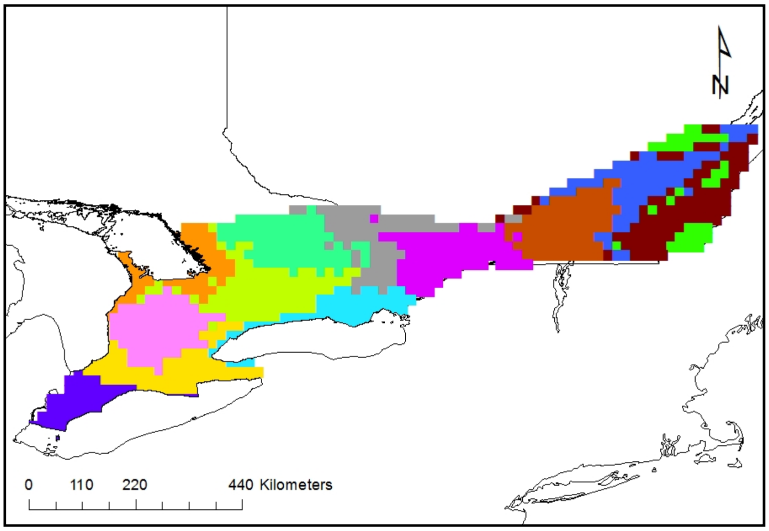


Figure 21: Precipitation regions delineated for the month of September for the Great Lakes study area.

The patterns of precipitation regions formed for October are identical to those shown in Figure 20 since the climate sites are also partitioned into 17 regions. In addition the regions in Figure 20 present similar patterns to January that has a preferred *c*-value of 20. The patterns of precipitation regions formed for March, April, May, June, July, September and December are similar to those presented in Figure 21 because the magnitudes of their *c*-values are in the range of 11 to 16. For the seasonal temporal resolution the DJF and MAM seasons have preferred *c*-values of 15 and 10 and therefore, their regional patterns are also similar to those presented in Figure 21. Climate sites are partitioned into 19 regions for the JJA season and the spatial distribution of regions is similar to those depicted in Figure 20. Finally, the *c*-value for the SON season is 27 so the precipitation regions form similar patterns to those shown in Figure 19. For the annual temporal resolution the number of precipitation regions is 13 and therefore, the spatial compositions of the precipitation regions are identical to those shown in Figure 21. The precipitation regions delineated for the maximum annual series of precipitation are very different from the other temporal scales as shown in Figure 22. The climate sites are partitioned into five large regions.

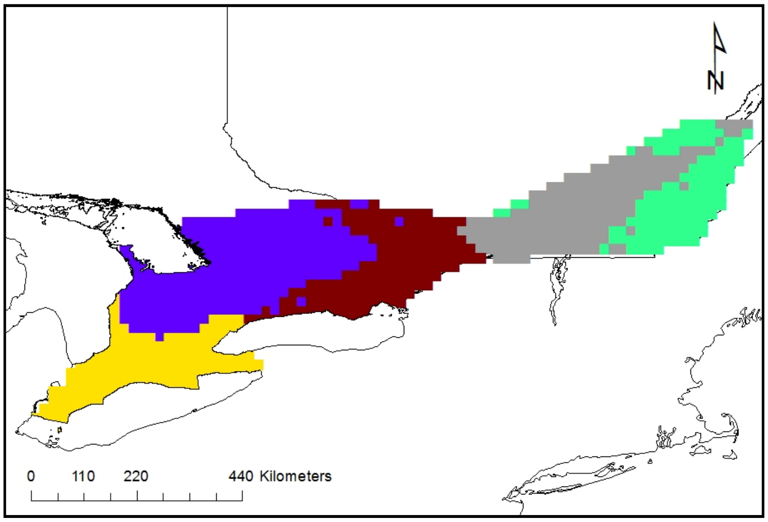


Figure 22: Precipitation regions delineated for the maximum annual series in the Great Lakes study area.

In the Prairie study area the patterns of the precipitation regions are similar for all temporal resolutions and time periods; although the *c*-values vary greatly from 36 to 76 over all scales. Figures 23 and 24 present the precipitation regions for the months of February and December in which the climate sites are partitioned into 36 and 76 regions (the limits of the *c*-value range), respectively. Evidently, for both figures, the majority of the regions are comparable in size and shape that is elongated in the horizontal direction. At the very West end of the study area the regions are non-contiguous because of the presence of the Canadian Rocky Mountains and the inclusion of elevation as a site attribute.

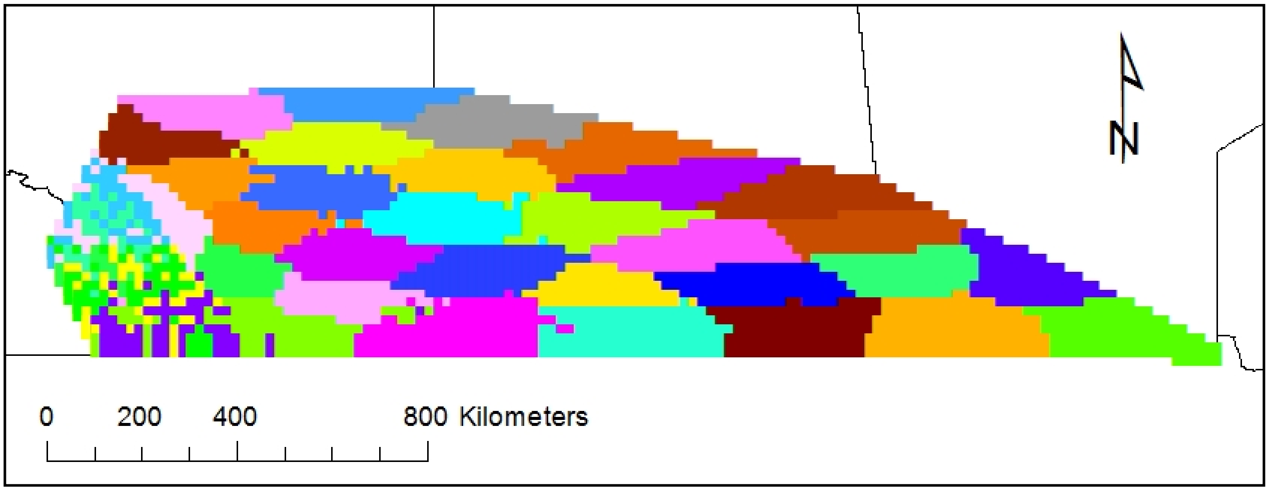


Figure 23: Precipitation regions delineated for the month of February in the Prairie study area.

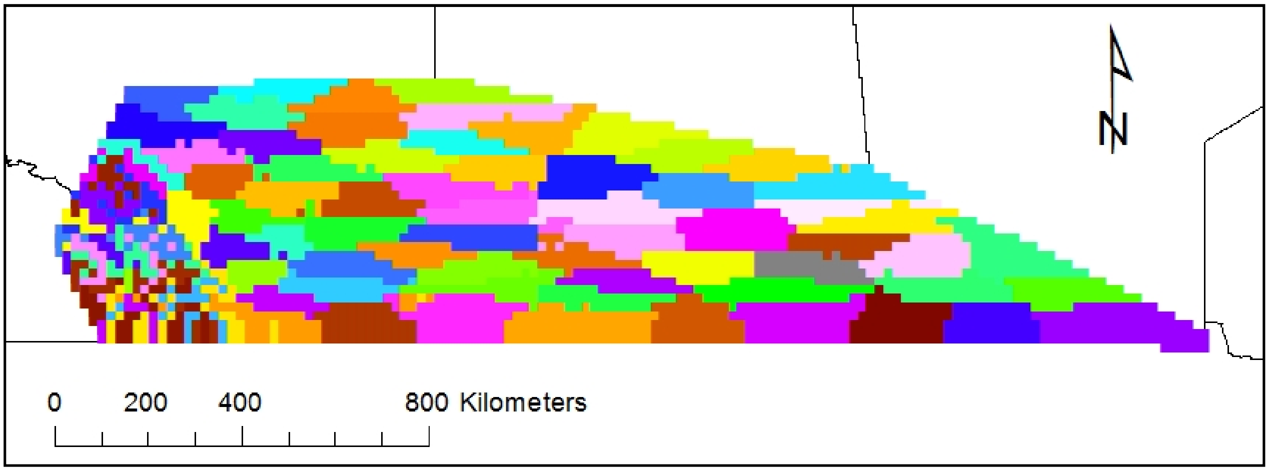


Figure 24: Precipitation regions delineated for the month of December in the Prairie study area.

6.3.3 Discussion (Temporal Resolution)

In this assessment location parameters including latitude, longitude, elevation and distance to water bodies are assigned as site attributes. Location parameters are temporally fixed and therefore, they form the same precipitation regions across all time scales for the employment of the same *c*-value. However, through the *c*-value selection process it is evident that the number of regions to which the climate sites are assigned change according to the precipitation variability of the implemented temporal resolution and time period (refer to section 6.2.4 for a detailed explanation of the relationship between the precipitation variability and the selection of the preferred number of regions).

A partitioning of climate sites that requires a large *c*-value to attain the minimum homogeneity criterion is indicative of high precipitation variability for the temporal resolution under investigation. For example, in the Great Lakes region, the *c*-values for the month of February and the SON season are relatively high as they are 29 and 27, respectively. The high variability of precipitation may be attributed to large isolated snowfall events for the month of February, and a combination of convective precipitation events that are typical of the late summer, early fall and lake effect precipitation that begins in November for the SON season. Conversely, the spatial variability of the maximum annual series of precipitation is very low in the Great Lakes climate region. The climate sites are partitioned into 5 large regions that consist of many sites with similar between site variability of the maximum annual series of precipitation.

The range of *c*-values for the Great Lakes and Prairie study areas and therefore, the differences in regional patterns, are greatest for the monthly temporal resolution (with the exception of the maximum annual series of precipitation in the Great Lakes region). Plausible reasons for this observation are as follows: (i) variations in the observed data become dampened at higher temporal scales (i.e., annual data have less variations when compared to monthly); (ii) higher frequency data (such as monthly) have more influence on the local climatic conditions when compared to lower frequency data (such as annual). Based on these observations it is inferred that the observed precipitation statistics measured at different temporal scales have significant effects on the regional compositions.

There does not appear to be a relationship between the precipitation regions formed under different temporal resolutions. For example, in the Great Lakes region the climate sites are partitioned into 13, 17 and 12 regions for the months of September, October and November and conversely, the sites are partitioned into 27 precipitation regions for the combination of these months in the SON season. In addition the magnitudes of the *c*-values are distributed from 10 to 29 for monthly, seasonal and annual temporal scales; however, the *c*-value for the maximum annual series is 5 that is well outside of the typical range. These observations may be attributed to changes in the statistical properties of the precipitation data when combined into different resolutions. As such, precipitation regions that are classified as homogeneous for one temporal resolution are not necessarily homogeneous for another for the study areas under investigation.

6.4 Proposed Regionalization Model Output

Model outputs from the main components of the proposed regionalization model (differential evolution and fuzzy Compromise programming) are presented in this section. The model is evaluated for its performance in four seasons (DJF, MAM, JJA and SON) at a monthly temporal resolution and for the preferences of three different decision makers. An assessment of the model is conducted in the Great Lakes study area only.

6.4.1 Differential evolution - Generation of alternative solutions

The differential evolutionary (DE) optimization algorithm is employed to derive set of optimal, alternative solutions that best satisfy the contrasting objective functions (maximization of the number of station-years in the regional precipitation record and maximization of the number of homogeneous precipitation regions). Figure 25 presents the values of the objective functions that correspond to optimal solutions. The figure contains four plots representing the model output for the DJF, MAM, JJA and SON precipitation datasets located on the top left, top right, bottom left and bottom right, respectively. The percentage of precipitation regions that are classified as homogeneous through the validation process and the minimum number of climate sites that belong to a region for a single partitioning (the objective functions) are plotted on the x and y axes, respectively.

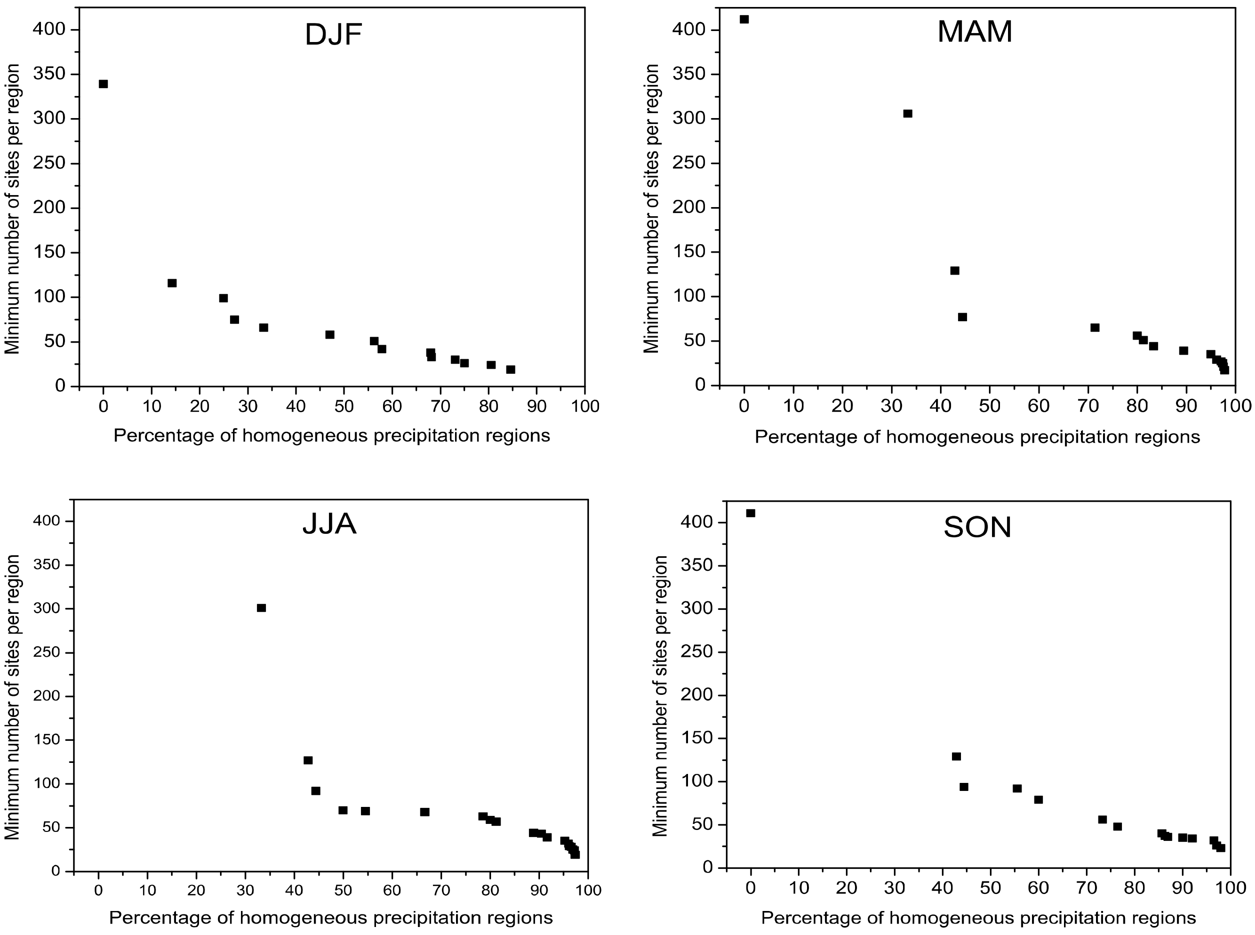


Figure 25: Values of the objective functions corresponding to the optimal solutions for the DJF, MAM, JJA and SON seasons.

Analysis of the results reveals a fairly even distribution of the points for the DJF season. However, the percentage of homogeneous precipitation regions does not exceed 85%. Alternatively, the points are more concentrated in the bottom right corners of the plots for the MAM, JJA and SON seasons; indicating that the majority of optimal solutions achieve between 80 to 100 percent regional homogeneity with a minimum of approximately 20 - 50 climate sites belonging to a region.

Similar to the evaluation of the climate site attributes, the current analysis finds that the climate sites must be partitioned into a much larger number of clusters in the DJF season as compared to the MAM, JJA and SON seasons in order for at least 80% of the regions to be classified as homogeneous. This observation is attributed to the large spatial variation of precipitation during the DJF season that is believed to be a result of isolated snowfall events and lake effect precipitation that are characteristic of the Great Lakes climate region. Evidently, the proposed systems model effectively captures the seasonal differences in the variation of precipitation statistics.

6.4.2 Fuzzy Compromise programming - Ranking of alternative solutions

Fuzzy Compromise programming is employed to evaluate and rank the alternative, optimal solutions in order of their proximity to the ideal solution in the objective space. During this process, criteria weights are assigned to the objective functions to incorporate decision maker preferences into the evaluation of the alternative solutions. Objectives that are assigned higher weights have stronger influences on the alternative rankings. Criteria weights are subjective parameters that change according to the priorities of the decision maker or the model application. For example, the model may be used for the regionalization of precipitation in order to obtain reliable estimates of the rainfall magnitudes corresponding to a low return period. In this case preference may be given to the solutions with a high percentage of homogeneous regions since a relatively low number of station-years of the regional precipitation record is required to capture the true statistics of the local precipitation.

The model output is evaluated for three sets of criteria weights: (i) Decision Maker 1 assigns equal criteria weights to the objective functions; (ii) Decision Maker 2 assigns a criteria weight of 1 to the percentage of regions with homogeneous precipitation statistics and a criteria weight of 2 to the minimum number of climate sites belonging to a region; and (iii) Decision Maker 3 allocates criteria weights that are opposite to Decision Maker 2. The results are presented as ranked alternative solutions in Figure 26 - 28 for Decision Maker 1, 2 and 3, respectively. The corresponding values of the objective functions are presented in Figures 29 - 31 for the respective decision maker preferences. Each figure contains four plots that correspond to the four seasons. The solution ranks are represented on the x axis and the line and bar plots correspond to the data on the primary (left) and secondary (right) y axes, respectively. For Figures 26 - 28 the primary and secondary y axes represent the values of *c* and the fuzzifier, respectively, and for Figures 29 - 31 the primary and secondary y axes represent the percentage of homogeneous precipitation regions and the minimum number of stations belonging to a region, respectively.

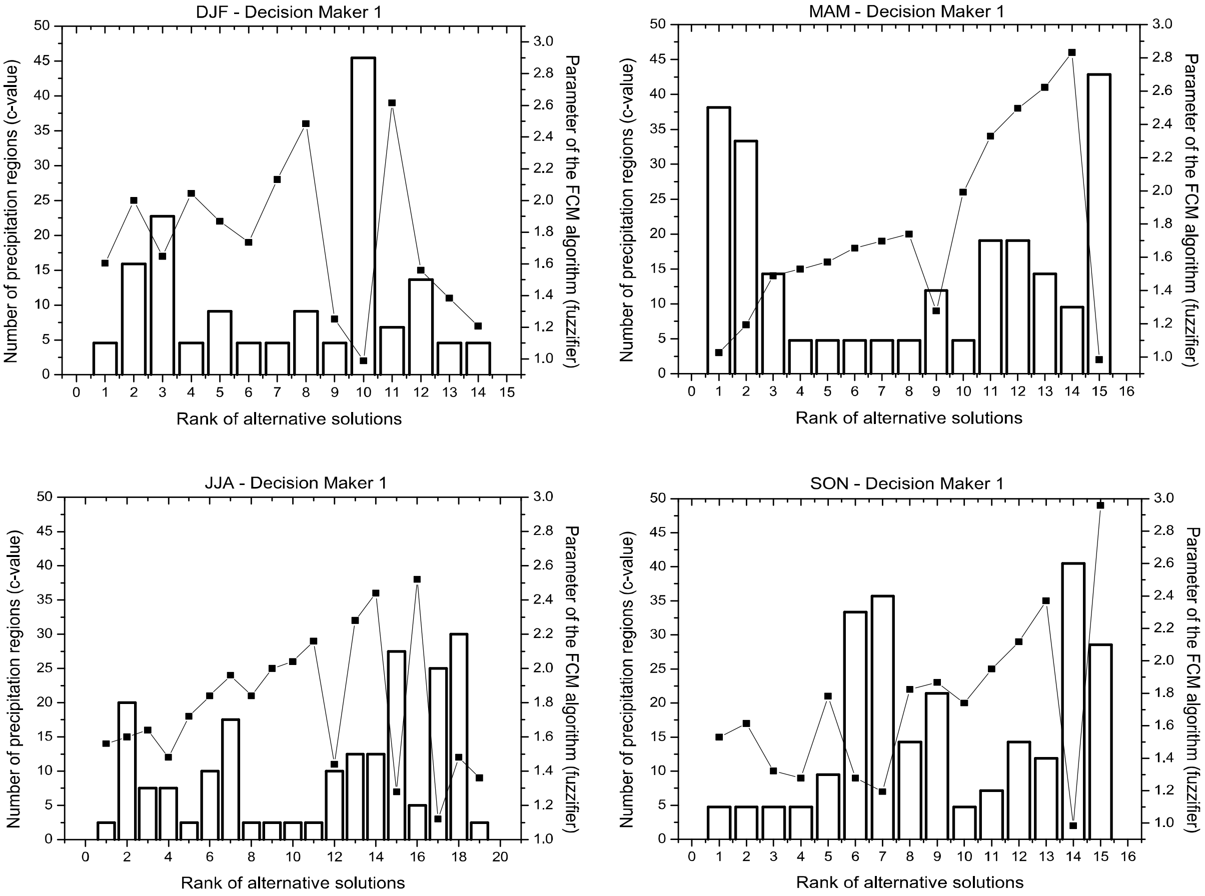


Figure 26: Ranked alternative solutions for the criteria weights set by Decision Maker 1 for the DJF, MAM, JJA and SON seasons. The line and bar plots correspond to the primary (left) and secondary (right) y axes, respectively.

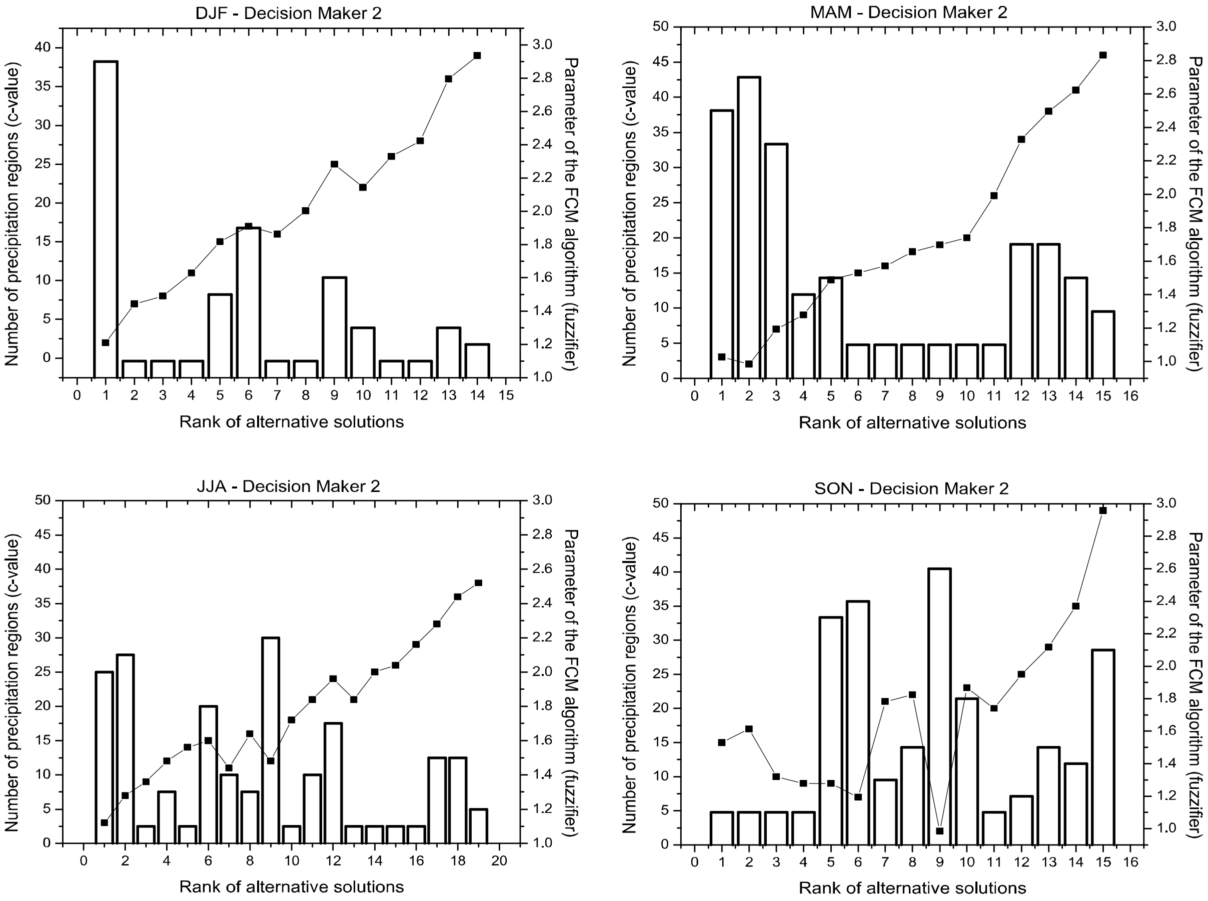


Figure 27: Ranked alternative solutions for the criteria weights set by Decision Maker 2 for the DJF, MAM, JJA and SON seasons. The line and bar plots correspond to the primary (left) and secondary (right) y axes, respectively.

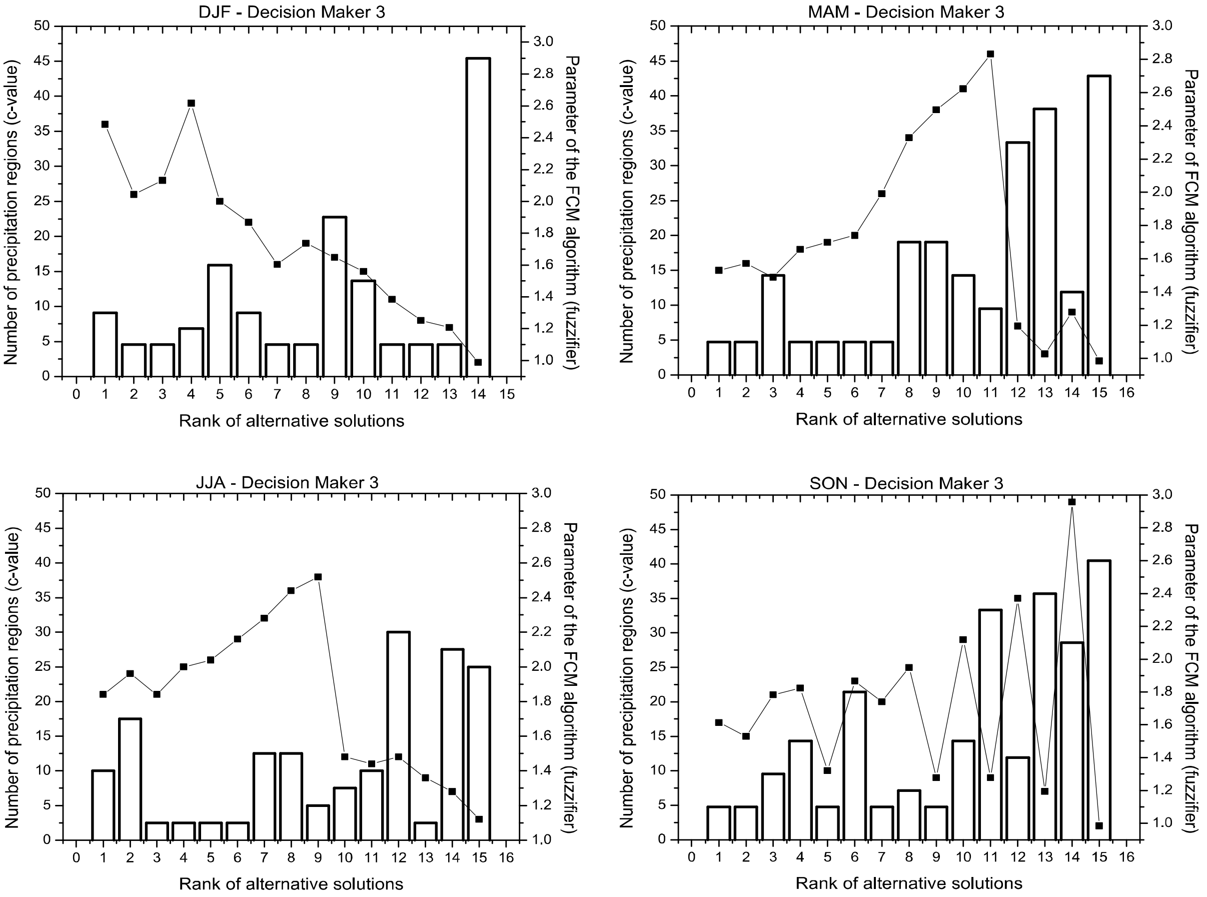


Figure 28: Ranked alternative solutions for the criteria weights set by Decision Maker 3 for the DJF, MAM, JJA and SON seasons. The line and bar plots correspond to the primary (left) and secondary (right) y axes, respectively.

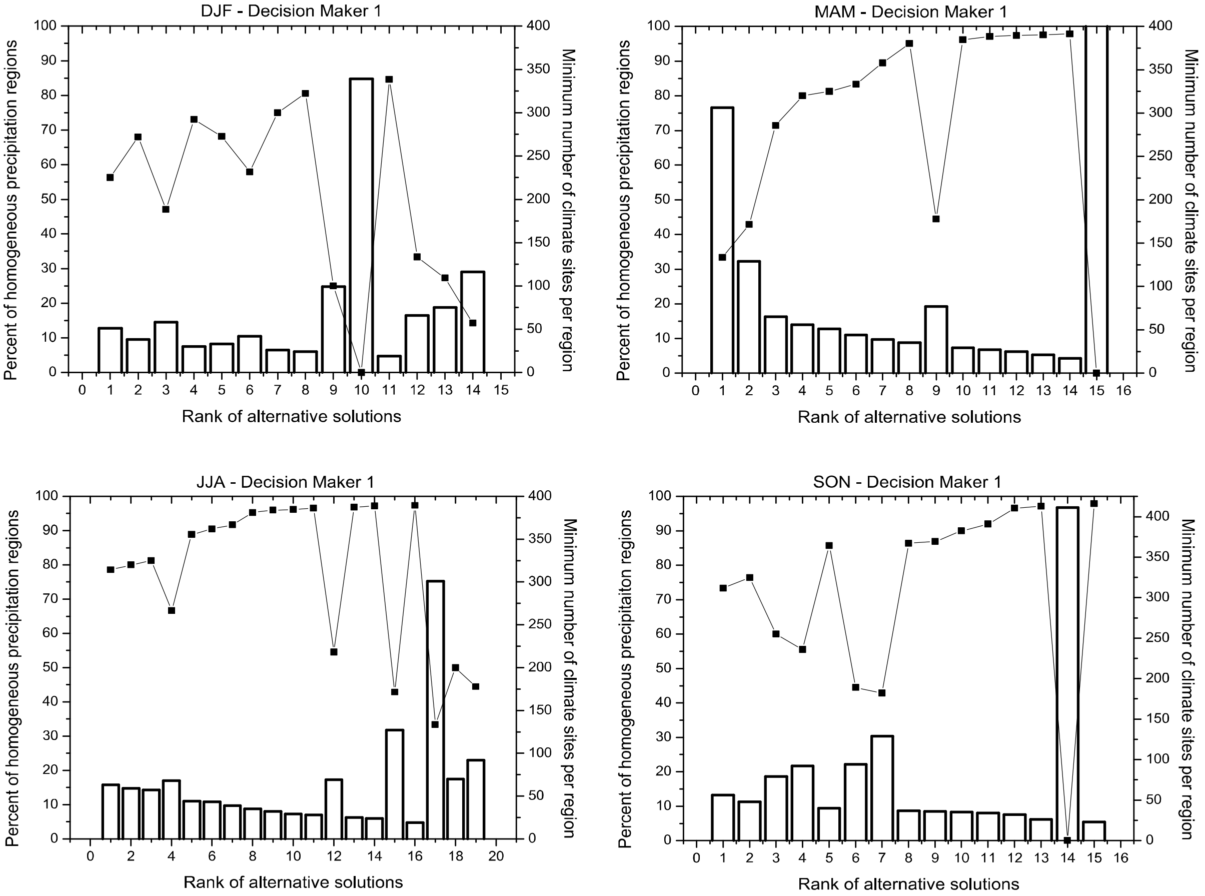


Figure 29: Objective function values of the ranked alternative solutions for the criteria weights set by Decision Maker 1 for the DJF, MAM, JJA and SON seasons. The line and bar plots correspond to the primary (left) and secondary (right) y axes, respectively.

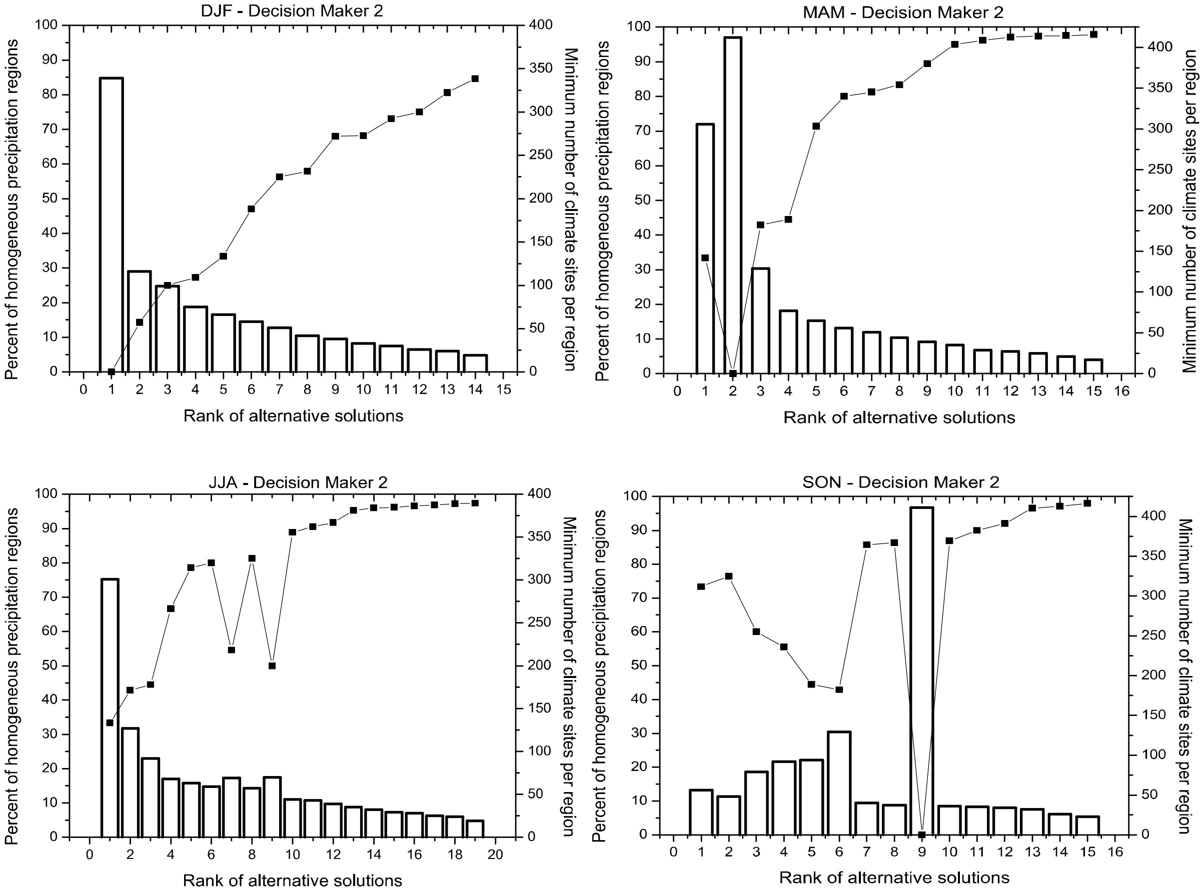


Figure 30: Objective function values of the ranked alternative solutions for the criteria weights set by Decision Maker 2 for the DJF, MAM, JJA and SON seasons. The line and bar plots correspond to the primary (left) and secondary (right) y axes, respectively.

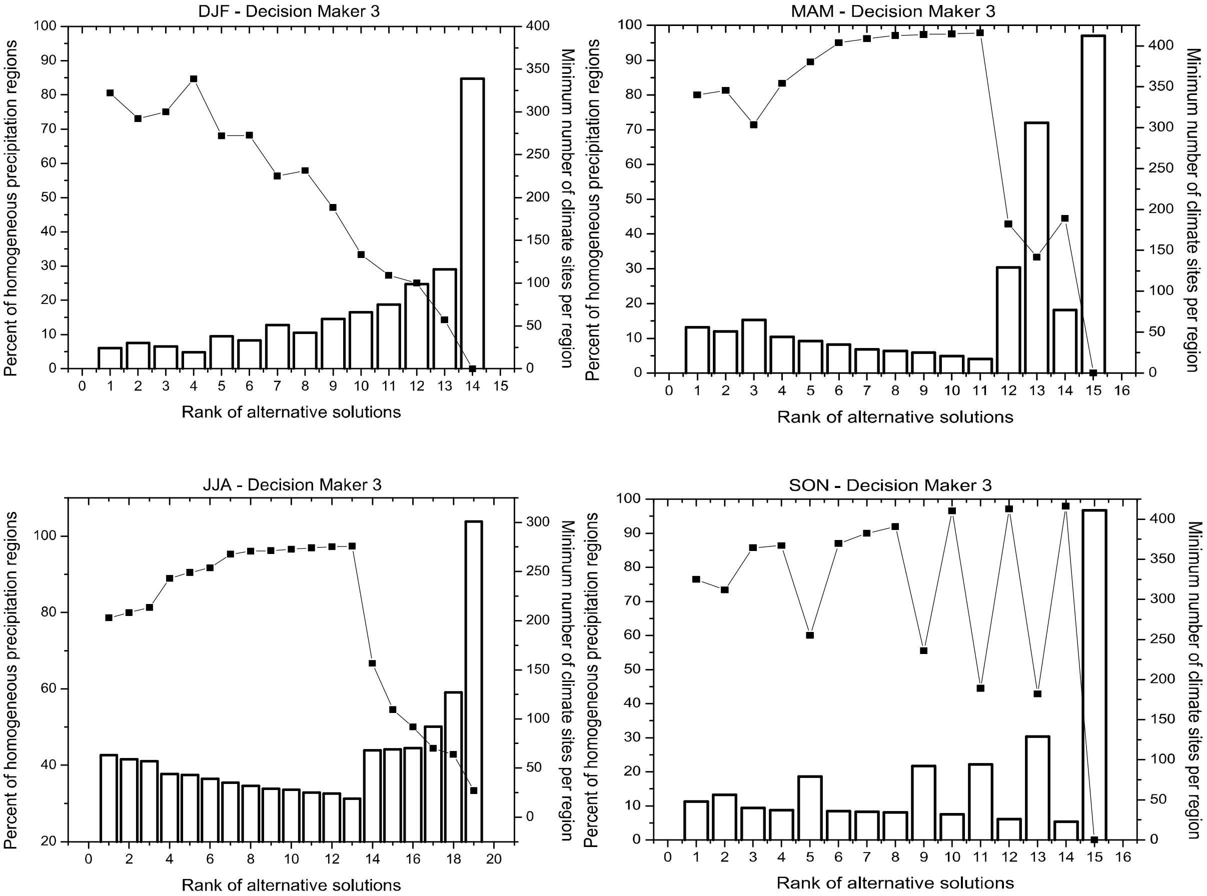


Figure 31: Objective function values of the ranked alternative solutions for the criteria weights set by Decision Maker 3 for the DJF, MAM, JJA and SON seasons. The line and bar plots correspond to the primary (left) and secondary (right) y axes, respectively.

In Figure 26 no trends are observed between the input parameters and solution ranks for Decision Maker 1 during DJF season. For the MAM, JJA and SON seasons the *c*-values increase for progressively lower ranks (where the highest rank is equal to 1); therefore the best solutions have low *c*-values. Analysis of the objective function outputs in Figure 29 does not reveal any significant trends. Since equal weights are assigned to the objectives there is an even tradeoff between the solution results and therefore, no strong trends are apparent.

The results corresponding to the preferences of Decision Maker 2 are shown in Figure 27. It is observed that the solutions' *c*-values and ranks form a strong relationship. The *c*-values increase for decreasing ranks such that the top ranked solution has the lowest *c*-value. No trends are observed between the solutions' ranks and the magnitudes of the fuzzifier. These observations are apparent for all seasons. In Figure 30 and for the DJF, MAM and JJA seasons the percentage of statistically homogeneous precipitation regions increases for progressively lower ranks. This observation is logical because the objectives are in contrast with one another and preference is given to the maximization of the minimum number of sites belonging to a region. No significant trends are observed for the results of the SON season.

In Figure 28 the model output corresponding to the preferences of Decision Maker 3 is presented. For the DJF season the *c*-values decrease for progressively lower ranks and therefore, the top ranked solution provides a large number of regions for the climate sites to be partitioned into. For the MAM and JJA seasons the *c*-values increase for decreasing ranks with the exception of the lowest ranked solutions that provide very small *c*-values. This observation is attributed to the relatively low spatial variation of precipitation during these seasons. Climate sites are not required to be partitioned into a large number of regions in order to attain a high percentage of homogeneous regions. By maintaining a relatively low value of *c,* the minimum number of sites belonging to a region is greater; thereby addressing both objectives. It is also observed that the value of the fuzzifier follows a weak increasing trend for decreasing solution ranks for all seasons. The best solutions provide low values of the fuzzifier and as such, the regional boundaries are more crisp. Clearly, Figure 31 shows that the top ranked solutions provide a high percentage of regions with uniform precipitation statistics; therefore reflecting the preference of the decision maker.

Note - a sample of the fuzzy distance metrics computed between the alternative and ideal solutions of the DJF season for Decision Maker 1 are presented Appendix B. The distance metric of the top ranked solution is highlighted. Through a visual inspection of the results it is evident that the centre of gravity of the top ranked solution is closest to the ideal solution that exists where the x-axis is equal to 0.

Chapter 7

7. Conclusion

Regional frequency analysis is used to obtain reliable estimates of local precipitation events. The general procedure involves: (i) the partitioning of climate sites into statistically homogeneous precipitation regions; and (ii) the combination of precipitation data that is recorded within the same region into a single frequency distribution from which local precipitation is estimated. The focus of the presented research is on the first step of the procedure; that is the formation of precipitation regions (also referred to as regionalization).

The compositions of the precipitation regions are sensitive to the selection of several subjective choices including the regionalization method, the number of regions to which the sites are assigned, the climate site attributes and the temporal resolution of the precipitation data. The primary objective of the presented research is to divide the regionalization procedure into its individual components and to assess their influences on the regional output. Summaries are provided for the assessments on the aforementioned regionalization components (the method, site attributes, temporal resolution); in addition to a review of the performance of a proposed regionalization model that uses optimization to select input parameters to the fuzzy *c*-means algorithm. Finally open issues for future works are suggested.

7.1 Selection of a Regionalization Method

Clustering is recognized as a suitable approach to the regionalization of precipitation because of its innate ability to identify underlying patterns in complex datasets. Of the available clustering algorithms, partitional clustering algorithms are preferred for their ability to iteratively update site membership values at each step in the algorithm. The most commonly used partitional clustering algorithm is the k-means technique. The fuzzy *c*-means algorithm, however, is becoming increasingly popular for the incorporation of its membership function that assigns partial membership values to the sites for each cluster. As such the fuzzy *c*-means algorithm is elected as the most suitable method to be used in the presented research.

To justify its application the performance of the k-means and fuzzy *c*-means algorithms are compared for the employment of three different attribute sets and monthly precipitation data that is organized into four seasons including: (i) DJF (December, January, February); (ii) MAM (March, April, May); JJA (June, July, August); and (iv) SON (September, October, November). In the Great Lakes region it is observed that the fuzzy *c*-means algorithm outperformed the k-means method for the majority of scenarios, and in the Prairie region the fuzzy *c*-means algorithm provided the best results in terms of the selection criteria for all scenarios.

7.2 Selection of Site Attributes

An investigation is conducted to determine the preferred attribute set to be employed to delineate precipitation regions for two study areas in Canada; the Prairie and the Great Lakes-St. Lawrence lowlands. The results are compared based on the percentage of large homogeneous precipitation regions formed and the computational efficiency of the regionalization and validation procedures. Location parameters including latitude, longitude, elevation and distance to major water bodies, and a complete set of atmospheric variables that are modeled at several pressure levels are considered as potential attributes.

Overall, it is recommended to use location parameters as attributes in the regionalization of precipitation. In this analysis, location parameters marginally outperform all attribute combinations in terms of regional homogeneity of precipitation and reduced computational time in two distinct study areas. Furthermore, they are easily attainable and can be applied in ungauged or data-sparse regions. It is not recommended to include site elevations in the Prairie climate region of Canada as it negatively impacts the results. For climate change applications, atmospheric variables should be considered as attributes. In this scenario attribute screening (linear correlation analysis) can be applied to reduce computational demands; however it does not provide any additional benefits in terms of the quality of the precipitation regions. Future values of atmospheric variables may be obtained from GCM projections.

7.3 Temporal Resolution of Precipitation Data

To study the effect of the temporal resolution (monthly, seasonal, annual and the maximum annual series) of the data on the formation of homogeneous precipitation regions an analysis is conducted in two diverse study areas; the Great Lakes-St. Lawrence lowlands and the Prairie climate regions of Canada. The fuzzy *c*-means algorithm and *L*-moment regional heterogeneity test are employed to delineate the regions. The goal of this investigation is to detect any relationships between the precipitation regions that are derived for different temporal resolutions such that the precipitation data can be accurately combined or disaggregated into different resolutions for applications in regional frequency analysis. This information is useful for studies that require precipitation to be recorded at a specific resolution that is limited or unavailable. Important observations are highlighted below.

It is recommended to perform the regionalization procedure for the temporal resolution that is most suited to the regional frequency analysis application for both study areas. The precipitation regions that are classified as homogeneous for one temporal resolution may not be for another. It is discovered that (i) the variations in the spatial distributions of the precipitation regions are greater for finer time scales (monthly) as opposed to coarser resolutions (seasonal and annual); and (ii) there does not appear to be a relationship between the regions formed for different temporal resolutions and time periods. For example, if precipitation regions are to be derived for their application in water infrastructure design in the Great Lakes region, the temporal resolution should be set to the maximum annual series (or daily, sub-daily) in order to obtain accurate results. Regions that are derived from the employment of the annual, seasonal and monthly temporal resolutions are vastly different than those formed for the maximum annual series and therefore, their application would lead to a deficient design. The regional frequency distribution that is derived from a heterogeneous region (for the temporal resolution of interest) does not accurately represent the local precipitation statistics.

7.4 Proposed Regionalization Model

A model is proposed for the regionalization of precipitation using a fuzzy clustering technique. It uses optimization to select input parameters to the fuzzy *c*-means clustering algorithm such that two objectives are achieved in the results: (i) maximization of the number of station-years in the regional precipitation record; and (ii) maximization of the number of precipitation regions that are classified as homogeneous through the validation procedure. The model is evaluated for its ability to generate and rank sets of alternative solutions that satisfy the objective functions. It is employed in the Great Lakes-St. Lawrence lowlands climate region of Canada to delineate monthly precipitation regions for four seasons (DJF, MAM, JJA and SON).

The model consists of two main components; differential evolution that performs optimization to derive a set of alternative, optimal solutions that best satisfy the conflicting objective functions simultaneously; and fuzzy Compromise programming that is subsequently employed to evaluate and rank the alternative solutions in order of their proximity to the ideal solution in the objective space. Through an analysis of the results it is observed that the model is capable of capturing the seasonal differences in the spatial variability of precipitation for the generation of alternative solutions. In addition, the fuzzy Compromise programming model effectively reflects the decision maker's preferences in the evaluation and ranking of the alternative solutions. The rank is greatly dependent upon the criteria weights that are assigned according to the application and the priorities of the decision maker. Overall, the proposed model incorporates an objective method for the selection of input parameters to the fuzzy *c*-means algorithm. It is capable of generating solutions that provide reasonable values for the defined objective functions and therefore, the outcomes of regionalization. Finally, the model has proven to be robust as it can adapt to different data inputs and decision maker preferences.

7.5 Recommendations for Future Works

Open issues to be addressed in future research are as follows:

(i) Test for non-linear relationships between atmospheric variables and local precipitation in the attribute selection process.

(ii) Improvement of the homogeneity of precipitation regions in the Great Lakes-St. Lawrence lowlands climate region using new attribute combinations, including near and remote teleconnection indices.

(iii) Regionalization of precipitation for finer temporal scales (daily, sub-daily); and study the effect of different storm durations on the regional formations.

(iv) Quantification of uncertainties in the regionalization process due to the non-stationarity of precipitation as a result of climate change; and identification of temporal shifts in precipitation regions under climate change scenarios.

(v) Addition of a third decision variable to the proposed fuzzy regionalization model that is the threshold (α - cut off) value for climate station membership in a precipitation region. The current version of the model uses a subjective equation to calculate threshold value (see section 5.3)

(vi) Improvement of the computational time required to run the proposed model.

References

Acreman MC, Wiltshire SE. 1987. Identification of regions for regional flood frequency analysis (abstract). *EOS Transactions American Geophysical Union AGU* **68(44)**: 1262.

Adelekan, IO. 1998 Spatio-temporal variations in thunderstorm rainfall over Nigeria, *International Journal of Climatology* **18(11):** 1273 - 1284.

Ashmore P, Church M. 2001. The impact of climate change on rivers and river processes in Canada. *Geological Survey of Canada Bulletin 555*. Ottawa, Canada.

Asong ZE, Khaliq MN, Wheater HS. 2015. Regionalization of precipitation characteristics in the Canadian Prairie Provinces using large-scale atmospheric covariates and geophysical attributes. *Stochastic Environmental Research Risk Assess* **29 (3):** 875-892. doi: 10.1007/s00477-014-0918-z.

Azad S, Vignesh TS, Narasimha R. 2010. Periodicities in Indian monsoon rainfall over spectrally homogeneous regions. *International Journal of Climatology* **30(15)**: 2289–2298.

Baeriswyl PA, Rebetez M. 1997. Regionalisation of Precipitation in Switzerland by means of Principal Component Analysis. *Theor. Appl. Climatol.* **58**: 31-41.

Bender MJ, Simonovic SP. 2000. A fuzzy compromise approach to water resource systems planning under uncertainty. *Fuzzy Sets and Systems* **115(1)**: 35-44.

Bezdek JC. 1981. Pattern Recognition with Fuzzy Object Function Algorithms. *Advanced Applications in Pattern Recognition.* Plenem Press: New York.

Bezdek JC, Pal NR. 1995. Two soft relatives of learning vector quantization. Neural Networks **8(5)**: 729–743.

Bhatia N, Irwin S, Srivastav RK, Simonovic SP. 2015. Delineation of Precipitation Regions: The Role of Cluster Validity Indices *(under review).*

Burn DH. 1990. Evaluation of regional flood frequency analysis with a region of influence approach. *Water Resources Research* **26 (10):** 2257-2265. doi. 10.1029/WR026i010p02257.

Burn DH. 1997. Catchment similarity for regional flood frequency analysis using seasonality measures. *Journal of Hydrology* **202(1-4):** 212-230. doi:10.1016/S0022-1694(97)00068-1.

Burn DH, Goel NK. 2000. The formation of groups for regional flood frequency analysis. *Hydrological Sciences Journal* **45 (1):** 97-112. doi.10.1080/02626660009492308.

Burn DH. 2014. A framework for regional estimation of intensity-duration-frequency (IDF) curves. *Hydrological Processes* **28**: 4209-4218.

Carter MM, Elsner JB. 1997. A statistical method for forecasting rainfall over Puerto Rico, *Weather and forecasting* **12(3)**: 515–525.

Castellarin A, Burn DH, Brath A. 2008. Homogeneity testing: How homogeneous do heterogeneous cross-correlated regions seem? *Journal of Hydrology* **360**: 67– 76.

Cavadias GS. 1990. The canonical correlation approach to regional flood estimation. *Regionalization in Hydrology*, (ed by MA Beran, M Brillym A Becker, O Bonacci) (Proc. Ljubljana Symp., April 1990), 171 - 178. IAHS Publ. no. 191.

Cavadias GS, Ouarda TBMJ, Bobée B, Girard C. 2001. A canonical correlation approach to the determination of homogeneous regions for regional flood estimation of ungauged basins. *Hydrological Sciences* **46**(4).

Chebana F, Ouarda TBMJ. 2008. Depth and homogeneity in regional flood frequency analysis. *Water Resources Research* **41(11)**.

Chen LJ, Chen DL, Wang HJ, Yan JH. 2009. Regionalization of precipitation regimes in China. *Atmospheric and Oceanic Science Letters* **2(5):** 301-307.

Chin, D.A. 2006. Water-Resources Engineering. Pearson Education, Inc. Upper Saddle River, New Jersey, 271-179 pp.

Coello Coello CA, Lamont GB, Veldhuizen DA. 2007. Evolutionary Algorithms for Solving Multi-Objective Problems. Berlin: Springer-Verlag.

Comrie AC, Glenn EC. 1998. Principal components-based regionalization of precipitation regimes across the Southwest United States and Northern Mexico, with an application to monsoon precipitation variability. *Clim. Res.***10:** 201-215. doi.10.3354/cr010201.

Dikbas F, Firat M, Koc AC, Gungor M. 2012. Classification of precipitation series using fuzzy cluster method. *Int. J. Climatol.* **32(10):** 1596–1603. doi. 10.1002/joc.2350.

Dikbas F, Firat M, Cem Koc A, Gungor M. 2013. Defining Homogeneous Regions for Streamflow Processes in Turkey Using a K-Means Clustering Method. *Arabian Journal for Science and Engineering* **38(6):** doi: 10.1007/s13369-013-0542-0.

Easterling DR. 1989. Regionalization of thunderstorm rainfall in the contiguous United States, *Int. J. Climatol.* **9(6):** 567–579. doi:10.1002/joc.3370090603.

Eum H-I, Dibike Y, Prowse T, Bonsal B. 2014. Inter-comparison of high-resolution gridded climate data sets and their implications on hydrological model simulation over the Athabasca Watershed, Canada. *Hydrological Processes* **28(14):** 4250-4271. doi. 10.1002/hyp.10236.

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

Gadgil S, Yadumani, Joshi NV. 1993. Coherent rainfall zones of the Indian region. *International Journal of Climatology* **13:** 547–566.

Gong X, Richman MB. 1995. On the Application of Cluster Analysis to Growing Season Precipitation Data in North America East of the Rockies. *Journal of Climate* **8**: 897-931.

Goyal MK, Gupta S. 2014. Identification of Homogeneous Rainfall Regions in Northeast Region of India using Fuzzy Cluster Analysis. *Water Resources Management* **28**: 4491-4511. doi. 10.1007/s11269-014-0699-7

Gurrutxaga I, Arbelaitz O, Muguerza J, Pe´rez J M, Perona I. 2013. An extensive comparative study on cluster validity indices. *Pattern Recognition* **46**: 243-256.

Hosking JRM, Wallis JR. 1997. Regional frequency analysis: an approach based on *L*-moments. Cambridge University Press, New York, USA.

Hosking JRM. 2013. Regional frequency analysis using *L*-moments, R package, version 205. Available from: http://CRAN.R-project.org/package=lmomRFA

Hopkinson RF, McKeney DW, Milewska EJ, Hutchinson MF, Papadopol P, Vincent LA. 2011. Impact of aligning climatological day on gridding daily maximum–minimum temperature and precipitation over Canada. *Journal of Applied Meteorology and Climatology* **50(8):** 1654–1665. doi. http://dx.doi.org/10.1175/2011JAMC2684.1.

Huang Z. 1998. Extensions to the k-means algorithm for clustering large data sets with categorical values. *Data Mining and Knowledge Discovery* **2**: 283-304.

Hutchinson MF, McKenney DW, Lawrence K, Pedlar JH, Hopkinson RF, Milewska E, 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(4):** 725–741. doi: http://dx.doi.org/10.1175/2008JAMC1979.1

Irwin S, Srivastav RK, Simonovic SP, Burn, DH. 2015. Delineation of Precipitation Regions using Location and Atmospheric Variables in Two Canadian Climate Regions: The Role of Attribute Selection. (Under review).

Irwin S. Srivastav RK, Simonovic SP, Burn DH. 2015. Delineation of Precipitation Regions using Location Parameters in Two Canadian Climate Regions: The Role of the Temporal Resolution. (Under review).

Jebari, S., Berndtson, R., Uvo, C. & Bahri, A. 2007 Regionalizing fine time-scale rainfall affected by topography in semi-arid Tunisia. Hydrol. Sci. J.52(6), 1199–1215.

Jiang, P., M. R. Gautam, J. Zhu, and Z. Yu 2013, How well do the GCMs/RCMs capture the multi-scale temporal variability of precipitation in the southwestern United States?, J. Hydrol.,479,75–85

Jaafar WZ, Liu J, Han D. 2011. Input variable selection for median flood regionalization. *Water Resources Research* **47(7**). doi. 10.1029/2011WR010436.

Johnson GL, Hanson CL. 1995. Topographic and atmospheric influences on precipitation variability over a mountains watershed. *J. Appl. Meteorol.* **34:** 67-87. doi: http://dx.doi.org/10.1175/1520-0450-34.1.68.

Kalkstein LS, Tan G, Skindlov JA. 1987. An Evaluation of Three Clustering Procedures for Use in Synoptic Climatological Classification. *Journal of Climate and Applied Meteorology* **26**: 717-730.

Kaufman L, Rousseeuw PJ. 1987. Clustering by means of Medoids. *Statistical Data Analysis Based on the L_1–Norm and Related Methods* 405–416.

Kim M, Ramakrishna R S. 2005. New indices for cluster validity assessment. *Pattern Recognition Letters* **26**: 2353-2363.

Kodinariya, T.M., Makwana, P.R. 2013. Review on determining number of cluster in k-means clustering. International Journal of Advanced Research in Computer Science and Management Studies, 1(6), ISSN: 2321-7782 (online).

Kulkarni A, Kripalani RH. 1998. Rainfall Patterns over India: Classification with Fuzzy *C*-Means Method. *Theoretical and Applied Climatology* **59 (3-4):** 137 – 146.

Lapen, D.R., Hayhoe, H.N. 2003 Spatial analysis of seasonal and annual temperature and precipitation normals in Southern Ontario, Canada. Journal of Great Lake Research, 29: 529–544

MacQueen JB, 1967. Some Methods for classification and Analysis of Multivariate Observations. *Proceedings of 5-th Berkeley Symposium on Mathematical Statistics and Probability*. Berkeley, University of California Press **1:** 281-29 7.

Matulla C, Penlap E K, Haas P, Formayer H. 2003. Comparative analysis of spatial and seasonal variability: Austrian precipitation during the 20th century. *International Journal of Climatology* **23**: 1577-1588.

McGinn SM. 2010. Weather and climate patterns in Canada's prairie grasslands., in Shorthouse, J.D. and Floate, K.D. (eds.) - Arthropods of Canadian Grasslands. *Ecology and Interactions in Grassland Habitats, Biological Survey of Canada (BSC)* **1 (5):** 105-119.

Michaelides CS, Constantinos S, Pattichis, Kleovoulou, G. 2001. Classification of rainfall variability by using artificial neural networks. *International Journal of Climatology* **21**(11).

Mills, FG. 1995. Principal component analysis of precipitaiton and rainfall regionalization in Spain. *Theoretical and Applied Climatology* **50**: 169-183.

Nathan RJ, McMahon TA. 1990. Identification of homogeneous regions for the purposes of regionalization. *Journal of Hydrology* **201(1-4):** 217-238. doi:10.1016/0022-1694(90)90233-N.

Ouarda TBMJ, Girard C, Cavadias GS, Bobe´e B. 2001. Regional flood frequency estimation with canonical correlation analysis. *Journal of Hydrology*. **254**: 157–173.

Pelchzer IJ, Cisneros-Iturbe HL. 2008. Identification of rainfall patterns over the Valley of Mexico. Proceedings of 11th International Conference of Urban Drainage, Edinburgh, Scotland, UK.

Price KV, Storn RM. 1997. Differential evolution - A simple evolution strategy for fast optimization. *Dr. Dobb's Journal* **22**: 18-24.

Rao AR, Srinivas VV. 2006. Regionalization of watersheds by hybrid-cluster analysis. *Journal of Hydrology* **318**: 37-56.

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

Saikranthi K., Rao T.N., Rajeevan M., and Rao S.V.B. 2013. Identification and validation of homogeneous rainfall zones in India using correlation analysis, Journal of Hydrometeorology, 14(1): 304–317, DOI: 10.1175/JHM-D-12–071.1.

Satyanarayana P, Srinivas VV. 2008. Regional frequency analysis of precipitation using large-scale atmospheric variables. *Journal of Geophysical Research* **113(24)**. doi. 10.1029/2008JD010412.

Satyanarayana P, Srinivas VV. 2011. Regionalization of precipitation in data sparse areas using large scale atmospheric variables - A fuzzy clustering approach. *Journal of Hydrology* **405(3-4):** 462-473. doi:10.1016/j.jhydrol.2011.05.044.

Schardong A, Simonovic SP. 2011. Multi-objective Evolutionary Algorithms for Water Resources Management. Water Resources Research Report no. 078, Facility for Intelligent Decision Support, Department of Civil and Environmental Engineering, London, Ontario, Canada, 167 pages. ISBN: (print) 978-0-7714-2907-1; (online) 978-0- 7714-2908-8

Scott R W, Huff FA. 1996. Impacts of the Great Lakes on regional climate conditions. Journal of Great Lakes Research **22**: 845–863.

Serra C, Fernandez Mills G, Periago M C, Lana X. 1996. Winter and Autumn Daily Precipitation Patterns in Catalonia, Spain. *Theoretical and Applied Climatology* **54**: 175-186.

Simonovic SP. 2009 Managing Water Resources - Methods and Tools for a Systems Approach. London: UNESCO Publishing.

Srinivas VV. Tripathi S, Rao AR, Govindaraju RS. 2008. Regional flood frequency analysis by combined self-organizing feature map and fuzzy clustering. *Journal of Hydrology* **348**: 148–166.

Sousounis PJ. 2001. Lake effect storms. *Encyclopedia of Atmospheric Sciences*. J. Holton, J. Pyle, and J. Curry, Eds., AcademicPress: 1104–1115.

Srinivas VV. 2013. Regionalization of Precipitation in India - A Review. *Journal of the Indian Institute of Science* **93(2)**: 153-162.

Statistics Canada. 2007. Weather conditions in capital and major cities (Precipitation). Retrieved September 2014 from http://www.statcan.gc.ca/tables-tableaux/sum-som/l01/cst01/phys08a-eng.htm.

Statistics Canada. 2012. Human Activity and the Environment publication. Retrieved 15 September 2014 from http://www.statcan.gc.ca/pub/16-201-x/2007000/5212632-eng.htm.

Storn RM, Price KV. 1995. Differential Evolution – A simple and efficient adaptive scheme for global optimization over continuous spaces. *Technical Report TR-95-012*, ICSI.

Unal Y, Kindap T, Karaca M. 2003. Redefining the climate zones of Turkey using cluster analysis. *International Journal of Climatology* **23**(9): 1045-1055.

USEPA. 2012. The Great Lakes: An Environmental Atlas and Resource Book. U.S. Environmental Protection Agency. Retrieved September 2014 from http://epa.gov/greatlakes/atlas/glat-ch1.html.

Wagener T, Wheater HS, Gupta HV. 2004. Rainfall-Runoff Modelling in Gauged and Ungauged Catchments. Imperial College Press, Singapore.

Xie XL, Beni G. 1991. A validity measure for fuzzy clustering. *IEEE Transactions on Pattern Analysis and Machine Intelligence* **13 (8)**: 841–847.

Zalik K R, Zalik B. 2011. Validity indices for clusters of different sizes and densities. *Pattern Recognition Letters* **32**: 221-234.

APPENDIX A - Correlation analysis results

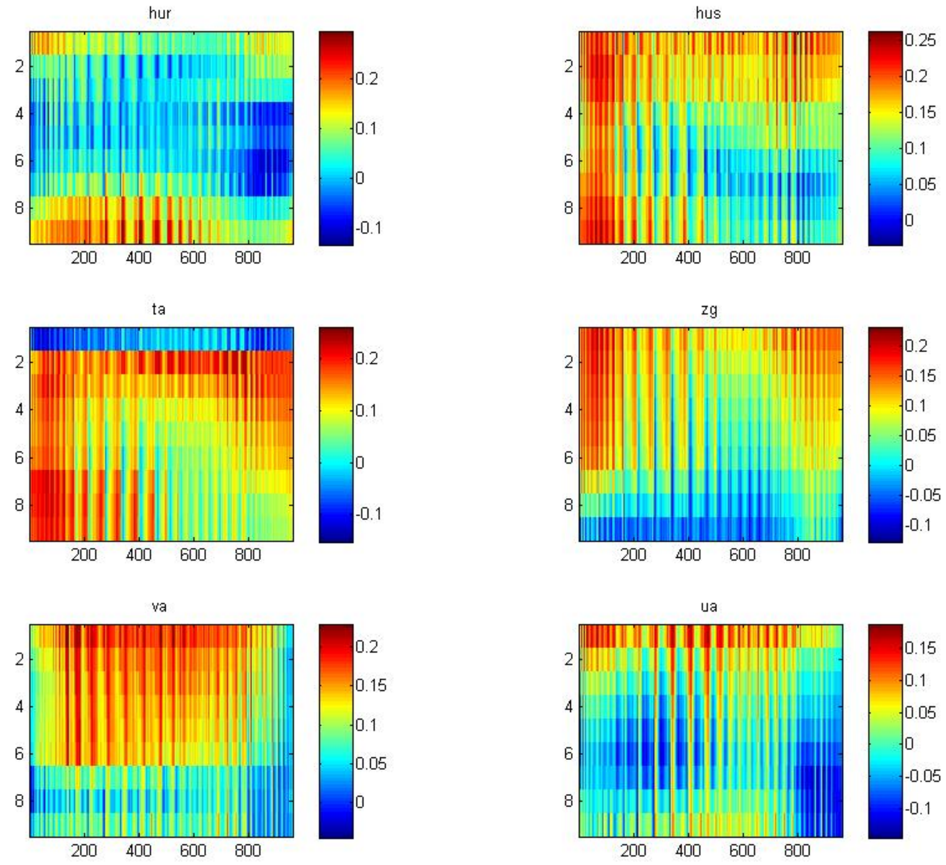


Figure A32: Correlation analysis results for the DJF season in the Great Lakes region. There is one plot per atmospheric variable: (i) relative humidity (hur); (ii) specific humidity (hus); (iii) air temperature (ta); (iv) geopotential height (zg); (v) Northward wind component (va); (vi) Eastward wind component (ua). The number of climate sites and pressure levels (x10 kPa) are plotted on the x and y axes. The colour scale corresponds to the correlation coefficient.

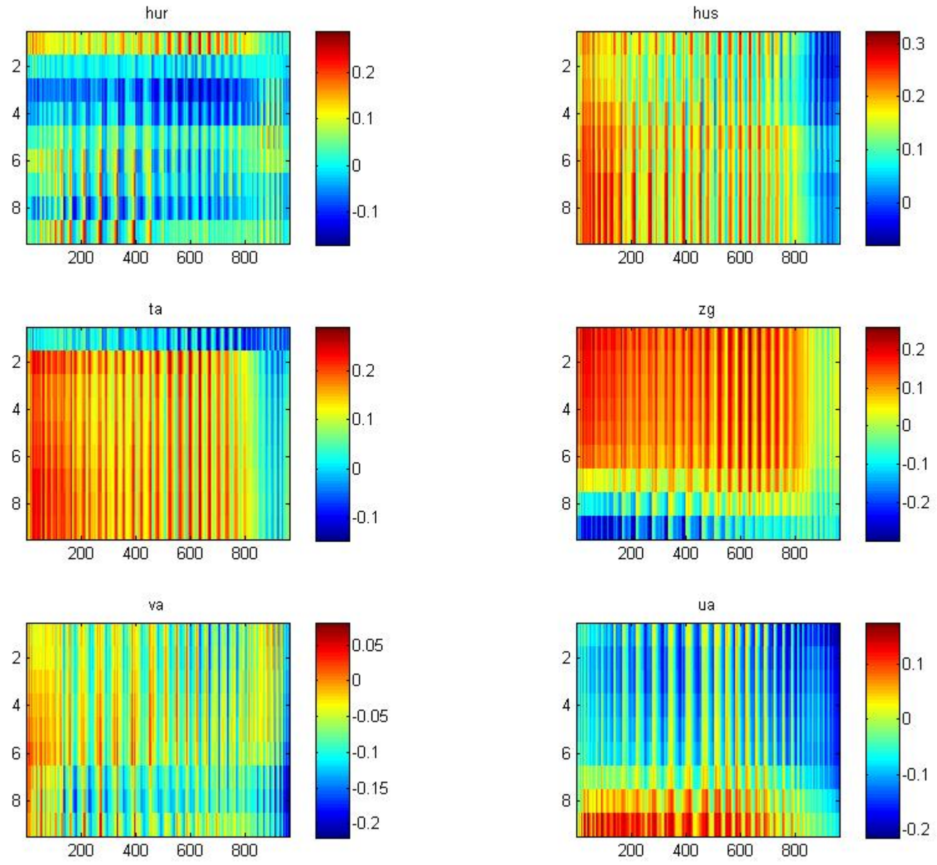


Figure A33: Correlation analysis results for the MAM season in the Great Lakes region. There is one plot per atmospheric variable: (i) relative humidity (hur); (ii) specific humidity (hus); (iii) air temperature (ta); (iv) geopotential height (zg); (v) Northward wind component (va); (vi) Eastward wind component (ua). The number of climate sites and pressure levels (x10 kPa) are plotted on the x and y axes. The colour scale corresponds to the correlation coefficient.

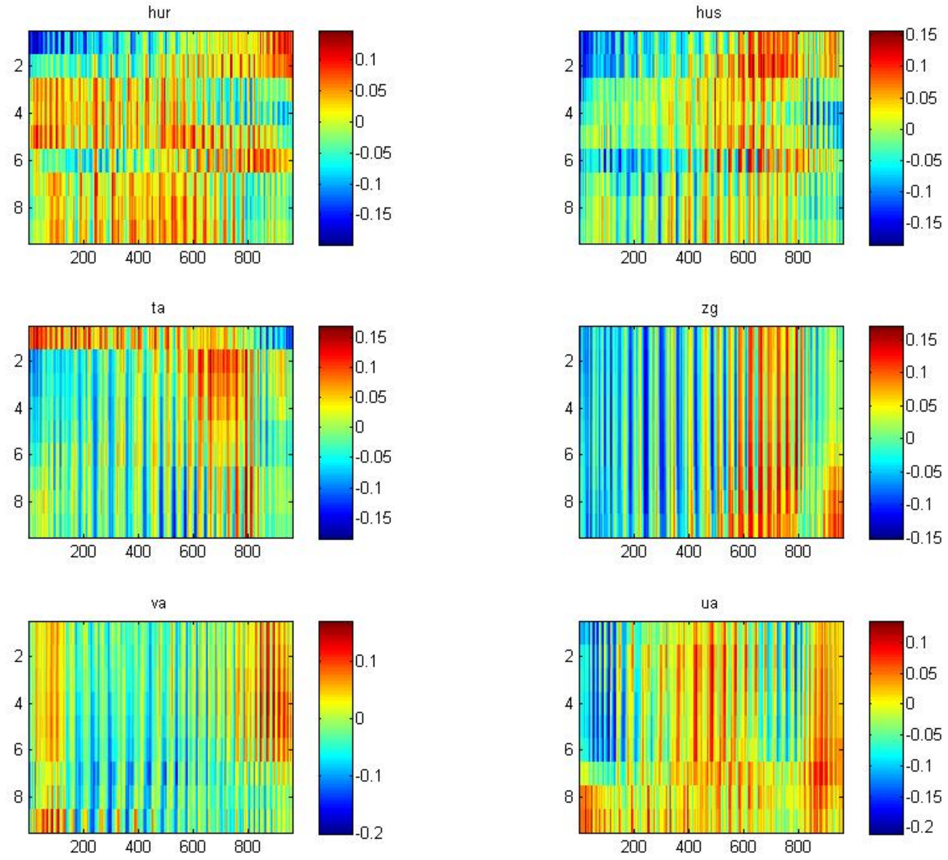


Figure A34: Correlation analysis results for the JJA season in the Great Lakes region. There is one plot per atmospheric variable: (i) relative humidity (hur); (ii) specific humidity (hus); (iii) air temperature (ta); (iv) geopotential height (zg); (v) Northward wind component (va); (vi) Eastward wind component (ua). The number of climate sites and pressure levels (x10 kPa) are plotted on the x and y axes. The colour scale corresponds to the correlation coefficient.

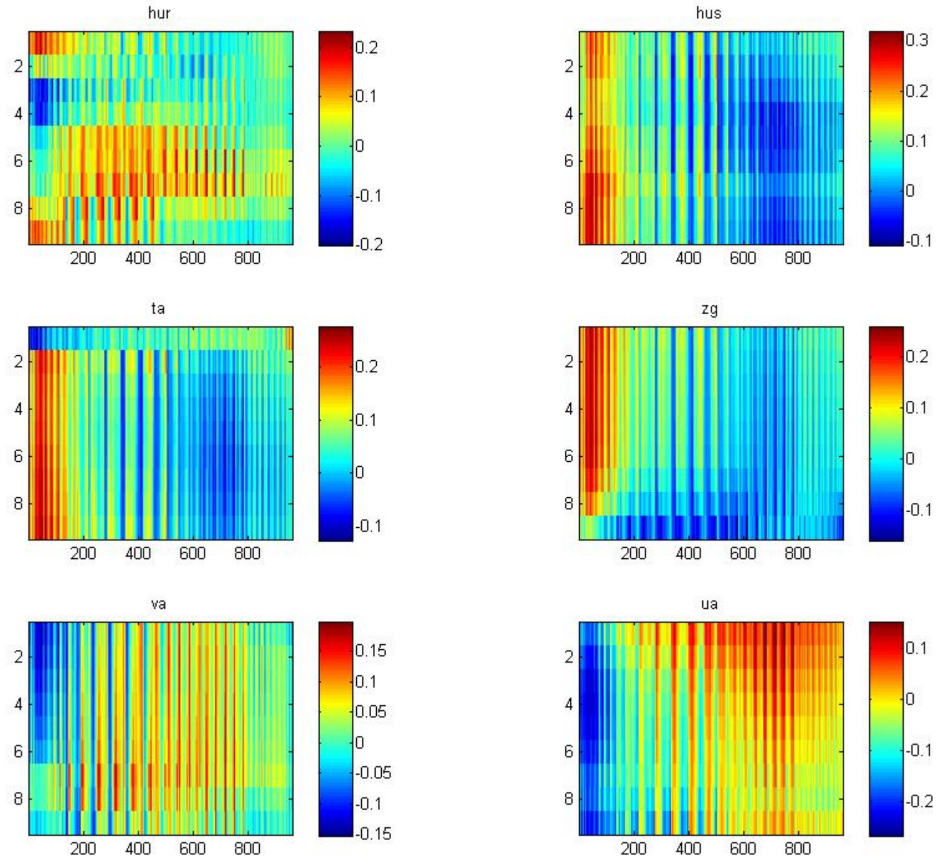


Figure A35: Correlation analysis results for the SON season in the Great Lakes region. There is one plot per atmospheric variable: (i) relative humidity (hur); (ii) specific humidity (hus); (iii) air temperature (ta); (iv) geopotential height (zg); (v) Northward wind component (va); (vi) Eastward wind component (ua). The number of climate sites and pressure levels (x10 kPa) are plotted on the x and y axes. The colour scale corresponds to the correlation coefficient.

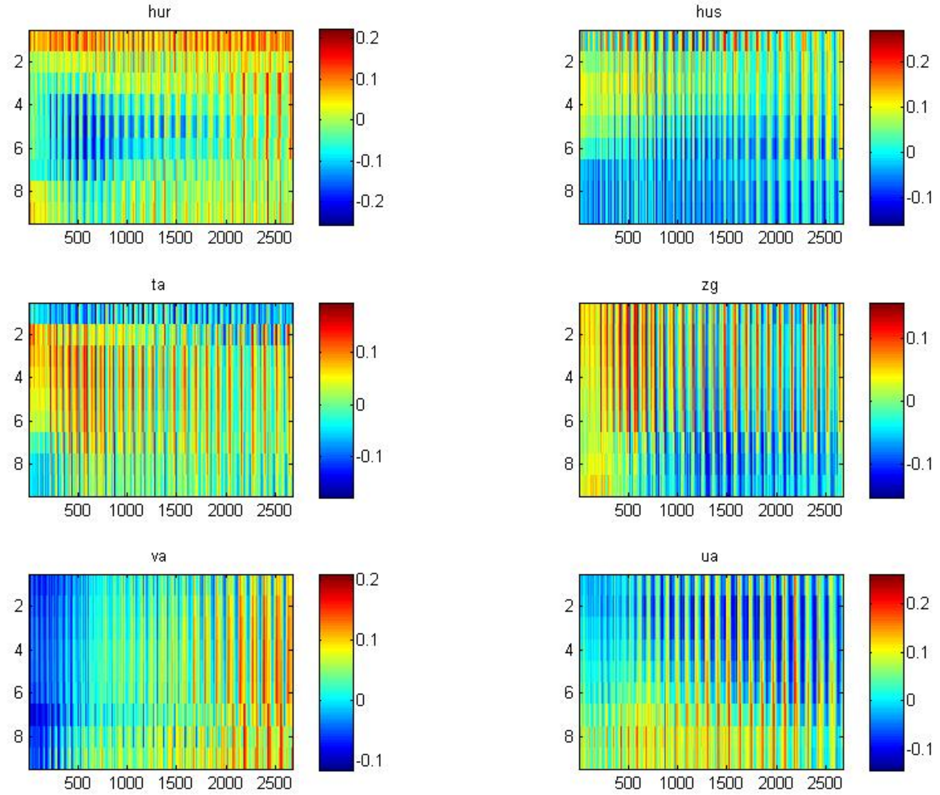


Figure A36: Correlation analysis results for the DJF season in the Prairie region. There is one plot per atmospheric variable: (i) relative humidity (hur); (ii) specific humidity (hus); (iii) air temperature (ta); (iv) geopotential height (zg); (v) Northward wind component (va); (vi) Eastward wind component (ua). The number of climate sites and pressure levels (x10 kPa) are plotted on the x and y axes. The colour scale corresponds to the correlation coefficient.

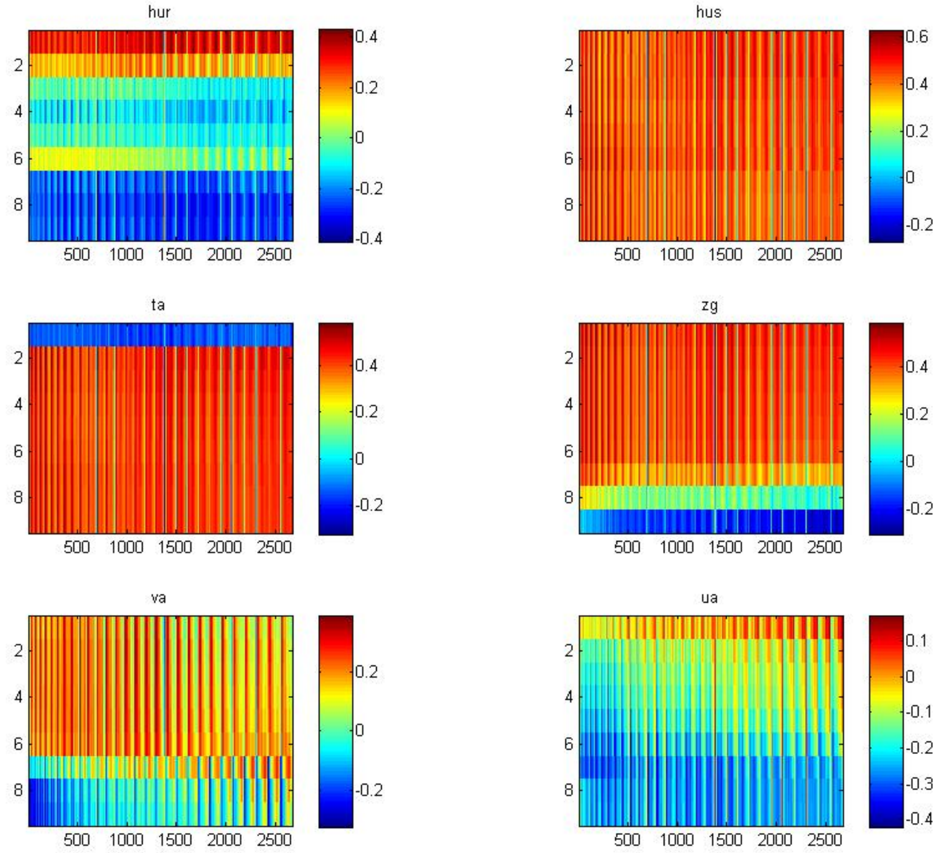


Figure A37: Correlation analysis results for the MAM season in the Prairie region. There is one plot per atmospheric variable: (i) relative humidity (hur); (ii) specific humidity (hus); (iii) air temperature (ta); (iv) geopotential height (zg); (v) Northward wind component (va); (vi) Eastward wind component (ua). The number of climate sites and pressure levels (x10 kPa) are plotted on the x and y axes. The colour scale corresponds to the correlation coefficient.

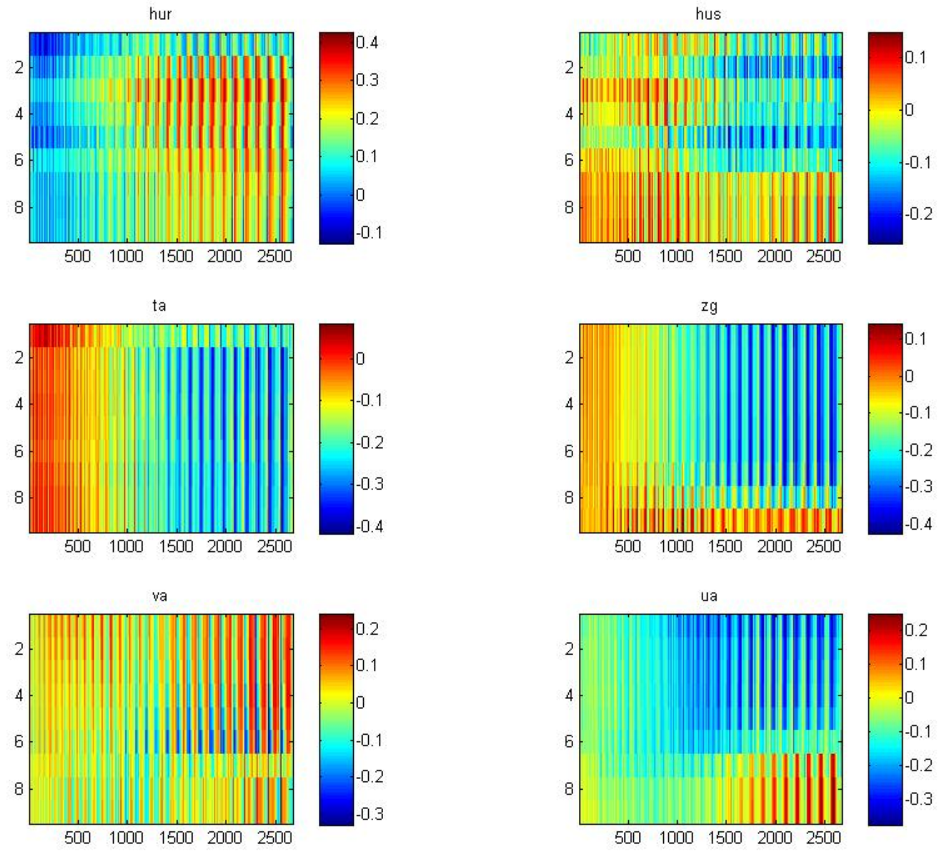


Figure A38: Correlation analysis results for the JJA season in the Prairie region. There is one plot per atmospheric variable: (i) relative humidity (hur); (ii) specific humidity (hus); (iii) air temperature (ta); (iv) geopotential height (zg); (v) Northward wind component (va); (vi) Eastward wind component (ua). The number of climate sites and pressure levels (x10 kPa) are plotted on the x and y axes. The colour scale corresponds to the correlation coefficient.

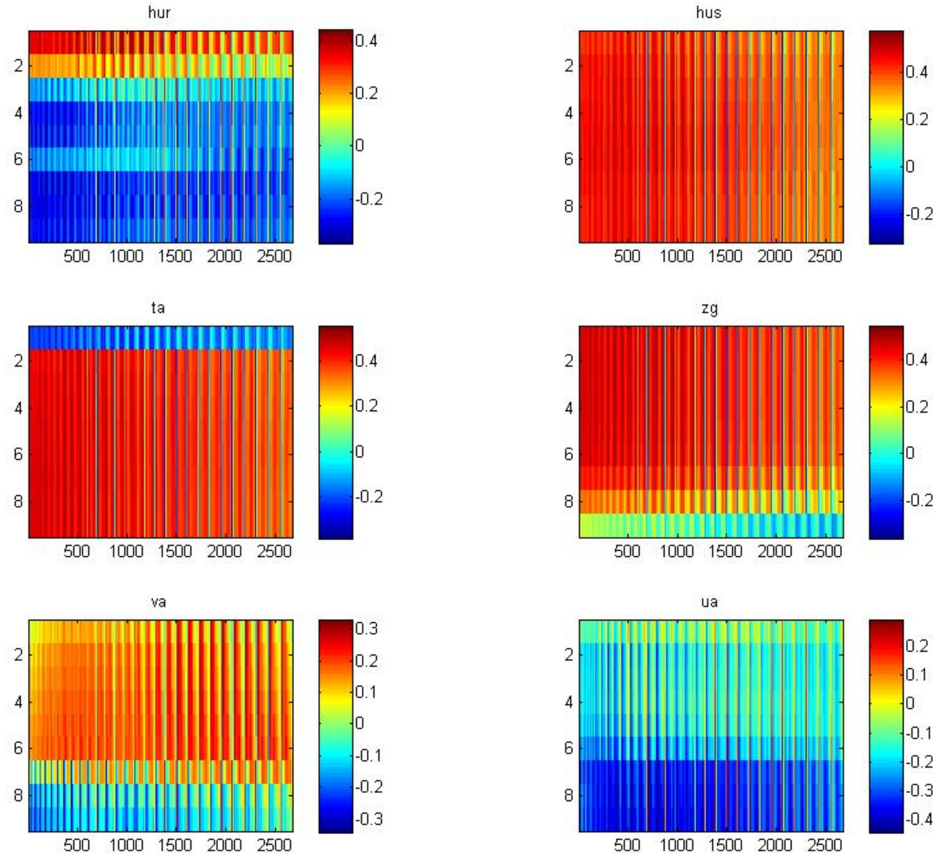


Figure A39: Correlation analysis results for the SON season in the Prairie region. There is one plot per atmospheric variable: (i) relative humidity (hur); (ii) specific humidity (hus); (iii) air temperature (ta); (iv) geopotential height (zg); (v) Northward wind component (va); (vi) Eastward wind component (ua). The number of climate sites and pressure levels (x10 kPa) are plotted on the x and y axes. The colour scale corresponds to the correlation coefficient.

APPENDIX B - Sample of the fuzzy distance metric membership functions

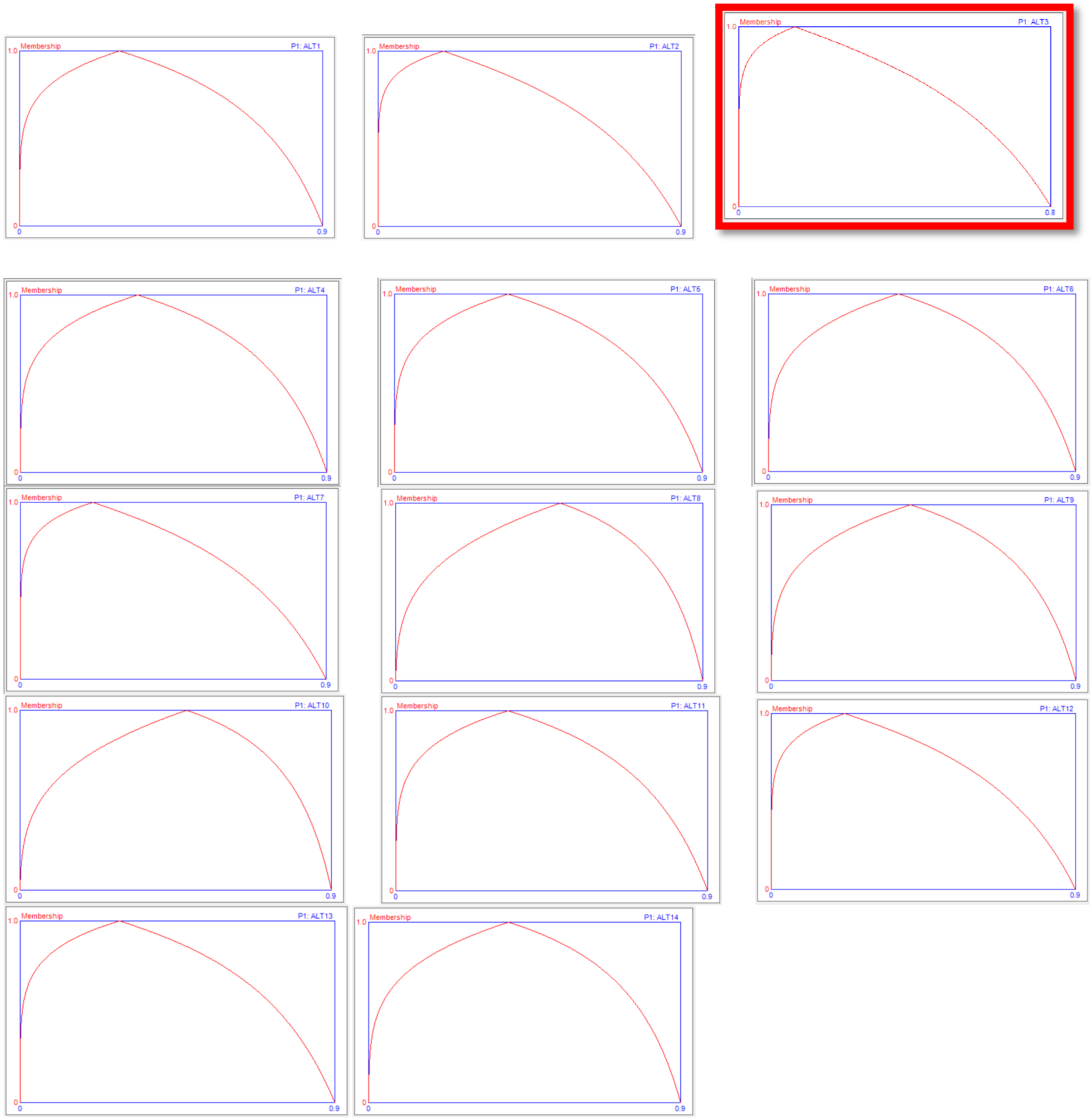


Figure B40: Fuzzy distance metrics of the alternative solutions generated for the DJF season and Decision Maker 1. The distance metric corresponding to the top ranked solution is highlighted in red. Distance is computed between the weighed centre of gravity of the membership function and where the x-axis equals zero (that represents the ideal solution).

APPENDIX C - Instructions for running the regionalization code

Note: Please contact Sarah Irwin at sirwin9@uwo.ca to request for a copy of the files required to run the regionalization code.

Required programs:

- MATLAB (R2011; R2012)
- R Statistical Software (R 3.1.0; R 3.1.1; R 3.1.2; download http://cran.r-project.org/bin/windows/base/old/3.1.1/)
- RStudio (download http://www.rstudio.com/products/rstudio/download/)

Content of Sarah_Run_Folder:

Scripts:

- cluster_fcmeans_ver1.m
- fun_feavec_ver5.m
- fun_validation_fcm_ver1.m
- RFA.R

Data files:

- distance2water.csv
- Coordinate_List_GLR_18km.csv
- Coordinate_List_prairie_18km.csv

- glr_precip_annual_18km.csv
- glr_precip_djf_18km.csv
- glr_precip_mam_18km.csv
- glr_precip_jja_18km.csv
- glr_precip_son_18km.csv
- glr_precip_djf_total_18km.csv
- glr_precip_mam_total_18km.csv
- glr_precip_jja_total_18km.csv
- glr_precip_son_total_18km.csv
- glr_precip_jan_18km.csv
- glr_precip_feb_18km.csv
- glr_precip_mar_18km.csv
- glr_precip_apr_18km.csv

- glr_precip_may_18km.csv
- glr_precip_jun_18km.csv
- glr_precip_jul_18km.csv
- glr_precip_aug_18km.csv
- glr_precip_sep_18km.csv
- glr_precip_oct_18km.csv
- glr_precip_nov_18km.csv
- glr_precip_dec_18km.csv
- ta_Amon_CanESM2_historical_r1i1p1_185001_200512_pelv_[pl]_glr18.csv
- hur_Amon_CanESM2_historical_r1i1p1_185001_200512_pelv_[pl]_glr18.csv
- hus_Amon_CanESM2_historical_r1i1p1_185001_200512_pelv_[pl]_glr18.csv
- zg_Amon_CanESM2_historical_r1i1p1_185001_200512_pelv_[pl]_glr18.csv
- va_Amon_CanESM2_historical_r1i1p1_185001_200512_pelv_[pl]_glr18.csv
- ua_Amon_CanESM2_historical_r1i1p1_185001_200512_pelv_[pl]_glr18.csv

- prairie_precip_annual_18km.csv
- prairie_precip_djf_18km.csv
- prairie_precip_mam_18km.csv
- prairie_precip_jja_18km.csv
- prairie_precip_son_18km.csv
- prairie_precip_djf_total_18km.csv
- prairie_precip_mam_total_18km.csv
- prairie_precip_jja_total_18km.csv
- prairie_precip_son_total_18km.csv
- prairie_precip_jan_18km.csv
- prairie_precip_feb_18km.csv
- prairie_precip_mar_18km.csv
- prairie_precip_apr_18km.csv
- prairie_precip_may_18km.csv
- prairie_precip_jun_18km.csv
- prairie_precip_jul_18km.csv
- prairie_precip_aug_18km.csv
- prairie_precip_sep_18km.csv
- prairie_precip_oct_18km.csv
- prairie_precip_nov_18km.csv
- prairie_precip_dec_18km.csv
- ta_Amon_CanESM2_historical_r1i1p1_185001_200512_pelv_[pl]_prairie18.csv
- hur_Amon_CanESM2_historical_r1i1p1_185001_200512_pelv_[pl]_prairie18.csv
- hus_Amon_CanESM2_historical_r1i1p1_185001_200512_pelv_[pl]_prairie18.csv

- zg_Amon_CanESM2_historical_r1i1p1_185001_200512_pelv_[pl]_prairie18.csv
- va_Amon_CanESM2_historical_r1i1p1_185001_200512_pelv_[pl]_prairie18.csv
- ua_Amon_CanESM2_historical_r1i1p1_185001_200512_pelv_[pl]_prairie18.csv

Procedure:

1) Install MATLAB, R-studio and R-statistical software to your personal computer.

2) Open R-Studio and install required packages: R.matlab, doParallel, lmomRFA (Hosking and Wallis, 2013)

   - Select *Install Packages* (bottom right window); input the names of the three required packages as listed above (one at a time); select *Install*

3) Open fun_validation_fcm_ver1.m in MATLAB

   - Ensure the correct R-version and file location are written in Line 43 that currently reads:

```
eval(['!C:/PROGRA~1/R/R-3.1.1/bin/Rscript ' CurrentDirectory
'/RFA.R'])
```

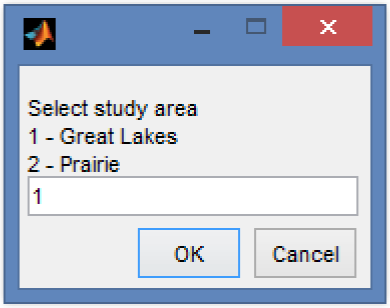
4) Open the MATLAB command window

   - Set the working directory to the Regionalization_Codes location

   - Type cluster_fcmeans_ver1 in the command window and select *Enter*

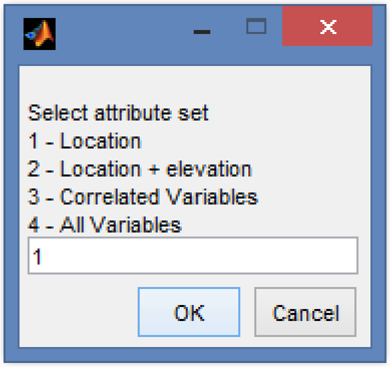
5) Enter values to the command prompts:

   i. Select the study area (data for the Great Lakes and Prairie region has been provided)



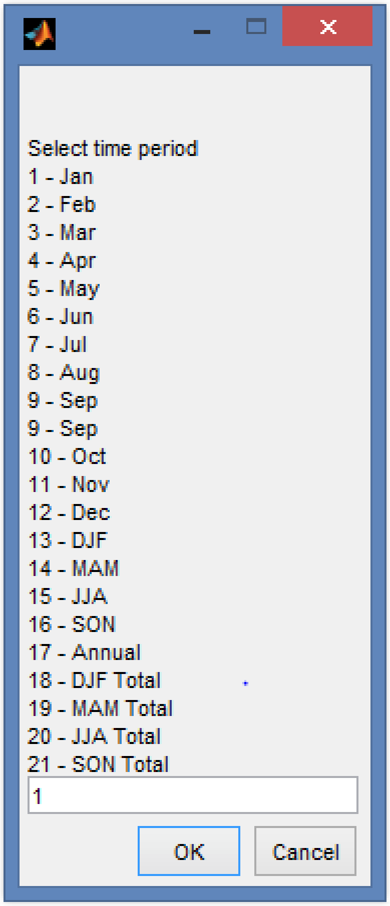
Select study area
1 - Great Lakes
2 - Prairie
1

ii. Select the set of climate site attributes:



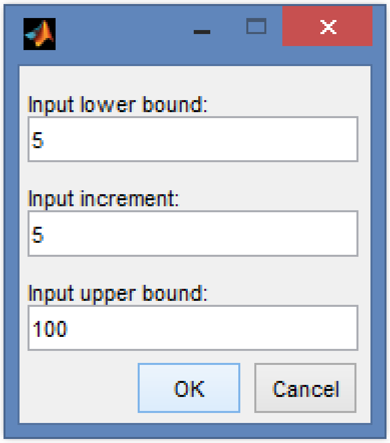
Select attribute set
1 - Location
2 - Location + elevation
3 - Correlated Variables
4 - All Variables
1

iii. Select the temporal resolution/period for precipitation data:


Select time period
1 - Jan
2 - Feb
3 - Mar
4 - Apr
5 - May
6 - Jun
7 - Jul
8 - Aug
9 - Sep
9 - Sep
10 - Oct
11 - Nov
12 - Dec
13 - DJF
14 - MAM
15 - JJA
16 - SON
17 - Annual
18 - DJF Total
19 - MAM Total
20 - JJA Total
21 - SON Total

1

iv. Select the range of *c*-values (numbers of clusters to which the sites are assigned):


Input lower bound:
5

Input increment:
5

Input upper bound:
100

6) The program stores the following output in the current directory: fcm_[study_area]_[temporal_resolution/period]_choice_[attribute_set].mat
    - The file saves two variables that are essential to the analysis:
    - **idx** contains the index values that represent the cluster to which each climate site belongs (climate sites are listed in rows, in the same order as the site list in the location file - Coordinate_List_[study_area]_18km.csv)
    - **tableH** provides the percentage of regions that are classified as homogeneous for each partitioning of the sites; this information is used directly in the figures and tables in the analysis
7) Precipitation region maps are created in ArcGIS 10.2 (http://resources.arcgis.com/ last accessed Nov, 2014) by plotting the climate site locations and colour coding them according to the index of the cluster to which they have the maximum membership.

APPENDIX D - Instructions for running the systems model

Note: Please contact Sarah Irwin at sirwin9@uwo.ca to request for a copy of the files required to run the systems model.

Required programs:

- MATLAB (R2011; R2012)
- R Statistical Software (R 3.1.0; R 3.1.1; R 3.1.2; download http://cran.r-project.org/bin/windows/base/old/3.1.1/)
- RStudio (download http://www.rstudio.com/products/rstudio/download/)
- Fuzzy Compromise Programming for Group Decision Making

Content of Run_Folder:

Scripts:

- run_DE.m
- DE.m
- regionalizaiton_DE.m
- my_creationfcn_de.m
- fun_validation_ver6.m
- RFA.R

Data files:

- glr_precip_djf_18km.csv
- glr_precip_mam_18km.csv
- glr_precip_jja_18km.csv
- glr_precip_son_18km.csv
- distance2water.csv

Procedure:

1) Install MATLAB, R-studio and R-statistical software to your personal computer.

2) Open R-Studio and install required packages: R.matlab, doParallel, lmomRFA (Hosking and Wallis, 2013)

- Select *Install Packages* (bottom right window); input the names of the three required packages as listed above (one at a time); select *Install*

3) Open fun_validation_ver6.m in MATLAB

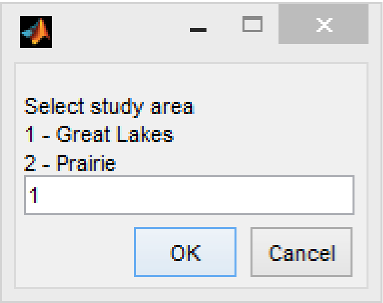
   - Ensure the correct R-version and file location are written in Line 48 that currently reads:

```
eval(['!C:/PROGRA~1/R/R-3.1.1/bin/Rscript ' CurrentDirectory
'/RFA.R'])
```

4) Open the MATLAB command window
   - Set the working directory to the Run_Folder location
   - Type run_DE in the command window and select *Enter*

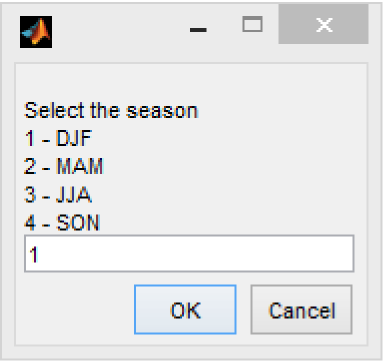
5) Enter values to the command prompts:

   i. Select the study area (data for the Great Lakes region has been provided):



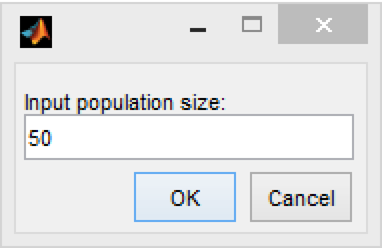
Select study area
1 - Great Lakes
2 - Prairie
1

   ii. Select the season for precipitation data: December, January, February (DJF); March, April, May (MAM); June, July, August (JJA); September, October, November (SON):

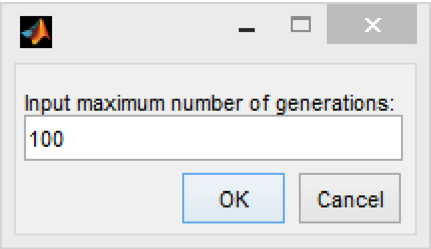


Select the season
1 - DJF
2 - MAM
3 - JJA
4 - SON
1

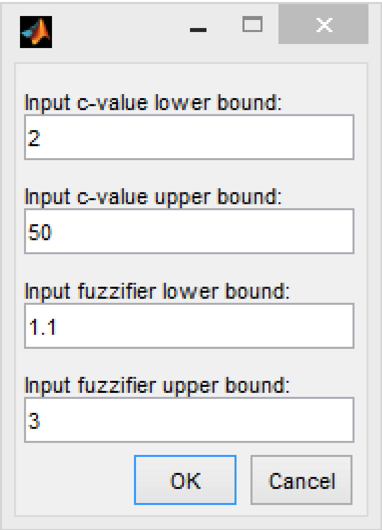
   iii. Input the size of the initial population of solutions (decision variables) for the Differential Evolutionary (DE) algorithm (recommended value is 50):


Input population size:
50

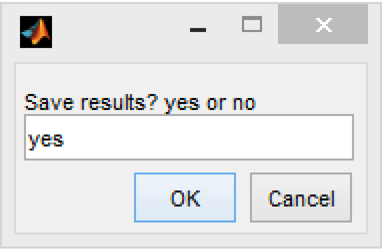
iv. Input the maximum number of generations for the DE algorithm (recommended value is equal to the number of decision variables x the population size that is 100):


Input maximum number of generations:
100

v. Define the constraints of the decision variables (recommended *c*-value limits are 2 to 50 and fuzzifier limits are 1.1 to 3):


Input c-value lower bound:
2
Input c-value upper bound:
50
Input fuzzifier lower bound:
1.1
Input fuzzifier upper bound:
3

vi. Indicate whether the to save the results:



Save results? yes or no

yes

6) The program stores the following output in the current directory:

   - **OUT_[date and time].m** - that contains several variables including the final output of the differential evolutionary algorithm that is the set of optimal, alternative solutions and the corresponding values of the objective functions. Note that the values of objective functions are negative because the script has been developed to minimize the objectives. Since the objectives of this model are to be maximized, their values are multiplied by a value of -1.

   - **fcpgdm_input_[study_area]_[season].xlsx** - that contains the input file to the Fuzzy Compro GDM program. Manually convert the XLS-file to a CSV-file so that it can be directly imported to Fuzzy Compro GDM.

7) Open Fuzzy Compro GDM

   - Select *File - New - Data Import* and load the input file created in the previous step.

8) Select *Computation - Do Rank*

   - The ranks of optimal, alternative solutions for Decision Maker 1, Decision Maker 2 and Decision Maker 3 (see section 6.4) are provided as the Distance Metric Value section of the user interface.

Curriculum Vitae

Name:	Sarah Irwin
Post-secondary Education and Degrees:	The University of Western Ontario London, Ontario, Canada 2008-2013 B.E.Sc
Honours and Awards:	Dean's List 2012-2013
Related Work Experience:	Undergraduate Research Assistant The University of Western Ontario 2011
	Engineering Intern Transportation Planning & Design City of London 2011-2012
	Teaching Assistant The University of Western Ontario 2014-2015

Publications:

King LM, Irwin S, Sarwar R, McLeod AI, Simonovic SP. 2011. The effects of climate change on extreme precipitation events in the Upper Thames River Basin: A comparison of downscaling approaches. *Canadian Water Resources Journal* **37(3)**

Irwin S, Sarwar R, King L, Simonovic SP. 2012. Assessment of Climatic Vulnerability in the Upper Thames River basin: Downscaling with LARS-WG. Water Resources Research Report no. 081, Facility for Intelligent Decision Support, Department of Civil and Environmental Engineering, London, Ontario, Canada, 80 pages. ISBN: (print) 978-0-7714-2964-4; (online) 978-0-7714-2965-1.

Irwin S, Srivastav RK, Simonovic SP. 2014. Instruction for Watershed Delineation in an ArcGIS Environment for Regionalization Studies. Water Resources Research Report no. 087, Facility for Intelligent Decision Support, Department of Civil and Environmental Engineering, London, Ontario, Canada, 45 pages. ISBN: (print) 978-0-7714-3071-8; (online) 978-0-7714-3072-5

Irwin S, Srivastav RK, Simonovic SP, Burn, DH. 2015. Delineation of Precipitation Regions using Location and Atmospheric Variables in Two Canadian Climate Regions: The Role of Attribute Selection - In review for *International Journal of Climatology.*

Irwin S. Srivastav RK, Simonovic SP, Burn DH. 2015. Delineation of Precipitation Regions using Location Parameters in Two Canadian Climate Regions: The Role of the Temporal Resolution - In review for *International Journal of Climatology.*