THE UNIVERSITY OF WESTERN ONTARIO
DEPARTMENT OF CIVIL AND
ENVIRONMENTAL ENGINEERING

Water Resources Research Report

Instructions for Operating the Proposed Regionalization Tool “Cluster-FCM” Using Fuzzy C-Means Clustering and L-Moment Statistics

By:
Sarah Irwin
Roshan Srivastav
and
Slobodan P. Simonovic

Report No: 092
Date: September 2015

ISSN: (print) 1913-3200; (online) 1913-3219
ISBN: (print) 978-0-7714-3101-2; (online) 978-0-7714-3102-9

INSTRUCTIONS FOR OPERATING THE PROPOSED REGIONALIZATION TOOL "CLUSTER-FCM" USING FUZZY C-MEANS CLUSTERING & L-MOMENT STATISTICS

By

Sarah Irwin

Roshan Srivastav

and

Slobodan P. Simonovic

Department of Civil and Environmental Engineering

Western University, Canada

September 2015

EXECUTIVE SUMMARY

This report introduces a tool for the delineation of precipitation regions known as Cluster-FCM. The tool employs the fuzzy *c*-means clustering algorithm and the *L*-moment regional heterogeneity test to form and validate the regions. The user is able to select from several subjective input parameters including: (i) the number of regions to which the sites are assigned; (ii) the site attributes and (iii) the temporal resolution of the precipitation data so that the outputs are best suited to the problem under investigation. The document explains the methodology used to develop the model. It also provides instructions for installing and operating the tool and presents sample outputs. The formation of homogeneous precipitation regions is an important component of the regional frequency analysis procedure that is used to obtain reliable estimates of local precipitation events for applications in water resources engineering.

TABLE OF CONTENTS

1. INTRODUCTION

1.1 General

Water resources engineers are responsible for developing plans, designs and operational procedures to manage uncertain hydrologic/precipitation events while balancing the needs of the natural and socio-economic environments. Precipitation is considered to be a random event and therefore, its occurrence is estimated using probabilistic (stochastic) methods. The accepted stochastic approach to estimating precipitation magnitudes is known as frequency analysis. Frequency analysis involves fitting a statistical probability distribution to the local precipitation record for a given duration. Precipitation estimates are extracted from the distribution and used in a variety of applications including the derivation of climate design values for water infrastructure design, the development of downscaling and forecasting models and the generation of input to hydrologic models, among others. It is important to obtain reliable estimates of the local precipitation to achieve social, environmental and economic objectives (overestimations can be costly and underestimations can lead to the failure of water infrastructure that can devastate the natural and built environments).

In the field of water resources engineering, plans and structures are designed to accommodate precipitation events that correspond to a certain return period. The return period is defined as the average number of years between precipitation events of a certain magnitude. For example, highway roads and related infrastructure are designed for return periods of 25 to 100 years. Bridge piers can be designed for return periods up to 500 years; meaning that they must be designed to accommodate rare, extreme precipitation events that have a probability of exceedance of 1 in 500 years. A major limitation of the traditional approach to frequency analysis is the lack of complete, sufficiently long precipitation records in Canada and around the world. Frequency distributions that are derived from records that are shorter than the return period required for design are unable to capture the true statistics of the local precipitation; therefore leading to deficient designs. To address this issue the regional frequency analysis (RFA) approach can be employed. RFA involves the combination of data from several sites that exhibit similar precipitation statistics into a single frequency distribution from which precipitation occurrence is more reliably estimated. RFA follows the identical distribution assumption that is also known as the rule of homogeneity; the at-site precipitation records must fit to approximately the same frequency distributions. To achieve spatial homogeneity, sites with statistically similar records can be grouped into precipitation regions in a process called regionalization; thus introducing the fundamental topic of the presented work.

The regionalization procedure involves the employment of a tool to partition the climate sites of a study area into homogeneous precipitation regions according to the similarity of their site attributes. Similarity is measured using correlation coefficients or distance metrics and attributes are typically the drivers of the local precipitation including atmospheric variables, location and topographic parameters. The resultant precipitation regions are subsequently validated for spatial homogeneity. Formations of precipitation regions are dependent upon several subjective selection criteria.

This document introduces a regionalization tool known as Cluster-FCM that allows the user to select certain subjective parameters: (i) the number of regions to which the climate sites are assigned; (ii) the site attributes; and (iii) the temporal resolution of the precipitation data. The tool employs the fuzzy *c*-means algorithm to delineate the precipitation regions (Bezdek, 1981). The spatial homogeneity of the regions is validated using the *L*-moment regional heterogeneity test developed by Hosking and Wallis (1997). The regionalization and validation methods are chosen based on the results of the literature review presented in Section 2. The document provides a manual for operating the Cluster-FCM tool. The model has been applied in previous studies by Irwin and Simonovic (2015) and Irwin *et al*. (under review) for assessments in the Prairie and Great Lakes-St. Lawrence lowlands climate regions of Canada. Results from these studies are presented in Section 4 to demonstrate the model output.

1.2 Organization of the Report

The remainder of the document is organized as follows: A literature review of the various regionalization components is provided in Section 2. The methodology used for model development is described in Section 3. Section 4 provides information for operating Cluster-FCM including: (i) a description for tool use; (ii) detailed instructions for running the tool; and (iii) a demonstration of the model for several different input choices. Finally, a summary of the results and concluding remarks are found in Section 5.

2. LITERATURE REVIEW

A review of the literature pertaining to the available regionalization methods, the choice of the number of precipitation regions, the site attributes and the temporal resolution of the precipitation data required for regional analysis is presented.

2.1 Regionalization Method

Many methods are available for the delineation of precipitation regions including: (i) correlation analysis (Cavadias, 1990; 2001; Unal, 2003); (ii) principal component analysis (Mills, 1995); (iii) site-focused pooling schemes (region-of-influence) (Burn, 1990; Burn and Goel, 2000; Gaal, 2008); and (iv) cluster analysis (Rao, 2006; Satyanarayana and Srinivas 2008; 2011; Srinivas 2013; Asong *et al.*, 2015), among others. Of the available methods clustering algorithms are preferred for their inherent ability to identify underlying patterns in complex datasets and as such, they are the most commonly used regionalization methods in climate literature. In general, clustering algorithms work by assigning sites to regions according to their similarity that is measured using a distance metric (Euclidean or Mahalanobis) in the attribute space. A summary of the various clustering algorithms is provided:

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 fundamental difference between most agglomerative algorithms is the means by which the distance metric is used to measure the similarity between sites and clusters (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 regions in the attribute space; thus, maximizing the within cluster similarity and likewise, the between cluster separation (Zalik and Zalik, 2011). The k-means clustering algorithm is a very common partitional algorithm (MacQueen, 1967) that calculates similarity as the distance between sites and the cluster centroids (the average attribute value of the member sites of a cluster) in the attribute space (Burn and Goel, 2000; Pelchzer, 2008; Satyanarayana and Srinivas, 2008; Dikbas, 2013). At each step in 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 be more similar to other clusters and therefore, site memberships may be reassigned. Other partitional algorithms include: (i) the k-mediods approach where similarity is measured between a climate site and median value of the member site attributes (Kaufman, 1987); and (ii) the k-modes algorithm where similarity is measured between the climate sites and mode of the member site attributes (Huang, 1998). The ability of the algorithms to update site membership 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 initial selection of 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, principal component analysis) methods and determined that the non-hierarchical algorithms provided more accurate results.

The aforementioned clustering algorithms form hard clusters such that each site belongs to only one region (Zalik and Zalik, 2011); therefore, implying that members 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 performs very similarly to the k-means method. The greatest distinction between the methods is that the former employs a membership function that 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 site 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. The membership function of the fuzzy *c*-means algorithm provides useful information for removing or relocating discordant sites, offering another advantage of the fuzzy clustering technique over the traditional clustering methods (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 each concluded that the former achieved a higher level of performance as it consistently delineated a greater number of homogeneous hydrologic regions. Although the analyses were conducted for the regionalization of flood quantiles, the findings can also apply to the formation of precipitation regions.

Evidently the fuzzy *c*-means algorithm has 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 requirement for the total number of clusters to be determined *a priori* is a disadvantage of all clustering algorithms; they are all incapable of establishing the number of clusters that provide for the natural partitioning of the sites in the attribute/precipitation space. Despite certain disadvantages the fuzzy *c*-means algorithm is gaining popularity as a regionalization method in climate literature because it produces a more accurate partitioning of the sites into clusters and therefore, it is elected as the most appropriate option for implementation in Cluster-FCM.

2.2 Number of Precipitation Regions

Clustering algorithms are recognized as preferred regionalization methods in climate literature. They are limited, however, by their inability to identify the number of clusters that provide for the natural partitioning of the climate sites in the attribute/precipitation space. Consequently, the number of clusters must be solved for prior to the employment of the algorithm (Gurrutxaga, 2013).

Cluster validity indices (CVIs) can be used to solve for the optimal number of clusters for the sites to be assigned to. CVIs are used to evaluate a partitioning of sites into precipitation regions for their compactness (similarity between 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 and together they produce a variety of outcomes. Currently, a single, universally accepted measure has not been identified. Satyanarayana and Srinivas (2011) employed five CVIs to determine input parameter values 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 to solve for input parameter values. 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.

A more reliable method is required for determining the optimal number of clusters for the sites to be assigned to in the precipitation space. At this time it is believed that trial and error is the only appropriate technique. Parameter values are varied for a range of magnitudes and used as input to the algorithm. The resultant cluster sets are validated for regional homogeneity and the number of clusters that achieves the desired outcome is retained and used in regional frequency analysis. The desired outcome depends on user preference that typically takes into account two criteria: (i) the numbers of sites assigned to the regions; and (ii) the percentage of homogeneous precipitation regions. Cluster-FCM employs the trial and error method where the fuzzy *c*-means algorithm can be employed for a range of numbers of clusters. The user can then select the preferred partitioning of sites according to these criteria or other factors.

2.3 Climate Site Attributes

Clustering algorithms partition sites into precipitation regions according to the similarity of their characteristics/attributes that are typical drivers of the local precipitation. The choice of the site attributes is a subjective one and the only requirement is that they are independent of the precipitation statistics that are used to test for regional homogeneity.

Site attributes should be selected such that they are physically meaningful to the problem under investigation. Common choices include the geographical site parameters such as latitude, longitude (Burn, 2014); topography including distance to major water bodies and elevation that contributes to orographic precipitation (Johnson and Hanson, 1995); large scale atmospheric variables recorded at several pressure levels (Satyanarayana and Srinivas, 2011); and seasonality that is defined as the timing of extreme precipitation events within the year (Comrie and Glenn, 1998).

Statistical analyses can be used to assess the significance of the relationship between the potential attributes and the local precipitation to reduce the computational time and improve regional homogeneity (Wagener, 2004; Jafaar, 2011). Asong *et al*. (2015) formed precipitation regions in the Canadian Prairie provinces; they considered geographical site parameters and suite of 21 atmospheric variables as potential attributes. A combination of principle component analysis and canonical correlation analysis were used to select the statistically significant attributes to be used in the regionalization procedure.

2.4 Temporal Resolution of Precipitation

Certain applications require precipitation estimates to be derived from regional frequency distributions of specific temporal scales. Several examples are provided below:

Climate forecasting applications form relationships between precipitation derived from regional frequency distributions and atmospheric variables to project weather patterns. Long-term precipitation (annual, seasonal, monthly resolutions) is projected for planning applications including the development of water budget (Johnson and Hanson, 1995; Saikranthi *et al*., 2012) Short-term precipitation projections (hourly and sub-hourly resolutions) are used in hydrologic model calibration and agricultural applications including the estimation of soil erosion and infiltration rates (Jebari, 2007). Analyses involving extreme hydrologic events including flooding and drought require a multi-temporal scale assessment to enhance the predictability of the hydrologic models (Jiang *et al*., 2013). Precipitation regions are delineated for extreme values such as the maximum annual series for the derivation of climate design values to be used in water infrastructure design (Burn, 2014).

3. METHODOLOGY

The methodology for the formation and validation of precipitation regions is explained in this section.

3.1 Delineation of Precipitation Regions

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).

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

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. 2], [Eq. 3] and [Eq. 4]:

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

$$u_{ik} = \frac{1}{\sum_{l=1}^{c} \frac{|z_i - C_k|}{|z_i - C_l|}^{2/(m-1)}} \tag{3}$$

$$C_k = \sum_{i=1}^{N} \frac{u_{ik}^{m} z_i}{\sum_{i=1}^{N} u_{ik}^{m}} \tag{4}$$

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.

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. 5] (Satyanarayana and Srinivas 2011):

$$T_i = \max\{\frac{1}{c}, \frac{1}{2}[max_{1 \le k \le c}(u_{ik})]\} \tag{5}$$

where T_i is the defined threshold value.

3.2 Validation of Regional Homogeneity

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:

$$t = \frac{l_2}{l_1} = \frac{(2b_1 - b_0)}{b_0} \tag{6}$$

$$t_3 = \frac{l_3}{l_2} = \frac{(6b_2 - 6b_1 + b_0)}{(2b_1 - b_0)} \tag{7}$$

$$t_4 = \frac{l_4}{l_2} = \frac{(20b_3 - 30b_2 + 12b_1 - b_0)}{(2b_1 - b_0)} \tag{8}$$

where,

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

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

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

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

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 *L*-Cv for all sites in the cluster, which is represented by V_1.

$$V_1 = \left\{\frac{\sum_{i=1}^{N_c} n_i (t_i - t^R)^2}{\sum_{i=1}^{N_c} n_i}\right\}^{1/2} \quad (9)$$

where N_c is the number of sites in a cluster; *n* is the site record length; t_i is *L*-Cv 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:

$$t^R = \frac{\sum_{i=1}^{N_c} n_i (t_i)}{\sum_{i=1}^{N_c} n_i} \quad (10)$$

$$t_3^R = \frac{\sum_{i=1}^{N_c} n_i(t_{3i})}{\sum_{i=1}^{N_c} n_i} \tag{11}$$

$$t_4^R = \frac{\sum_{i=1}^{N_c} n_i(t_{4i})}{\sum_{i=1}^{N_c} n_i} \tag{12}$$

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:

$$H_1 = \frac{(V_1 - \mu_V)}{\sigma_v} \tag{13}$$

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.

$$H_{1,adj} = H_1 + 0.122\ x\ \overline{p^2}(N_c - 1) \tag{14}$$

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.

4. OPERATING CLUSTER-FCM

In this section the application of Cluster-FCM is explained, instructions for the installation and operation of the tool are provided and finally, the tool is demonstrated in the Prairie and Great Lakes-St. Lawrence lowland climate regions of Canada.

4.1 Application of Cluster-FCM

Cluster-FCM allows the user to select certain input parameters to the fuzzy c-means algorithm including the number of regions for the sites to be partitioned into as well as site attributes and temporal resolutions of precipitation data from a depository of variables described below:

Available site attributes include location parameters (latitude, longitude), topographic variables (distance to major water bodies, elevation) and a complete set of atmospheric variables (air temperature, geopotential height, humidity, Northward and Eastward wind components) for a range of pressure levels between 20 to 100 kPa. Seasonality is not an effective attribute choice for smaller study areas where the timing of precipitation events are spatially uniform and therefore, it is not included as an option in the tool. The tool also does not include a component for assessing the statistical significance of the relationship between precipitation and the potential attributes; however the user may wish to perform such analyses independently. Options for the temporal resolutions include annual, seasonal, monthly resolutions as well as the maximum annual series of the precipitation data that is the maximum precipitation event recorded within the year.

4.2 Instructions for Operating Cluster-FCM

The procedures for installing and using the regionalization tool are presented in this section of the report.

4.2.1 Preparing to use Cluster-FCM

The Cluster-FCM requires the following programs to be installed on the computer system:

- 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/)

The tool is requires the following scripts to be copied into its working directory:

- cluster_fcmeans_ver1.m
- fun_feavec_ver5.m
- fun_validation_fcm_ver1.m

- RFA.R

The regionalization tool also requires several input files to be stored into the working directory.

- Location parameters: This file stores the latitude and longitude that correspond to the sites for a specified study area and spatial resolution. Sample data is currently available for an 18 x 18 km spatial resolution for application in the Great Lakes (glr) and Prairie (prairie) Canadian climate regions.
    - location_[study_area]_[spatial_resolution]km.csv
    - ex. location_glr_18km.csv

- Topographic parameters: Site elevation and distance to major water body parameters are computed for each site. Sample files are stored in a topographic information file for each study area. The parameter values correspond to the climate sites that are listed in the same order as the location parameter file.
    - topography_[study_area]_[spatial_resolution]km.csv
    - ex. topography_prairie_18km.csv

- Atmospheric variables: Six large scale atmospheric variables including relative humidity (hur), specific humidity (hus), air temperature (ta), geopotential height (zg), Northward wind component (va) and Eastward wind component (ua) recorded at nine pressure levels (20, 30, 40, 50, 60, 70, 85, 92.5 and 100kPa) can be accepted as input to the regionalization tool. The values of the atmospheric variables for each pressure level are interpolated to the list of climate sites that is consistent with all datasets. Rows of the input file correspond to the site/grid point (in the same order as the location and topographic parameter files) and the columns correspond to the monthly time step.

  - [atm_var]_Amon_CanESM2_historical_r1i1p1_185001_200512_plevel_[plevel]_[study_area][spatial_resolution].csv
  - ta_Amon_CanESM2_historical_r1i1p1_185001_200512_plevel_20000_prairie18.csv

- Precipitation data: All sample precipitation data is extracted from the ANUSPLIN gridded dataset. It is the other datasets (topographic parameters, atmospheric variables) that are calculated for and interpolated to the spatial grid of the precipitation data. Precipitation data is available for several temporal resolutions including monthly (Jan, Feb, Mar, Apr, May, Jun, Jul, Aug, Sep, Oct, Nov, Dec), seasonal (DJF-Dec, Jan, Feb; MAM-Mar, Apr, May; JJA-Jun, Jul, Aug; SON-Sep, Oct, Nov), annual and the maximum annual series of the data. Rows of the input file correspond to the site/grid point (in the same order as the location parameter file) and the columns correspond to the monthly, seasonal, annual and maximum annual series time steps.
  - [study_area]_precip_[temporal_resolution]_[spatial_resolution]km.csv
  - ex. glr_precip_jan_18km.csv (monthly resolution)
  - ex. glr_precip_djf_18km.csv (seasonal resolution)
  - ex. glr_precip_annual_18km.csv (annual resolution)
  - ex. glr_precip_mas_18km.csv (maximum annual series of the precipitation data)

Note that all scripts that are required to run Cluster-FCM may be downloaded from the FIDS website (http://www.eng.uwo.ca/research/iclr/fids/). Sample datasets are also available so that the user can reproduce the results presented in Section 4.3.

4.2.2 Procedure for Running Cluster-FCM

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'])
```

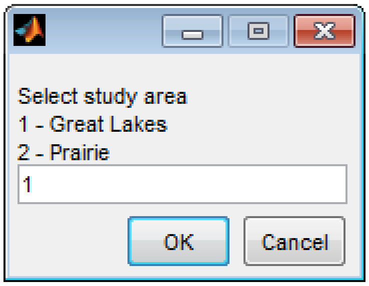
4) Open the MATLAB command window

   - Set the working directory to the appropriate location

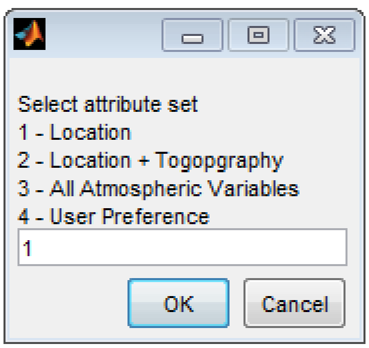
   - Type cluster_fcm_ver1 in the command window and select *Enter*

5) Enter values to the command prompts:

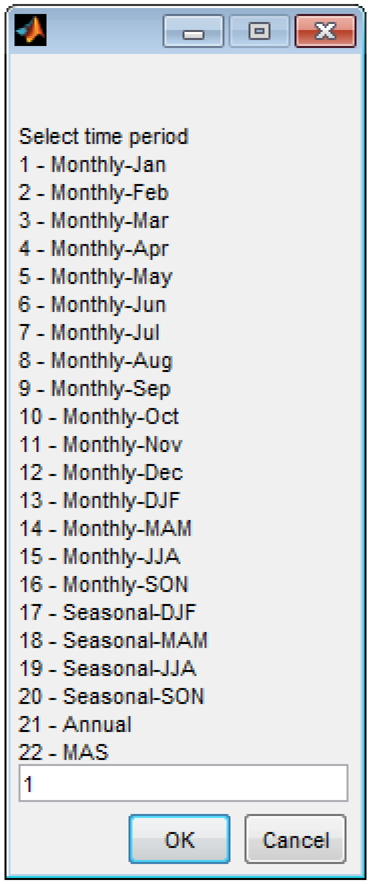
i. Select the study area (data for the Great Lakes and Prairie region has been provided)



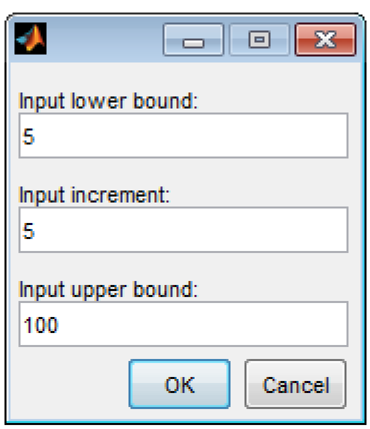
ii. Select the set of climate site attributes:



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



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



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 - location_[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.

4.3 Demonstration of Cluster-FCM

The current version of Cluster-FCM has been used in studies by Irwin and Simonovic (2015) and Irwin *et al*., (under review) where it was applied in two Canadian climate regions including the Prairie and the Great Lakes-St. Lawrence lowlands. First a description of the study area is provided, followed by an explanation of the data used as tool input. Results from the aforementioned studies for the Great Lakes region are then presented to demonstrate the use of the Cluster-FCM.

4.3.1 Study Area

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 and it is shown as the south-most purple region in **Figure 1**. 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 (Lapen and Hayhoe, 2003; Sousounis, 2001). 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).

Figure 1: Map depicting the Canadian Climate Regions (Source: Statistics Canada, 2012).

4.3.2 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 topographic 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.

4.3.3 Sample Output – Site Attribute Selection

First the sensitivity of the formations of the precipitation regions to the selection of site attributes is demonstrated. **Figures 2** plots the relationships between the number of regions to which the sites are assigned and the percentage of regions that are classified as homogeneous on the x and y axes, respectively. (This information is stored in variable **tableH** of the output file). The coloured lines with the circular and upward, downward and left facing triangles represent the output corresponding to different combinations of site attributes. The site attribute combinations considered in this analysis are: Attribute Set 1 (AS-1) that consists of location parameters (latitude, longitude); AS-2 includes location and topographic parameters (site elevation and distance to major water bodies); AS-3 is composed of the atmospheric variables that form significant linear correlations with the local precipitation; and AS-4 is comprised of a comprehensive set of atmospheric variables recorded at a range of pressure levels. AS-1, AS-2 and AS-4 are standard site attribute combinations that may be selected from the Cluster-FCM. The variables that form AS-3 change according to the season and study area and therefore, they must be selected manually using the *User Preference* option of the attribute selection window. For the selection of atmospheric variables as site attributes, the temporal resolution of the precipitation data is automatically set to monthly and organized into four seasons (this is because the atmospheric data is available for a monthly temporal resolution). The model output for the December-January-February (DJF), March-April-May (MAM), June-July-August (JJA) and September-October-November (SON) seasons are plotted on the top left, top right, bottom left and bottom right, respectively.

The black horizontal line indicates the number of regions required to achieve approximately 80% homogeneity. The location at which the coloured lines cross the black, threshold line is approximately at the plot elbow. The elbow represents a suitable tradeoff between the criteria for selecting a preferred number of regions for the sites to be partitioned into. It is the point at which further improvement (increase) in the percentage of homogeneous regions requires a relatively large increase in the number of regions to which the sites are assigned; effectively compromising the number of station-years in the regional precipitation record. More station-years that are used to fit the regional frequency distribution provides for more robust estimates of the true, local precipitation. Alternatively, precipitation regions that are composed of fewer, similar sites (in terms of precipitation statistics) are more likely to be homogeneous thereby satisfying the identical distribution assumption. For each season and combination of site attributes, the preferred number of regions is selected as the lowest number that achieves a minimum of 80% regional homogeneity. **Table 1** summarizes the preferred numbers of regions/clusters (*c* parameter of the fuzzy *c*-means algorithm) for all scenarios. The regionalization tool is employed again using these *c*-values as model input. The **idx** variable of the output file is then used to plot the resultant precipitation regions. Note that *c*-values could not be derived for AS-3 in the JJA and SON seasons because no linear correlations were observed between the atmospheric variables and local precipitation.

Table 1: The c-values for all combinations of attribute sets and seasons

	AS-1	AS-2	AS-3	AS-4
	Great Lakes	Great Lakes	Great Lakes	Great Lakes
DJF	30	37	36	34
MAM	15	19	17	15
JJA	12	20	-	15
SON	15	19	-	19

Analysis of **Figure 2** reveals that the number of regions required to achieve the 80% regional homogeneity criterion is similar between attribute sets for the same season. The seasonal difference between the *c*-values is more significant; for example the *c*-values range between 30 to 37 for the DJF season and 12 to 19 for all other seasons.

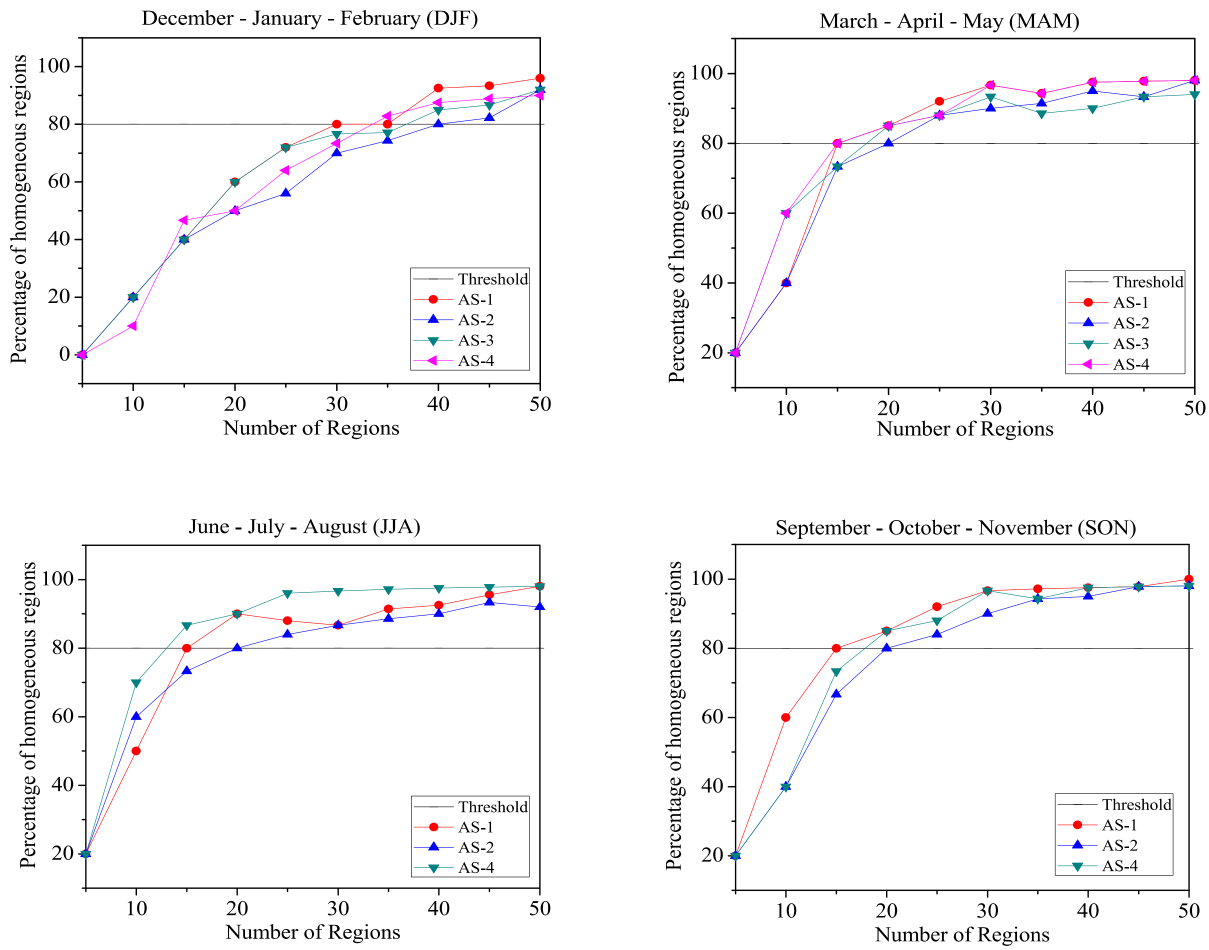


Figure 2: Relationship between the number of clusters to which the climate sites are assigned (*c*-parameter) and the percentage of regions that are classified as homogeneous through the validation procedure for the Great Lakes study area.

The site index values (that indicate the region to which the sites belong) are stored in variable **idx** of the output file. The **idx** variable consists of a list (column) of values that are in the same order as the site's coordinates in the location input file; allowing the site index values to be copied beside their latitude/longitude and plotted directly in an ArcGIS environment. **Figures 3-6** are examples of precipitation regions delineated using different combinations of site attributes and seasons.

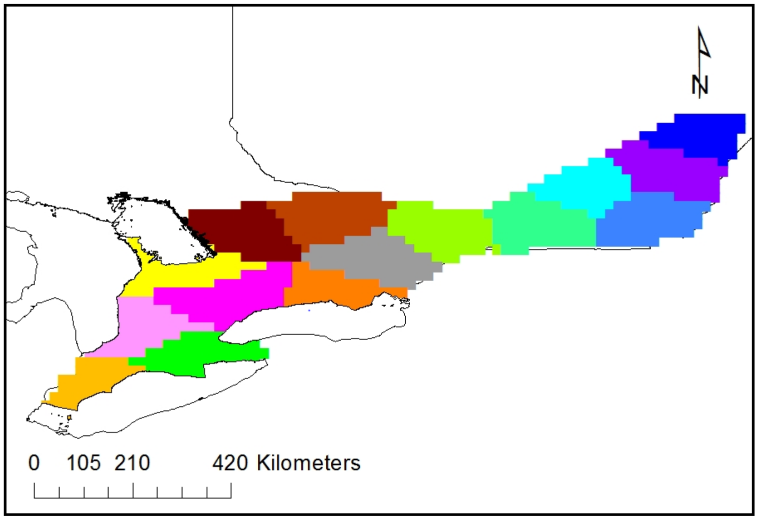


Figure 3: Precipitation regions formed by AS-1 for the MAM season in the Great Lakes region.

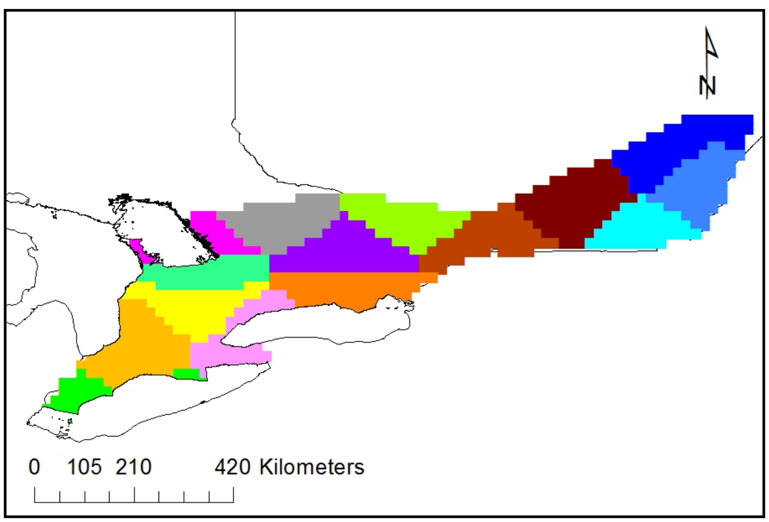


Figure 4: Precipitation regions formed by AS-4 for the MAM season in the Great Lakes study area.

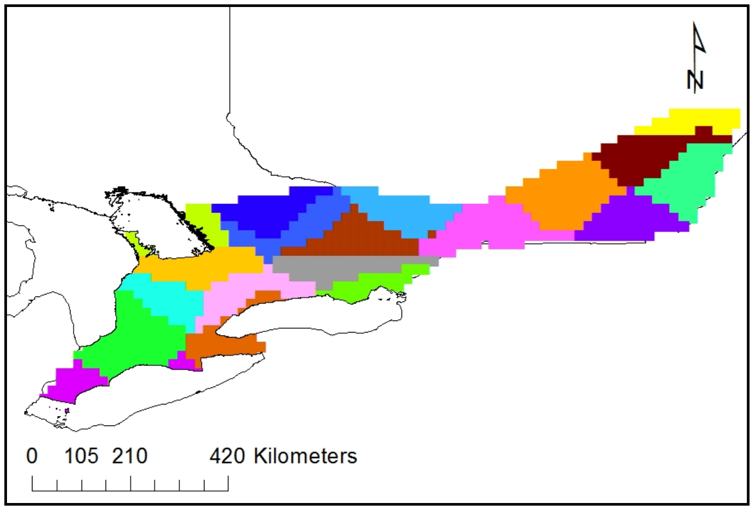


Figure 5: Precipitation regions formed by AS-4 for the SON season in the Great Lakes study area.

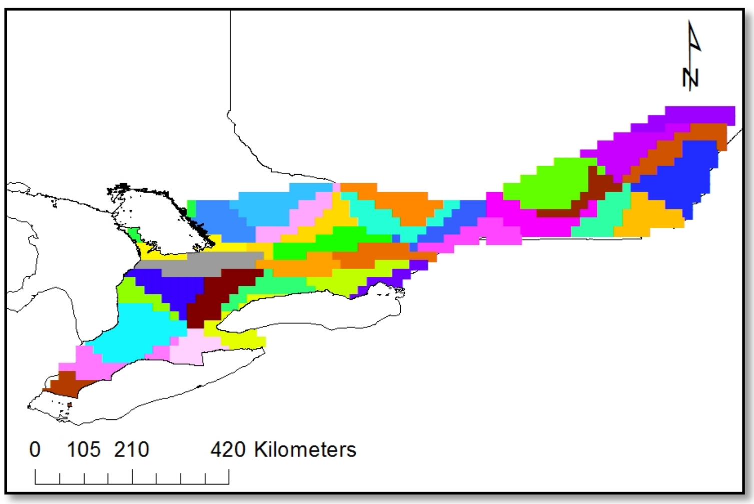


Figure 6: Precipitation regions formed by AS-4 for the DJF season in the Great Lakes study area.

Evidently, the shapes of the precipitation regions change between AS-1 (**Figure 3**) and AS-3/AS-4 (**Figure 4-5**). The regions formed using the third and fourth attribute set are similar in shape and size for the same season; likely because they are formed using different combinations of atmospheric variables. Additionally, the seasonal difference between precipitation regions is observed. **Figure 6** presents the regions that are delineated using AS-4 in the DJF season and evidently, the climate sites of the study area are partitioned into a greater number of clusters that are smaller in size.

4.3.4 Sample Output – Temporal Resolution Selection

Here the sensitivity of the formations of the precipitation regions to the selection of the temporal resolution of the data is demonstrated. **Figure 7** plots the relationships between the number of regions to which the sites are assigned and the percentage of regions that are classified as homogeneous on the x and y axes, respectively. Again output for the DJF, MAM, JJA and SON seasons are plotted on the top left, top right, bottom left and bottom right of the figure, respectively. The coloured lines, however, represent the outputs for individual months as well as for their respective season. The left-most plot of **Figure 8** presents the results for the annual and maximum annual series of the data for the Great Lakes study area.

It is observed that sometimes the seasonal resolution calls for fewer precipitation regions to attain 80% regional homogeneity; therefore outperforming the monthly temporal resolutions (fewer regions means more sites per region and more station-years of data to fit the regional frequency distribution). This is true for the DJF and MAM seasons; however for the JJA and SON seasons the monthly resolutions call for the lowest number of regions for the sites to be partitioned into. The plots reveal that there are moderate differences between the magnitudes of the c-values with the exception of the maximum annual series. For the annual, seasonal and monthly resolutions the values of c range from 10 to 29; however the sites are only required to be assigned to 5 regions to attain 80% regional homogeneity for the employment of the maximum annual series of the data.

Again **Tables 2-3** summarize the preferred numbers of regions/clusters for the seasonal, annual and maximum annual series of the data (**Table 2**) and the monthly resolution (**Table 3**). Cluster-FCM is employed once more using these *c*-values (that achieve the defined selection criteria) as input. The **idx** variable of the output file is used to plot the resultant precipitation regions. **Figures 9-11** show the precipitation regions formed and validated using different temporal resolutions of the data. AS-1 (location) and AS-2 (location and topography) are the only options for site attributes for all temporal resolutions; therefore AS-2 is used to form the regions presented in these figures.

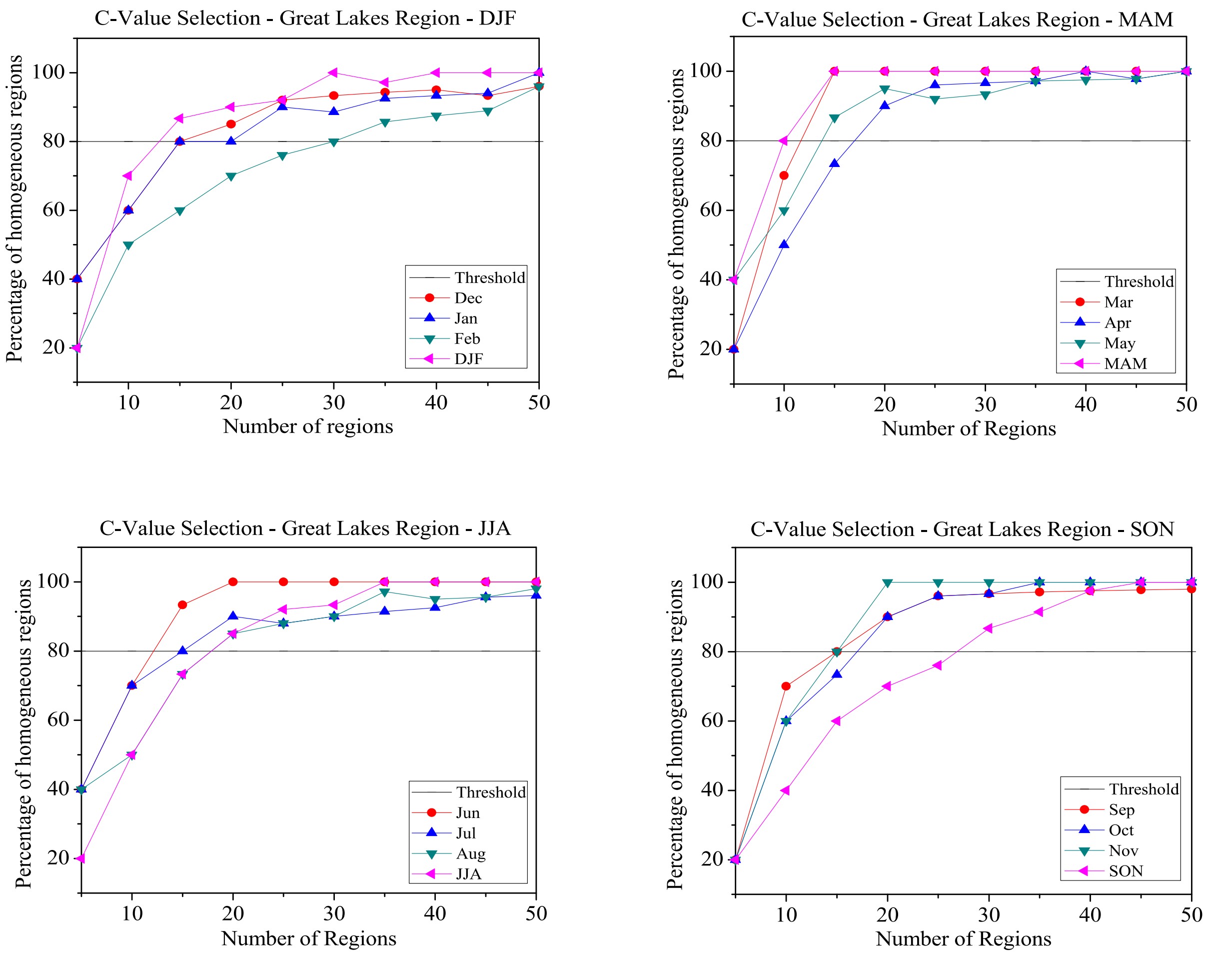


Figure 7: Relationship between the number of clusters to which the climate sites are assigned (c-parameter) and the percentage of regions that are classified as homogeneous through the validation procedure for the Great Lakes study area.

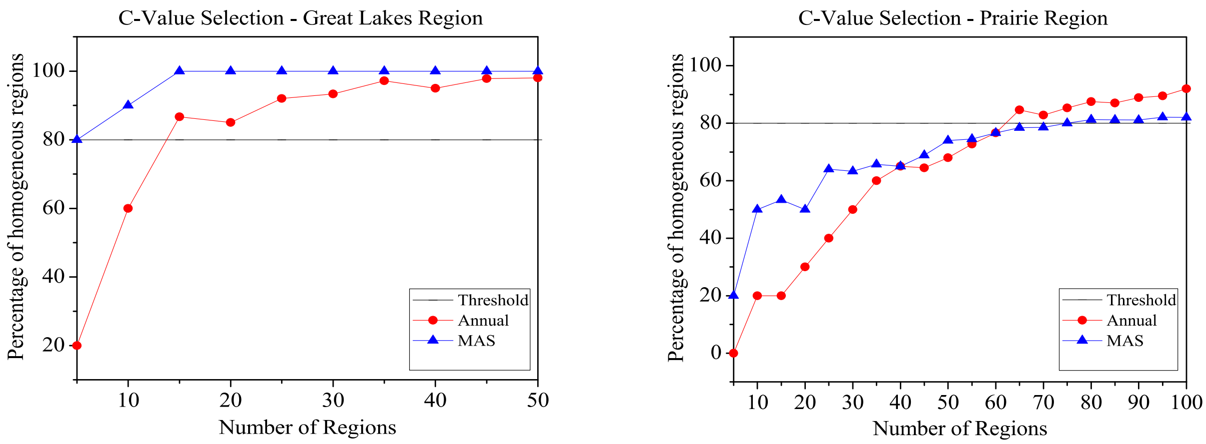


Figure 8: Relationship between the number of clusters to which the climate sites are assigned (c-parameter) and the percentage of regions that are classified as homogeneous through the validation procedure in the Great Lakes (left) and Prairie (right) study area

Table 2: Final c-values for the annual, maximum annual series and seasonal temporal resolutions in the Great Lakes climate region.

Region	Annual	MAS	DJF	MAM	JJA	SON
GLR	13	5	15	10	19	27

Table 3: Final *c*-values for the monthly temporal resolution in the Great Lakes climate region.

Region	Jan	Feb	Mar	Apr	May	Jun	Jul	Aug	Sep	Oct	Nov	Dec
GLR	20	29	10	16	13	11	15	17	13	17	12	15

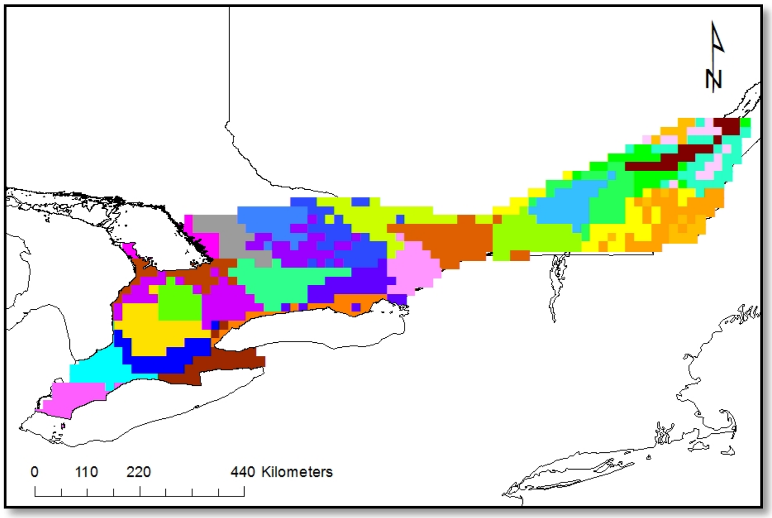


Figure 9: Precipitation regions delineated for the month of February in the Great Lakes study area.

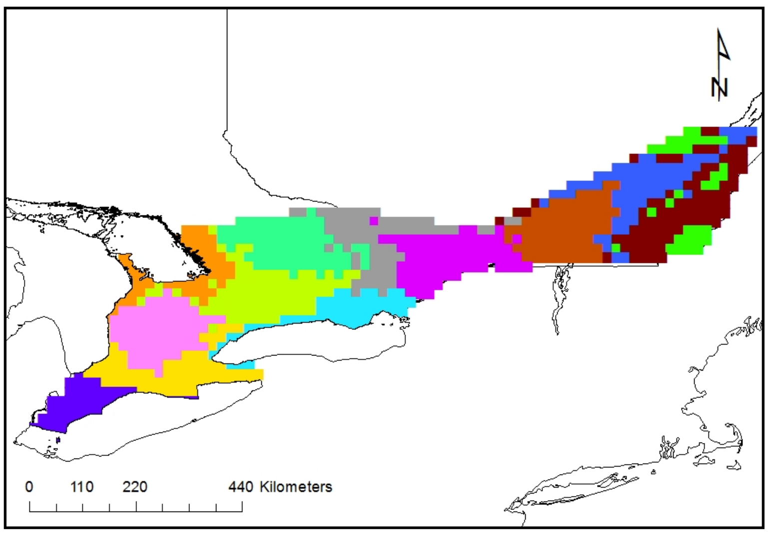


Figure 10: Precipitation regions delineated for the month of September for the Great Lakes study area.

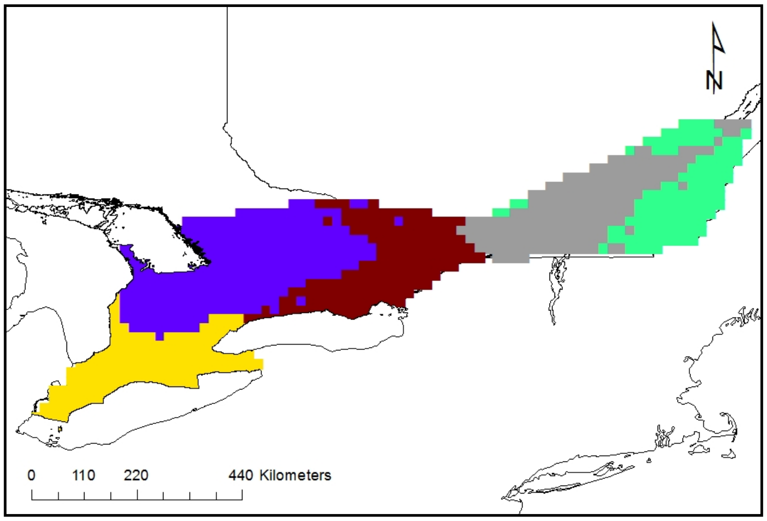


Figure 11: Precipitation regions delineated for the maximum annual series in the Great Lakes study area.

It is evident that the choice of the temporal resolution of the data has a considerable effect on the formation of the precipitation regions in the Great Lakes study area. In this assessment location and topographic parameters are assigned as site attributes for the employment of all temporal resolutions. These parameters are temporally fixed and therefore, when the sites are partitioned into the same number of regions the regional patterns are identical. However, through the technique used to determine the preferred *c*-value (that provides for a suitable tradeoff between the maximization of regional homogeneity and the maximization of the regional precipitation record length) it is evident that the number of regions to which the sites are assigned changes according to the precipitation variability of the chosen temporal resolution. This is because precipitation variability is an important influencing factor in the classification of regional homogeneity that is used to determine the preferred number of regions.

5. 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 number of regions into which the sites are partitioned, the climate site attributes and the temporal resolution of the precipitation data (as is observed in Section 4.3). This report introduces a regionalization tool that employs the fuzzy *c*-means clustering algorithm (Bezdek, 1981) and the *L*-moment regional heterogeneity test (Hosking and Wallis, 1997) to delineate and validate precipitation regions for regional frequency analysis. The tool allows the user to choose from several subjective input parameters: (i) the number of regions to which the sites are assigned; (ii) the site attributes (atmospheric variables recorded at a range of pressure levels, location and topographic parameters); and (iii) the temporal resolution of the precipitation data (monthly, seasonal, annual and the maximum annual series of the data). The tool is developed for application in the Prairie and Great Lakes-St. Lawrence lowlands climate regions of Canada; however it may be applied in other locations. This document explains the methodology used in model development and it also provides instructions for installing and operating the tool, and sample outputs are presented.

Cluster-FCM performs successfully; however, it is restrictive to certain combinations of input parameters. There are several ways in which the model can be improved:

(i) Integrate all combinations of site attributes and temporal resolutions into the model. Currently atmospheric variables can only be employed as site attributes for the monthly temporal resolution that is organized into four seasons. This also implies that only location and topographic parameters can be assigned as site parameters for all temporal resolutions;

(ii) Adjust the model to include finer temporal scales (daily, sub-daily) that have significant applications in agriculture and water resources engineering;

(iii) Incorporate a statistical analyses methods (principle component analysis; linear/non-linear regression analyses) to the attribute selection component in order to reduce the computational time and to improve the quality of the output in terms of regional homogeneity and lengthy regional precipitation records;

(iv) Add more options for site attributes including near and remote teleconnection indices;

(v) Re-write the code for the regionalization tool in R-statistical software alone; currently the script passes information between R and MATLAB resulting in large computational demands.

REFERENCES

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.

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

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

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

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

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

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.

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.

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.

Hutchinson MF, McKenney DW, Lawrence K, Pedlar JH, Hopkinson RF, Milewska E, Papadopol P. 2009. Development and testing of Canadawide 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, Simonovic S. 2015. Assessment of the Regionalization of Precipitation in Two Canadian Climate Regions: A Fuzzy Clustering Approach. The University of Western Ontario Electronic Thesis and Dissertation Repository.

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.

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

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.

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.

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.

Rao AR, Srinivas VV. 2005. 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.

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

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. 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.

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.

LIST OF PREVIOUS REPORTS IN THE SERIES

ISSN: (Print) 1913-3200; (online) 1913-3219

In addition to 69 previous reports (No. 01 – No. 69) prior to 2011

Tarana A. Solaiman and Slobodan P. Simonovic (2011). Quantifying Uncertainties in the Modelled Estimates of Extreme Precipitation Events at Upper Thames River Basin. Water Resources Research Report no. 070, Facility for Intelligent Decision Support, Department of Civil and Environmental Engineering, London, Ontario, Canada, 167 pages. ISBN: (print) 978-0-7714-2878-4; (online) 978-0-7714-2880-7.

Tarana A. Solaiman and Slobodan P. Simonovic (2011). Assessment of Global and Regional Reanalyses Data for Hydro-Climatic Impact Studies in the Upper Thames River Basin. Water Resources Research Report no. 071, Facility for Intelligent Decision Support, Department of Civil and Environmental Engineering, London, Ontario, Canada, 74 pages. ISBN: (print) 978-0-7714-2892-0; (online) 978-0-7714-2899-9.

Tarana A. Solaiman and Slobodan P. Simonovic (2011). Development of Probability Based Intensity-Duration-Frequency Curves under Climate Change. Water Resources Research Report no. 072, Facility for Intelligent Decision Support, Department of Civil and Environmental Engineering, London, Ontario, Canada, 89 pages. ISBN: (print) 978-0-7714-2893-7; (online) 978-0-7714-2900-2.

Dejan Vucetic and Slobodan P. Simonovic (2011). Water Resources Decision Making Under Uncertainty. Water Resources Research Report no. 073, Facility for Intelligent Decision Support,

Department of Civil and Environmental Engineering, London, Ontario, Canada, 143 pages. ISBN: (print) 978-0-7714-2894-4; (online) 978-0-7714-2901-9.

Angela Peck, Elisabeth Bowering, and Slobodan P. Simonovic (2011). City of London: Vulnerability of Infrastructure to Climate Change, Final Report. Water Resources Research Report no. 074, Facility for Intelligent Decision Support, Department of Civil and Environmental Engineering, London, Ontario, Canada, 66 pages. ISBN: (print) 978-0-7714-2895-1; (online) 978-0-7714-2902-6.

M. Khaled Akhtar, Slobodan P. Simonovic, Jacob Wibe, Jim MacGee and Jim Davies (2011). An Integrated System Dynamics Model for Analyzing Behaviour of the Social-Energy-Economy-Climate System: Model Description. Water Resources Research Report no. 075, Facility for Intelligent Decision Support, Department of Civil and Environmental Engineering, London, Ontario, Canada, 211 pages. ISBN: (print) 978-0-7714-2896-8; (online) 978-0-7714-2903-3.

M. Khaled Akhtar, Slobodan P. Simonovic, Jacob Wibe, Jim MacGee and Jim Davies (2011). An Integrated System Dynamics Model for Analyzing Behaviour of the Social-Energy-Economy-Climate System: User's Manual. Water Resources Research Report no. 076, Facility for Intelligent Decision Support, Department of Civil and Environmental Engineering, London, Ontario, Canada, 161 pages. ISBN: (print) 978-0-7714-2897-5; (online) 978-0-7714-2904-0.

Nick Millington, Samiran Das and Slobodan P. Simonovic (2011). The Comparison of GEV, Log-Pearson Type 3 and Gumbel Distributions in the Upper Thames River Watershed under Global Climate Models. Water Resources Research Report no. 077, Facility for Intelligent Decision Support, Department of Civil and Environmental Engineering, London, Ontario, Canada, 53 pages. ISBN: (print) 978-0-7714-2898-2; (online) 978-0-7714-2905-7.

Andre Schardong and Slobodan P. Simonovic (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.

Samiran Das and Slobodan P. Simonovic (2012). Assessment of Uncertainty in Flood Flows under Climate Change. Water Resources Research Report no. 079, Facility for Intelligent Decision Support, Department of Civil and Environmental Engineering, London, Ontario, Canada, 67 pages. ISBN: (print) 978-0-7714-2960-6; (online) 978-0-7714-2961-3.

Rubaiya Sarwar, Sarah E. Irwin, Leanna King and Slobodan P. Simonovic (2012). Assessment of Climatic Vulnerability in the Upper Thames River basin: Downscaling with SDSM. Water Resources Research Report no. 080, Facility for Intelligent Decision Support, Department of Civil and Environmental Engineering, London, Ontario, Canada, 65 pages. ISBN: (print) 978-0-7714-2962-0; (online) 978-0-7714-2963-7.

Sarah E. Irwin, Rubaiya Sarwar, Leanna King and Slobodan P. Simonovic (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.

Samiran Das and Slobodan P. Simonovic (2012). Guidelines for Flood Frequency Estimation under Climate Change. Water Resources Research Report no. 082, Facility for Intelligent Decision Support, Department of Civil and Environmental Engineering, London, Ontario, Canada, 44 pages. ISBN: (print) 978-0-7714-2973-6; (online) 978-0-7714-2974-3.

Angela Peck and Slobodan P. Simonovic (2013). Coastal Cities at Risk (CCaR): Generic System Dynamics Simulation Models for Use with City Resilience Simulator. Water Resources Research Report no. 083, Facility for Intelligent Decision Support, Department of Civil and Environmental Engineering, London, Ontario, Canada, 55 pages. ISBN: (print) 978-0-7714-3024-4; (online) 978-0-7714-3025-1.

Roshan Srivastav and Slobodan P. Simonovic (2014). Generic Framework for Computation of Spatial Dynamic Resilience. Water Resources Research Report no. 085, Facility for Intelligent Decision Support, Department of Civil and Environmental Engineering, London, Ontario, Canada, 81 pages. ISBN: (print) 978-0-7714-3067-1; (online) 978-0-7714-3068-8.

Angela Peck and Slobodan P. Simonovic (2014). Coupling System Dynamics with Geographic Information Systems: CCaR Project Report. Water Resources Research Report no. 086, Facility for Intelligent Decision Support, Department of Civil and Environmental Engineering, London, Ontario, Canada, 60 pages. ISBN: (print) 978-0-7714-3069-5; (online) 978-0-7714-3070-1.

Sarah Irwin, Roshan Srivastav and Slobodan P. Simonovic (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.

Andre Schardong, Roshan K. Srivastav and Slobodan P. Simonovic (2014).Computerized Tool for the Development of Intensity-Duration-Frequency Curves under a Changing Climate: Users Manual v.1 Water Resources Research Report no. 088, Facility for Intelligent Decision Support, Department of Civil and Environmental Engineering, London, Ontario, Canada, 68 pages. ISBN: (print) 978-0-7714-3085-5; (online) 978-0-7714-3086-2.

Roshan K. Srivastav, Andre Schardong and Slobodan P. Simonovic (2014). Computerized Tool for the Development of Intensity-Duration-Frequency Curves under a Changing Climate: Technical Manual v.1 Water Resources Research Report no. 089, Facility for Intelligent Decision Support, Department of Civil and Environmental Engineering, London, Ontario, Canada, 62 pages. ISBN: (print) 978-0-7714-3087-9; (online) 978-0-7714-3088-6.

Roshan K. Srivastav and Slobodan P. Simonovic (2014). Simulation of Dynamic Resilience: A Railway Case Study. Water Resources Research Report no. 090, Facility for Intelligent Decision Support, Department of Civil and Environmental Engineering, London, Ontario, Canada, 91 pages. ISBN: (print) 978-0-7714-3089-3; (online) 978-0-7714-3090-9.

Nick Agam and Slobodan P. Simonovic (2015). Development of Inundation Maps for the Vancouver Coastline Incorporating the Effects of Sea Level Rise and Extreme Events. Water Resources Research Report no. 091, Facility for Intelligent Decision Support, Department of Civil and Environmental Engineering, London, Ontario, Canada, 107 pages. ISBN: (print) 978-0-7714-3092-3; (online) 978-0-7714-3094-7.