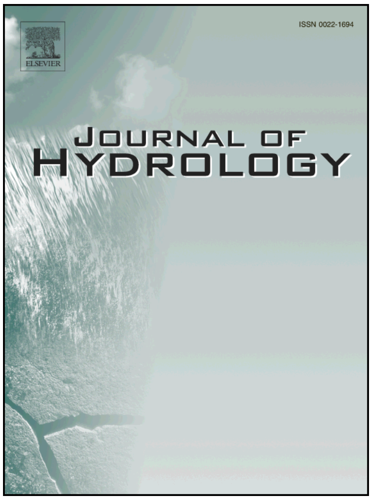
Research papers

Automatic Feature Selection and Weighting for the Formation of Homogeneous Groups for Regional IDF Estimation

Zhe Yang, Donald H. Burn



PII: S0022-1694(19)30460-3
DOI: https://doi.org/10.1016/j.jhydrol.2019.05.015
Reference: HYDROL 23748

To appear in: *Journal of Hydrology*

Received Date: 18 May 2018
Revised Date: 2 May 2019
Accepted Date: 5 May 2019

Please cite this article as: Yang, Z., Burn, D.H., Automatic Feature Selection and Weighting for the Formation of Homogeneous Groups for Regional IDF Estimation, *Journal of Hydrology* (2019), doi: https://doi.org/10.1016/j.jhydrol.2019.05.015

Automatic Feature Selection and Weighting for the Formation of Homogeneous Groups for Regional IDF Estimation

Zhe Yang[1], Donald H. Burn[1]

[1]Department of Civil and Environmental Engineering, University of Waterloo, Waterloo, ON, Canada N2L 3G1.

Abstract: The intensity-duration-frequency (IDF) curve has been used as an effective tool to quantify the risk associated with the impact of extreme rainfall on civil infrastructure. However, recent changes in the rainfall climatology caused by climate change and urbanization have made estimates in the stationary environment provided by the traditional regional IDF approach increasingly inaccurate. This inaccuracy is mainly caused by the lack of consideration for the temporal and spatial difference in the selection of similarity indicators (attributes that are used to measure similarity of extreme rainfall patterns among different stations), resulting in ineffective formation of a homogeneous group (a group of stations that share similar extreme rainfall patterns) at various regions. To consider the temporal differences of similarity indicators, including meteorological factors, topographic features and urban impact indicators, a three-layer design is proposed based on the three stages in extreme rainfall formation: cloud formation, rainfall generation and change of rainfall intensity over an urban surface. During the process, the impacts from climate change and urbanization on extreme rainfall patterns are considered through the inclusion of potential features that relate to the rainfall mechanism at each layer. The spatial differences of similarity indicators for Homogeneous Group Formation (HGF) at various regions is resolved by using an automatic feature selection and weighting algorithm, specifically the hybrid searching algorithm of Tabu Search, Lagrange Multiplier and Fuzzy C-means clustering, to select the optimal combination of features for HGF based on the uncertainty in the regional estimates of the rainfall quantiles for a specific site. The proposed methodology fills the gap of including the urbanization impacts on the extreme rainfall patterns during HGF process and challenges the traditional assumption that the same set of features can be equally effective in generating the optimal homogeneous group in regions with different geographic and meteorological characteristics.

Key words: Regional frequency analysis; IDF curve; Feature selection and weighting; Climate change; Urbanization; Homogeneous group formation

1. Introduction

Extreme rainfall events can result in devastating impacts. Under the influence of climate change, more frequent and intensive extreme rainfall events have been observed around the world. In 1998, a series of extreme rainfall events on the Yangtze River affected the living situation for more than 223 million people, which caused the estimated economic damages at 166,600 million Yuans (Zong & Chen, 2000). In 2005, Canada experienced one of the most devastating floods in the province of Alberta and large areas in southern Ontario, and suffered CAD 800 million insured damages (Sandink, 2013). In the same year, Hurricane Katrina, one of the deadliest and costliest storms in the history of the United States, resulted in nearly 1,500 fatalities (Boyd, 2010). In 2013, extreme floods in Central Europe, which were caused by long periods of heavy rains, resulted in €12 billion overall losses (Khazai et al., 2013). Under the current trend of climate change, the frequency of extreme rainfall events is expected to increase in some parts of world (Goswami et al, 2006; Cai et al., 2014). Considering the potentially increasing damages that can be caused by future extreme rainfall, accurate intensity-duration-frequency (IDF) curves are needed for updating the drainage infrastructure especially in urban areas.

Regional frequency analysis, which extends the data records from one site by gathering the observations from sites that share similar rainfall characteristics, is widely used for IDF estimation at specific sites. However, the estimates of design rainfall using traditional regional IDF curves are becoming increasingly inaccurate in the nonstationary environment. One of the major problems that causes this inaccuracy is the ineffective formation of the homogeneous region. Normally, the stations in the homogeneous group are selected based on their rainfall similarity to the target site. The similarity indicators can be divided into the categories of site characteristics and at-site statistics (Hosking & Wallis, 1997). Site characteristics, which are the physical representations of the weather stations (such as the geographic location, the orientation of the landscape, etc.), are used in the traditional geographic HGF process under the assumption that geographic proximity indicates rainfall similarity (Hosking & Wallis 1997; Ahmad et al. 2013; Burn, 2014; Haddad et al. 2015). However, extreme rainfall events that are assumed to remain stationary have been altered by climate change across Canada since 1950, and the patterns of their changes are not spatially uniform (Warren & Lemmen, 2014). At a large scale, a large increase in precipitation totals is observed in the Arctic, while decreasing trends have been detected in the Prairies region in southern Canada (Zhang et al, 2000). As for the rainfall patterns at a local scale, many papers have stated that daily rainfall is affected by the urban surface heat fluxes (Shepherd, 2005; Miao et al, 2011; Li et al., 2013). Since the constant site characteristics, specifically the geographic features, cannot reveal the possible changes in the rainfall patterns, their effectiveness as similarity indicators in the HGF process needs to be questioned.

Two potential solutions have been proposed to solve this problem. One is to ignore the regionalization step, and directly use remote sensing rainfall records at regional scale for the regional IDF estimation (Marra et al., 2017). However, this approach is used to provide areal information of the IDF curves, and the high uncertainty associated with remote sensing rainfall records can still be a huge problem. The other is to search for the more effective similarity indicators, which will be the focus in this study. To adjust HGF process under climate change, the at-site statistics, such as the mean annual precipitation, the mean number of wet days, the ratio of minimum average 2-month precipitation to maximum average 2-month precipitation, the parameters of hydrologic distributions etc. (Easterling, 1989; Hosking & Wallis, 1997; Gaál & Kyselý, 2009; Yang et al., 2010), have been used as the additional indicators in the process. However, the at-sites statistics are only obtainable at the gauged stations and can only achieve reliable conclusions when long rainfall records are available. To resolve this issue, atmospheric variables have been used as the new similarity indicators. Satyanarayana and Srinivas (2008) considered the region's hydro-meteorology through including monthly mean series of the large-scale atmospheric variables such as the specific humidity, temperature, precipitable water, wind velocity and wind direction, together with the geographical factors as the indicators for the HGF. Gabriele and Chiaravalloti (2013) used the Convective Available Potential Energy and the Q vector Divergence as indicators in the homogeneous macro region formation, and the Vertically Integrated Moisture Flux for the formation of homogeneous sub-regions. These methods are conducted under the assumption that the same set of features can serve as the optimal similarity indicators to detect rainfall patterns at different stations. Asong et al. (2015) advanced the former approach by using canonical correlation analysis to select atmospheric variables, teleconnection indices and geographical site attributes that could explain the spatial patterns of precipitation of rainfall events in Canadian Prairie Provinces. However, the impacts of these selected indicators on the monthly precipitation were not consistent at different locations over the study area. Since there is a spatial difference among the extreme rainfall patterns at stations with different meteorological and geographic characteristics, should this difference be considered when selecting the similarity indicators for representing these different extreme rainfall patterns at various stations? This paper proposes an objective-oriented automatic feature selection and weighting algorithm to select the optimal feature combination for describing the rainfall patterns at the target site, which is used to form the homogeneous group that has the lowest uncertainty in the quantile estimates.

In the context of the above research gaps, this proposed methodology seeks to answer the following questions:

1) Do the potential similarity indicators respond to extreme rainfall events on the same temporal scale?
2) Is there a spatial difference in the similarity indicators that are used to distinguish rainfall patterns among the input stations at various locations?
3) Can urbanization alter the extreme rainfall regime, thus changing the frequency and magnitude of extreme rainfall events for an urban area?

These questions are explored through the proposed searching algorithm in a three-layer framework. The remainder of the paper is organized as follows. Section 2 presents the basic procedures for conducting regional IDF analysis. Section 3 describes the automatic feature selection and weighting algorithm in a three-layer design. Sections 4 and 5 present the data used in the study and the results from the analysis. The paper concludes in section 6 with conclusions and a discussion of potential avenues for future work.

2. Regional IDF estimation

An Intensity-Duration-Frequency (IDF) curve is used to describe the occurrence probability of certain rainfall intensity and can provide essential information for civil infrastructure design to reduce the potential damage caused by heavy rainfall. To quantify the associated risk of these extreme rainfall events and capture the changes of the rainfall temporal distribution, extreme events are commonly described by their return periods (T). In this research, IDF curves are estimated using the annual maximum rainfall series (AMS), which contains independent observations generated from different hydro-climatological events.

Regional frequency analysis is performed under the assumption that rainfall series from all of the stations in the homogeneous region share the same frequency distribution. The quantile estimate can be described by the following index-event equation (Dalrymple, 1960):

$$\hat{R}_{T,D} = \bar{R}_D \hat{x}_{T,D} \quad (1)$$

where D indicates the duration of the rainfall events under consideration, $\hat{R}_{T,D}$ is the regional estimate of T-year event for D duration at a specific site, $\bar{R}_D$ is the site's index event at D duration and $\hat{x}_{T,D}$ is the regional estimate of the dimensionless growth curve. The estimates can be obtained through the following steps (Burn, 2014):

1) Data screening. To satisfy the stationary requirement in the traditional frequency analysis, the rainfall series with a change point or a trend identified through Pettitt test or Mann–Kendall nonparametric test are removed.
2) Attribute selection. Geographic distance is used to identify stations sharing similar rainfall patterns with the target site.
3) HGF. Homogeneous groups can be formed using the selected indicators through different pooling or clustering approaches.
4) Evaluation of the homogeneous region. A heterogeneity measure is evaluated for the initial pooling group and revisions of this group will be conducted if the heterogeneity measure does not meet the criteria of homogeneity.

5) Estimation of quantiles. The scaled time series from the group obtained from step (4) is used to identify the appropriate distribution and the corresponding quantile function. The rainfall quantile at a target site can be estimated by applying Equation (1) to the distribution obtained from step (5).
6) Uncertainty quantification. Parametric samplings through 1000 simulations are used to quantify the uncertainty in the rainfall quantile estimates (Hosking & Wallis, 1997)

3. Methodology

To address the research questions and consider the possible alterations in the rainfall regime caused by climate change and urbanization, this paper explores the possibility of including relevant meteorological factors, topographic features and urban impact indicators as the similarity measures for HGF through the feature selection and weighting process in a three-layer framework. The procedures in the proposed algorithm can be briefly summarized as: 1) Based on the rainfall mechanism, potential similarity indicators at all three layers are selected. 2) The relationships between the most responsive indicators to the extreme rainfall events and less responsive ones are analyzed to determine their appropriate temporal resolutions (TRs). 3) Homogeneous groups are formed using the selected similarity indicators and the uncertainty in quantile estimates at the target site is quantified. During the HGF process, feature weighting is used to reduce the impact of feature correlation on the formation process. 4) Step (3) is repeated multiple times using a search algorithm, and the optimal combination of similarity indicators is determined based on the quantified uncertainties.

The details of the proposed methodology are described in the following sections. Based on the mechanism of extreme rainfall events, a three-layer design is proposed in Section 3.1, and potential features that can affect extreme rainfall events at each layer are identified. To obtain the optimal TRs for the potential features at each layer, entropy differences and correlation coefficients between the most responsive indicators and less responsive ones at different TRs are used as the measures in Section 3.2. In Section 3.3 and 3.4, Fuzzy C mean clustering is proposed as the tool for HGF, and Lagrange Multiplier is used to reduce the impact of feature correlation on the clustering results. Tabu Search, which is used as the search algorithm, is described in Section 3.5.

3.1 Three-layer design

Three stages can be identified in the process of extreme rainfall formation: cloud formation, rainfall generation and the change of rainfall intensity above the urban surface. Based on the possible heights of these stages, a three-layer design is proposed as the framework in which related features at each stage are selected as shown in Figure 1. The primary goal of this design is to distinguish the possible temporal differences among the features for the three layers.

The first layer of the proposed framework includes the atmosphere beyond the Planetary Boundary Layer (PBL), and is designed to consider the climate change impacts on the formation of the clouds that produce the extreme rainfall events at a large scale. Cloud is a suspended water or crystal mass in the air, which forms when the water vapor in the atmosphere condenses on the nuclei such as dust or ice. Cloud formation is affected by several atmospheric variables including air temperature, geopotential height, specific humidity, U-component and V-component of wind velocity. Thus these features at three pressure levels (300hPa, 500hPa and 700hPa) are included as potential similarity indicators at this layer. In contrast to the rainfall events that are caused by low level clouds, extreme rainfall events are normally caused by two types of clouds located at middle or high levels of the troposphere: cumulonimbus and nimbostratus, which are corresponding to the production of convective and stratiform rainfall (Houze, 1989; Houze Jr., 1997). Cumulonimbus is a multi-level cloud that starts as a low level cloud (cumulonimbus calvus) and expands to the middle (cumulonimbus capillatus) even high level (cumulonimbus incus) due to the unstable atmosphere. This cloud type is capable of producing severe convective rainfall events in a short period. Nimbostratus forms in the middle level of the troposphere because of the rising warm air, and occurs along a warm or occluded front where the stratiform rainfall can be produced. To describe the atmosphere instability in the process of these two cloud formations, the Convective Available Potential Energy (CAPE) index, which represents the amount of energy that would be needed to lift a parcel of air through a certain distance in the atmosphere (Moncrieff & Miller, 1976), is selected and acts as the most responsive similarity indicator to the extreme rainfall events in the first layer (Gabriele & Chiaravalloti, 2013).

After a cloud is formed, rainfall can be generated when the requirement of cooling is met, which happens within the PBL, i.e. in the Urban Mixing Layer (UML) which is the second layer of the framework. As the linkage between the first and third layer, the UML is under the influence from both climate change and urbanization at a regional scale (Collier, 2006). Based on the cloud types that can be used to generate the extreme rainfall events, the types of rainfall can be divided into the convective rainfall, and stratiform rainfall type which can be further divided into relief and frontal rainfall depending on the underlying geographic topography. Thus the geographic attributes including latitude and longitude of the potential stations, and the same category of atmospheric variables used in the first layer but at lower pressure levels (850hPa and 925hPa) are collected as potential indicators. What makes the extreme rainfall events different from the normal rainfall events is the abundance of water supply to feed the events (Gabriele & Chiaravalloti, 2013). Thus the vertical integral of divergence of moisture flux, which describes the transport of the net atmosphere moisture flux per unit volume, is selected as the most responsive indicator in this layer.

The final layer is the Urban Surface Layer (USL), where rainfall intensity can be affected by the urban surface energy in sub-regional or even local scale. Many papers have stated that the local rainfall climatology in the urban environment has been altered due to the following mechanisms (Li et al., 2013): 1) Urban heat island effects. The presence of high buildings can increase surface roughness, and also affect the local energy fluxes or disrupt the thermal balance. This results in the locally enhanced convergence and increased surface temperature, which causes changes in the rainfall patterns in the urban area (Roth & Oke, 1993; Roth, 2000; Bornstein & Lin, 2000; Shepherd, 2005); 2) Urban canopy effects. The lack of the canopy covers and the altered underlying surface characteristics can affect the surface temperatures or even divert precipitating systems (Shepherd, 2005; Loughner et al., 2012); 3) Urban aerosol effects. The increasing accumulation of aerosols in the atmosphere can both decrease and increase rainfall amount because of their radiative and cloud condensation nuclei (CCN) activities (Diem & Brown, 2003; Kaufman & Koren, 2006; Jin et al. 2010; Rosenfeld et al., 2014): a) radiation effects. On one hand, the aerosols impede the process of solar radiation reaching the land surface thus reduce the amount of available water for evaporation and cloud formation. On the other hand, aerosols can absorb solar radiation and cause the warming in the lower atmosphere, which can strengthen Asia monsoon circulation and increase local precipitation. b) CCNs activities. While the added CNNs can accelerating the process of cloud drops turning into raindrops, they can also cause the adverse effects if the cloud drops are too small. Thus relevant features are selected as rainfall similarity indicators in the last layer: The urban energy flux including the surface sensible heat flux and the surface latent heat flux (Miao et al., 2011), Photosynthetically Active Radiation index (PAR), Surface Net Solar Radiation (SNSR), Surface Net Thermal Radiation (SNTR) and the Surface Roughness (SR). Potential features at all three layers are listed in Table 1.

3.2 TR detection of the features series

Based on the description in Section 3.1, the potential features at each layer are under the influence of various phenomena and thus can be related to the rainfall events at different temporal scales. It is necessary to obtain representative values of those features at the optimal TRs that can best describe the rainfall related patterns at the different layers. The correlations and entropy differences between the most responsive features and the less responsive ones at different TRs are used as indicators to measure their similarities and determine the optimal TRs at the three layers. While correlation coefficient works best among linear correlated features, the entropy difference may provide better explanations when the input features are nonlinearly correlated (Quiroz et al., 2011).

The discrete wavelet decomposition (DWT) is used to obtain the orthonormal bases of feature values at different TRs. In the discrete form, the wavelet transform of a function $f(t)$is defined as the integral transform as (Kumar & Foufoula-Georgiou 1997):

$$Wf\left(m_w,n\right)=\lambda_0^{-m_w/2}\int_{-\infty}^{\infty} f\left(t\right)\psi\left(\lambda_0^{-m_w}t-nt_0\right)dt \tag{2}$$

$$\psi_{m_w,n}\left(t\right)=\frac{1}{\sqrt{\lambda_0^{m_w}}}\psi\left(\frac{t-nt_0\lambda_0^{m_w}}{\lambda_0^{m_w}}\right)=\lambda_0^{-m_w/2}\psi\left(\lambda_0^{-m_w}t-nt_0\right) \tag{3}$$

where $f\left(t\right)$ indicates the one-dimensional time series; λ is the scale parameter; t is a location parameter (indicates time in time series); $\overline{\psi}_{\lambda,t}\left(\mu\right)$is the complex conjugate of $\psi_{\lambda,t}\left(\mu\right)$ (in this study, the Least Asymmetric wavelet); m_w and n are an integer and λ_0 is a fixed dilation step greater than 1. Details of the multiresolution analysis can referred to Mallat (1989) and Daubechies (1990).

To focus the methodology solely on the annual extreme rainfall events, the feature values should match the date of occurrence for the annual extreme rainfall events and be extracted at the optimal TRs at each layer.

3.3 Homogeneous group formation

The primary goal of this study is to obtain homogeneous groups that can generate the lowest uncertainty in the quantile estimates. Normally, HGF can be conducted using two approaches: pooling and clustering. The pooling method , such as the region of influence (ROI) approach, gathers the stations centred around the pre-set target site based on their similarities (Burn, 1990). However, the problem of determining the number of stations in the target group to achieve the optimal balance between the number of stations in the pooling group and its homogeneity is still unsolved (Burn, 2014). Here, the clustering, specifically the fuzzy c means clustering is used for the HGF (Bezdek et al, 1984).

Unlike non-overlapping clustering, Fuzzy clustering provides the possibility for one site having partial membership in more than one cluster, which has the advantage to form the overlapping clusters for undistinctive data with vague boundaries (Bezdek et al, 1984). Furthermore, in the situation when only limited number of input stations are available, the possibility of non-target site being the cluster centre allows the clustering method to form better groups to generate the lowest uncertainty in quantile estimates.

Fuzzy c means clustering is conducted through the minimization of the following objective function (Bezdek et al, 1984):

$$J_m = \sum_{i=1}^{N} \sum_{j=1}^{C} u_{ij}^{m} \left\| x_i - c_j \right\|^2, \qquad 1 \le m < \infty \tag{4}$$

$$u_{ij} = \frac{1}{\sum_{k=1}^{C}\left[\frac{\left\|x_i - c_j\right\|}{\left\|x_i - c_k\right\|}\right]^{\frac{2}{m-1}}} \qquad (5)$$

where N is the number of stations; C is the number of clusters; u_{ij} is the membership value of station i in the cluster j, and can be calculated using Equation (5) ; x_i is the value of data points in station i; c_j is the center value of the cluster j; and m is the weighting exponent in the fuzzy c means cluster.

3.4 Feature weighting

To deal with the possible issues caused by feature correlation during clustering, the Lagrange Multiplier is used (Borgelt, 2008). The objective function that is used in Lagrange Multiplier can be described as:

$$\varepsilon = \sum_{V=1}^{V}(w_V)^p \sum_{i=1}^{N}\sum_{k=1}^{M}\delta_{ik}\left\|\vec{x}_i^{\,V} - \vec{m}_k^{\,V}\right\|^2 \qquad (6)$$

where p is the exponential parameter that is used to control the sparsity of the feature weightings. Research has shown that the clustering results remain stable when p increases to a certain value (Xu, et al. 2014), and is set to 6 in this study; V is the number of variables in the cluster; w_V is the weighting value for each variable; N is the number of stations involved in the cluster; M is the number of center groups obtained from clustering; δ_{ik} is the belonging value of each station to each cluster; $\vec{x}_i^{\,V}$ is the time series vector for a variable for station i; $\vec{m}_k^{\,V}$ is the time series vector of certain variable for each cluster center k .

Based on the objective function in Equation (6), the Lagrange formula that will be applied to obtain the weighting value for each variable is (Xu, et al., 2014):

$$L(\varepsilon,\lambda)=\varepsilon(w)+\lambda\left(\sum_{V=1}^{V}w_V-1\right)\qquad subject\ to:\sum_{V=1}^{V}w_V=1 \tag{7}$$

Based on Equation (7), the corresponding weightings can be derived from the derivative and are listed as (Xu, et al., 2014):

$$w_V=\frac{1}{\sum_{V'=1}^{V}\left(\frac{D_V}{D_{V'}}\right)^{1/(p-1)}},\qquad P>1;\quad D_V=\sum_{i=1}^{N}\sum_{k=1}^{M}\delta_{ik}\left\|x_i^V-m_k^V\right\|^2 \tag{8}$$

3.5 Search algorithm

To address the first research question and select the optimal combination of similarity indicators for HGF, Tabu Search, which is regarded as the most effective tool to obtain the optimal subsets of relevant attributes in the feature selection domain, is used (Zhang & Sun, 2002). The Tabu Search is a combinatorial optimization algorithm that avoids the possibility of the searching process being stuck in a local optimum by crossing the boundaries of local optimality through using the inferior objective values in the process (Glover, 1986).

Three types of memory structures are involved in the search process: short, intermediate and long-term. The short-term memory or the regency-based memory, which is constantly modified in the search, is used to record the Tabu list and guide the search process (Glover, 1989; Zhang & Sun 2002). It is constructed by labeling the attributes that have been visited recently as the Tabu-active attributes, and these attributes will not be used again in the following search (Glover, 1989). While Tabu list is recorded to reduce the visiting of pre-visited attributes, the aspiration criteria is used to restrict the pre-visited moves (i.e. the Tabu move), and can be updated if the objective function value resulted from using the solution in consideration is better than the current aspiration value (Glover, 1989). In summary, the short-term Tabu Search can be carried out through the steps showing in Figure 2 (Sait & Youssef, 1999; Zhang & Sun, 2002). The intermediate memory and long term memory are used in the intensification and diversification process separately to either re-evaluate historically good solutions or incorporate features that have not been previously included as new solutions (Glover, 1989).

Considering the large number of potential features involved in Tabu Search, the process is established under the assumption that the best feature combinations that generates the optimal objective value at each layer can result in the best final outcome for the whole process. During the process, the searching for the optimal feature combination at each layer will be conducted separately, as shown in Figure 3. To avoid over-clustering caused by conducting the search at every layer, the comparison of the objective values between the groups obtained from the higher layer and the lower layer are used as the criteria to determine the necessity of conducting the procedure at the lower layer. If the lower layer generates a better objective function value compared to that from the higher layer, the search results from the lower layer is acceptable. If this condition is not met, the proposed procedure will not be conducted in the lower layer, and the algorithm will directly move to the next layer instead. The objective value being used in the searching process is the uncertainty of quantile estimates generated from the formed group, i.e., the average widths of confidence interval (CI) for the rainfall quantile estimates at different return periods.

3.6 Automatic feature selection and weighting algorithm

Based on the above sections, the automatic feature selection and weighting algorithm can be conducted through the following procedures. The flow chart of the proposed search algorithm is illustrated in Figure 3.

1) Original Feature Gathering. The time series of the potential rainfall-related features for the initial group of stations at all three layers are extracted at available spatial resolutions.
2) Temporal Scaling. Based on the method presented in Section 3.2, wavelet decomposition is used to extract the feature values at different TRs.
3) Feature Scaling. The method of scaling to unit length is used to standardize the range of each feature value, can be achieved by using equation $x' = \frac{x}{\|x\|}$, where x is the feature vector before scaling, $\|x\|$ is the Euclidean length of the vector, and x' is the vector after scaling.
4) Feature selection and weightings at the first layer or the higher layer. The hybrid searching algorithm can be conducted through the following procedures:
   (a) Initial selection of features. Randomly select certain number of features from the potential feature datasets in the higher layer, and record them in the Tabu list.
   (b) Initial equal weightings for the selected features.
   (c) Fuzzy c mean clustering. Use the weightings from step (b) and the selected features from step (a) in the fuzzy c mean clustering to form the homogeneous group for the target site.
   (d) Features weightings recalculation. Based on the clustering results from step (c), re-calculated the features weightings using Lagrange Multiplier.
   (e) Repeat step (c) and (d). Use the feature weightings from step (d) as the new weightings, and repeat the steps (c) and (d). This process continues until the weighting obtained from

step (d) is close to the weightings used in step (c). Then use these final weightings in step (d) to obtain the final homogeneous group for the target site.

(f) Objective value calculation. Use the formed group obtained from step (e) for the objective value calculation, i.e., the average CI widths for the quantile estimates at different return periods. Record this objective value in the searching process.

(g) Repeat steps (a) – (f) certain times. Based on the recorded memory of the objective values and Tabu list, select the next possible combination of features, and repeat the above process.

(h) Searching stop. After the pre-set number of iterations in the search process, select the optimal set of features based on the recorded memory of objective values.

5) Comparison. Compare the formed group from step 4) with the original input group for uncertainty in quantile estimates of the target site through the following steps: a) Uncertainty quantification. To quantify the uncertainty in the quantile estimates, parametric sampling is used to calculate the CI widths (Hosking & Wallis, 1997). b) Uncertainty Comparison. Calculate the ratio of the CI widths between the formed homogeneous groups from lower layer and the original input group from the higher layer. c) Boxplots. Repeat step b) 1000 times and box plots are created based on these calculated ratios. If the median values of the ratio in this boxplot was less than one at most of the return periods, the formed group at the lower layer would be considered as a better homogeneous group and accepted as the input group for repeating step 4) at the next layer. If not, the original group of stations would be used as the input for repeating step 4) at the next layer.

6) Repeat steps 4) and 5) at the second layer using the resulting group from previous layer as the input.

7) Repeat steps 4) and 5) at the third layer using the resulting group from previous layer as the input.
8) Final homogeneous group is obtained. The resulting group from step 7) will be used as the final homogeneous group for the quantile estimation at target site.

4. Data and study area

4.1 Study area

The proposed methodology is tested in four regions in Canada. Across Canada, there are eight different climate regions: Arctic, Pacific Maritime, Cordilleran, Taiga, Boreal, Prairie, Southeastern and Atlantic Maritime. To verify the assumption of the proposed methodology that different optimal feature combinations can be used as the similarity indicators for HGF with different input stations at various regions, four regions with a large number of rainfall stations were selected as the tested regions:

1) Region 1 is along the west coast and includes the IDF stations from Yukon Territory, Northwest Territories and intensive urbanized area in British Columbia. Due to the close proximity to the ocean and unique geographic characteristic in the mountain area, this region mainly experiences the Pacific Maritime and Cordilleran climate type; thus stratiform rainfall will be the main rainfall type throughout the year.

2) Region 2 is in the west and specifically includes weather stations from the Prairie and the adjacent Boreal climate region, where convective extreme rainfall is the main extreme rainfall type in the summer.

3) Region 3 is in the Boreal and Southeastern regions and includes stations from Ontario and Quebec, both of which have some intensive urbanized areas. While convective extreme rainfall

events may be the main rainfall type in the Boreal region, stratiform rainfall is perceived to be more common in the Southeastern region due to its closeness to large water bodies.

4) Region 4 is in the Atlantic region and includes the stations from Nova Scotia, New Brunswick, Prince Edward Island and Newfoundland, where rainfall events are affected by cyclonic storms from the Atlantic Ocean. Due to the limited number of IDF stations in the Atlantic region, some nearby sites from Quebec are also included.

4.2 Extreme rainfall datasets

Based on the methodology presented in Section 3, two categories of information are required before conducting the searching process: the values of the extreme rainfall at different durations and their corresponding dates of occurrence. The first category of information is used for the rainfall quantile estimation during the uncertainty comparison and can be retrieved from the Engineering Climate Datasets provided by Environment and Climate Change Canada (ECCC) in the form of the annual maximum rainfall series (AMS) at nine durations ranging from 5 min to 24 h (5, 10, 15, and 30 minutes and 1, 2 6, 12, and 24 h). If the data records were not complete during the analysis period, the historical dataset also provided by ECCC could be used as an alternative source for providing missing AMS data points for the duration of 24h. For the second category, the corresponding dates of occurrence for the annual extreme rainfall events were obtained based on the information provided from both the historical dataset on ECCC website and the total precipitation from ERA-Interim database provided by European Centre for Medium-Range Weather Forecasts (ECMWF) at the resolution of 0.125 degrees.

4.3 Potential features

Based on the methodology presented in Section 3.1, the potential features that on a yearly basis correspond to the date of occurrence for the AMS are extracted at different TRs from two sources: NOAA Global Ensemble Forecast System Reforecast (GEFS/R) and ERA-Interim Database from ECMWF (Berrisford et al., 2011;Hamill et al., 2013).

a) GEFS/R: The geopotential height, temperature, U-component and V-component of the wind and the specific humidity were extracted at five pressure levels (925hPa, 850hPa, 700hPa, 500hPa, 300hPa), plus the convective available potential energy, were collected at a resolution of 1 degree. All the data were extracted at the 3-hourly temporal interval basis.

b) ERA-Interim: The second layer indicator Vertical Integral of Divergence of Moisture Flux (VIMF) at a resolution of 1 degree, and the third layer indicators including Photosynthetically Active Radiation index (PAR), Surface Net Solar Radiation (SNSR), Surface Net Thermal Radiation (SNTR) and the Surface Roughness (SR) were extracted at a resolution of 0.125 degree. All of these indicators are extracted at the 6-hourly temporal interval basis.

4.4 Data Screening

Based on the availability of the observations from the above sources, the period for the search algorithm was set from 1985 to 2004. However, in the uncertainty comparison among different formed groups, all the available data points in the stationary rainfall period from the Engineering Climate Datasets were used considering the statistical accuracy that can be achieved with a larger sample size. Thus during the process of data screening, the selected stations in each region should have stationary extreme rainfall time series for two periods (the period from 1985 to 2004 and the whole available time period obtained from Engineering Climate Datasets), since the extreme rainfall series from shorter period could be regarded as the representatives for the longer period if they were both stationary. This objective was achieved through applying Pettitt test and Mann–Kendall nonparametric test with the block bootstrap resampling (Burn, 2014). The final number of stations with stationary rainfall series for each region is: 86 stations in Region 1 with VANCOUVER INTL A being the target site; 78 stations in Region 2 with the CALGARY INT'L being the target station; 162 stations in Region 3 with TORONTO CITY being the target station; and 72 stations in Region 4 with GANDER AIRPORT CS being the target station.

5. Application

To demonstrate the effectiveness of the proposed methodology and address all the research questions, the analysis in the following sections were conducted: 1) to answer the first research question, the methodology in Section 3.2 was applied in Section 5.1 to distinguish the optimal TRs for the features at different layers. 2) To answer the second question, all correlation coefficients and entropy differences between the most responsive indicators and less responsive ones at the first and second layers for all the stations were compared in Section 5.1 to demonstrate the spatial differences among their extreme rainfall pattern indicators. 3) To prove the necessity of conducting the algorithm at all three layers and answer the third question, the uncertainty in the quantile estimates from the homogeneous groups that were generated by using the algorithm at lower layer were compared to that in the formed groups from the higher layer in Section 5.2. 4) To illustrate the effectiveness of the proposed methodology, the algorithm was applied to form the homogeneous groups for rainfall events with different durations at four different regions in Section 5.2 and 5.3, and the uncertainty in their quantile estimates were compared to that from using at-site analysis and also that in the formed groups generated by using a geographic approach in Section 5.4.

5.1 Temporal scaling for the potential features

Both entropy differences and correlation coefficients between the most responsive feature and less responsive ones at different TRs are used to determine the optimal TRs at first and second layers. Since the relationship among the potential features at each layer may not be linear, entropy differences may provide more reliable conclusions than correlation coefficients. In the first layer, correlations between the CAPE and the rest of the first layer features at all stations reach the highest values at the 128- day TR (level 10 decomposition), and their corresponding entropy differences are reasonably small at this level, thus 128 days is selected as the optimal TR for the features at the first layer. This 128-day TR is reasonable for the following reasons: 1) While precipitation process is a mesoscale atmospheric processes, it can be affected by the large scale phenomena including large clouds formation, warm and cold fronts collision etc. 2) In the three-layer design, the features in the first layer are used as the similarity indicators to gather stations that share similar extreme-rainfall related atmospheric characteristics at large scales. 3) Studies shows that extreme events-related atmospheric movements, specific CAPE-related features, are strong seasonal correlated with the seasonal migration of monsoon over West Africa and India (Murugavel et al., 2012; Meukaleuni et al., 2016).

The situation is more complicated at the second layer. For the majority of the stations, the correlation coefficients between the VIMF and the rest of features at the second layer increase steadily as the decomposition level increases until it reaches either level 10 or 9 (64-day TR) and the declines in their corresponding entropy differences become barely noticeable after level 9. Thus the 64-day TR is used as the optimal TR for extracting feature values at the second layer. In the third layer, the optimal TR is set to the duration of the extreme rainfall events of concern (i.e., 24-hour in this case) for two reasons: 1) No universal TR for all the stations can be obtained based on the values of entropy differences and correlation coefficients. 2) To meet the needs of considering extreme events at different durations for different construction purposes.

The spatial difference of the correlations between potential features for the stations in Region 1 is presented in Figure 4. Although at each station, the highest correlation coefficient values between the most responsive features and less responsive ones were all reached at level 10 or 9 decomposition for the first or second layers, major differences are still observed among these correlation values across different stations. At each station (which is described as each column in Figure 4), all the potential features at first (Figure 4(a)) and second (Figure 4(b)) layers have different levels of correlations to the extreme rainfall mechanisms. For each potential feature (which is described as each row in Figure 4), its correlations to extreme rainfall events vary largely among different stations at both layers. Thus the assumption of spatial difference for similarity indicators existing at different stations is valid. The same situation can be applied in Regions 2, 3 and 4.

However the approach of determining similarity indicators by choosing the ones that have higher correlation values at the target site is not suggested for the following reasons: 1) The correlation or entropy differences obtained is not equivalent to the representation of the mutual information among different features. 2) Even if the mutual information were used as a potential measure, the approach of choosing the features that are highly relevant to the most responsive features might result in biased clustering as these features may not represent all aspects in the complex rainfall system. 3) In the stations where low correlations have been detected across all the features, such as the first column in Figure 4 (a), the approach of choosing comparatively higher correlated features as the similarity indicators may result in inferior HGF.

5.2 HGF in Tabu Search

The cluster number was set to 2 in the fuzzy c mean clustering for the following reasons: 1) the larger the formed group the lower the uncertainty in the quantile estimates. The goal of this study is to form homogeneous groups with the lowest uncertainty in the quantile estimates not to find the optimal partition of the input stations in the clustering. In the formed group, both group heterogeneity and small group size can increase the uncertainty in the final quantile estimates. But the influence from group size normally outweighs the impact from heterogeneity as the geographic proximity of the original input group guarantees a certain level of group homogeneity. 2) The pre-knowledge of the cluster number for the first two layers. The clustering at the first layer is constructed to distinguish the stations whose extreme rainfall events are mainly caused by cumulonimbus or nimbostratus clouds. However, the extreme rainfall events at certain durations, especially the longer ones, are likely caused by the combination of different cumulonimbus or nimbostratus clouds. For those stations whose rainfall events are caused by the combination of cumulonimbus and nimbostratus, the second layer is used to further separate the input stations into the convective and stratiform rainfall type. For those stations whose rainfall events are caused by the nimbostratus, the clustering at the second layer is used to further separate rainfall type into relief and frontal rainfall. 3) The validity measures including partition coefficient (PC) index, partition entropy (PE) and the modified Xie–Beni index are not very effective in determining the optimal cluster number in this study (Xie & Beni, 1991; Wang & Zhang, 2007; Satyanarayana & Srinivas, 2011). As a conclusion, to gather a large number of stations to form the homogeneous group and reduce the computational cost without jeopardizing the search for the optimal feature combination, the search for the optimal feature combination was conducted with the fixed value of the weighting exponent (*m=2*) and the cluster number (*n=2*) for the clustering at all layers.

Based on the methodology presented in Section 3.4 and 3.5, the automatic feature selection and weighting algorithm is conducted consecutively in three layers. At each layer, the number of iterations is set to 60 with 3 repetitions in Tabu Search.

Two possible issues should be noted when selecting the optimal feature combination based on the memories recorded in Tabu Search: 1) Due to the stochastic characteristics in the estimation for the objective function values for each homogeneous group formed, one choice of feature combination can repeat several times in the Tabu Search with different objective function values. Thus the frequency of the feature combinations appearing in the top 100 in the recorded memory is also taken into consideration during the determination of optimal feature selection. 2) The high correlations among the input features can cause many sets of feature combinations, with different weightings, to generate the same clusters. Thus, the combinations with the minimal number of features that can be used to form the homogeneous group is selected as the final choice. The selected feature combination with the corresponding weightings are presented in Table 2. The homogeneous groups at each layer in four regions can then be generated using the selected features; the groups are presented in Figure 5.

According to the graphic displays in Figure 5 and the boxplots for comparison of uncertainty in the quantile estimates in Figure 6, several findings can be observed: 1) The algorithm at first layer serves its purpose. As shown in Figure 5, all of the stations from each final homogeneous group (Groups 1, 2 and 3) in Regions 1, 2 and 3 lie within the same climate region and its adjacent area. In Region 4, the limited number of rainfall stations in the climate region are insufficient to produce accurate estimates and stations from other climate regions are also included in Group 4 to improve the accuracy of the estimates. 2) The implementation of judging criterion is effective. During the searching process in Region 3, the formed group from the second layer is not admissible as the uncertainty in its quantile estimates is higher than that from the formed group generated from the first layer as shown in the second plot in Figure 6 Region 3. This forces the search process to skip the procedure in the second layer and step directly to the third layer. 3) The three-layer framework is effective. The uncertainty in the quantile estimates from the formed groups is decreasing after each layer even when the number of stations in these group is also decreasing. 4) The third layer in the proposed algorithm can be equally effective in the less urbanized area. In Regions 2 and 4 where most of the stations are located in the less urbanized area, only a small number of stations are removed from the formed group generated from the procedure at the second layer, and the final homogeneous region resulted from the procedure at third layer can still generate the most accurate quantile estimates among all the formed groups. 5) Within each homogeneous group, the gathered stations may be located far away from the target site for the following reasons: a) the geographic location is just one potential similarity indicators used in HGF in the second layer. The final formed group does not entirely relied on the geographic proximity. b) The Fuzzy C-mean clustering allows for the possibility of non-target site being the centre of the formed group since the generated clusters are heavily depending on the information from the input group. In the case where some stations locating away from the target site, these stations may still have the feature values that are similar to the cluster centre. c) Even under the circumstance that the selected stations are comparatively far away from the target station, all the stations in the formed homogeneous group are still located in the same climate region.

5.3 Homogeneous group modification for rainfall series in short durations

For the rainfall events that have durations shorter than 24h, the three layer algorithm is simplified to the procedure at the first layer, since information about the date of occurrence for short duration events is not available. The reasons to only apply the first layer of proposed methodology in the

HGF are: 1) the temporal step selected to extract feature values is 128-day, nearly a four-month period. Considering the extreme rainfall patterns before 2007, there is a great chance that the date of occurrence of the short and long extreme rainfall events happen in the same four-month period. 2) The first layer is used to gather stations that share similar extreme rainfall condition above the PBL. No matter the length of concerned duration, the pre-conditions for the occurrence of extreme rainfall events especially above the PBL are similar, even when the long- and short- duration events do not happen in the same time frame. Thus the approach of using the same feature values as the similarity indicators can be justified. 3) To factor the difference in the rainfall amounts among different durations into the HGF for the extreme rainfall events at short durations, the annual maximum rainfall series at short durations will be used to calculate the objective values in the new searching process.

Based on the assumption that the shorter duration events occurred in the same four-month time window as the 24h duration events, this simplified procedure can be conducted. The homogeneous groups for the rainfall events at shorter durations are shown in Figure 7, and the following findings can be observed:

1) Region 1: The homogeneous group for the rainfall events in all of shorter durations is the same and consists of 38 stations in one climate region shown as the red dots (Group 1) in Figure 7(a) and 7(b). For the 24 hour rainfall series, the homogeneous group obtained from the procedure at the first layer consists of 53 stations, and has the same level of uncertainty in the quantile estimates as that which resulted from this 38-station group. Following the principle of gathering as many stations as possible without increasing the uncertainty in the quantile estimate, the 53-station group is regarded as the better homogeneous group for the quantile estimates of 24h duration rainfall events. Due to the proximity to the ocean, most of extreme rainfall events in

this region are caused by stratiform rainfall, which can last for a long duration and increases the possibility that rainfall at all durations are caused by the same rainfall event and thus share a similar homogeneous region.

2) Region 2: The homogeneous groups for the rainfall events that have durations shorter than 1h share the same group which consists of 31 stations shown as the yellow dots (Group 2) in Figure 7(a). For the rainfall at all longer durations, the homogeneous region consists of 42 stations shown as the yellow dots (Group 2) in Figure 7(b). Region 2 is in the Prairies, where convective extreme rainfall events occur during shorter time windows. Extreme rainfall events at longer durations are likely caused by the stratiform rainfall events or the combination of convective and stratiform rainfall events.

3) Region 3: The same homogeneous group with 93 stations (shown as green dots of Group 3 in Figure 7(a) and 7(b)) for the rainfall events at all durations are identified and are located inside the Southeastern region. The main rainfall type in the Southeastern region is stratiform rainfall that can last for long durations.

4) Region 4: The homogeneous group allocation in this region is similar to the situation in Region 2. In summary, two homogeneous groups for the rainfall events are identified: 27-station group for the rainfall events with the durations less than 1h shown as the blue dots in Group 4 from Figure 7(a), 29-station group for the events with durations between 1h to 12h shown as the blue dots in Group 4 from Figure 7(b). For the homogeneous region for the 24h duration rainfall events, the group of 35 stations obtained from the procedure at the first layer has the same level of uncertainty as the 29-station group. This situation can be explained by the same reasons noted above for Region 2.

5.4 Comparison between proposed approach and traditional geographic approach

To prove the homogeneity of the formed homogeneous groups in the selected regions, the homogeneity (H) test which can be constructed by using L-CV, L-skew or L-kurtosis was used (Hosking & Wallis, 1997). If one of these three calculated H measures for the formed group is less than 2, the formed groups can be regarded as homogeneous. The HGFs for the AMS at 24-h duration in Region 1, 3 and 4 meet above requirement as their lowest positive H measures equal to 0.16, 1.80 and 1.18, respectively. For the homogeneous group in Region 2, all of the H measures are negative because of the positive correlation among the included sites (Castellarin, et al., 2008).

To demonstrate the effectiveness of the proposed methodology, the obtained homogeneous regions are compared with the pooling groups generated from the traditional geographic approach in which the geographic distance from the target stations is used as the similarity indicator in the process (Burn, 2014). To reduce the possible impacts from sample size on the uncertainty in the quantile estimates, the formed groups from the proposed method and the groups generated using the geographic approach have the same number of stations.

The boxplots for this comparison of uncertainty in the quantile estimates for the rainfall series at 24-h duration are shown in Figure 8. The median values of the ratios in all boxplots at all return periods are less than one, which indicates that the selected features at all three layers are more effective than the geographic distance as similarity indicators to form the homogeneous regions. In terms of the applicability of the first layer algorithm of proposed methodology for the rainfall events at shorter durations, the same procedure is conducted to obtain the boxplots for comparison, and the same conclusions are reached.

To further validate the proposed methodology, two additional comparisons are conducted using the AMS at 24-h duration: 1) the comparison of CI widths in the quantile estimates from using the at-site approach and the proposed regional methodology. Based on the results shown in Figure 9, the formed homogeneous groups at four target sites generate narrower CI widths than that from using at-site approach for the longer return periods, and generally perform equally or inferior for the 2-year return period in terms of the uncertainty in the estimates. 2) The comparison of root mean square error (RMSE) in the growth curves of the HGFs from using proposed method and the traditional geographic approach. In this study, the RMSEs are the average differences between at-site and regional growth curves at different return periods, and can be calculated by using the procedures introduced by Mostofi Zadeh, et al. (2019). Based on the results shown in Table 3, the HGFs from using the proposed methodology generally generate lower RMSEs than that from using traditional geographic approach at all return periods.

6. Conclusions

A HGF process based on the feature selection and weighting algorithm in a three-layer design is proposed to consider the spatial difference of similarity indicators at various regions and reduce the uncertainty for the quantile estimation in regional IDF estimation. Based on the stages of the extreme rainfall formation at different heights, the process extracted both atmospheric and geographic features to consider the possible alteration of rainfall patterns caused by climate change and urbanization at certain locations. During the process, the hybrid searching, which is the combination of Tabu Search, fuzzy c mean clustering and Lagrange multiplier is used to select the optimal feature combination and form the homogeneous group that generates the lowest uncertainty in the quantiles estimates.

The advantages of the proposed methodology can be summarized into the following points: our method 1) considers the possible changes in the homogeneous region for the rainfall events causing by climate change and urbanization. 2) Considers spatial difference of the similarity indicators at different regions for conducting more effective HGF. 3) Allows the possibility of letting the data itself to drive the group centre and best utilize the most information provided from the available weather stations. The proposed method is applied in four regions across Canada with different rainfall types. The results are evidently superior in comparison with the traditional geographic approach.

Acknowledgement

This work was supported by the Natural Science and Engineering Research Council (NSERC) Canadian FloodNet (Grant number: NETGP 451456). The authors are grateful to Dr. Nadav Peleg and two anonymous reviewers whose comments and suggestions contributed to the improvement of the manuscript.

References

Ahmad, N. H., Othman, I. R., & Deni, S. M. (2013). Hierarchical Cluster Approach for Regionalization of Peninsular Malaysia based on the Precipitation Amount. *Journal of Physics: Conference Series*, *423*, 012018. http://doi.org/10.1088/1742-6596/423/1/012018

Asong, Z. E., Khaliq, M. N., & Wheater, H. S. (2015). Regionalization of precipitation characteristics in the Canadian Prairie Provinces using large-scale atmospheric covariates and geophysical attributes. *Stochastic Environmental Research and Risk Assessment*, *29*(3), 875–892. http://doi.org/10.1007/s00477-014-0918-z

Berrisford, P., Dee, D. P., Poli, P., Brugge, R., Fielding, M., Fuentes, M., … Simmons, A. (2011). The ERA-Interim archive Version 2.0, (1), 23. Retrieved from https://www.ecmwf.int/node/8174

Bezdek, J. C., Ehrlich, R., & Full, W. (1984). FCM: The fuzzy c-means clustering algorithm. *Computers & Geosciences*, *10*(2–3), 191–203. http://doi.org/10.1016/0098-3004(84)90020-7

Borgelt, C. (2008). Feature weighting and feature selection in fuzzy clustering. *Fuzzy Systems, 2008. FUZZ-IEEE 2008.(IEEE World Congress on Computational Intelligence). IEEE International Conference On*, 838–844. http://doi.org/10.1109/FUZZY.2008.4630468

Bornstein, R., & Lin, Q. (2000). Urban heat islands and summertime convective thunderstorms in Atlanta: Three case studies. *Atmospheric Environment*, *34*(3), 507–516. http://doi.org/10.1016/S1352-2310(99)00374-X

Boyd, E. C. (2010). Estimating and Mapping the Direct Flood Fatality Rate for Flooding in Greater New Orleans Due To Hurricane Katrina. *Risk, Hazards & Crisis in Public Policy*, *1*(3), 87–110. http://doi.org/10.2202/1944-4079.1017

Burn, D. H. (1990). Evaluation of regional flood frequency analysis with a region of influence approach.

Water Resources Research, *26*(10), 2257–2265. http://doi.org/10.1029/WR026i010p02257

Burn, D. H. (2014). A framework for regional estimation of intensity-duration-frequency (IDF) curves. *Hydrological Processes*, *28*(14), 4209–4218. http://doi.org/10.1002/hyp.10231

Cai, W., Borlace, S., Lengaigne, M., Van Rensch, P., Collins, M., Vecchi, G., … Jin, F. F. (2014). Increasing frequency of extreme El Niño events due to greenhouse warming. *Nature Climate Change*, *4*(2), 111–116. http://doi.org/10.1038/nclimate2100

Castellarin, A., Burn, D. H., & Brath, A. (2008). Homogeneity testing: How homogeneous do heterogeneous cross-correlated regions seem? *Journal of Hydrology*, *360*(1–4), 67–76. http://doi.org/10.1016/j.jhydrol.2008.07.014

Collier, C. G. (2006). The impact of urban areas on weather. *Quarterly Journal of the Royal Meteorological Society*, *132*(614), 1–25. http://doi.org/10.1256/qj.05.199

Dalrymple, T. (1960). Flood Frequency Analyses. *Geological Survey Water-Supply Paper*, 1543–A.

Daubechies, I. (1990). The Wavelet Transform , Time-Frequency Localization and Signal Analysis. *IEEE Transactions on Information Theory*, *36*(5), 961–1005.

Diem, J. E., & Brown, D. P. (2003). Anthropogenic Impacts on Summer Precipitation in Central Arizona, U.S.A. *The Professional Geographer*, *55*(3), 343–355. http://doi.org/10.1111/0033-0124.5503011

Easterling, D. R. (1989). Regionalization of thunderstorm rainfall in the contiguous United States. *International Journal of Climatology*, *9*(6), 567–579. http://doi.org/10.1002/joc.3370090603

Gaál, L., & Kyselý, J. (2009). Comparison of region-of-influence methods for estimating high quantiles of precipitation in a dense dataset in the Czech Republic. *Hydrology and Earth System Sciences*, *13*(11), 2203–2219. http://doi.org/10.5194/hess-13-2203-2009

Gabriele, S., & Chiaravalloti, F. (2013). Searching regional rainfall homogeneity using atmospheric fields. *Advances in Water Resources*, *53*, 163–174. http://doi.org/10.1016/j.advwatres.2012.11.002

Glover, F. (1986). Future Paths for Integer Programming and Links to Artificial Intelligence. *Computers and Operations Research*, *13*(5), 533–549.

Glover, F. (1989). Tabu Search-Part I. *ORSA Journal on Computing*, *1*(3), 190–206.

Goswami, B. N., Venugopal, V., Sengupta, D., Madhusoodanan, M. S., & Xavier, P. K. (2006). Increasing Trend of Extreme Rain Events Over India in a Warming Environment. *Science*, *314*(5804), 1442–1445. http://doi.org/10.1126/science.1132027

Haddad, K., Johnson, F., Rahman, A., Green, J., & Kuczera, G. (2015). Comparing three methods to form regions for design rainfall statistics: Two case studies in Australia. *Journal of Hydrology*, *527*, 62–76. http://doi.org/10.1016/j.jhydrol.2015.04.043

Hamill, T., Bates, G., Whitaker, J., Murray, D., Fiorino, M., & Galarneau, T. (2013). A Description of the 2nd Generation NOAA Global Ensemble Reforecast Data Set. *NOAA Earth System Research Lab, Physical Sciences Division Bouder, Colorado, USA*, 10. Retrieved from https://www.esrl.noaa.gov/psd/forecasts/reforecast2/README.GEFS_Reforecast2.pdf

Hosking, J. R. M., & Wallis, J. R. (1997). *Regional Frequency Analysis: An approach based on L-moments.* Retrieved from https://books.google.com.pe/books?hl=es&lr=&id=gurAnfB4nvUC&oi=fnd&pg=PP1&dq=Regional+frequency+analysis+an+approach+based+on+l-moments&ots=7Re17uu4PZ&sig=cQloBXfu6O-1BS3wGAj_pUvSJYI#v=onepage&q&f=false

Houze Jr., R. a. (1997). Stratiform precipitation in the tropics: A meteorological paradox? *Bull. Amer. Meterol. Soc.*, *78*(10), 2179–2196. http://doi.org/10.1175/1520-

0477(1997)078<2179:SPIROC>2.0.CO;2

Houze, R. A. (1989). Observed structure of mesoscale convective systems and implications for large-scale heating. *Quarterly Journal of the Royal Meteorological Society*, *115*(487), 425–461. http://doi.org/10.1002/qj.49711548702

Jin, M., Shepherd, J. M., & Zheng, W. (2010). Urban Surface Temperature Reduction via the Urban Aerosol Direct Effect: A Remote Sensing and WRF Model Sensitivity Study. *Advances in Meteorology*, *2010*, 1–14. http://doi.org/10.1155/2010/681587

Kaufman, Y. J., & Koren, I. (2006). Smoke and pollution aerosol effect on cloud cover. *Science*, *313*(5787), 655–658.

Khazai, B., Bessel, T., Möhrle, S., Dittrich, A., Schröter, K., Mühr, B., … Trieselmann, W. (2013). June 2013 Flood in Central Europe - Focus Germany Report 1 – Update 2: Preconditions, Meteorology, Hydrology, *1*(June), 1–13.

Kumar, P., & Foufoula-georgiou, E. (1997). Wavelet analysis for geophysical applications. *Reviews of Geophysics*, *35*(4), 385–412.

Li, D., Bou-Zeid, E., Baeck, M. L., Jessup, S., & Smith, J. a. (2013). Modeling Land Surface Processes and Heavy Rainfall in Urban Environments: Sensitivity to Urban Surface Representations. *Journal of Hydrometeorology*, *14*(4), 1098–1118. http://doi.org/10.1175/JHM-D-12-0154.1

Loughner, C. P., Allen, D. J., Zhang, D. L., Pickering, K. E., Dickerson, R. R., & Landry, L. (2012). Roles of urban tree canopy and buildings in urban heat island effects: Parameterization and preliminary results. *Journal of Applied Meteorology and Climatology*, *51*(10), 1775–1793. http://doi.org/10.1175/JAMC-D-11-0228.1

Mallat, S. G. (1989). A theory for multiresolution signal decomposition: the wavelet representation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *11*(7), 674–693. http://doi.org/10.1109/34.192463

Marra, F., Morin, E., Peleg, N., Mei, Y., & Anagnostou, E. N. (2017). Intensity–duration–frequency curves from remote sensing rainfall estimates: comparing satellite and weather radar over the eastern Mediterranean. *Hydrology and Earth System Sciences*, *21*(5), 2389–2404. http://doi.org/10.5194/hess-21-2389-2017

Meukaleuni, C., Lenouo, A., & Monkam, D. (2016). Climatology of convective available potential energy (CAPE) in ERA-Interim reanalysis over West Africa. *Atmospheric Science Letters*, *17*(1), 65–70. http://doi.org/10.1002/asl.601

Miao, S., Chen, F., Li, Q., & Fan, S. (2011). Impacts of urban processes and urbanization on summer precipitation: A case study of heavy rainfall in Beijing on 1 August 2006. *Journal of Applied Meteorology and Climatology*, *50*(4), 806–825. http://doi.org/10.1175/2010JAMC2513.1

Moncrieff, M. W., & Miller, M. J. (1976). The dynamics and simulation of tropical cumulonimbus and squall lines. *Quarterly Journal of the Royal Meteorological Society*, *102*(432), 373–394. http://doi.org/10.1002/qj.49710243208

Mostofi Zadeh, S., Durocher, M., Burn, D. H., & Ashkar, F. (2019). Pooled flood frequency analysis: a comparison based on peaks-over-threshold and annual maximum series. *Hydrological Sciences Journal*, *64*(2), 121–136. http://doi.org/10.1080/02626667.2019.1577556

Murugavel, P., Pawar, S. D., & Gopalakrishnan, V. (2012). Trends of Convective Available Potential Energy over the Indian region and its effect on rainfall. *International Journal of Climatology*, *32*(9), 1362–1372. http://doi.org/10.1002/joc.2359

Quiroz, R., Yarlequé, C., Posadas, A., Mares, V., & Immerzeel, W. W. (2011). Improving daily rainfall estimation from NDVI using a wavelet transform. *Environmental Modelling & Software*, *26*(2), 201–209. http://doi.org/10.1016/j.envsoft.2010.07.006

Rosenfeld, D., Rosenfeld, D., Lohmann, U., Raga, G. B., Dowd, C. D. O., Kulmala, M., … Andreae, M. O. (2014). Flood or Drought : How Do Aerosols Affect Precipitation ? *Science*, *1309*(2008), 1309–1314. http://doi.org/10.1126/science.1160606

Roth, M. (2000). Review of atmospheric turbulence over cities. *Quarterly Journal of the Royal Meteorological Society*, *126*(564), 941–990. http://doi.org/10.1002/qj.49712656409

Roth, M., & Oke, T. R. (1993). Turbulent transfer relationships over an urban surface. I. Spectral characteristics. *Quarterly Journal of the Royal Meteorological Society*, *119*(513), 1071–1104. http://doi.org/10.1002/qj.49711951311

Sait, S. M., & Youssef, H. (1999). *Iterative Computer Algorithms with Applications in Engineering: Solving Combinatorial Optimization Problems* (1st ed.). Los Alamitos, CA, USA: IEEE Computer Society Press.

Sandink, D. (2013). *Urban flooding in Canada: Lot-side risk reduction through voluntary retrofit programs , code interpretation and by-laws*.

Satyanarayana, P., & Srinivas, V. V. (2008). Regional frequency analysis of precipitation using large-scale atmospheric variables. *Journal of Geophysical Research*, *113*(D24), 1–16. http://doi.org/10.1029/2008JD010412

Satyanarayana, P., & Srinivas, V. V. (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. http://doi.org/10.1016/j.jhydrol.2011.05.044

Shepherd, J. M. (2005). A review of current investigations of urban-induced rainfall and recommendations for the future. *Earth Interactions*, *9*(12). http://doi.org/10.1175/EI156.1

Tahir, M. A., Bouridane, A., & Kurugollu, F. (2007). Simultaneous feature selection and feature weighting using Hybrid Tabu Search/K-nearest neighbor classifier. *Pattern Recognition Letters*, *28*(4), 438–446. http://doi.org/10.1016/j.patrec.2006.08.016

Wang, W., & Zhang, Y. (2007). On fuzzy cluster validity indices. *Fuzzy Sets and Systems*, *158*(19), 2095–2117. http://doi.org/10.1016/j.fss.2007.03.004

Warren, F. J., & Lemmen, D. S. (2014). Synthesis. *Canada in a Changing Climate: Sector Perspectives on Impacts and Adaptation*, 1–18. Retrieved from http://www.nrcan.gc.ca/environment/resources/publications/impacts-adaptation/reports/assessments/2014/16309

Xie, X. L., & Beni, G. (1991). A validity measure for fuzzy clustering. *IEEE Transactions on Pattern Analysis and Machine Intelligence*. http://doi.org/10.1109/34.85677

Xu, Y. M., Wang, C. D., & Lai, J. H. (2014). Weighted Multi-view Clustering with Feature Selection. *Pattern Recognition*, *53*, 25–35. http://doi.org/10.1016/j.patcog.2015.12.007

Yang, T., Shao, Q., Hao, Z.-C., Chen, X., Zhang, Z., Xu, C.-Y., & Sun, L. (2010). Regional frequency analysis and spatio-temporal pattern characterization of rainfall extremes in the Pearl River Basin, China. *Journal of Hydrology*, *380*(3–4), 386–405. http://doi.org/10.1016/j.jhydrol.2009.11.013

Zhang, H., & Sun, G. (2002). Feature selection using tabu search method. *Pattern Recognition*, *35*(3), 701–711. http://doi.org/10.1016/S0031-3203(01)00046-2

Zhang, X., Vincent, L. A., Hogg, W. D., & Niitsoo, A. (2000). Temperature and precipitation trends in Canada during the 20th century. *Atmosphere - Ocean*, *38*(3), 395–429.

http://doi.org/10.1080/07055900.2000.9649654

Zong, Y., & Chen, X. (2000). The 1998 flood on the Yangtze, China. *Natural Hazards*, *22*(2), 165–184. http://doi.org/10.1023/A:1008119805106

Figures

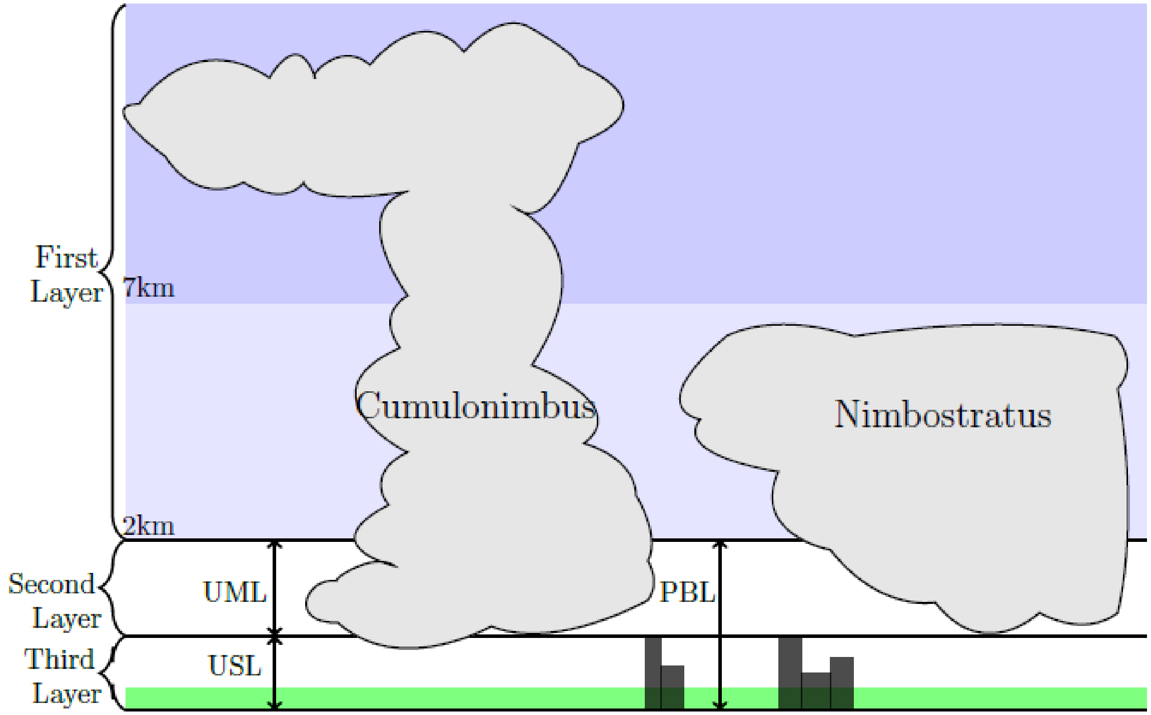
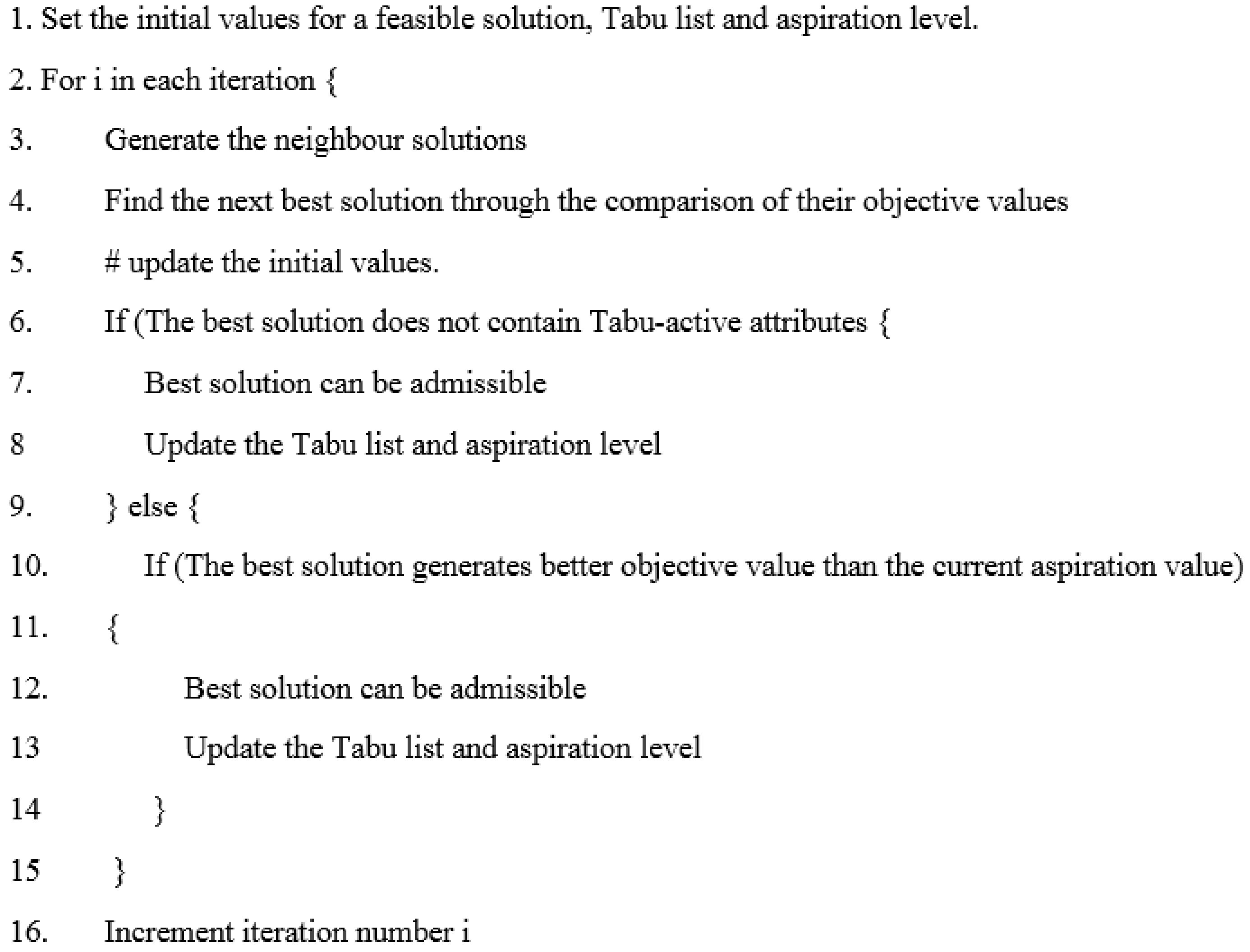


Figure 1 The vertical structure of the three-layer design: Planetary Boundary Layer (PBL), Urban Mixing Layer (UML), Urban Surface Layer (USL). In the figure, the green area indicate the height of vegetation cover, the dark grey boxes indicate buildings (Adapted from Figure 1 in Shepherd, 2005)



```
Set the initial values for a feasible solution, Tabu list and aspiration level.
For i in each iteration {
    Generate the neighbour solutions
    Find the next best solution through the comparison of their objective values
    # update the initial values.
    If (The best solution does not contain Tabu-active attributes {
        Best solution can be admissible
        Update the Tabu list and aspiration level
    } else {
        If (The best solution generates better objective value than the current aspiration value)
    {
            Best solution can be admissible
            Update the Tabu list and aspiration level
        }
     }
    Increment iteration number i
```

Figure 2 Algorithmic description of Tabu Search (Adapted from Figure 3 in Tahir et al, 2007)

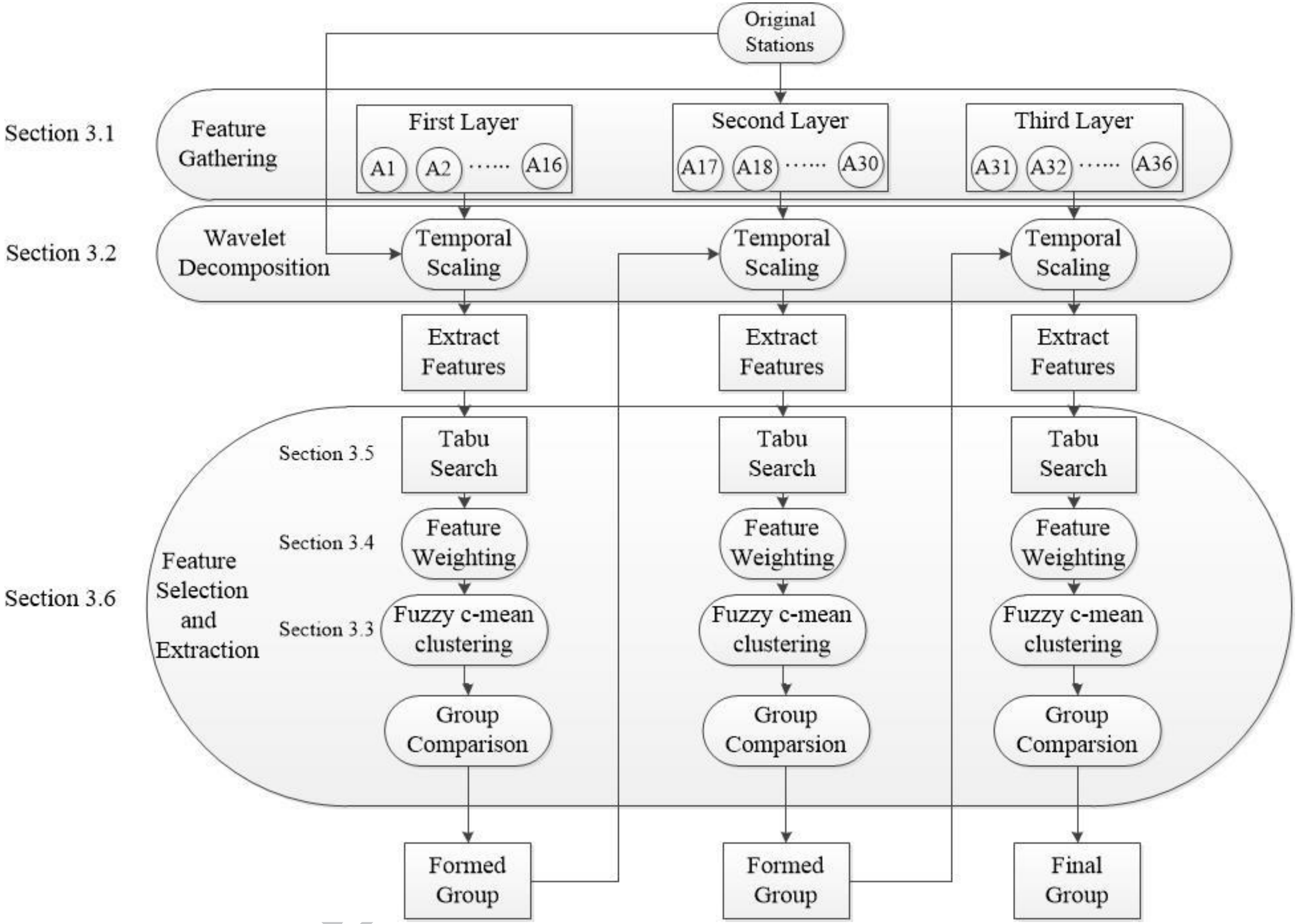


Figure 3 The flowchart of Automatic Feature Selection and Weighting algorithm

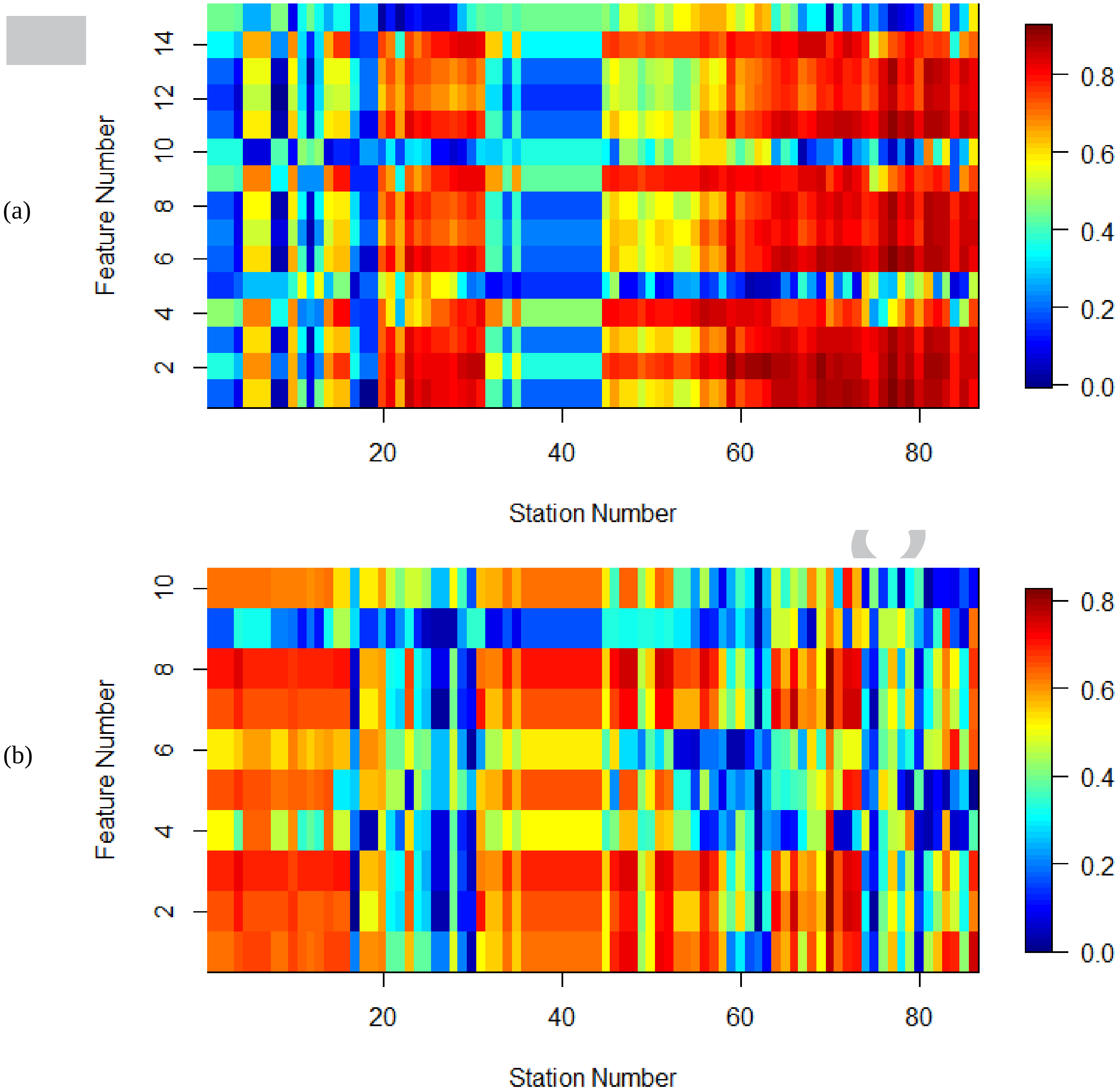


Figure 4 Graphic display of the correlation coefficient values between the most responsive features and the less responsive ones in the first (a) and second (b) layers at level 10 and 9 decomposition from the stations in Region 1. The column number indicates the stations in Region 1, and the row number indicates the potential features in first and second layers. Each grid represents correlation value between the most responsive feature and one of the less responsive ones at one station.

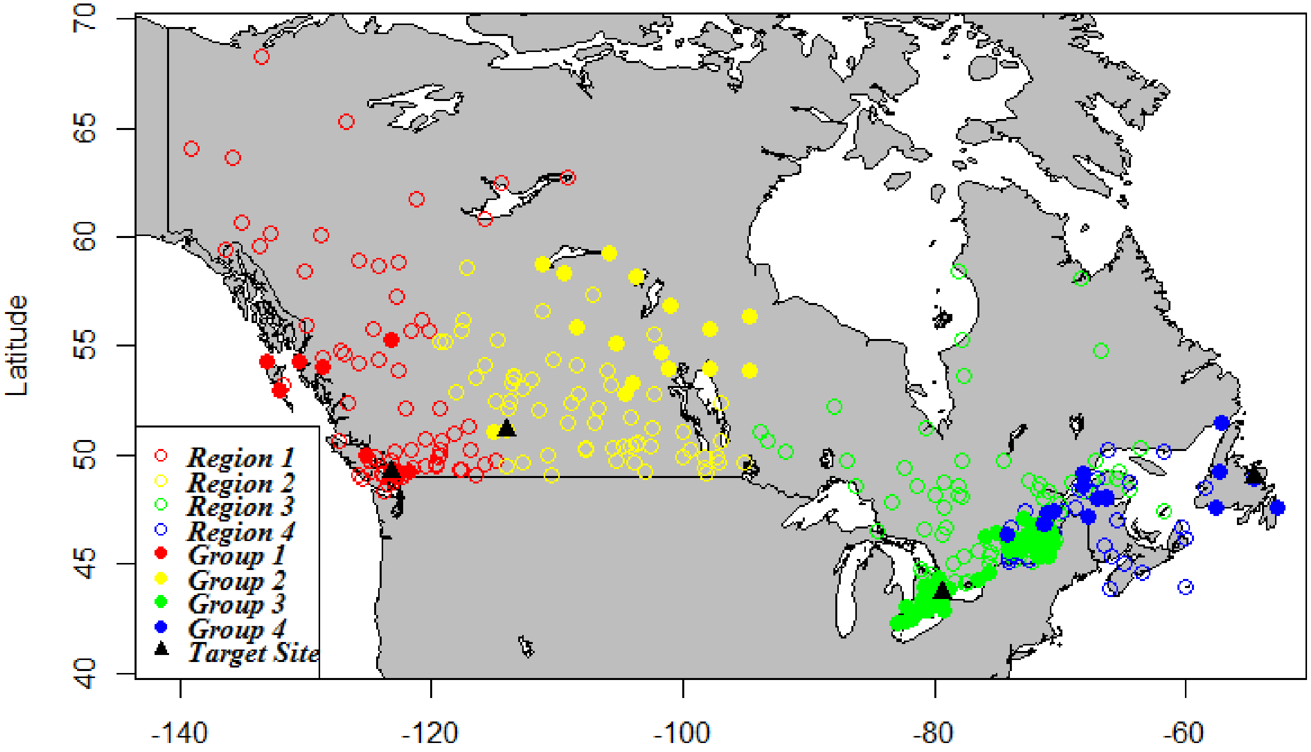
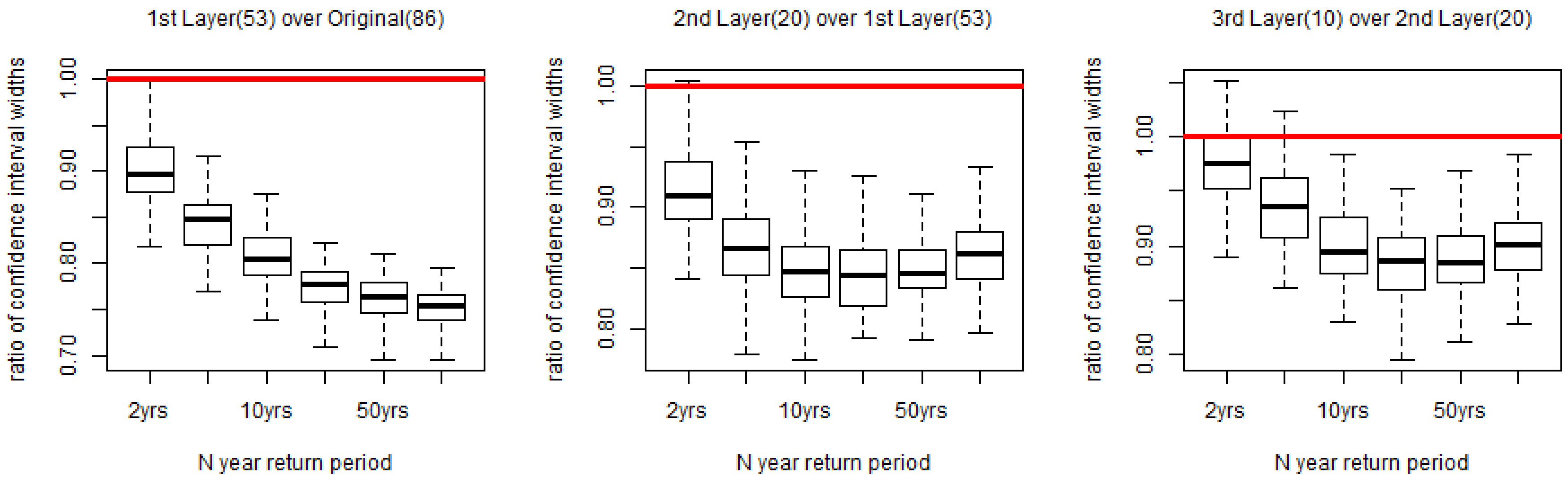
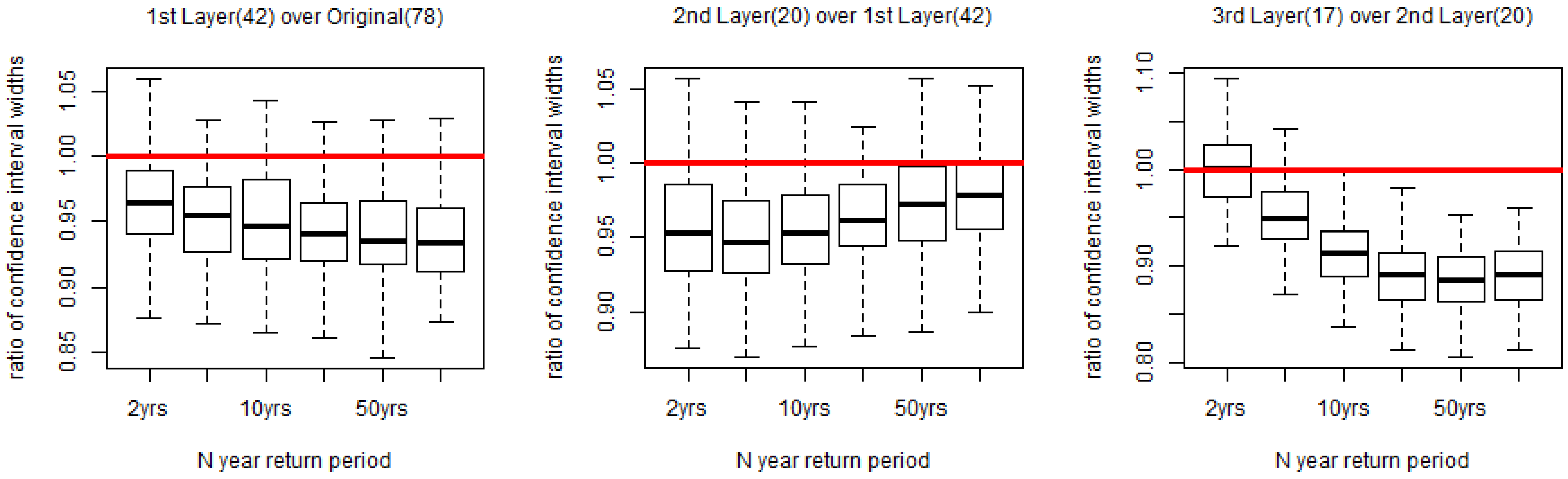


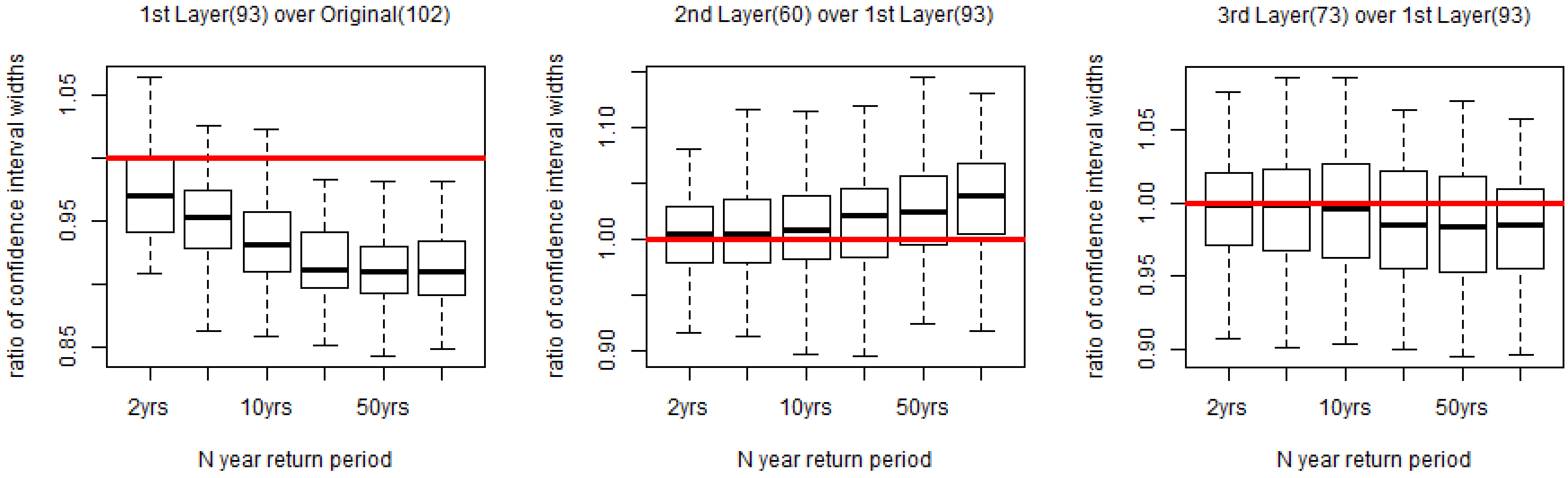
Figure 5 Graphic display of homogeneous groups and original input groups for extreme rainfall events at 24h duration in Region 1, 2, 3 and 4. The circle dots in red, yellow, green and blue in the graph indicate the original input stations, while the solid dots at four colors represent the homogeneous groups in four regions. The black triangles represent the target sites in four different regions.



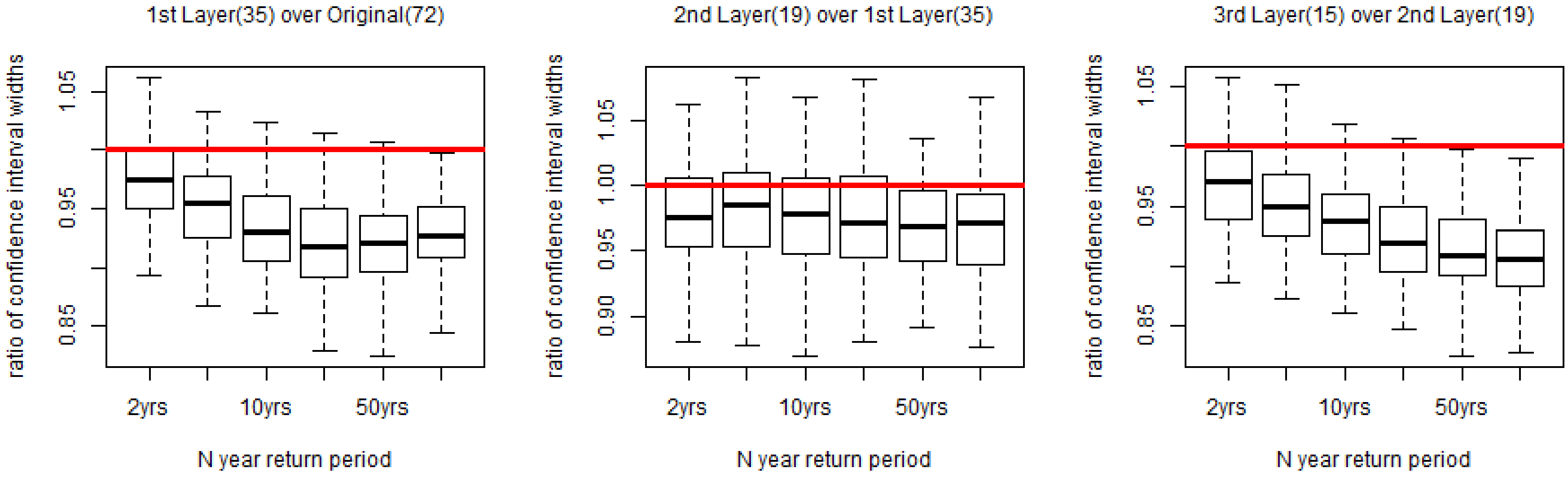
Region 1



Region 2



Region 3



Region 4

Figure 6 Boxplots of the ratio of CI widths among different formed groups in Regions 1, 2, 3 and 4 using the AMS at 24-h duration. The boxplots compare the homogeneous group from procedures at higher layer to that from previous lower layer, and numbers in the brackets indicate the number of stations in the formed homogeneous groups. The red line in every boxplot indicates where the value equals to 1.

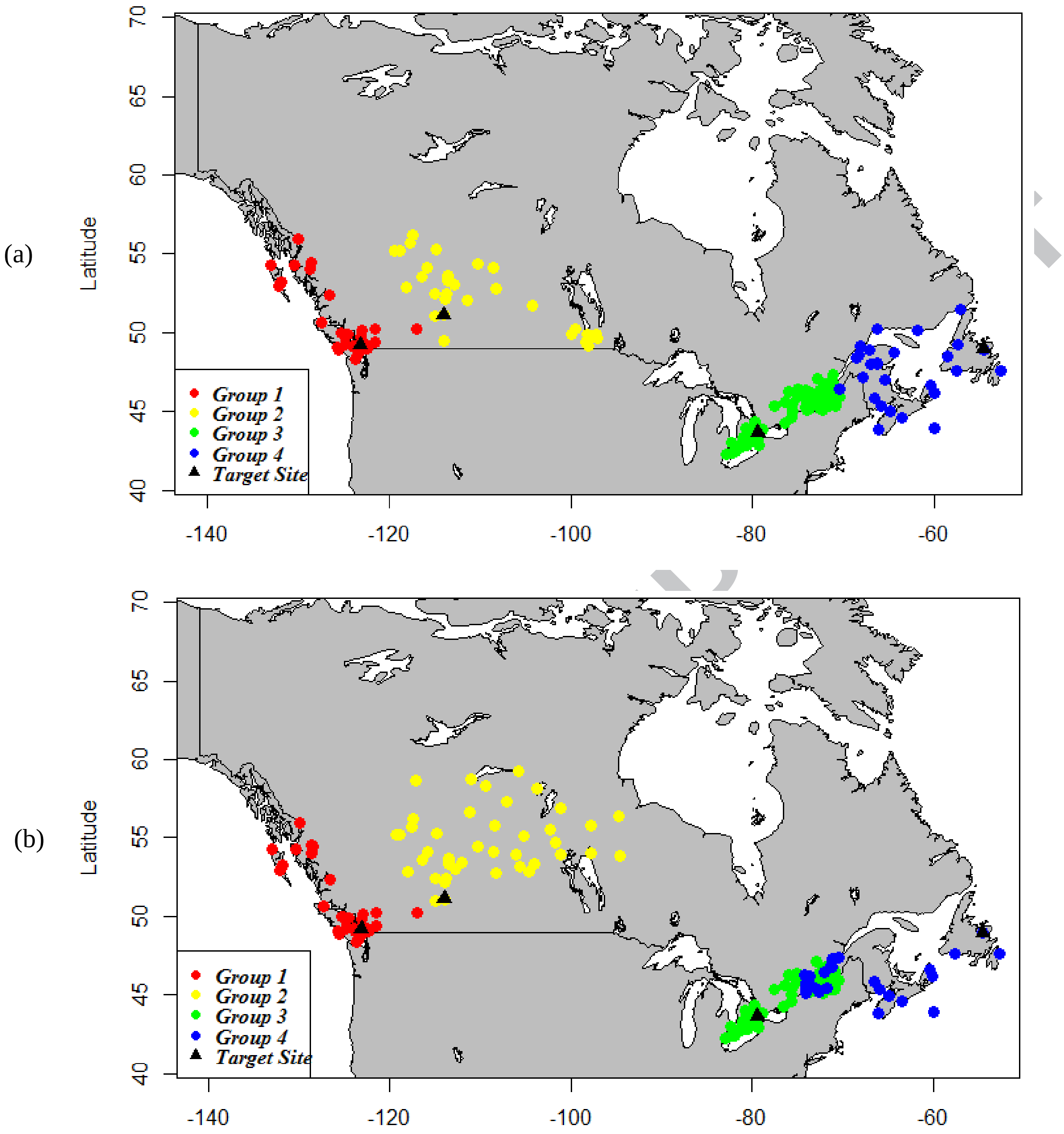
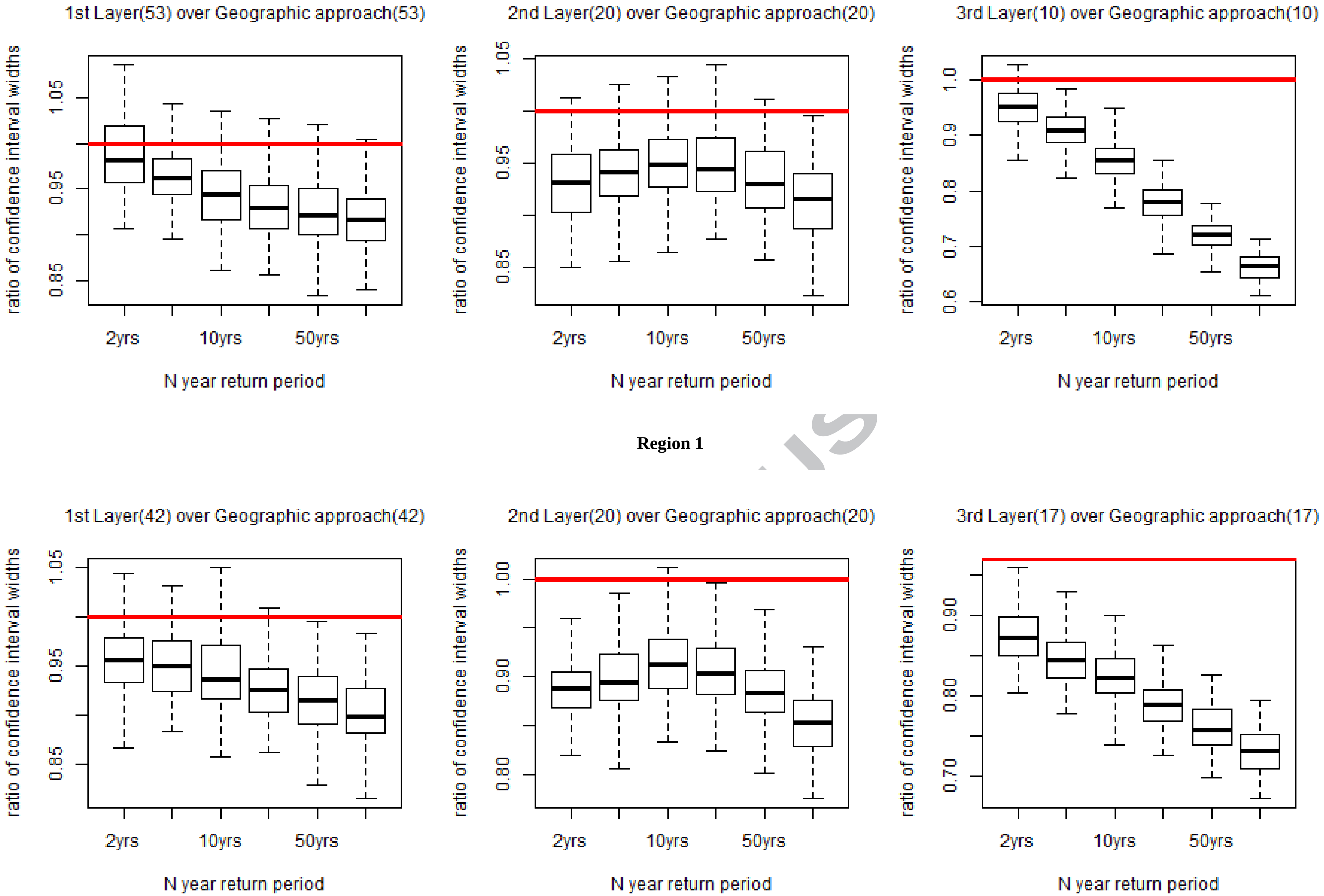
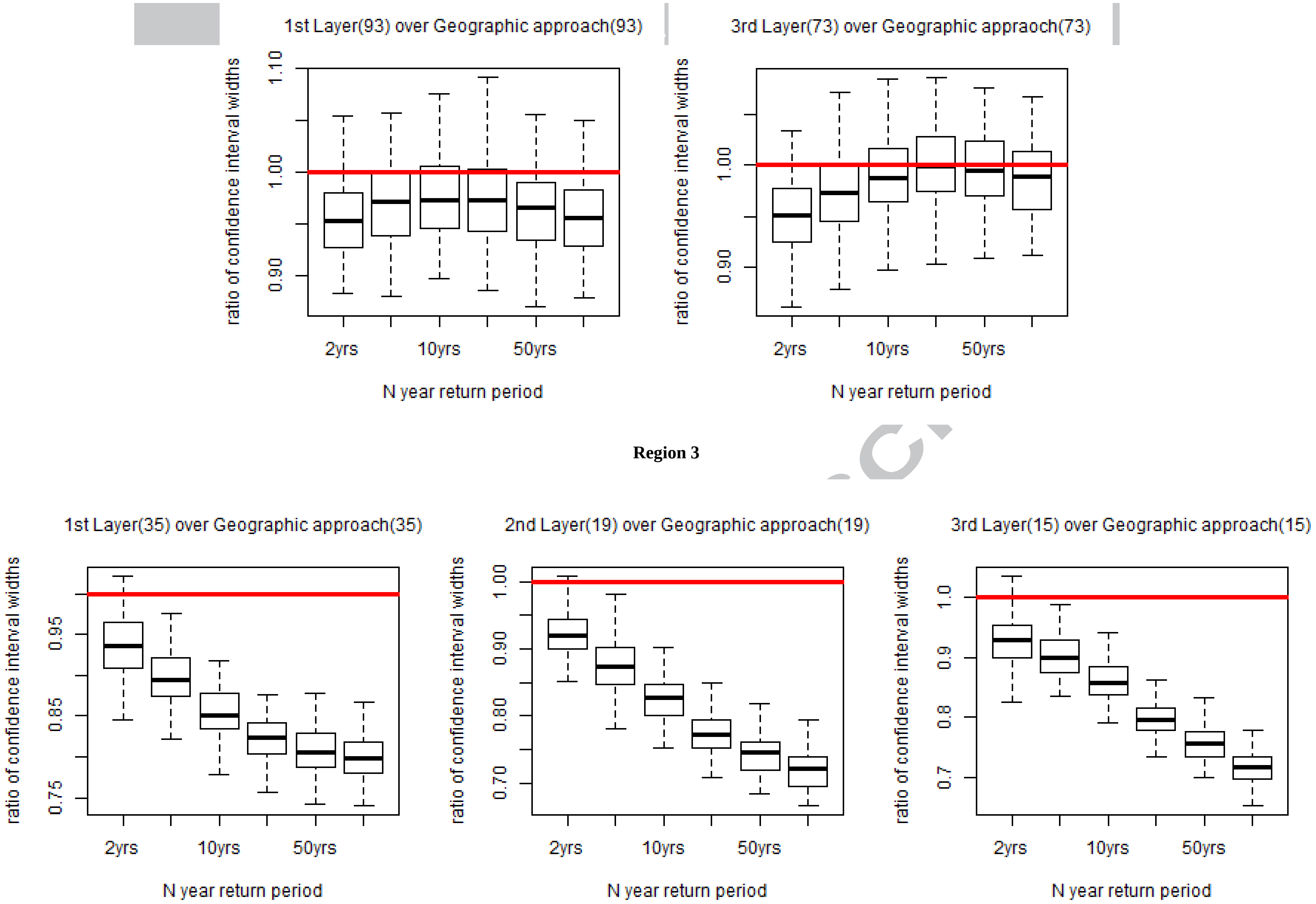


Figure 7 Graphic display of homogeneous groups for extreme rainfall events at duration 5 min, 15 min, and 30 min in (a) and at duration 1h, 2h, 6h and 12h in (b) for Region 1, 2, 3 and 4 The solid dots in red, yellow, green and blue represent the homogeneous groups in four regions. The black triangles represent the target sites in four different regions.



Region 1

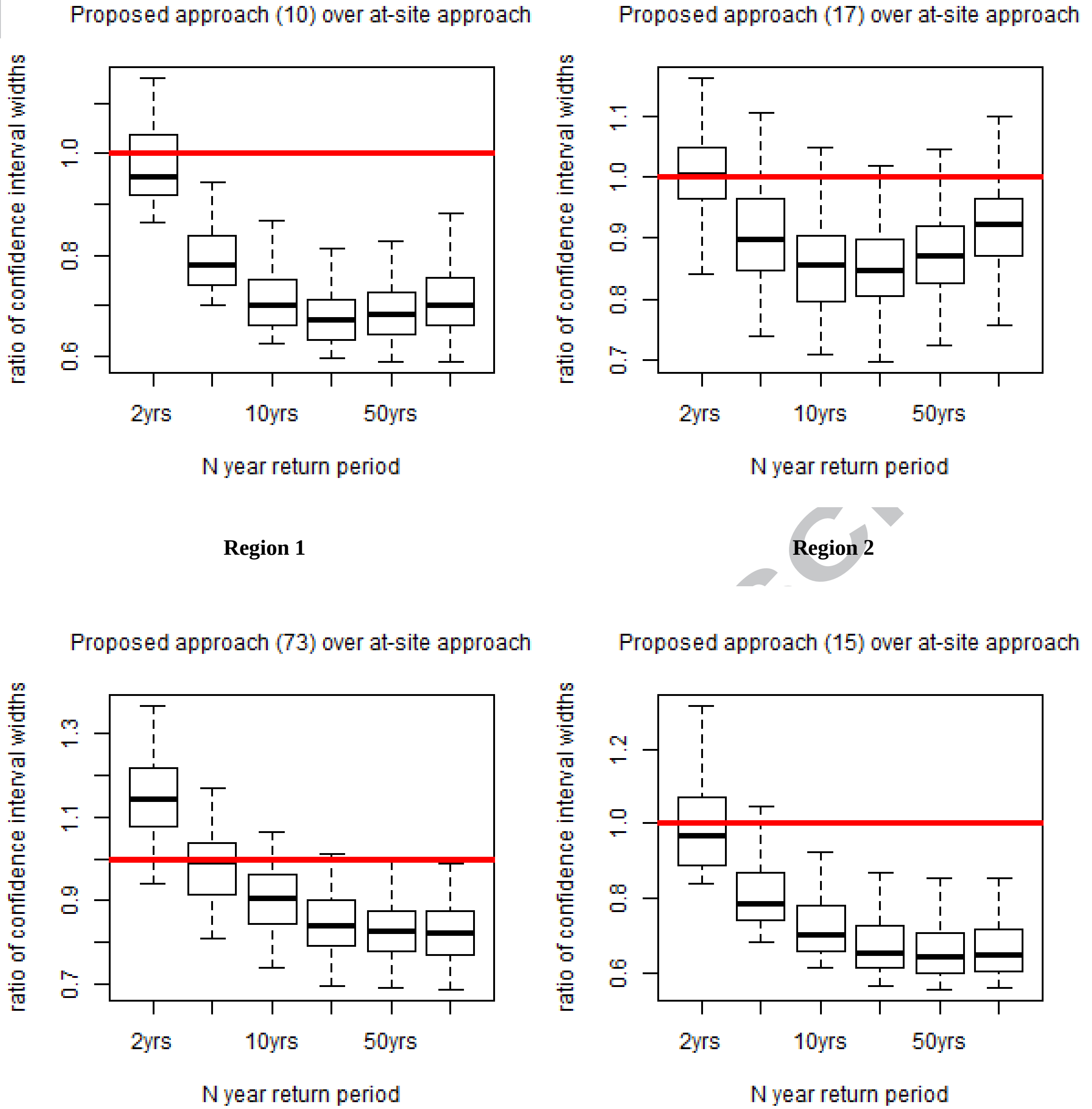
Region 2



Region 3

Region 4

Figure 8 Boxplots of the ratio of CI widths between formed groups obtained from using the proposed approach and the geographic approach in Regions 1, 2, 3 and 4 using the AMS at 24-h duration. The boxplots are conducted at all three layers in which the proposed homogeneous groups are compared with the groups that have the same number of stations formed using geographic approach. The numbers in the brackets indicate the number of stations in the formed homogeneous groups. The red line in every boxplot indicates where the value equals to 1.



Region 1

Region 2

Region 3

Region 4

Figure 9 Boxplots of the ratio of CI widths between proposed approach and at-site approach for four target sites in the selected regions using the AMS at 24-h duration. Each target site contains more than 50 data points.

Tables

Table 1 Potential feature dataset for each layer at the three layer design

Target layer	Potential features
First layer	Air temperature (Air), Geopotential height (Geo), Specific humidity (Sphu), U-component (Uwind) and V- component (Vwind) of the wind velocity (at the 300hPa, 500hPa and 700hPa pressure level), Convective Available Potential Energy (CAPE) and Q vector Divergence (QD)
Second layer	Air temperature (Air), Geopotential height (Geo), Specific humidity (Sphu), U-component (Uwind) and V- component (Vwind) of the wind velocity (at the 850hPa and 925hPa pressure level), Vertical Integral of Divergence of Moisture Flux (VIDWV), Latitude, Longitude, Elevation
Third layer	Urban Surface Sensible Heat Flux (SHTFL), Urban surface Latent Heat Flux (LHTFL), Photosynthetically Active Radiation index (PAR), Surface Net Solar Radiation (SNSR), Surface Net Thermal Radiation (SNTR), Surface Roughness (SR)

Table 2 Feature selection and weighting results from Three-layer search algorithm

(a) results from the procedure at first layer

		air300	air500	air700	geo300	geo500	geo700	vw300	vw500	vw700	uw300	uw500	uw700	sphu300	sphu500	sphu700	CAPE
Region 1	Selection	0	0	1	0	1	0	0	0	0	1	0	1	0	1	0	0
	Weightings			0.243		0.239					0.167		0.166		0.184		
Region 2	Selection	1	0	0	1	0	0	0	0	0	0	0	1	1	1	0	0
	Weightings	0.242			0.236								0.166	0.177	0.178		
Region 3	Selection	0	0	0	0	0	1	0	0	0	0	1	0	0	0	1	1
	Weightings						0.308					0.233				0.245	0.214
Region 4	Selection	0	0	0	1	1	1	0	0	0	0	1	0	0	1	1	0
	Weightings				0.190	0.188	0.185					0.140			0.148	0.149	

(b) results from the procedure at second layer

		air850	air925	geo850	geo925	vw850	vw925	uw850	uw925	sphu850	sphu925	VIMF	Latitude	Longitude	Elevation
Region 1	Selection	1	0	1	0	0	1	0	0	1	1	0	0	0	0
	Weightings	0.236		0.235			0.166			0.181	0.181				
Region 2	Selection	1	0	1	0	0	0	0	1	0	0	0	0	0	0
	Weightings	0.366		0.370					0.264						
Region 3	Selection														
	Weightings														
Region 4	Selection	0	0	0	0	0	0	0	0	1	0	0	0	0	0
	Weightings									1					

(c) results from the procedure at third layer

		LHTFL	SHTFL	PRS	SSR	SNR	SRH
Region 1	Selection	0	0	0	1	0	0
	Weightings				1		
Region 2	Selection	1	1	1	1	0	0
	Weightings	0.239	0.248	0.259	0.254		
Region 3	Selection	1	0	0	0	0	1
	Weightings	0.510					0.490
Region 4	Selection	0	0	0	1	0	1
	Weightings				0.520		0.480

Table 3 Summary of RMSEs from using different HGF approaches with the AMS at 24-h duration

	Return Period						
Testing Regions	**Approach**	**2-year**	**5-year**	**10-year**	**25-year**	**50-year**	**100-year**
Region 1	Proposed method	0.016	0.012	0.013	0.030	0.047	0.065
	Traditional geographic method	0.014	0.018	0.036	0.063	0.084	0.107
Region 2	Proposed method	0.010	0.006	0.011	0.016	0.019	0.021
	Traditional geographic method	0.017	0.027	0.050	0.078	0.099	0.120

Region 3	Proposed method	0.009	0.020	0.030	0.042	0.051	0.059
	Traditional geographic method	0.016	0.019	0.039	0.067	0.089	0.113
Region 4	Proposed method	0.017	0.019	0.031	0.051	0.067	0.085
	Traditional geographic method	0.011	0.020	0.037	0.055	0.068	0.081