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

Zhe(Emma) Yang and Donald H. Burn

Department of Civil & Environmental Engineering, University of Waterloo, Waterloo, Ontario, Canada

Abstract

The IDF curve has been widely accepted as an effective tool to provide the essential hydrological information for urban planning. However, the impacts from recent climate change and urbanization have caused the traditional IDF estimates to become increasingly inaccurate. Under the non-stationary environment, an approach for automatic feature selection and weighting for the homogeneous group formation at a specific region is proposed in this research to improve the current IDF estimation. According to the vertical structure of the planetary boundary, the time series of multiple features in three successive sub-layers will be selected to consider the impacts of urbanization and climate change on the rainfall climatology. Then, the hybrid searching algorithm of Tabu search and supervised clustering will be applied sequentially at different layers to obtain the optimal subsets of attributes and their corresponding weights. The three-layer hierarchical searching is designed to save computational time and accommodate the need of separating possible stationary and nonstationary features during clustering. The results demonstrate the effectiveness of the approach for extreme rainfall quantile estimation in the catchments under study. This approach fills the gap of including the urbanization impacts in the pooling group formation; furthermore, it challenges the traditional assumption that the same set of features can be equally effective in generating the optimal homogeneous group for regions with different geographic and meteorological characteristics.

Objectives

- ❖ consider the impacts from climate change and urbanization in the homogeneous group formation by incorporating the three layer design with time series datasets as the inputs.
- ❖ Automatic feature selection and Weighting for the formation of homogeneous group at specific region.

Methodology

The three-layer design of the PBL

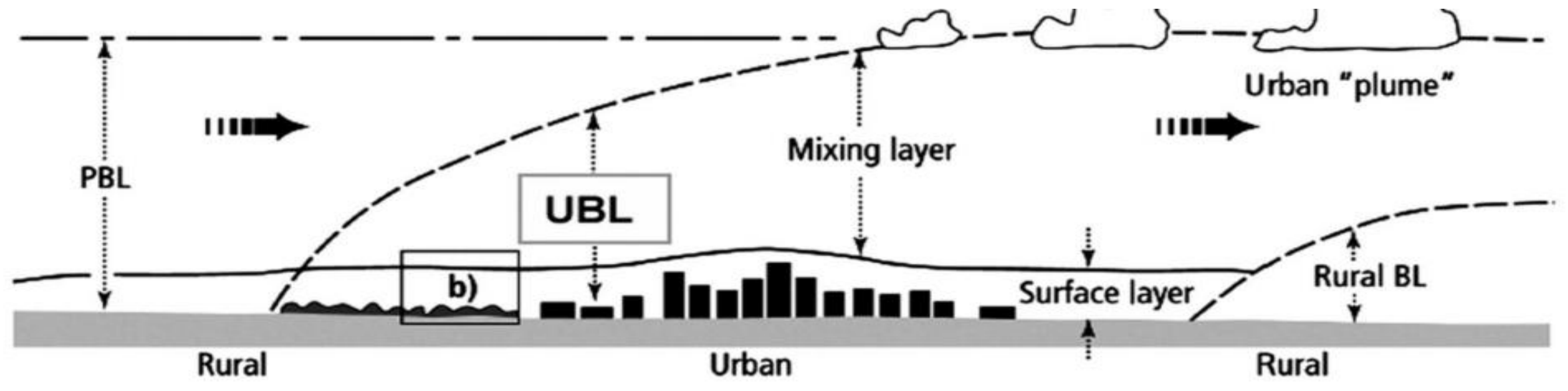


Figure 1: the vertical structure of the Planetary Boundary Layer(Collier, 2006)

Table 1: Potential feature dataset

The target layer	Potential features
The PBL(Planetary Boundary Layer)	Air temperature, Geopotential height, Specific humidity, U-component and V- component of the wind velocity(at the 300hPa, 500hPa and 700hPa pressure level), and Convective Available Potential Energy (CAPE)
The UML(Urban Mixed Layer)	Air temperature, Geopotential height, Specific humidity, U-component and V- component of the wind velocity (at the 850hPa and 925hPa pressure level), and Vertical integral of water vapor(VIWV).
The USL(Urban Surface Layer)	The urban surface sensible heat flux (SHTFL), the urban surface latent heat flux (LHTFL), The normalized difference vegetation index(NDVI) and Leaf area index (LAI).

The three layer design is proposed to incorporate the climate change impact on homogeneous group formation through using time series data instead of the site characteristics, and consider the possible urban heat effects and urban canopy effects on the local rainfall climatology(Li, Bou-Zeid, Baeck, Jessup, & Smith, 2013).

Temporal scaling of the time series

To increase the clustering accuracy, the temporal scales of the input features at first two layers are obtained through the correlation analysis at different levels of the wavelet decompositions(Brunsell & Young, 2008).

$$WTS(\lambda,\tau)=\int_{-\infty}^{\infty} S(t)\Psi_{\lambda,\tau}(t)dt \qquad \psi_{\lambda,\tau}(t)=\frac{1}{\sqrt{\lambda}}\psi\left(\frac{t-\tau}{\lambda}\right)$$

At the first layer, the correlation is established between CAPE and the remaining factors. CAPE, which can be used to describes convective rainfall that caused by the large buyout of accent energy, is the ideal indicator for the type of the rainfall(Gabriele & Chiaravalloti, 2013). Thus the obtained values at generated temporal scales will be effective in the clustering of different rainfall types. At the second layer, the selected features have great influences on the potential rainfall amount of each rainfall event. Thus the correlation is established between the Vertical integral of water vapor and other selected features.

The Hybrid Searching Algorithm

❖ Semi-supervised fuzzy c-means cluster

The clustering is conducted to divide the original group into two clusters. The traditional fuzzy cluster was modified by adding two constraints, one is to narrow the distance between the target site and its belonging center, the second is to enlarge the target site memberships' difference. So the objective function can be described as:

$$J_m=\sum_{i=1}^{N}\sum_{j=1}^{C}u_{ij}^{m}\left\|x_i-c_j\right\|^2+\left(dist\left(x_{t\arg et},c_{t\arg et}\right)-abs\left(u_{tagret2}-u_{t\arg et1}\right)\right)$$

❖ Lagrange multiplier

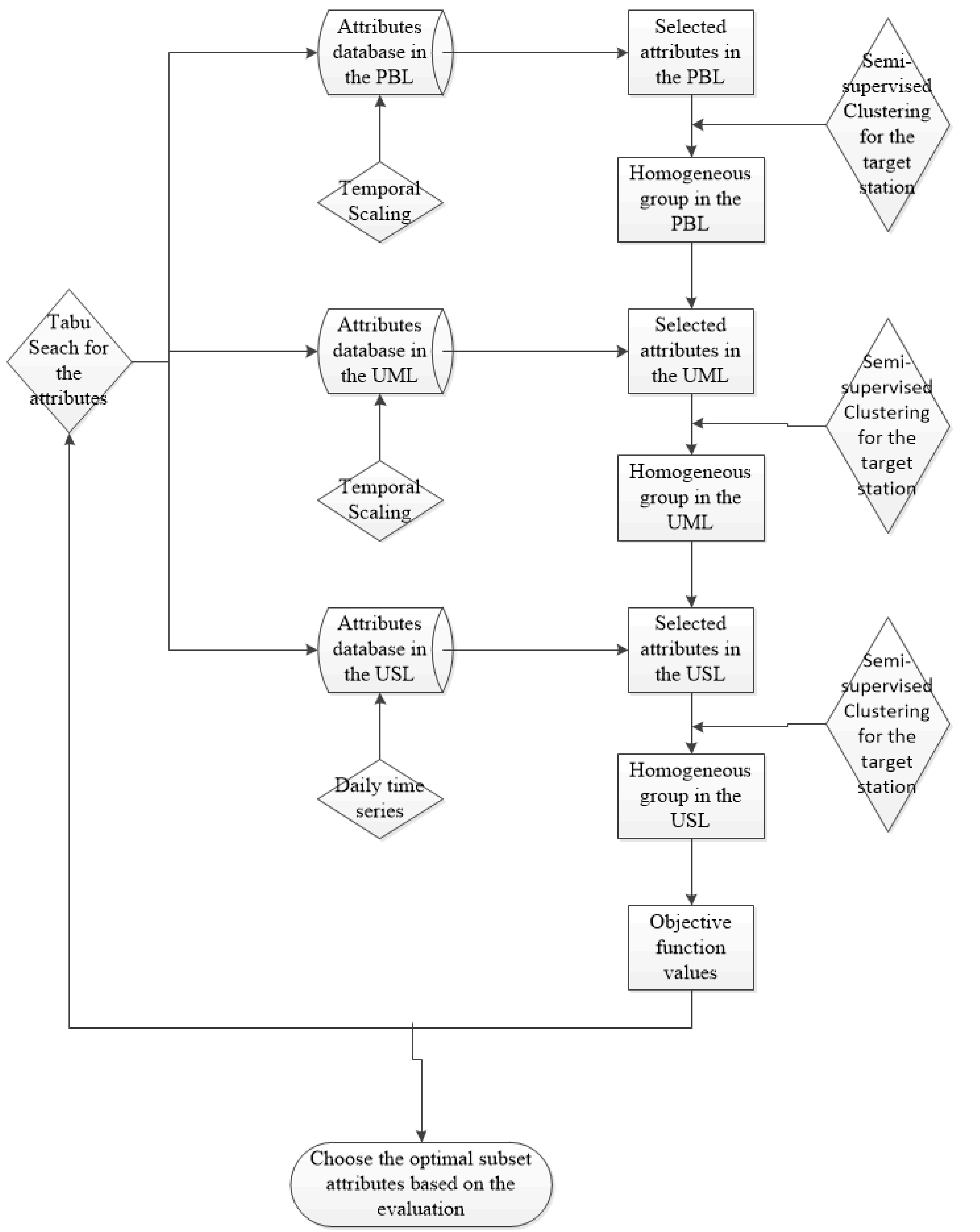
After the initial cluster was obtained, Lagrange multiplier will be applied to obtain the initial set of the feature weighting, then a loop is introduce start multiply the obtained weight in the clustering stage, proceed with this procedure until the nearly constant weightings have reached(Xu, Wang, & Lai, 2014).

$$L(\varepsilon,\lambda)=\varepsilon(w)+\lambda\left(\sum_{V=1}^{V}w_V-1\right) \Rightarrow \omega_V=\frac{1}{\sum_{V'=1}^{V}\left(\frac{D_V}{D_{V'}}\right)^{1/(p-1)}},\quad p>1.$$

❖ Tabu search

Tabu search has been regarded as one of the most effective search methods in the features selection approach(Zhang & Sun, 2002). It can be viewed as the iterative algorithm that visits different parts of the search space repeatedly, which includes the following stages: Initialize the settings, generating the initial neighborhood and searching the local optimal for the objective function, move the searching space into the new neighborhood and compare the new local optimal with the previous one, and then terminate the searching if the criteria had met(Zhang & Sun, 2002). The objective function in the tabu search is the confidence interval width of the regional daily rainfall analysis.

Framework of the Searching Algorithm


Tabu Search for the attributes
Attributes database in the PBL
Temporal Scaling
Selected attributes in the PBL
Semi-supervised Clustering for the target station
Homogeneous group in the PBL
Attributes database in the UML
Temporal Scaling
Selected attributes in the UML
Semi-supervised Clustering for the target station
Homogeneous group in the UML
Attributes database in the USL
Daily time series
Selected attributes in the USL
Semi-supervised Clustering for the target station
Homogeneous group in the USL
Objective function values
Choose the optimal subset attributes based on the evaluation

Case Study

❖ Study area

This application is conducted based on the 63 IDF stations in the Ontario province. The target station is the City of Toronto.

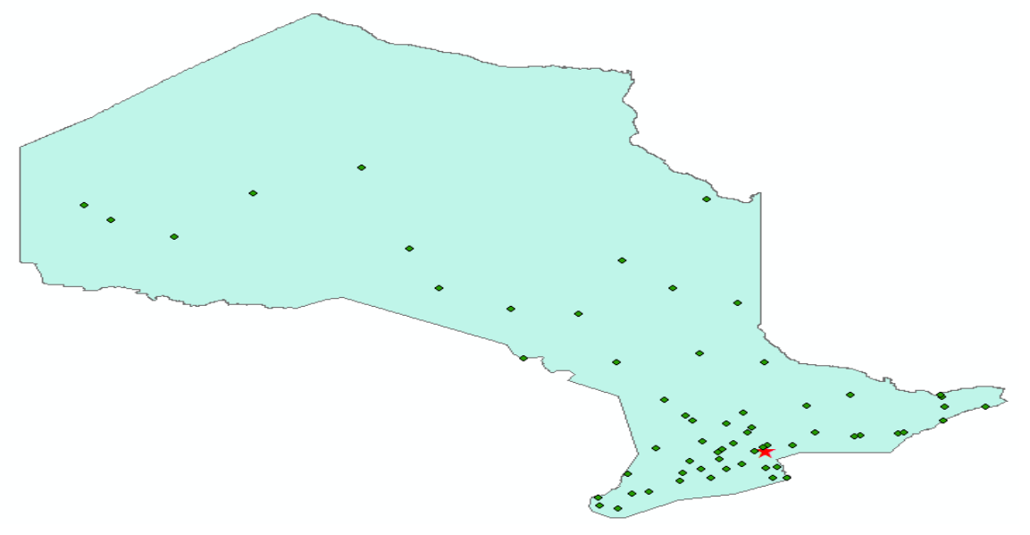


Figure 2. Location map showing the 63 IDF sites

❖ Dataset

Spatial resolution

1)the feature values at the PBL are the ensemble mean data obtained from the 2^{nd-}generation NOAA Global Ensemble Forecast System Reforecast (GEFS/R) at the resolution of 1 degree.

2) In the UML, other than the VIWV, which were obtained from the ERA-Interim Database, the remaining features are obtained from the GEFS/R Database. All of the features are at 1 degree resolution.

3)The features at USL are obtained from GEFS/R and MODIS at the 0.5 degree resolution.

Temporal steps and range

1) Other than the MODIS datasets, all the input features are 6-hour time series.
2) In the MODIS Datasets, NDVI is 16-day time series and LAI is 4-day time series.
3) all of the time series are obtained from 2000 to 2007

IDF datasets : 63 IDF datasets are obtained from Environment Canada

❖ Feature selection and weighting result

To reduce the computation time, the Hybrid Searching Algorithm will be applied sequentially at each layer with the same objective function.

1) Features in the PBL: the temporal step is 64 days.
the homogeneous group of 50 stations is obtained by applying the hybrid searching algorithm in this layer.

Table2 , feature selection and weighting results in PBL

Feature	air300	air500	air700	geo300	geo500	geo700	vw300	vw500	vw700	uw300	uw500	uw700	sphu300	sphu500	sphu700	CAPE
selection	0	1	1	1	1	1	1	1	1	0	0	1	1	0	1	1
weighting		0.1420	0.1427	0.1405	0.1383	0.1354	0.0568	0.0093	0.0108			0.0084	0.0830		0.0846	0.0483

2) Features in the UML: the temporal step is 8 days.
the homogeneous group of 31 stations is obtained by applying the hybrid searching algorithm in this layer.

Table3 , feature selection and weighting results in UML

Feature	air850	air925	geo850	geo925	vw850	vw925	uw850	uw925	sphu850	sphu925	VIWV
Selection	1	1	1	0	1	0	0	0	0	0	1
weighting	0.2562	0.2476	0.2339		0.1066				.		0.1557

3) Features in the USL: the temporal step is 1 day.
the homogeneous group of 20 stations is obtained by applying the hybrid searching algorithm in this layer.

Table4 , feature selection and weighting results in USL

Feature	LHTFL	SHTFL	NDVI	Lai	Evlation	Latitude	Longtitude
Selection	0	1	0	1	0	1	1
weighting		0.154924		0.162286		0.3483	0.33449

❖ Regional frequency analysis

Based on the homogeneous group of 20 stations, we can get the Box plots of the ratio of confidence interval widths for the target stations at different durations:

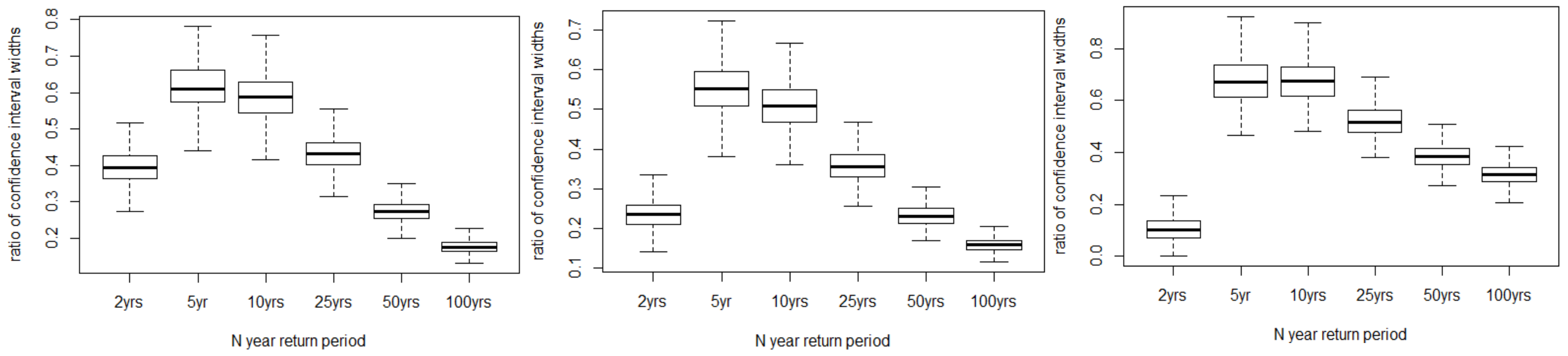


Figure3 :Box plots of the ratio between the confidence interval widths of reginal estimator and that of the single site analysis in the Toronto City site. Each graph contains results for one rainfall duration(24h , 6h and 15min) start from the left .

Figure 3 shows there is a consistent change of the return period in different durations. Regional frequency analysis can be effective in the estimation for return periods of 5 years and 10 years.

Conclusions & Future work

The proposed hybrid searching algorithm has generated desired results in the PBL and the UML as the difference between each station's membership values in two clusters can be easily detected, while this is not the case in the clustering in the USL . Future work will be explored in the following aspects to improve this situation:

1) In the feature selection stage, different combination of the potential features can generate the same clustering results. This is probably due to the correlations among the selected feature. In the future, the feature extraction approaches will be applied after the potential features being selected to reduce this possibility.
2) Due to the limit of the available datasets, the homogenous group in the UML was formed based on the daily time series of the input features for all the durations. In the future, a more advanced approach will be proposed to form different pooling group for each duration.

References

Brunsell, N. a., & Young, C. B. (2008). Land surface response to precipitation events using MODIS and NEXRAD data. *International Journal Of Remote Sensing*, *29*(7), 1965–1982. doi:10.1080/01431160701373747

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

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

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. doi:10.1175/JHM-D-12-0154.1

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

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

Contact Information

Zhe(Emma) Yang, PhD Candidate
Department of Civil & Environmental Engineering
University of Waterloo, Waterloo, Ontario, Canada
Tel: (519) 721-9755
Email: z232yang@uwaterloo.ca