Regional Hydrological Frequency Analysis at Ungauged Sites with Random Forest Regression

Shitanshu Desai[1,*], Taha B. M. J. Ouarda[1**]

[1] Canada Research Chair in Statistical Hydro-climatology, Centre Eau-Terre-Environnement, Institut National de la Recherche Scientifique, INRS-ETE, 490 De la Couronne, Québec (QC), Canada. G1K 9A9.

* Corresponding author: shitanshu.desai@ete.inrs.ca; situdesai@gmail.com

** Co-corresponding author: taha.ouarda@ete.inrs.ca

November 2020

ABSTRACT:

Flood quantile estimation at sites with little or no data is important for the adequate planning and management of water resources. Regional Hydrological Frequency Analysis (RFA) deals with the estimation of hydrological variables at ungauged sites. Random Forest (RF) is an ensemble learning technique which uses multiple Classification and Regression Trees (CART) for classification, regression, and other tasks. The RF technique is gaining popularity in a number of fields because of its powerful non-linear and non-parametric nature. In the present study, we investigate the use of Random Forest Regression (RFR) in the estimation step of RFA based on a case study represented by data collected from 151 hydrometric stations from the province of Quebec, Canada. RFR is applied to the whole data set and to homogeneous regions of stations delineated by canonical correlation analysis (CCA). Using the Out-of-bag error rate feature of RF, the optimal number of trees for the dataset is calculated. The results of the application of the CCA based RFR model (CCA-RFR) are compared to results obtained with a number of other linear and non-linear RFA models. CCA-RFR leads to the best performance in terms of root mean squared error. The use of CCA to delineate neighborhoods improves considerably the performance of RFR. RFR is found to be simple to apply and more efficient than more complex models such as Artificial Neural Network-based models.

Keywords:

Random Forest Regression, Canonical Correlation Analysis, Regional Flood Frequency Analysis, Ungauged basin, Machine Learning, Regional estimation.

Highlights:

- Random Forest Regression (RFR) is used for regional flood frequency analysis (RFA).
- RFR is also combined with Canonical Correlation Analysis (CCA): CCA-RFR.
- The two techniques are compared to other linear and non-linear RFA models.
- CCA-RFR leads to the best performance in terms of root mean squared error.
- RFR is simple to apply and more efficient than more complex models.

LIST OF ABBREVIATIONS

RFA : Regioanl Frequency Analysis

CCA : Canonical Correlation Analysis

ANN : Artificial Neural Network

GAM : Generalized Additive Model

RF : Random Forest

RFR : Random Forest Regression

CART : Classificationa and Regression trees

CCA-RFR : Random Forest Regression with Canonical Correlation Analysis

OOB : out-of-bag

SANN : Single Artificial Neural Network

EANN : Ensemble Artificial Neural Network

CCA-SANN : Single Artificial Neural Network with Canonical Correlation Analysis

CCA-EANN : Ensemble Artificial Neural Network with Canonical Correlation Analysis

CCA-GAM : Generalized Additive Model with Canonical Correlation Analysis

RMSE : Root mean squared error

NASH : Nash Sutcliffe model efficiency criterion

RMSEr : Relative Root Mean Squared Error

BIAS : Mean Bias

BIASr : Relative Mean Bias

k-fold CV : K-fold Cross Validation

Area : Basin Area

MBS : Basin Mean Slope

FAL : Fraction of Basin Area Occupied by Lakes

AMP : Annual Mean Total Precipitation

AMD : Annual Mean Degree-days above 0°

q100, q50 and q10 : Specific Flood quantiles corresponding to 100, 50 and 10 year return periods

MDI : Mean Decrease in Impurity

1. Introduction

Floods represent one of the most commonly occurring natural disasters (Stefanidis and Stathis, 2013). Floods cause significant environmental, economic and social damages. In spite of all flood protection measures being taken, from 1990 to 2013, floods have caused damages of about 600 billion US dollars and close to 7 million deaths worldwide (Wang et al., 2015). Thus, it is of the utmost importance to adequately predict the characteristics of such events at all sites.

However, hydrological information may not be available at certain sites of interest. At these "ungauged sites", Regional Frequency Analysis (RFA) can be used to develop estimates of flood characteristics. RFA allows transfer of information from gauged sites to the ungauged site of interest. RFA usually consists of two main steps. The first step is the delineation of homogeneous regions. In this step, sites that are similar according to some homogeneity criteria are grouped together. The rationale here is that as the sites within a given homogenous region are similar, information can reasonably be transferred from gauged to ungauged sites. The second step is the application of a regional estimation model within each delineated region (Ouarda, 2013; Wazneh et al., 2015). The regional estimation models are then trained to establish functional relationships between physio-meteorological basin characteristics and flow characteristics at ungauged basins.

Delineation can be done on the basis of geographical proximity, but that does not guarantee that such regions are homogenous in regards to their hydrologic response. In contrast, "Site focused" regionalization techniques (also called neighborhood-based techniques) have received much attention due to their effectiveness (Ouarda, 2016; Rahman et al., 2019). In "Site focused" techniques, each site has a prospective set of catchments which form a homogenous region for that particular site. One such technique is the Region of Influence (ROI) approach which identifies sites in a homogeneous region based on the distances in aS multidimensional space of catchment attributes from the target site to the contributing catchments. Haddad et al. (2012) showed that the ROI approach leads to more efficient and accurate flood quantile estimates compared to the fixes regions approach. Another such technique, Canonical Correlation Analysis (CCA), has been used for delineating homogenous regions in a number of studies (See for instance Ouarda et al., 2000; Han et al., 2020). In the present study, CCA is used to delineate homogenous regions as Ouarda et al. (2008) indicated that it leads to superior performances.

Among the large number of RFA estimation methods proposed in the literature, linear models and their variants are commonly adopted because of their simplicity and the speed in which they can be trained as well as deployed. However, hydrological systems are characterized by complex processes and it is unrealistic to assume a linear relationship between physio-meteorological basin characteristics and flow characteristics. Sivakumar and Singh (2012) showed that the relationship between these variables is characterized by dominant non-linear relationships. Pandey and Nguyen (1999) and Grover et al. (2002) showed that non-linear regression models provide better performances for RFA.

Several non-linear techniques have been proposed in the literature. An Artificial Neural Network (ANN), a non-linear and a non-parametric approach modelled on the neurons present in the human brain, was used for solving several hydrological problems such as regional flood frequency analysis, streamflow forecasting, rainfall-runoff modelling, flood forecasting, etc. (Aziz et al., 2014; Chokmani et al., 2008; Huo et al., 2012; Khalil et al., 2011; Kumar et al., 2015; Ouarda and Shu, 2009; Tiwari and Chatterjee, 2018). Generalized Additive Models (GAM) due to their considerable flexibility, are used in regional flood frequency analysis, water quality estimation, river discharge modeling, etc. (Chebana et al., 2014; Iddrisu et al., 2017; Morton and Henderson, 2008; Ouarda et al., 2018; Rahman et al., 2017). Other non-linear approaches used RFA include Projection Pursuit Regression (Durocher et al. (2015), Non-Linear CCA Ouali et al. (2015), and Adaptive Neuro-Fuzzy Inference Systems (ANFIS) (Shu and Ouarda, 2008).

Random Forest (RF), first proposed by Breiman (2001), is one such non-linear and non-parametric technique. It is a popular technique for classification, regression, variable selection, outlier detection and variable importance. When random forest is used for the purpose of function approximation or regression, it is called Random Forest Regression (RFR) or Regression Forests. In RFR, from a given set of data, multiple samples are randomly drawn and Classification and Regressions Trees (CART) are built. Eventually, the results of all such trees are combined and an estimate of target variables is obtained by averaging the outputs of individual trees.

A number of studies have been conducted in the field of hydrology using RFs. Chen et al. (2012) used RF to build a drought forecast model. Nguyen et al. (2015) used RF to forecast daily water levels. Monira et al. (2010) and Taksande and Mohod (2015) respectively used

RF for daily and monthly rainfall forecasting. Wang et al. (2015) developed a flood hazard risk assessment model based on RF. RF represents a good alternative to Support Vector Machines (Meyer et al., 2003; Verikas et al., 2001) and possesses a number of advantages including a reasonable amount of tolerance towards noise and outliers, high accuracy in forecasting and no overfitting problems.

The aim of the present study is to introduce the RF technique for regional flood quantile estimation. RFR is used to establish non-linear relationships between physio-meteorological basin characteristics and flow characteristics, and to estimate flood characteristics at ungauged sites. RFR is also applied to hydrological neighborhoods derived using CCA (CCA-RFR) for flood quantile estimation. A comparative analysis is carried out with several other approaches based on the application to a case study of data derived from the Province of Quebec, Canada.

The paper is organized as follows. In section 2, the theoretical background of RFR and CCA is presented along with the evaluation procedure and brief information about the models to be compared. The case study is presented in section 3 and the results are presented and discussed in section 4. Finally, the conclusions and recommendations for further research are presented in section 5.

2. Methodology

2.1. Random Forest Regression

2.1.1. RFR Principle

Random Forest is an ensemble learning technique proposed by Breiman (2001). RF is one of the most accurate general-purpose learning algorithms. Random Forest has been shown to give a very good performance while using few computational resources. RF exhibits great performance improvement over single tree algorithms like CART. It is fast and has error rates comparable to more traditional and resource intensive algorithms.

In Random forest for regression, the tree predictors $h(x, \theta_k)$, k = 1….K take on numerical values depending on the random vectors $\{\theta_k\}$ (Breiman, 2001). It is important to note that $\{\theta_k\}$ are identically distributed and independent random vectors. The training data is randomly and independently drawn from a joint distribution of (X, Y), where the random vector X is the observed input and the random vector Y is the expected numerical output. Individual trees are grown using the Classification and Regression Trees (CART) algorithm. Below is the algorithm for Random forest for regression as presented in Trevor et al. (2009).

---

(1) *For* $b = 1$ *to* B*:*

(a)Draw a bootstrap sample Z^* *of size N from training data.*

(b)Grow a random-forest tree T_b *to the bootstrapped data by recursively repeating the following steps for each terminal node of the tree, until the minimum node size* n_{min} *is reached.*

(i) Select m variables at random from p variables.

(ii) Pick the best variable/split-point among the m.

(iii) Split the node into two daughter nodes.

---

(2) *Output the ensemble of trees* $\{T_b\}_1^B$

- *To make a prediction at a new point x:*

$$\bar{f}_{rf}^{B} = \frac{1}{B}\sum_{b=1}^{B} T_b(x)$$

RFR possesses two important features, out-of-bag error rate, and variable importance. Generally, we use about two third of the data in a bootstrap sample and the rest one third are left out. These are known as out-of-bag (OOB) samples. The error estimated on these left out samples is known as OOB-error rate. OOB error rate can be used for validation purposes as well as for the calculation of the optimum number of trees required. Variable importance is a measure of which predictors are most useful for predicting the response variable. Variable importance can be computed using RF by recording improvements, at each node in every tree in the forest.

Another advantage of using RFR is that it possesses an 'acceptable' tolerance to noise and outliers, as the input training sets are drawn by random bootstrap sampling, and as the nodes to be split are selected randomly. Also, as there is no correlation between individual trees and as each tree is allowed to grow to its maximum size, there is no overfitting of data. Consequently, the only parameter to be tuned is the number of trees or estimators.

2.1.2 Classification and Regression Trees (CART)

CART decision tree is a binary recursion partitioning scheme which is capable of processing continuous and nominal attributes for regression and classification. In the present study, we use CART trees for regression. Regression trees are a nonparametric regression method that approximates real-valued functions. A regression tree is built using binary partitioning, where each node is iteratively split into two partitions or branches. Initially, all input variables are grouped into the same partition. Then mean squared error (mse) is calculated and a split decision is taken. The split decision is taken based on Greedy minimization. The split which minimizes the mse is selected and further that node is split into two off-springs. The splitting rule is then applied to each of the new offsprings. Each tree is grown to the largest possible extent which aids in better regression accuracy.

2.2 CCA approach in RFA

This section contains a brief discussion about CCA and its connection to the delineation step of RFA. Let X = {X_1, X_2 … X_r} be a random variable containing basin meteorological and physiographical variables, for eg. basin area, etc. and Y = {Y_1, Y_2 … Y_r} be a random variable containing basin hydrological variables like flood quantiles.

Consider linear combinations V and W of the variables X and Y:

$$V = a_1X_1 + a_2X_3 + \cdots + a_rX_r = a'X \quad (1)$$

$$W = b_1Y_1 + b_2Y_2 + \cdots + b_rY_r = b'Y \quad (2)$$

where a' and b' are transposes of vector a and b respectively. CCA enables identifying vectors a and b such that $corr(V, W)$ is maximum with vectors V and W having unit variances. For each basin B_k, where $k = 1, 2 \dots K$ from the set B of basins, $v_{i,k}$ and $w_{i,k}$ are corresponding values of V_i and W_i. We have the values of vector v_0 and our aim is to estimate the unknown vector w_0, where v_0 and w_0 represent the canonical scores of physio-meteorological and hydrological variables respectively.

The approximation of the w_0 vector can be obtained from a $100(1 - \alpha)\%$ confidence interval about λv_0 by constituting all the realizations w of W where:

$$(w - \lambda v_0)'(I_p - \lambda^2)^{-1}(w - \lambda v_0) \leq \chi^2_{\alpha,p}, \quad (3)$$

is conditional on $\chi^2_{\alpha,p}$ being $P(\chi^2 \leq \chi^2_{\alpha,p}) = 1 - \alpha$. For more detailed information concerning the algorithm, the reader is referred to (Ouarda et al., 2001).

2.3. Selection of Methods for Comparison

The RFR and CCA-RFR models are used to estimate the 100, 50 and 10-year flood quantiles. To evaluate the relative performances of these two approaches, they are compared to the following models:

- Canonical Correlation Analysis-Multiple linear regression model (CCA-MLR) (Ouarda et al., 2001). After selecting the optimal hydrological neighborhoods for each site using CCA analysis, multiple regression is used for regional flood estimation.

- Single Artificial Neural Network (SANN) (Shu and Burn, 2004). A single ANN is used to identify a functional relationship between physio-meteorological variables and flood quantiles.
- Ensemble ANN (EANN) (Shu and Burn, 2004). An ANN ensemble is created by bagging several single ANNs. This helps in improving the generalization ability of the SANN model. The final output is generated by taking the mean of the outputs of individual ANNs.
- Canonical Kriging Model (CCA-Kriging) (Chokmani and Ouarda, 2004). The physiographical space defined by CCA is used by the Kriging model to obtain regional flood estimates by interpolating data over that physiographic space. This method was shown to lead to comparable results to the traditional CCA model but is computationally less complicated.
- Single Artificial Neural Network in CCA physiographical space (CCA-SANN) (Shu and Ouarda, 2007). CCA is used to form the canonical physiographical space and then single ANN is applied to the data to form functional relationships between physiographical variables and flood quantiles.
- Ensemble ANN in CCA physiographical space (CCA-EANN) (Shu and Ouarda, 2007). In the CCA-EANN model, each component uses the same configuration as a Single ANN but the CCA-EANN is trained on bootstrapped sample data and the results are averaged out.
- Generalized Additive Model in conjunction with CCA (CCA-GAM) (Chebana et al., 2014). In the CCA-GAM approach, firstly backward stepwise selection is used to select the variables to be used in the model. Then GAM is applied to the neighborhoods delineated by CCA.

2.4. Evaluation Metrics

The following metrics are used to assess the quality of our regional flood analysis models. They are NASH (Nash Criterion), RMSE (Root mean squared error), RMSEr (Relative Root Mean Squared Error), BIAS (Mean Bias) and BIASr (Relative Mean Bias).

$$NASH = 1 - \frac{\sum_{i=1}^{n}(o_i - s_i)^2}{\sum_{i=1}^{n}(o_i - \overline{o})^2} \tag{4}$$

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(o_i - s_i)^2} \tag{5}$$

$$RMSEr = \sqrt{\frac{1}{n}\sum_{i=1}^{n}\left(\frac{o - s_i}{o_i}\right)^2} \tag{6}$$

$$BIAS = \frac{1}{n}\sum_{i=1}^{n}(o_i - s_i) \tag{7}$$

$$BIASr = \frac{1}{n}\sum_{i=1}^{n}\left(\frac{o_i - s_i}{o_i}\right) \tag{8}$$

where, o_i is the observed value at site i, s_i is the simulated value using the model for site i, $\overline{o}$ is the mean of observed at-site values and n is the number of sites.

2.5. Evaluation Procedure

K-fold Cross Validation (k-fold CV) is used as the model validation technique in this work. In k-fold CV the data is split into k small and equal sets. A model is trained using k – 1 folds as training data and then the model is validated using the remaining data. The performance thus reported by k-fold CV is the mean of the values computed in the loop.

The reason for using k-fold CV in the present study is that models trained with k-fold CV have lower variance than models trained with the jackknife validation procedure. In jackknife validation, there is more overlap between training folds as only one sample is omitted which means that almost the entire dataset is used for training. While in k-fold CV there is less overlap between training folds and thus it leads to smaller variability. Therefore, results obtained with jackknife might be better but the results obtained using k-fold CV are more robust.

3. Case Study

The dataset used in the present study consists of 151 hydrometric stations located in the southern part of the province of Quebec (between 45° and 55°N), Canada. The stations are operated by the Ministry of Environment of Quebec. The adopted dataset has been used in a number of previous RFA studies (Chebana and Ouarda, 2008; Shu and Ouarda, 2007) making it convenient for comparison of the results with those obtained with other methodologies.

On the basis of the work of Chokmani and Ouarda (2004) with the same database, a total of five physio-meteorological variables are selected, of which three are physiographical and two are meteorological variables. These variables are the basin area (Area), the mean basin slope (MBS), the fraction of basin area occupied by lakes (FAL), the annual mean total precipitation (AMP) and the annual mean degree-days above 0° (AMD), respectively.

A number of statistics of these data, like the minimum, mean, maximum and standard deviation are presented in table 1.

The database compiled by (Kouider et al., 2002) is used to extract at-site flood estimates for all of the 151 gauging stations in the study area. The most appropriate statistical distribution is used to get flood quantile estimates for each site by fitting the distribution to observed flood data. To avoid negative scale effects, specific quantiles (quantiles divided by basin areas) are used. The 100-year, 50-year, and 10-year quantiles (q100, q50, and q10 respectively) are the three specific flood quantiles used in the present study.

The reader is directed to (Shu and Ouarda, 2007) for more details concerning the dataset, such as scatter plots of basins in canonical space and geographical location of stations, to avoid redundancy. According to the recommendations of Shu and Ouarda (2007), the logarithmic transformation is applied to the variables q10, q50, q100, Area, MBS, AMP and AMD and a square root transformation is applied to FAL.

4. Results

In the present study, Scikit-learn module of Python is used to obtain the results (Pedregosa et al., 2011). In RF the size of the dataset, the number of trees (n_estimators) and the number of variables at each split have a huge impact on the error rate. According to Breiman (2001), the number of variables at each split should be taken as the square root of the total number of variables, i.e. 2 in this study. As the size of the dataset is not a tunable parameter, only the number of trees is tuned in this study.

Figure 1 illustrates that the OOB error rate decreases as the number of trees increases. At around 30 trees the value levels off and there is almost no improvement after this point by increasing the number of trees. Therefore, the number of trees is fixed at 30 for the present study. It is also important to note that all the trees were allowed to grow to the maximum extent without pruning.

The results of the application of the two models RFR and CCA-RFR along with the models described in Section 0 to the dataset described in Section 0 are illustrated in Table 2. The bold font describes the best approach for that particular flood quantile and the particular evaluation metric. Results indicate that CCA-RFR either outperforms or is comparable to other models in all the metrics except the NASH criterion. Also, CCA-RFR outperforms RFR in every metric other than NASH.

Figure 2 illustrates the relative errors associated with quantiles q50 estimated using RFR and CCA-RFR. Figure 2 indicates that CCA-RFR performs better than RFR for large basins, while RFR outperforms CCA-RFR for very small basins. These smaller basins are associated with larger specific quantiles. Therefore we can attribute the low NASH scores associated to CCA-RFR to these smaller sites Similarly, according to McCuen et al. (2006), the NASH criterion is sensitive to a number of factors including sample size and outliers.

In CCA-RFR, as only the stations in the hydrological neighborhoods are considered for the prediction and training, the sample size is considerably smaller than the complete original dataset. Also, the NASH criterion is heavily influenced by the model used (Schaefli and Gupta, 2007). RFR provides a reasonable tolerance to outliers which can be seen in the RFR NASH values. However, as we use just the neighborhoods for CCA-RFR, the sample size is small and thus outliers have more effect than in the basic RFR model which leads to lower NASH values.

Although we have low values for the NASH criterion for both RFR and CCA-RFR in comparison to other models, we can observe that CCA-RFR leads to the best RMSE and RMSEr values among all the models studied in this work. RMSE provides an evaluation of prediction accuracy in the absolute scale while RMSEr does the same in relative terms. CCA based RFR provides better generalization ability than the basic RFR model. As RFRs are nonparametric data-driven approaches, they have limited scope for extrapolation beyond the observed data. Therefore, the combination of RFR along with CCA, a parametric model helps the performance of RFR. Consequently, even though the NASH value for CCA-RFR is lower than other models the prediction accuracy is not compromised and is rather improved.

The BIAS and BIASr are evaluation criteria used to determine whether the model overestimates or underestimates the various quantiles. In general, CCA-RFR has the lowest BIAS of all the models considered and BIASr is also comparable with CCA-EANN and

CCA-GAM which have the best BIASr value. It is also important to point out that, in terms of BIAS, CCA-RFR overestimates flood quantiles while RFR underestimates them. However, when BIASr is used, all the models underestimate the flood quantiles.

Overall, it can be concluded that applying RFR to CCA delineated neighborhoods improves the results in comparison to RFR applied to the whole set of stations. This is consistent with the results of previous studies, such as Chokmani and Ouarda (2004) and Shu and Ouarda (2007), which indicated that applying other estimation techniques to CCA delineated neighborhoods leads to better performances for the estimation of flood quantiles than their application to the whole set of stations in the database.

The scatter plots of regional estimates using RFR and CCA-RFR are shown in Figure 3 and Figure 4, respectively. As would be expected, we observe that the estimation error and bias are positively correlated with the return period. With the increase in return periods, bias and estimation error increase simultaneously. Also, the low NASH scores can be explained by high variation as seen in Figure 4. It is clear from the results that all models underestimate flood quantiles at sites with higher specific quantiles. These sites can be associated with smaller basins which have large specific quantiles (Shu and Ouarda, 2007).

An additional experiment is conducted to identify the importance of individual predictor variables for flood quantile estimation. In the python implementation of RFR, "Mean decrease in Impurity (MDI)" or "Gini importance" is used to calculate the importance of each variable on the accuracy of the model. MDI is defined as "total decrease in node impurity averaged over all the trees. Node impurity is weighted by the probability of reaching that node (which is approximated by the proportion of sample reaching that node)"(Brieman et al., 1984). The results are illustrated in Table 3. Basin Area (Area) is shown to be by far the most important physio-meteorological variable. Annual mean total precipitation (AMP) and Annual mean degree days over 0° C (AMD) are distant second and third, respectively. Mean Basin Slope (MBS) is fourth while the Fraction of Area covered by lakes (FAL) is the least important of all physio-meteorological variables.

5. Conclusions

RF has been commonly used in gene classification, banking, medicine, and E-commerce. However, so far it has not found much application in the field of hydrology and especially in RFA. Most common studies in RFA establish linear relationships between physio-meteorological variables and flood quantiles. However, these models do not generally explain the complex relationships between the response variable and the explanatory variables. Random forest, a non-linear and a non-parametric data-driven approach, is one such technique which has shown good performances in other fields in explaining such complex relationships. This method is very easy to apply in practice as it does not require specific subjective choices by the user. The purpose of this study is to first introduce RFR in RFA and then apply RFR to neighborhoods delineated by CCA.

The number of trees in the RF for this study was fixed at 30. Also, all the trees were allowed to grow to their maximum potential without pruning. The comparison with other models indicates that, although CCA-RFR has a lower NASH score, it is more accurate than the other models. RFR is particularly more advantageous because of its low computational cost and high prediction quality. The results further indicate that the Random Forest, when used in conjunction with CCA, provides more robust and accurate results.

The research presented in this work is based on the introduction of the RF approach to RFA. The use of Extremely Randomized Trees and other variants of RF in RFA should also be attempted in the future. Future research activities should also focus on the use of RF in conjunction with other delineation techniques such as the Region of Influence approach, statistical depth functions, or projection pursuit regression. The effectiveness of the same techniques should also be investigated in the future using other data sets from different climates and different parts of the world to check the generality of the results obtained in this study. The efficiency of the technique should especially be examined for case studies with a higher level of heterogeneity in the physiographical variables. Future efforts should also investigate the use of the RF approach in the case of partially gauged sites and in the context of the use of procedures for the combination of local and regional information (see Seidou et al., 2006, for instance). The extension of the approach to the nonstationary case and for other hydrological variables such as low flows or suspended sediments should also be considered.

6. Acknowledgments. Financial support for this study was graciously provided by the Natural Sciences and Engineering Research Council (NSERC) of Canada. The authors would like to thank the Ministry of Sustainable Development, Environment, and Fight Against Climate Change of the Province of Quebec (MDDELCC) for the employed datasets. The authors would like to thank the Editor, Prof. Andras Bardossy, the Associate Editor, Prof. Félix Francés, and two anonymous reviewers for all their efforts and for their comments which helped improve the quality of the manuscript.

7. References

Aziz, K., Rahman, A., Fang, G., Shrestha, S., 2014. Application of artificial neural networks in regional flood frequency analysis: a case study for Australia. Stochastic Environmental Research and Risk Assessment, 28(3): 541-554. DOI:10.1007/s00477-013-0771-5

Breiman, L., 2001. Random forests. Machine Learning, 45(1): 5-32. DOI:10.1023/a:1010933404324

Brieman, L., Friedman, J., Olshen, R., Stone, C., 1984. Classification and Regression Trees. Wadsworth. Inc, Pacific Grove, CA.

Chebana, F., Charron, C., Ouarda, T.B.M.J., Martel, B., 2014. Regional Frequency Analysis at Ungauged Sites with the Generalized Additive Model. Journal of Hydrometeorology, 15(6): 2418-2428. DOI:10.1175/jhm-d-14-0060.1

Chebana, F., Ouarda, T.B., 2008. Depth and homogeneity in regional flood frequency analysis. Water resources research, 44(11).

Chen, J., Li, M., Wang, W., 2012. Statistical uncertainty estimation using random forests and its application to drought forecast. Mathematical Problems in Engineering, 2012.

Chokmani, K., Ouarda, T.B.M.J., Hamilton, S., Ghedira, M.H., Gingras, H., 2008. Comparison of ice-affected streamflow estimates computed using artificial neural networks and multiple regression techniques. Journal of Hydrology, 349(3-4): 383-396. DOI:10.1016/j.jhydrol.2007.11.024

Chokmani, K., Ouarda, T.B.M.J., 2004. Physiographical space-based kriging for regional flood frequency estimation at ungauged sites. Water Resources Research, 40(12). DOI:10.1029/2003wr002983

Durocher, M., Chebana, F., Ouarda, T.B.M.J., 2015. A Nonlinear Approach to Regional Flood Frequency Analysis Using Projection Pursuit Regression. Journal of Hydrometeorology, 16(4): 1561-1574. DOI:10.1175/jhm-d-14-0227.1

Grover, P.L., Burn, D.H., Cunderlik, J.M., 2002. A comparison of index flood estimation procedures for ungauged catchments. Canadian Journal of Civil Engineering, 29(5): 734-741. DOI:10.1139/l02-065

Han, X., Ouarda, T.B.M.J., Rahman, A., Haddad, K., Mehrotra, R., Sharma, A., 2020. A Network Approach for Delineating Homogeneous Regions in Regional Flood Frequency Analysis. Water Resources Research, 56(3): e2019WR025910. DOI:10.1029/2019wr025910

Haddad, K., Rahman, A., 2012. Regional flood frequency analysis in eastern Australia: Bayesian GLS regression-based methods within fixed region and ROI framework – Quantile Regression vs. Parameter Regression Technique. Journal of Hydrology, 430-431: 142-161. DOI:10.1016/j.jhydrol.2012.02.012

Huo, Z., Feng, S., Kang, S., Huang, G., Wang, F., Guo, P., 2012. Integrated neural networks for monthly river flow estimation in arid inland basin of Northwest China. Journal of Hydrology, 420-421: 159-170. DOI:10.1016/j.jhydrol.2011.11.054

Iddrisu, W.A., Nokoe, K.S., Luguterah, A., Antwi, E.O., 2017. Generalized Additive Mixed Modelling of River Discharge in the Black Volta River. Open Journal of Statistics, 07(04): 621-632. DOI:10.4236/ojs.2017.74043

Khalil, B., Ouarda, T.B.M.J., St-Hilaire, A., 2011. Estimation of water quality characteristics at ungauged sites using artificial neural networks and canonical correlation analysis. Journal of Hydrology, 405(3–4): 277-287. DOI:10.1016/j.jhydrol.2011.05.024

Kouider, A., Gingras, H., Ouarda, T., Ristic-Rudolf, Z., Bobée, B., 2002. Analyse fréquentielle locale et régionale et cartographie des crues au Québec.Research report (R619). INRS-Eau, Terre et Environnement, Québec.

Kumar, R., Goel, N.K., Chatterjee, C., Nayak, P.C., 2015. Regional Flood Frequency Analysis using Soft Computing Techniques. Water Resources Management, 29(6): 1965-1978. DOI:10.1007/s11269-015-0922-1

McCuen, R.H., Knight, Z., Cutter, A.G., 2006. Evaluation of the Nash–Sutcliffe efficiency index. Journal of Hydrologic Engineering, 11(6): 597-602. DOI:doi:10.1061/(ASCE)1084-0699(2006)11:6(597)

Meyer, D., Leisch, F., Hornik, K., 2003. The support vector machine under test. Neurocomputing, 55(1-2): 169-186. DOI:10.1016/s0925-2312(03)00431-4

Monira, S.S., Faisal, Z.M., Hirose, H., 2010. Comparison of artificially intelligent methods in short term rainfall forecast, Computer and Information Technology (ICCIT), 2010 13th International Conference on. IEEE, pp. 39-44. DOI:10.1109/ICCITECHN.2010.5723826

Morton, R., Henderson, B.L., 2008. Estimation of nonlinear trends in water quality: An improved approach using generalized additive models. Water Resources Research, 44(7). DOI:10.1029/2007wr006191

Nguyen, T.-T., Huu, Q.N., Li, M.J., 2015. Forecasting time series water levels on Mekong river using machine learning models, Knowledge and Systems Engineering (KSE), 2015 Seventh International Conference on. IEEE, pp. 292-297. DOI:10.1109/KSE.2015.53

Ouali, D., Chebana, F., Ouarda, T.B.M.J., 2015. Non-linear canonical correlation analysis in regional frequency analysis. Stochastic Environmental Research and Risk Assessment, 30(2): 449-462. DOI:10.1007/s00477-015-1092-7

Ouarda, T.B.M.J., 2013. Hydrological Frequency Analysis, Regional, Encyclopedia of Environmetrics. DOI:10.1002/9780470057339.vnn043

Ouarda, T.B.M.J., 2016. Regional flood frequency modeling, Chap. 77, Chow's Handbook of Applied Hydrology, 2nd Edn., edited by Singh, V. P. Mc-Graw Hill, New York, pp. 77.1–77.8, ISBN 978-0-07-183509-1

Ouarda, T.B.M.J., Ba, K.M., Diaz-Delgado, C., Carsteanu, A., Chokmani, K., Gingras, H., Quentin, E., Trujillo, E., Bobee, B., 2008. Intercomparison of regional flood frequency estimation methods at ungauged sites for a Mexican case study. Journal of Hydrology, 348(1-2): 40-58. DOI:10.1016/j.jhydrol.2007.09.031

Ouarda, T.B.M.J., Charron, C., Hundecha, Y., Saint-Hilaire, A., Chebana, F., 2018. Introduction of the GAM model for regional low-flow frequency analysis at ungauged basins and comparison with commonly used approaches. Environmental Modelling & Software, 109: 256-271. DOI:10.1016/j.envsoft.2018.08.031

Ouarda, T.B.M.J., Girard, C., Cavadias, G.S., Bobée, B., 2001. Regional flood frequency estimation with canonical correlation analysis. Journal of Hydrology, 254(1-4): 157-173. DOI:10.1016/s0022-1694(01)00488-7

Ouarda, T.B.M.J., Haché, M., Bruneau, P., Bobée, B., 2000. Regional flood peak and volume estimation in northern Canadian basin. Journal of cold regions engineering, 14(4): 176-191. DOI:10.1061/(ASCE)0887-381X(2000)14:4(176)

Ouarda, T.B.M.J., Shu, C., 2009. Regional low-flow frequency analysis using single and ensemble artificial neural networks. Water Resources Research, 45(11): W11428. DOI:10.1029/2008wr007196

Pandey, G., Nguyen, V.-T.-V., 1999. A comparative study of regression based methods in regional flood frequency analysis. Journal of Hydrology, 225(1-2): 92-101. DOI:10.1016/S0022-1694(99)00135-3

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., 2011. Scikit-learn: Machine learning in Python. Journal of machine learning research, 12: 2825-2830.

Rahman, A., Charron, C., Ouarda, T.B.M.J., Chebana, F., 2017. Development of regional flood frequency analysis techniques using generalized additive models for Australia. Stochastic Environmental Research and Risk Assessment, 32(1): 123-139. DOI:10.1007/s00477-017-1384-1

Rahman, A., Haddad, K., Kuczera, G., Weinmann, E., 2019. Regional flood methods. Australian Rainfall and Runoff: A Guide To Flood Estimation. Book 3, Peak Flow Estimation: 105-146.

Schaefli, B., Gupta, H.V., 2007. Do Nash values have value? Hydrological Processes, 21(15): 2075-2080. DOI:10.1002/hyp.6825

Seidou, O., Ouarda, T.B.M.J., Barbet, M., Bruneau, P., Bobée, B., 2006. A parametric Bayesian combination of local and regional information in flood frequency analysis. Water Resources Research, 42(11): W11408. DOI:10.1029/2005wr004397

Shu, C., Burn, D.H., 2004. Artificial neural network ensembles and their application in pooled flood frequency analysis. Water Resources Research, 40(9). DOI:10.1029/2003wr002816

Shu, C., Ouarda, T.B.M.J., 2007. Flood frequency analysis at ungauged sites using artificial neural networks in canonical correlation analysis physiographic space. Water Resources Research, 43(7). DOI:10.1029/2006wr005142

Shu, C., Ouarda, T.B.M.J., 2008. Regional flood frequency analysis at ungauged sites using the adaptive neuro-fuzzy inference system. Journal of Hydrology, 349(1-2): 31-43. DOI:10.1016/j.jhydrol.2007.10.050

Sivakumar, B., Singh, V.P., 2012. Hydrologic system complexity and nonlinear dynamic concepts for a catchment classification framework. Hydrology and Earth System Sciences, 16(11): 4119-4131. DOI:10.5194/hess-16-4119-2012

Stefanidis, S., Stathis, D., 2013. Assessment of flood hazard based on natural and anthropogenic factors using analytic hierarchy process (AHP). Natural Hazards, 68(2): 569-585. DOI:10.1007/s11069-013-0639-5

Taksande, A.A., Mohod, P., 2015. Applications of data mining in weather forecasting using frequent pattern growth algorithm. IJSR, 4(6): 3048-51.

Tiwari, M.K., Chatterjee, C., 2018. Flood Forecasting and Uncertainty Assessment Using Wavelet- and Bootstrap-Based Neural Networks, Handbook of Research on Predictive Modeling and Optimization Methods in Science and Engineering. Advances in Computational Intelligence and Robotics, pp. 74-93. DOI:10.4018/978-1-5225-4766-2.ch004

Hastie, T., Tibshirani, R., Friedman, J., 2009. The elements of statistical learning: data mining, inference, and prediction. Springer, New York, NY.

Verikas, A., Gelzinis, A., Malmqvist, K., 2001. Using unlabelled data to train a multilayer perceptron. Neural Processing Letters, 14(3): 179-201. DOI:10.1023/A:1012707515770

Wang, Z., Lai, C., Chen, X., Yang, B., Zhao, S., Bai, X., 2015. Flood hazard risk assessment model based on random forest. Journal of Hydrology, 527: 1130-1141. DOI:10.1016/j.jhydrol.2015.06.008

Wazneh, H., Chebana, F., Ouarda, T.B.M.J., 2015. Delineation of homogeneous regions for regional frequency analysis using statistical depth function. Journal of Hydrology, 521: 232-244. DOI:10.1016/j.jhydrol.2014.11.068

LIST OF TABLES AND FIGURE CAPTIONS

Table 1 : Descriptive Statistics of physio-meterological and Hydrological Variables.

Table 2: NASH, RMSE, RMSEr, BIAS and BIASr values for all the models. Best values for each quantile for the corresponding metrics are marked in bold.

Table 3: Feature Importance of Five Input Variables used for Specific Flood Quantile Estimation.

Figure 1: Number of trees (n_estimators) vs OOB error rate for 10, 50 and 100-year flood quantiles.

Figure 2: Relative errors associated with at-site quantile q50 calculated using RFR and CCA-RFR (the sites are ordered according to their area)

Figure 3: A) q10, B) q50 and C) q100 estimation using RFR approach.

Figure 4: A) q10, B) q50 and C) q100 estimation using CCA-RFR approach.

Table 1: Descriptive Statistics of physio-meterological and Hydrological Variables.

Variables	Minimum	Mean	Maximum	Standard deviation
q10 ($m^3/s.km^2$)	0.03	0.31	0.94	0.20
q50 ($m^3/s.km^2$)	0.03	0.28	0.77	0.18
q100 ($m^3/s.km^2$)	0.03	0.22	0.53	0.13
Area (km^2)	208	6255	96600	11716
MBS (%)	0.96	2.43	6.81	0.99
FAL (%)	0.00	7.72	47.00	7.99
AMP (mm)	646	988	1534	154
AMD (degree day)	8589	16346	29631	5382

Table 2: NASH, RMSE, RMSEr, BIAS and BIASr values for all models. Best values for each quantile for the corresponding metrics are marked in bold.

	Hydrological Variables	CCA-SANN	CCA-EANN	CCA-Kriging	CCA-MLR	SANN	EANN	CCA-GAM	RFR	CCA-RFR
NASH	q10	0.82	**0.84**	0.78	0.78	0.75	0.78	0.82	0.721	0.577
	q50	0.78	**0.8**	0.72	0.72	0.69	0.72	0.76	0.657	0.532
	q100	0.77	**0.78**	0.7	0.68	0.66	0.69	0.67	0.644	0.507
RMSE	q10	0.053	0.05	0.05	0.059	0.06	0.058	0.054	0.063	**0.049**
	q50	0.082	0.079	0.093	0.094	0.098	0.093	0.087	0.089	**0.07**
	q100	0.095	0.093	0.11	0.112	0.115	0.109	0.115	0.099	**0.08**
RMSEr	q10	38	37	51	43	47	44	33.7	80.74	**29.44**
	q50	44	43	64	49	55	53	43.5	93.39	**33.27**
	q100	46	45	70	51	64	60	37.0	96.45	**35.02**
BIAS	q10	0.006	0.005	-0.004	**0.001**	0.006	0.004	0.009	-0.0013	0.002
	q50	0.009	0.009	-0.007	0.005	0.01	0.009	**-0.003**	-0.0073	**0.003**
	q100	0.013	0.012	-0.008	0.007	0.015	0.013	0.043	-0.019	**0.004**
BIASr	q10	-5	-5	-16	-9	-7	-7	**-3.5**	-21.12	-6.64
	q50	-7	**-5**	-21	-11	-8	-8	-11.4	-25.97	-8.14
	q100	-7	-6	-23	-11	-11	-10	**3.4**	-27.85	-8.89

Table 3: Feature Importance of Five Input Variables used for Specific Flood Quantile Estimation.

Input Variables	Relative Importance, %		
	q10	q50	q100
Area	87.17	88.53	78.25
MBS	1.39	0.65	0.99
FAL	1.10	0.70	0.57
AMP	8.86	7.71	17.89
AMD	1.46	2.38	2.27

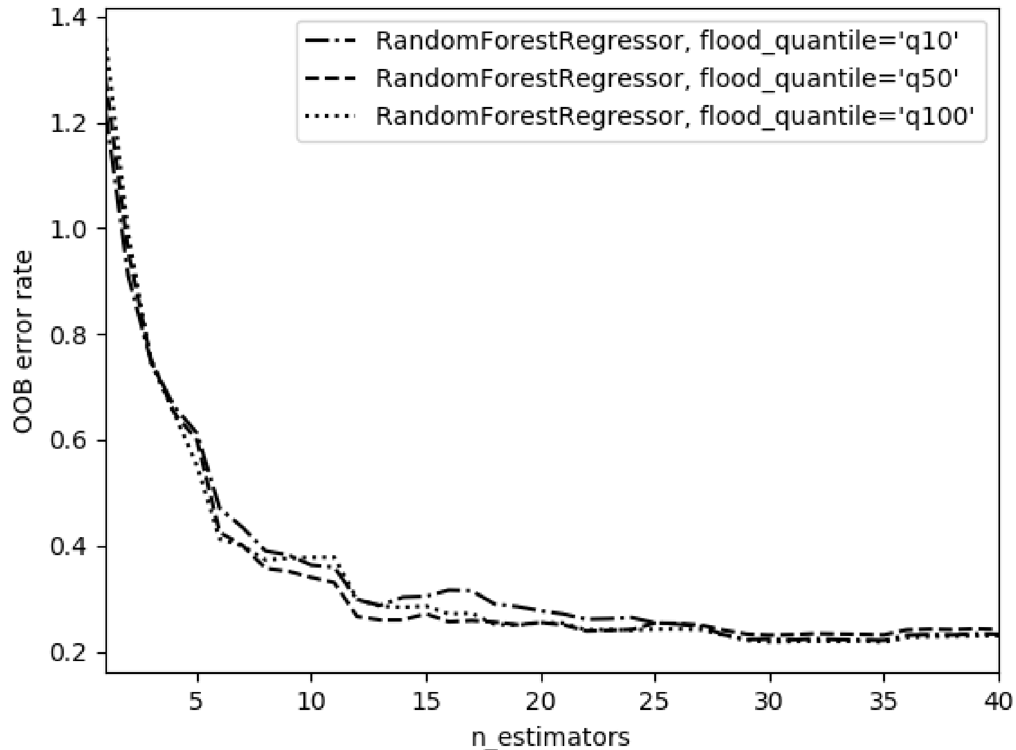


Figure 1: Number of trees (n_estimators) vs OOB error rate for 10, 50 and 100-year flood quantiles.

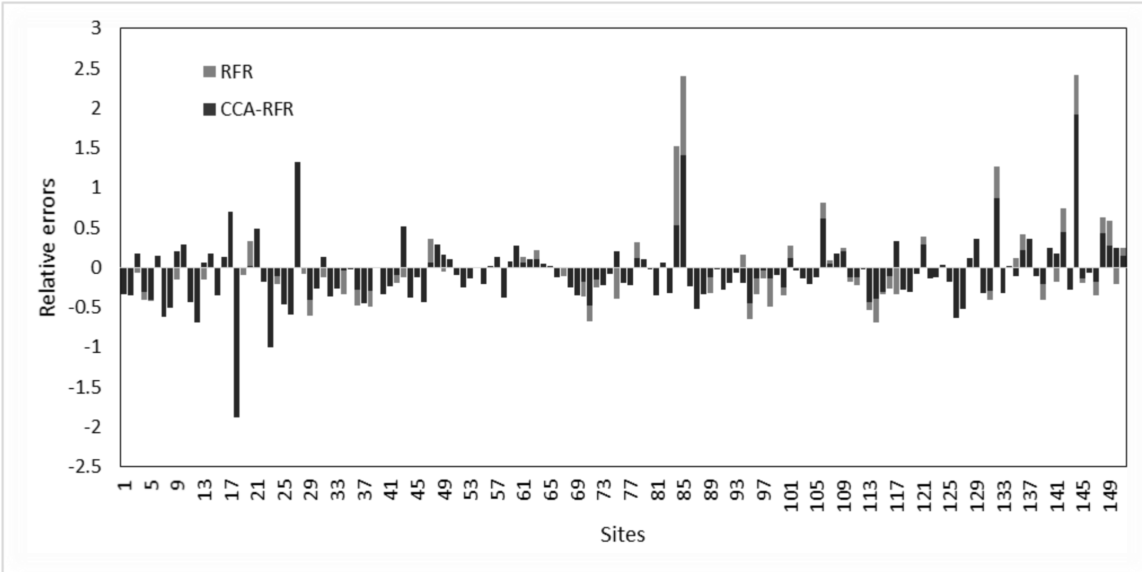
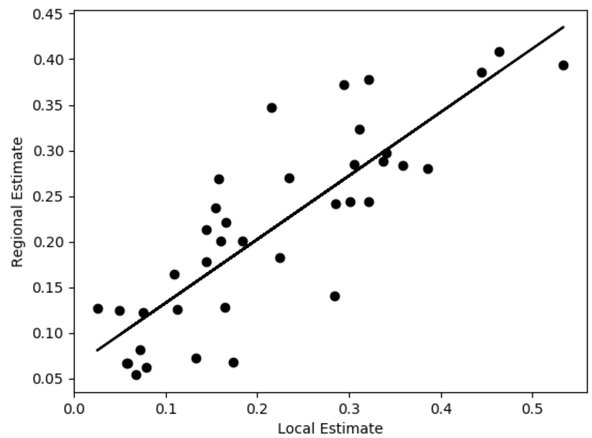
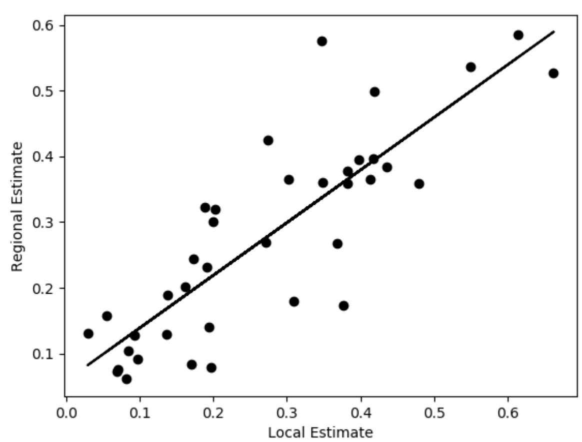


Figure 2: Relative errors associated with at-site quantiles q50 calculated using RFR and CCA-RFR (the sites are ordered according to the increasing area)

A) q10 estimation



B) q50 estimation



C) q100 estimation

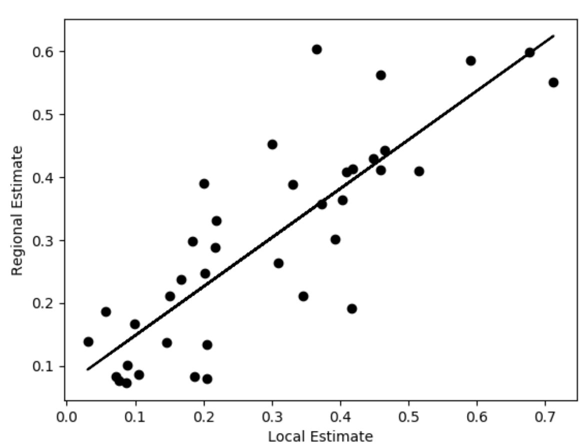
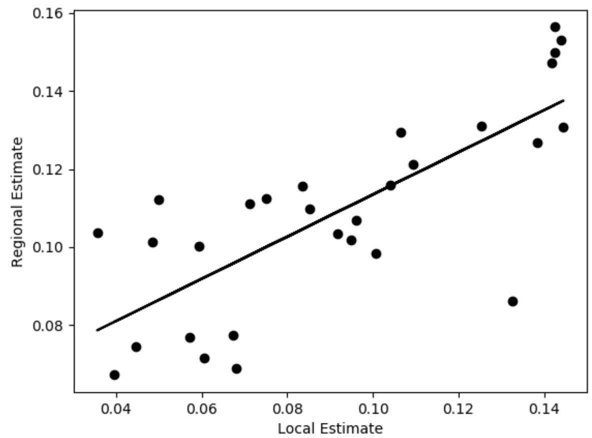
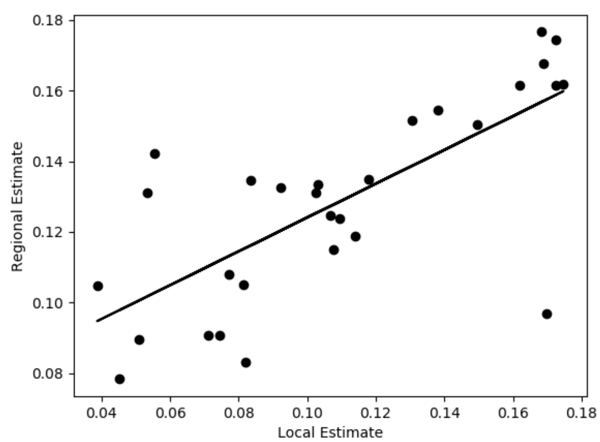


Figure 3: Estimation using the RFR approach

A) q10 estimation



B) q50 estimation



C) q100 estimation

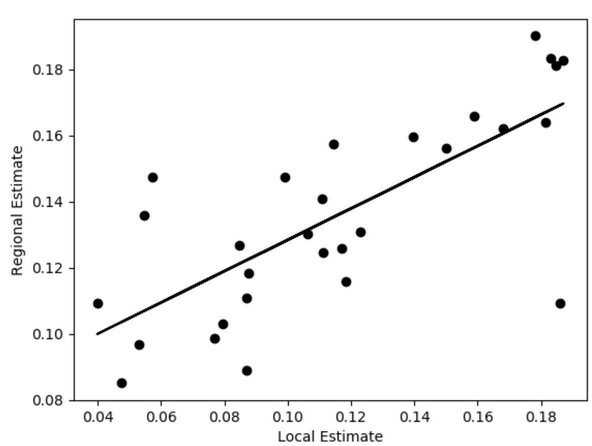


Figure 4: Estimation using the CCA-RFR approach