Delineation of homogenous regions using hydrological variables predicted by projection pursuit regression

Martin Durocher, Fateh Chebana, and Taha B. M. J. Ouarda

Manuscript: hess-2016-123

Dear Editor,

Please find herewith the response to all comments made by the two reviewers concerning the manuscript "Delineation of homogenous regions using hydrological variables predicted by projection pursuit regression" in the interactive discussion on the website of Hydrology and Earth System Science (HESS).

The authors gratefully acknowledge the helpful comments that have contributed to the improvement of the paper. Detailed replies to these comments are provided below.

Should you have any questions or require further information concerning this revision, please contact me.

Sincerely,
Martin Durocher

Reply to Reviewer 1 comments

The authors are grateful to the reviewer for his comments which contributed to improving the quality of the paper. The authors provide hereafter the answers to the reviewer's comments. Please note that the numbering of the figures has been changed, which explains the difference between the reviewer's comments and the provided answers.

General comments

Reviewer: The article "Delineation of homogeneous regions using hydrological variables predicted by projection pursuit regression" by Durocher et al. describes an improvement of existing techniques for the regional estimation of flood quantiles. The topic is very relevant, but I found the manuscript not completely clear in some parts and with some methodological flaws. While readability (see comment n. 1) can be improved with a revision of the text, some methodological issues would require a complete reanalysis of the work. My main concerns (see comments n.2 and 3) are basically related to the use of a very complex procedure that is not justified by the results. This finding (i.e., that the procedure does not really produce improvements) would be a result itself, but the authors seem to overlook it to support their initial hypotheses. For these reasons, I suggest a rejection of manuscript.

Answer: The authors would like to thank the Reviewer for the thorough review and the constructive comments. In the following we will provide a response to the comments formulated by the reviewer. Please note that the response to the general comments (mainly the justification for a complex model) is in fact available in the response to the major comments below.

Major comments

(1.1) Reviewer: "The point list on page 6, and in particular the step i) (which is the main focus of the paper) should be supported by a quantitative example to make the procedure easier to understand. For instance, the plots in figure 3 could be used in this part of the manuscript to better describe how the procedure works (and not only from page 11 to comment results)."

Answer: The authors agree that the illustration provided by Figure 4 (initially Figure 3) is useful to understand this step of the methodology. Figure 1 was initially intended to provide a schematic illustration of this step. Figure 1 was improved in the revised manuscript in order to be more similar in its interpretation. The following sentences were modified in agreement with the new figure:

> P7-L19, "Figure 1 illustrates a region with several sites where two neighborhoods are resulting from the RVN method with different predicted centers. The target site is illustrated as a green filled circle and neighborhood is formed of the 10 nearest sites indicated by small empty circles. The other sites are designated by crosses. The red and blue neighborhoods are delineated by circles where the radius is selected to include the 10 nearest sites. The predicted center of the red neighborhood is closer to the target site. Consequently, it can be seen that except from one site, the same sites as the target neighborhood are included (empty circles). On the other hand, the blue neighborhood has a predicted center further to the target site and hence a lower proportion of the sites truly

> closer to the target are found. It shows the importance of correctly predicting neighborhood centers in order to identify sites that are truly similar to the target site."

(1.2) Reviewer: "Moreover, step i) seems a kind of "preliminary" regionalization of the L-moments of the target site. Such L-moments are then used to support the delineation of the region. Why such preliminary estimates cannot be directly used in the prediction of flood quantiles? This point should be discussed by the authors, highlighting the possible differences with the direct estimation of the quantiles based on preliminary L-moments."

Answer: In a situation where the target distribution is assumed to be known (i.e. the same family of distribution is assumed everywhere) the preliminary prediction could be used indeed to deduce the parameter of the target distribution from the L-moments. However, when the distribution is unknown, the L-moments cannot be used directly to estimate the flood quantile without specifying a family of distributions. This is achieved here by using the information of the relevant neighborhoods.

Moreover, the approach of using the preliminary L-moments is not generally applicable to the present RVN methodology as it fixed the reference variables, while the present methodology includes a stepwise procedure to perform this selection. The following sentences are added to the revised manuscript:

> P12-L12, "At this point, the steps 1-4 of the RVN methodology are performed and the neighborhoods are identified. Notice that for the RVN-LM method, the reference variables include the first three L-moments, which could be used as a moment estimator to deduce the target distribution. This approach is, however, not generally applicable to the present methodology as the reference variables are selected by a stepwise procedure. Moreover, it is necessary to identify a proper family of distributions from regional information, which is achieved here by analyzing the distribution of the gauged sites inside the neighborhoods."

(2.1) Reviewer: "Figure 5 tells me that there is no significant difference between the methods, so I would use the simpler one. Of course, by computing an error metric over the whole set of residuals one may obtain slightly better performances of the RVN models (but this is not reported; a summary table would be appreciated). "

Answer: Notice that two error metrics are actually used in the present study: RRMSE and NHS. The authors agree that an additional table would help summarize more directly the information provided in the text and in Figures 5 and 8 (initially 4 and 6) for the calibrated model. In this sense, Table 1 below is added to revised manuscript:

Table 1: Evaluation criteria for the RVN method for optimal neighborhood sizes.

	Model	Size	RRMSE	NHS	AHM	ARE
Index-Flood						
	ROI	30	46.2	86.5	72.8	57.3
	CCA	28	45.4	86.2	41.7	42.9
	RVN-LM	29	45.0	**87.1**	**14.5**	**36.9**
	RVN-HYB	24	**40.1**	86.2	16.5	43.1
Regression-based						
	ROI	30	44.9	86.9	72.8	64.7
	CCA	28	43.5	86.1	41.7	**30.6**
	RVN-LM	39	41.7	**87.6**	17.9	39.8
	RVN-HYB	24	**39.5**	86.2	**16.5**	42.5

Best criteria in bold

(2.2) Reviewer: "The authors state on page 12, line 25 onwards, that improvement is effective for sites with largest discrepancies. This seems not true except for two point in figure 5a and one point in figure 5b (all the points in the bottom-left corner of each panel). Figures 5c and 5d show points equally distributed around the bisector also in the bottom-left corner. Hence, is the complexity of the RVN model justified by a so small performance improvement?"

Answer: The authors want to make the precision that the comment "improvement is effective for sites with largest discrepancies" applied only to the relative residuals as this comment is made while describing Figures 6a and 6b (initially 5a,b). To improve clarity, the term "largest relative discrepancies" is used in the revised version of the manuscript:

> P13-L31, "However, Figures 6a,b show that the RVN-HYB specifically improves the prediction of the sites with the lowest and largest relative discrepancies as the red line is clearly located under the $y = x$ lines, which explains the improved RRMSE in Table 1."

The additional complexity of the RVN method in comparison to the traditional ROI method concerns only the preliminary step, where the missing hydrological information is substituted by predicted values. The question is thus if the addition of the preliminary step is justified. Notice that RRMSE and NHS are cross-validated criteria and hence they are criteria that "penalize" excessive complexity in the prediction. Consequently, a better RRMSE implies that a model with a higher criterion truly brings additional information on the quantiles. The overall improvement in terms of RRMSE for the RVN-HYB method with respect of the ROI and CCA methods is respectively 6.1% and 5.3% (see Table 1). Although of moderate amplitude, the authors believe that these improvements are nonetheless important as several references publications in the field had been published with improvements in the range of 5%. The following sentences are modified in the revised manuscript:

> P13-L11, "Hence, the calibrated models are set according to the RRMSE criterion and are represented by circles in Figure 5 and are summarized in Table 1. RVN-HYB, with a RRMSE of 40.1% outperforms the other methods. In particular, a difference of 6.1% and 5.3% is observed respectively with the traditional ROI and CCA methods."

The authors agree with the reviewer that the red lines in the bottom-left corner of Figures 6a,b are mostly influenced by few points. Nevertheless, in panel a, the 4 largest relative discrepancies are better predicted by RVN-HYB and the two best relative improvements are of 77.2% and 68.5%. In these figures, small changes in average can be difficult to assess by the naked eye. The red lines are smooth curves fitted between the residuals of the two models, and are there to indicate which models in average have locally lower residuals. It can be seen that for the residuals approximately above 0.2 (i.e. 20% of the observed values) the red line is distinctively under the bisector (upper right corner) and this is true for several points. This means that the sites that are overestimated are on average less overestimated by the RVN-HYB method. Actually, 8 out of the 9 residuals located at the most right are better predicted by the RVN-HYB model. Similar behavior is also noticeable in upper-left corner of Figure 5b.

In the revised version of the manuscript the following lines were modified to better explain the implication of the result shown in Figure 6:

> P13-L26, "As mentioned in section 4.2, previous studies have identified few problematic stations in the considered dataset. Figure 6 presents the residuals between different methods. As it may be difficult to see small improvements by uniquely observing points around the $y = x$ lines, the visualization of Figure 6 is helped by adding a flexible fit of the point cloud, using a standard smoothing spline approach. The resulting red lines indicate, in average, if close to x the residuals are lower for one of the two methods. In general, the points associated to the largest relative discrepancies are close to the $y = x$ line, which indicates that the sites that are difficult to predict are essentially the same for all methods. However, Figures 6a,b show that the RVN-HYB specifically improves the prediction of the sites with the largest relative discrepancies as the red line is clearly located under the $y = x$ lines (left and right), which explains that this method leads to the best RRMSE. On the other hand, Figures 6c,d demonstrate that at the logarithmic scale, the RVN-LM method achieved predicted values that are mostly similar to the ROI and CCA methods, which explains the similarity of the NHS criteria for all the compared methods.
>
> The present case study is an example of a region where some sites are problematic for likely any method. In practice, the residuals are not known, consequently we do not know if the target sites of interest will be "problematic" or not. Globally, what Figure 6a indicates is that the RVN-HYB model is more robust (in a certain way), because for the sites that are well predicted by simpler models, such ROI, RVN-HYB will perform in average similarly. However, if the target site is predicted less accurately, the RVN-HYB model will (in average) be better in terms of RRMSE. Consequently, the overall gain may seem of moderate magnitude, but for some problematic stations the gain could be more substantial. In particular, the red lines in the left part of Figure 6a appears mostly influenced by two points, but the two improvements are of 77.2% and 68.5%, which is considerable."

(3) Reviewer: "On page 10, line 29, the nonlinear relationship between the (transformed) predictors and the (transformed) L-moment is mentioned and the authors say that it is shown in figure 2. This non-linear relationship would justify the use of a spline interpolator, but actually this is a questionable point. In fact, figure 2 tells a different story. Panel a clearly show a linear relationship (in the transformed variables; this is expected as often the mean value can be linearized with log transformations). In the b, c and d panels there is a much larger scattering, which does not allow to identify a clear complex pattern, even if all the plots show an increasing trend. Looking at the scatter plots I believe that most of the people would adopt a simple linear regression (said with 2 parameters) which is much more stable and robust. My personally idea is that the choice of the authors is not justified and that a linear model should be at least compared to the spline interpolator."

Answer: The authors understand the concern of the reviewer in the use of the PPR in Figure 3 (initially Figure 2). Comparison with linear model are made. For L1 and LSK the difference between the NHS criteria is about 1 %, which the authors agree is very small. However, the gains for the LCV and LKT are respectively of 5.1% and of 7.6%, which is substantial. Notice that the NHS is a cross-validation criterion and hence this result does not represent a form of overfitting in favor of the PPR as the predictions are obtained without the use of the predicted sites. Unnecessary complexity is then «penalized» by such criteria. The fitting of LCV and LKT are cases of mild nonlinearity, but such mild nonlinearity can be adjusted here because 151 is a reasonable number of sites. The authors agree that the situation would be different if for instance only 20 sites where available. Therefore, in the present case study, it is not true that linear models are more stable and robust than PPR as they are assessed by cross-validation. The following changes are made to the revised manuscript:

> P11-L26, "Figure 3a shows a strong linear relationship between L1 and the predictor $\alpha'\mathbf{X}$. Conversely, Figures 3b,c,d show mild nonlinearity and hence indicate the need for more flexible models, such as PPR. The predictive performances of the reference variables are evaluated by the NHS criterion with values 91.5%, 33.3%, 6.7% and 55.7% respectively for L1, LCV, LSK and LKT. These results show that L1 is accurately predicted by the site characteristics, while a poor fit is associated to LSK. Indeed, Figure 3c suggests that apart from a few sites on the right of the curve, LSK appears not highly related to the predictor $\alpha'\mathbf{X}$. In comparison, linear models applied on the same reference variables lead to NHS criterion: 90.9%, 28.2%, 7.8% and 48.1% respectively. Remark that NHS criterion is calculated by cross-validation, consequently even though the improved performances by the PPR method appear moderate this represent true fitting improvements."

Minor comments

(1) Reviewer: "P11 L5-7 I found quite strange that the L-kurtosis performs much better that the L-skewness as in general the prediction ability deteriorates with increasing order of L-moments. The authors should investigate in more detail this issue."

Answer: The authors agree with the reviewer that L-skewness is expected to be, in general, better predicted than the L-kurtosis. The data have been investigated for data manipulation errors and nothing have been found. It appears to be a legitimate exception to the rule.

(2) Reviewer: "P7 L6 Please, give a more detailed description of "true neighborhood" meaning."

Answer: To clarify the term "true neighborhood" the following lines are modified in the revised manuscript:

> P7 L16, "If the hydrological variables $\mathbf{t}_0$ were known at the target location, the distance h_i would be available and the neighborhood that truly regroups the most hydrologically similar sites to the target location can be identified. However, in practice this true neighborhood is unknown. Using instead the estimate $f(\mathbf{x}_0)$ has the effect that some sites are falsely suggested as more hydrologically similar than other sites."

(3) Reviewer: Figure 5. Not clear which kind of information is provided by the "smooth fitting of the residuals". Also in this case, the smooth fitting seems too complex tool which does not add any further information.

Answer: As briefly discussed in the major comment (2.2), at the proximity of a point say (X,X) in Figure 6 (initially Figure 5), several points maybe under or above the bisector, which makes it difficult to see small advantages for a given method. The smooth fitting, provided by the red lines, indicates specifically the average between the residuals of the two models in the proximity of a point X. The authors agree with the reviewer that parsimony is important. The smoothing splines procedure used here as a visual guide is widely available in most numerical software's and the "degree of complexity" is controlled by the generalized cross-validation criteria (GCV), which is a widely used criterion to avoid overfitting. Hence, with the information from 151 sites, the authors believe that the smooth fitting is not "too complex", but it is simply "as complex" as the residuals suggest it. The following precisions are added to the revised manuscript:

> P13 L27, "As it may be difficult to see small improvements by uniquely observing points around the $y = x$ lines, the visualization of Figure 6 is helped by adding a flexible fit of the point cloud, using a standard smoothing spline approach. The resulting red lines indicate if close to x the residuals are in average lower for one of the two methods."

Reply to Reviewer 2 comments

The authors are grateful to the reviewer for his comments which contributed to improving the quality of the paper. The authors provide hereafter the answers to the reviewer's comments. Please be aware that the numbering of the figures has change, which explained difference between the reviewer's comments and the answers provided by the authors.

Major comments

(1) Reviewer: The literature is not complete and does not state what other researchers have done in order to improve the flood estimation at ungauged sites. So, the authors should improve the description of the existing literature on the topic investigated. In particular, the manuscript should elaborate a little bit better on the evolution of the ROI method as the study focuses on the neighborhood approach for homogenous region delineation.

Answer: The authors would like to thank the Reviewer for the thorough review and the constructive comments. In the following we will provide a response to the comments formulated by the reviewer.

The authors agree with the reviewer and the revised version of the manuscript includes an improved description of the existing literature on ROI. Notice, however, that much effort in the recent literature on the ROI method deals with the problem of estimating the model by generalized least squares to account for different aspects, which is not the problem addressed in the present study. The main focus is more about the improvement of the ideas behind the CCA method, for which the recent developments are included in the introduction. The sentences below are added to the revised manuscript:

> P2-L24, "The traditional approach, based on the distance between site characteristics, is commonly referred to as the Region of Influence (ROI) model (Burn, 1990), which received a particular attention in the hydrological literature. The focus was mainly on the estimation of the model parameters, where for instance the generalized least-squares were used to account for unequal variability in the at-site estimations (e.g. Griffis and Stedinger, 2007; Stedinger and Tasker, 1985) and to deal with the presence of spatial correlation (e.g. Kjeldsen and Jones, 2009). "
>
> Kjeldsen, T.R., Jones, D.A., 2009. An exploratory analysis of error components in hydrological regression modeling. Water Resour. Res. 45, n/a–n/a. doi:10.1029/2007WR006283
>
> Griffis, V., Stedinger, J., 2007. The use of GLS regression in regional hydrologic analyses. J. Hydrol. 344, 82–95.
>
> Stedinger, J., Tasker, G., 1985. Regional hydrologic analysis: 1. Ordinary, weighted, and generalized least squares compared. Water Resour. Res. 21, 1421–1432. doi:10.1029/WR021i009p01421

(2.1) Reviewer: The methodology is blurred, difficult to follow, and contains some odd judgements. For instance, LSK was maintained because it is associated to better predictive performance, however, it is poorly predicted by the site characteristics (P11 L5 – 12).

Answer: The authors believe that the proposed methodology has relatively simple steps. The authors agree that the 3 initial main steps could be decomposed in more direct steps as above. In the revised version of the manuscript, the results section provides clearer indications of which step the discussion deals with.

The authors did not intend to impose a specific procedure for choosing the reference variables in the methodology. In the result section, the choice is made to adopt a backward stepwise selection procedure. This procedure is commonly used in regression modeling to select the explanatory variables. The effect of a reference variable on the final prediction is not straightforward and depends of the "interaction" with the other variables. For instance, two reference variables can be very well predicted, but if both contain the same information, one will be rejected by the procedure. In the present situation, the LSK is not well predicted, but it still appears to bring few information that is not contained in the other RV, which improves the final prediction. Therefore, LSK is included.

The part below is changed in the revised manuscript:

> P6-L21, "The general procedure can be described by the steps below:
>
> 1. Select the reference variables
> 2. If necessary, predict the reference variables that are not available at the target site
> 3. Calculate the distance between the reference variables
> 4. Form the neighborhood based on the previous distance
> 5. Fit a regional model on the neighborhood
> 6. Predict the target site
>
> In step 1, the selection of a set of the reference variables can be subjective and depend on the problem at hand. In the present study, backward stepwise selection procedure is considered to remove from an initial set of references variables those that are not contributing to the prediction power of the model. This selection procedure is more objective and depends on performance criteria that will be described in section 3.2.
>
> Step 2 is required only if some reference variables are unknown at the target sites, otherwise, if a target location designated by $i=0$, the radius of the neighborhood used in step 3 can be computed as $h_i = d\left(\mathbf{t}_i, \mathbf{t}_0\right)$ where d is a metric and $\mathbf{t}_i{}' = \left(t_{i,1}, \ldots, t_{i,q}\right)$ are the reference variables of the *i*th site. For simplicity, the Euclidian metric d is considered throughout the present study, but other metrics or dissimilarity measures could be employed as well. In particular, the Mahalanobis distance, the weighted distance and the depth function could be considered (Chebana and Ouarda, 2008; Cunderlik and Burn, 2006; Ouarda et al., 2000).

> If some hydrological information is unavailable at the target location, the estimation of the hydrological reference variables is necessary to produce an estimate $\mathbf{t}_0 = f(\mathbf{x}_0)$ in step 2 from site characteristics $\mathbf{x}_0$ at the target location. This substitution leads in step 3 to the distance $h_{(i)} = d\left[\mathbf{t}_i, f(\mathbf{x}_0)\right]$, which may be seen as an approximation of the true distance h_i. This study considers PPR models in order to fit every hydrological reference variable as described in section 2.3. The motivations for adopting PPR are that it does not require a prior delineation of regions, it accounts for nonlinear relationships, it has good predictive performances and it leads to a straightforward interpretation of the reference variables when a few directions α_k are necessary (Durocher et al., 2015)."

(2.2) Reviewer: Also, the authors did not show some details such as the additional translation which necessary to avoid numerical difficulties of LSK and LKT due to negative values (P10 L19).

Answer: The distributions of LSK and LKT are skewed but not positive. Hence the logarithm is used on a translated variable instead. In the revised version of the manuscript, the following sentences are modified to provide a mathematical formulation of the actual transformation:

> P11-L7, "These reference variables are transformed and standardized to obtain zero mean and unit variance. More precisely, the transformation for L1 and LCV is the logarithm and for LSK and LKT, the transformation is $\log(x - m_x + 1)$, where m_x is the minimum of the reference variables."

(3.1) Reviewer: Although the authors introduced a complicated methodology, they did not make enough efforts to clarify the description of the results; such as confusing explanation of Fig. 4 (P12 L8 – 17) (e.g., why 80 sites? in P12 L12), and unclear Fig. 5 and its explanation (P12 L26 – 33).

Answer: Figure 5 (initially Figure 4) presents the 4 cross-validation criteria with respect to the different possible neighborhood sizes. The main point of this approach is to show that selecting a calibrated size implies a trade-off between the different criteria. Additionally, to the Figures 5 and 8, the authors include Table 1 in the revised manuscript that summarized these criteria.

The authors agree that best NHS for RV-HYB at 80 sites is surprising. However, Figures 5a and 5b mostly show that regionalization is not very useful in terms of NHS, but it is important in terms of RRMSE. The authors use RRMSE as a calibration criterion.

In Figure 5, small changes in average can be difficult to assess by the naked eye. The red lines are smooth curves fitted between the residuals of the two models and are there to indicate which models in average have locally lower residuals. The explanation below is added to the revised manuscript:

> P13 L26, "As mentioned in section 4.2, previous studies have identified few problematic stations in the considered dataset. Figure 6 presents the residuals between different methods. As it may be difficult to see small improvements by uniquely observing points around the $y = x$ lines, the visualization of Figure 6 is helped by adding a flexible fit of the point cloud, using a standard smoothing spline approach. The resulting red lines indicate if close to x the residuals are lower in average for one of the two methods. In general, the points associated to the largest relative discrepancies are close to the $y = x$ line, which indicates that the sites that are difficult to predict are essentially the same for all methods. However, Figures 6a,b show that the RVN-HYB specifically

improves the prediction of the sites with the largest relative discrepancies as the red line is clearly located under the $y = x$ lines (left and right), which explains that this method leads to the best RRMSE. On the other hand, Figures 6c,d demonstrate that at the logarithmic scale, the RVN-LM method achieved predicted values that are mostly similar to the ROI and CCA methods, which explains the similarity of the NHS criteria for all the compared methods.

The present case study is an example of a region where some sites are problematic for likely any methods. In practice, the residuals are not known, consequently we do not know if the target sites of interest will be "problematic" or not. Globally, what Figure 6a indicates is that the RVN-HYB model is more robust (in a certain way), because for the sites that are well predicted by simpler models, such ROI, RVN-HYB will perform in average similarly. However, if the target site is predicted less accurately, the RVN-HYB model will (in average) be better in terms of RRMSE. Consequently, the overall gain may seem of moderate magnitude, but for some problematic stations the gain could be more substantial. In particular, the red lines in the left part of Figure 6a appears mostly influenced by two points, but the two improvements are of 77.2% and 68.5%, which is considerable."

(3.2) Reviewer: Furthermore, the presentation of the results of the regression-based model needs improvements to be clearer (P12 L35 – P13 L10). I recommend using the simple Q-Q plot to assess the compared methods regarding the estimation of regional flood quantile.

Answer: The authors want to highlight that the description of the steps of the regression-based model in the result section is voluntarily short because they are the same as the index-flood model. Based on the new 6 steps of the methodology (see comments 2.1), only the steps 5-6 change, which consists to fit a common linear model on the at-site quantile. The sentences below are added to the revised manuscript:

P14-L8, "Prediction of Q100 at the target location is also performed by the regression-based model using the same delineation methods as with the index-flood model, but with potentially different calibration values for the neighborhood sizes. Consequently, the description of steps 1-4 (in section 3.1) are identical to those of the index-flood approach and are not repeated here."

The authors agree with the reviewer and QQ plots (Figure 7) are included in the revised version of the manuscript to improve the analysis of the regression-based model:

P14-L10, "The fit of the regression-based model is graphically assessed in Figure 7 by Quantile-Quantile plots. It is showed that for all delineation approach the regression-based models correctly predict the flood quantile Q100 at target."

(3.3) Reviewer: Also, the results should contain numerical tables to quantitatively clarify the differences between the considered methods. The authors can find a close example for the presentation of such results in the reference Gado and Nguyen (2016). Finally, comparing the results of the index flood and the regression methods would be valuable here.

Answer: The added Table 1 now compared the index-flood and the regression-based model and the following sentences are added to the revised manuscript

P14 L24: "Table 1 provides also a comparison between the performance of the index-flood and the regression-based model. In terms of RRMSE and NHS criterion, the two approaches lead to very

> similar results, which is coherent with what it is reported in other studies (GREHYS, 1996a, 1996b; Haddad and Rahman, 2012). Therefore, similar conclusions can be draw from the two approaches. For instance, in both cases, the RVN-HYB leads to the best results in terms of RRMSE.“
>
> GREHYS, 1996. Presentation and review of some methods for regional flood frequency analysis. Journal of Hydrology 186, 63–84.
>
> GREHYS, 1996. Inter-comparison of regional flood frequency procedures for canadian rivers. Journal of hydrology(Amsterdam) 186, 85–103.
>
> 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

(4.1) Reviewer: The authors support the idea of using the estimation of hydrological variables, instead of site characteristics, to delineate homogeneous regions. Yet, the estimation of hydrological variables is based on subjective selections of site characteristics and subject to model errors.

The authors want to clarify that the idea itself of using hydrological variables, called here reference variables (RV), was not proposed in the present study. Implicitly, the traditional CCA method has suggested already to delineate homogenous regions using flood quantiles as reference variables. More precisely, the idea supported in the present paper is that a larger class of reference variables could be considered as well as different estimation methods.

The authors agree that there is uncertainty in the reference variables, but it does not represent an additional uncertainty for the proposed method. The ROI method has implicitly these uncertainties. If the predicted reference variables can be taken as the average of the ROI neighborhoods, as it is done with the index-flood model, then these predicted reference variables are not predicted without error. This fact is illustrated in Figures 4b,c,d of the revised manuscript where the average LSK and LCV are not centered at the target location. It is actually an advantage of the proposed method to explicitly model that uncertainty.

> P7-L28, “The errors related to prediction of the hydrological reference variables suggest that the RVN method may include an additional source of uncertainty, which is not accurate. Indeed, the same source of uncertainty is present among the sites of a neighborhood delineated on the basis of known site characteristics (i.e that the average of the hydrological variables in the neighborhood is not a perfect predictor). This can be seen as an advantage of the RVN method since it directly assesses this source of uncertainty and tries to reduce it.”

(4.2) Reviewer: Moreover, since, homogeneity tests (e.g. Hosking and Wallis 1997) are generally based on hydrological variables (e.g., L1, LCV), these variables should not be used in delineating homogeneous regions. In other words, the same information should not be used for both delineating the homogeneous regions and testing the homogeneity of such regions.

The authors understand the concern of the reviewer and agree that in the framework proposed by Hosking and Wallis (1997) the same hydrological variables could not have been used for delineating and testing the homogenous regions. However, the present methodology does not perform any homogeneity test. The criteria used for selecting the size of the neighborhood is the RRMSE and is based on cross-validation, which tends to optimize the prediction corresponding to a specific return period. Consequently, the L-moments are not the variables used in the calibration of the neighborhood.

(4.3) Reviewer: And this clarifies the good results regarding the improvements of the homogenous properties (i.e., the results of AHM and ARE) of the resulted neighborhoods by the new method, while the improvements in the results of the regional flood estimations are insignificant (i.e., the results of RMSE and NHS)

The authors thank the reviewer for raising this question, which gives them the opportunity to clarify this point. However, the authors do not agree that the terminology "insignificant improvements" properly describes the results of the present study. The authors have already argued that the results in terms of RRMSE are in fact substantial. Briefly, the improvement for the RVN-HYB method in comparison to the ROI method is of 6.1% in terms of the RRMSE and it is shown that the improvements are more important for sites that have larger discrepancies (See comment 2.2 of the first reviewer).

Indeed, better performances in terms of AHM and ARE are a direct consequence of using reference variables, which is the main purpose of the proposed methodology. The authors believe that the important point is not if these criteria are better, but how much better they are. In the revised version of the manuscript the author added the explanations below:

> P13-L15, "Figures 5c,d present respectively the AHM and the ARE criteria obtained from the considered methods. The AHM criterion indicates that the ROI and the CCA methods have in general lower heterogeneity than the whole dataset, but are outperformed by the RVN-LM and RVN-HYB methods especially for smaller neighborhoods. This quantifies the intuitive assumption that the regional LCV is calculated with less uncertainty when the L-moments are directly considered instead of other reference variables. In particular, the AHM of the ROI method is 72.8% with the optimal neighborhood size of 30. In comparison, the AHM of the RVN-LM method is 14.5% with the optimal neighborhood size of 29 sites, which is considerably lower. Figure 5c shows that the AHM criterion of the RVM-LM method does not reach a similar level to the ROI method until using as much as 120 sites. These results indicate that even for relatively small neighborhoods, the ROI method identifies regions that are only slightly less hydrologically heterogeneous than all sites pooled together. This suggests that, in the present case study, the ROI method has difficulties identifying sites that are similar to the target site in terms of LCV."

Minor comments

(1) Reviewer: P1 L16 – 17. Which properties does the hydrological information in Regional Frequency Analysis enforce for a group of gauged stations? I suggest to add "desired properties".

Answer: The modification is done in the revised version of the manuscript:

> P1-16, "This study investigates the utilization of hydrological information in Regional Flood Frequency Analysis (RFFA) to enforce desired properties for a group of gauged stations."

(2) Reviewer: P1 L18. Ungauged sites can be defined by site characteristics in the neighborhood delineation methods (e.g., ROI). Therefore, there is no a challenge for using neighborhoods in RFA regarding the unavailable hydrological information at ungauged sites.

Answer: The study of Oudin et al. (2010) shows that pooling sites together based on the similarity between the physiographical variables (i.e. site characteristics) does not necessarily lead to the same group of sites as if hydrological similarity was considered. Their study reports a case of 60% overlapping sites. In other words, it means that if one would like to pool together sites based on hydrological similarity to extrapolate the behavior of another site, but that one substitutes it by physiographical information, then 40% the identified sites would not be the ones that will have been chosen if the hydrological information was available.

Therefore, by "challenge" the authors mean that physiographical cannot replace the missing hydrological information completely. In the revised version of the manuscript, the following sentence :

> P1-L18, "A challenge for using neighborhoods in RFFA is that hydrological information is not available at target locations and it cannot be completely replaced by the available physiographical information."

Oudin, L., Kay, A., Andréassian, V., Perrin, C., 2010. Are seemingly physically similar catchments truly hydrologically similar? Water Resources Research 46, doi:10.1029/2009WR008887

(3) Reviewer: P1 L23. The regional frequency analysis can be applied for flood or extreme rainfall or any other extreme events. Hence, it should be stated that the case study is for regional flood estimation.

Answer: The authors agree to specify the scope of the study. In the revised version of the paper, the Acronym RFA for "Regional Frequency analysis" is replaced by RFFA for "Regional Flood Frequency Analysis".

(4) Reviewer: P2 L21. "the distance between hydrological variables". The distance is between locations not variables.

Answer: The authors would like to thank the reviewer for pointing out this blunder. The authors agree that the formulation needs to be changed and distances remain between locations not between variables. The sentence below is modified accordingly in the revised version of the manuscript:

> P2-L19, "To identify the most similar gauged sites in terms of hydrological properties, a notion of distance is needed to evaluate the proximity, or relevance, of each gauged site to the target location and identify the most hydrologically similar gauged sites. However, when the target location is ungauged, this distance cannot be directly calculated due to the missing hydrological information."

(5) Reviewer: P3 L4. "as an estimation model"

Answer: The modification is done in the revised version of the manuscript.

> P3-L6: "This is achieved by replacing CCA in the prior analysis of hydrological variables by Projection Pursuit Regression (PPR), a nonparametric regression method recently considered as an estimation model in RFFA (Durocher et al., 2015)."

(6) Reviewer: P9 L4. Please, define NH in equation 14.

Answer: The authors agree that the notation is not clear. In Eq. (14), NH corresponds to the product of the variables N and H, which are defined. N corresponds to the number of gauged sites in the neighborhoods and H is the heterogeneity measure in Eq. (13) calculated on all n available gauged sites.

In the revised version of the manuscript a product symbol is added to Eq. (14) to clarify that it is the product of two separate variables:

$$\mathrm{AHM} = \frac{1}{N \cdot H} \sum_{j=1}^{n} H_{(j)}$$

(7) Reviewer: P9 L14 – 15. How can a regression model fitted on two different neighborhoods, for the same target location, obtain identical values?

Answer: A simple illustration would be to consider sites with value: {1,2,3,4,5}. Two delineation methods lead to the group {1,3,5} and {2,3,4}. Here, both groups have predicted value 3 as their mean, but the first group has a variance of 4 and the second has a variance of 1.

The following sentence will be modified in the revised version of the manuscript to clarify that "identical values" stands for "similar predicted values":

> P10-L3, "Notice that a regression model fitted on two different neighborhoods (for the same target location) can lead to very similar predictions, but with different levels of variance."

(8) Reviewer: P10 L2. I don't believe that 15 years of data are enough to get statistically reliable results, why did authors choose 15 years as the minimum time series used in the study.

Answer: This choice has not been made in the present study. We are using in this study a database that has been used in a number of previous studies. This allows to compare the results with those obtained with other methods that are now commonly accepted. This is explained in the revised manuscript:

> P10-L20, "Only a brief description of the data and the at-site frequency analysis is provided since the elements were presented in details in previous studies (e.g., Chokmani and Ouarda, 2004)."

Notice also that 15 years is a minimum. Most time series are much longer. The average length is added to the revised version of the manuscript to highlight this point:

> P10-L19, "Each site has at least 15 years of data available, with an average length of 31 years."

(9) Reviewer: P10 L3. I think you should have at least a map showing the locations of the selected stations in the case study (Quebec).

Answer: The authors agree with the reviewer and a new map (Figure 2) is added to the revised version of the manuscript:

> P10-L18, "The analysis is performed on 151 sites located in Southern Quebec, Canada, which are presented in Figure 2."

(10) Reviewer: P10 L7. Using the maximum likelihood for parameter estimation with small time series (e.g., 15 years) may cause convergence problems, I would recommend using L-moments instead.

Answer: As mentioned in the minor comment (8), the at-site frequency analysis was performed and validated in previous studies, where the necessary precautions were taken to ensure reliable estimates. Moreover, notice the use of the terms "including" and "In general" in the following sentences of the revised manuscript to underline that only a brief description of the full methodology is provided:

> P10-L23, "The at-site distributions are selected among several families including: generalized extreme values (GEV), Pearson type III (P3), generalized logistic (GLO) and log-normal with 3 parameters (LN3). In general, the estimation of the at-site distribution was achieved by maximum likelihood and the final choices of distributions are based on the Akaike information criterion."

(11) Reviewer: P11 L16. What does HYB denote for in "RVN-HYB"?

Answer: The acronym "HYB" stands for hybrid, because hydrological and physiographical variables are used as reference variables. The following sentences will be added to the revised version of the manuscript to clarify this point:

> P11-L4, "The first group is based on L-moments only and the second is based on the combination of L-moments and site-characteristics. The acronym LM for L-moment and HYB for Hybrid are used to identify the two groups."

(12) Reviewer: P11 L19. "One of the objectives of RFA is to identify a proper family of distributions from regional information" This is not an objective of the RFA. I suggest to write one of the main steps.

Answer: In the revised version of the manuscript the sentence has been modified as follows:

> P12-12, "Moreover, it is necessary to identify a proper family of distributions from regional information, which is achieved here by analyzing the distribution of the gauged sites inside the neighborhoods."

(13) Reviewer: Please, define here the Q(r) as the regional quantile. The authors defined the Q(r) on P11 L29 but it still needed to be defined immediately after the equation in P11 L23.

Answer: The author agrees with the reviewer. The change indicated by the reviewer is done in the revised version of the manuscript:

> P12-L18, "In this model, the regional quantile $Q_i(r) = \mu_i Q(r)$ corresponding to a return period r at a target location i, where μ_i is the index-flood."

(14) Reviewer: The second part of the title of figure 3: (b), (c), and (d) Regional L-moments based on the 15 nearest gauged sites for 3 selected target locations.

The correction is made in the revised version of the manuscript. The authors reiterate their thanks to the reviewer.

> "Figure 4: L-moments ratio diagram for index-flood model. (a) Regional L-moments for RVN-LM with 29 gauged sites. (b),(c) and (d) Regional L-moments based on the 15 nearest gauged sites for 3 selected target locations. "

Delineation of homogenous regions using hydrological variables predicted by projection pursuit regression

Martin Durocher [1], Fateh Chebana [2], Taha B. M. J. Ouarda[2,3]

[1]Université du Québec à Trois-Rivières, University of Quebec,
3351, boul. des Forges, C.P. 500, Trois-Rivières, G9A 5H7, Canada

[2]Institut National de Recherche Scientifique (INRS-ETE), University of Quebec,
490 de la Couronne, Québec G1K 9A9, Canada

[3] Institute Center for Water Advanced Technology and Environnemental Research (iWater),
Masdar Institue of Science and Technology, P.O. Box 54224, Abu Dhabi, UAE

[*]*Corresponding author:* Martin Durocher (martin.durocher@uqtr.ca)

Abstract

This study investigates the utilization of hydrological information in Regional Flood Frequency Analysis (RFFA) to enforce desired properties for a group of gauged stations. Neighborhoods are a particular type of regions that are centered on target locations. A challenge for using neighborhoods in RFFA is that hydrological information is not available at target locations and it cannot be completely replaced by the available physiographical information. Instead of using known site characteristics (not hydrological) to define the center of a target location, this study proposes to introduce estimates of (hydrological) reference variables to ensure better homogeneity. These reference variables represent nonlinear relations with the site characteristics obtained by projection pursuit regression; a nonparametric regression method. The resulting neighborhoods are investigated in combination with common regional models: the index-flood model and the regression-based models. The complete approach is illustrated on a real-world case study with gauged sites located in Southern Quebec, Canada, and is compared with the traditional approaches "Region of Influence" and "Canonical Correlation Analysis". The evaluation focuses on the neighborhood properties as well as prediction performances, with special attention to problematic stations. Results show clear improvements in neighborhood definitions and quantile estimates.

Keywords: Index-flood model, Regional frequency analysis, Ungauged site, Region of influence, Projection pursuit regression, Canonical Correlation Analysis.

1. Introduction

Accurate estimates of the risk of occurrence of extreme hydrological events are necessary for the minimization of the impacts of these events and for the optimal design and management of water resource systems. However, necessary information is not always available at the sites of interest. Hence, it is necessary to develop procedures to transfer, or to regionalize, the information available at existing gauged sites to the ungauged ones. Regional Flood Frequency Analysis (RFFA) represents a large class of techniques commonly used in water sciences to evaluate the risk of occurrence of extreme hydrological phenomena of rare magnitudes at ungauged locations (Haddad and Rahman, 2012; Hosking and Wallis, 1997; Laio et al., 2011; Pandey, 1998; Reis et al., 2005).

RFFA methods are usually composed of two main steps. The first step is the formation of homogenous regions. This step aims at pooling together sites that are approximately similar according to homogenous criteria. Inside these homogenous regions, it is assumed that hydrological information can be reasonably transferred from gauged to ungauged locations (Cunnane, 1988). The second step, the estimation of flood quantiles, consists in the calibration of a regional model that characterizes the interrelation between hydrological variables of interest and explanatory physio-meteorological variables corresponding to known site characteristics. Consequently, RFFA is used to study unobserved hydrological behaviour from available hydrological and physio-meteorological information.

Neighborhoods are specific forms of regions inside which gauged sites are not classified into fixed regions, but are composed of gauged sites that are the most similar to a given target. Hence, two distinct target locations have their own neighborhoods that may overlap. Comparative studies showed that neighborhoods lead to better regional estimates than fixed regions (Burn, 1990; Ouarda et al., 2008; Tasker et al., 1996). To identify the most similar gauged sites in terms of hydrological properties, a notion of distance is needed to evaluate the proximity, or relevance, of each gauged site to the target location and identify the most hydrologically similar gauged sites. However, when the target location is ungauged, this distance cannot be directly calculated due to the missing hydrological information. Physio-meteorological information is hence used for similarity evaluation. The traditional approach, based on the distance between site characteristics, is commonly referred to as the Region of Influence (ROI) model (Burn, 1990), which received a particular attention in the hydrological literature. The focus was mainly on the estimation of the model parameters, where for instance generalized least-squares were used to account for unequal variability in the at-site estimations (e.g. Griffis and Stedinger, 2007; Stedinger and Tasker, 1985) and to deal with the presence of spatial correlation (e.g. Kjeldsen and Jones, 2009).

Alternatively, Ouarda et al. (2001) used Canonical Correlation Analysis (CCA) to build neighborhoods from a canonical distance that accounts for the interrelation between flood quantiles and site characteristics. For this method, neighborhoods are formed by gauged sites that are the most similar to the target location, according to the distance between vectors of flood quantiles corresponding to different return periods. Due to the missing hydrological information, the CCA method in RFFA estimates the unavailable hydrological variables as linear combinations of site characteristics. Consequently, the available site characteristics are transformed into more meaningful "hydrological" quantities for the purpose of delineating neighborhoods. However, the CCA method suffers from some limitations, such as linearity and normality assumptions (He et al., 2011). Subsequent studies aimed to improve the CCA method by improving the CCA technique itself (Chebana and Ouarda, 2008; Ouali et al., 2015). However, little attention has been paid to the importance of properly choosing the hydrological quantities in the delineation step whereas much effort has been devoted to the modeling step. Indeed, Chebana and Ouarda (2008) employed an iterative linear procedure to estimate neighborhood centers and they showed that the quality of these centers' estimates is the crucial element to improve the final model performance.

This study aims to provide a general framework with more flexibility regarding the linearity and normality assumptions. This is achieved by replacing CCA in the prior analysis of hydrological variables by Projection Pursuit Regression (PPR), a nonparametric regression method recently considered as an estimation model in RFFA (Durocher et al., 2015). The present study is also interested in validating the advantages of employing hydrological variables other than the at-site flood quantiles in prior modeling as well as considering a combination of these hydrological variables with site characteristics.

L-moments have already been used in RFFA to test the homogeneity of fixed regions when the target site is gauged (Chebana and Ouarda, 2007; Hosking and Wallis, 1997). In the present study, the prediction of the L-moments at ungauged sites is also considered to improve the delineation of the neighborhoods by reducing uncertainties. Moreover, a conceptual advantage of using L-moments conversely to at-site flood quantiles is that the L-moments do not depend on the subjective selection of at-site distributions.

The present paper is organized as follows. Section 2 presents the background for the techniques commonly used in RFFA. Section 3 elaborates on the prior analysis of hydrological variables and their integration with the techniques presented in Section 2 to form a complete procedure. Section 3 suggests criteria for the evaluation of the predictive performances and the neighborhood properties. Section 4 illustrates the application of the method on a case study. Traditional ROI and CCA methods serve as references in order to evaluate the relative performance of the investigated method. Finally, concluding remarks are provided in the last section.

2. Background

2.1 Delineation of neighborhoods

In RFFA, neighborhoods are used to identify gauged sites from which information is transferred to the target location. A neighborhood is characterized by a center and a radius that delimits an area (not necessary in the geographical sense). Gauged sites inside the area delineate a region that includes relevant sites to the target location. At each site $i = 1,\ldots,n$, p characteristics $\mathbf{x}_i = \left(x_{i,1},\ldots,x_{i,p}\right)$ are available. Typically, the ROI method forms neighborhoods according to a radius based on a metric d:

$$d\left(\mathbf{x}_i,\mathbf{x}_j\right) = \sqrt{\sum_{k=1}^{p} \frac{(x_{i,k} - x_{j,k})^2}{\sigma_k{}^2}} \qquad (1)$$

where σ_k is the standard deviation of $\left\{x_{i,k}\right\}_{i=1}^{n}$ the kth site characteristic (Eng et al., 2005).

Alternatively, CCA is a multivariate technique used to unveil the interrelation between two groups of variables. Let Y and X be normally distributed random vectors with zero means. The CCA method defines canonical pairs $\left(U_k, V_k\right)$ as linear combinations of the original random variables:

$$U_k = a_k X \tag{2}$$

$$V_k = b_k Y \tag{3}$$

where the correlations $\rho_k = corr\left(U_k, V_k\right)$ are sequentially maximal for $k = 1, \ldots, K$ under the conditions $corr\left(U_k, U_l\right) = corr\left(V_k, V_l\right) = 0$ for $k \neq l$. Only the canonical pairs $\left(U_k, V_k\right)$ with unit variances are considered.

To delineate neighborhoods, the CCA approach considers the canonical scores $\mathbf{u}_i = \left(a_1, \ldots, a_r\right)'\mathbf{x}_i$ and $\mathbf{v}_i = \left(b_1, \ldots, b_r\right)'\mathbf{y}_i$ that are respectively linear combinations of site characteristics $\mathbf{x}_i$ and flood quantiles corresponding to different return periods $\mathbf{y}_i$ for site *i*. Due to the missing hydrological information at the ungauged location denoted $i = 0$, the flood quantiles $\mathbf{y}_0$ and the corresponding linear combination $\mathbf{v}_0$ are unknown. Nevertheless, CCA provides a linear estimate $\mathbf{v}_0 \approx \Lambda \mathbf{u}_0$, where $\Lambda = \mathrm{diag}(\rho_1, \ldots, \rho_K)$. Accordingly, a neighborhood is delineated in the canonical space according to the distance:

$$d(\mathbf{v}_i, \Lambda \mathbf{u}_0) = \left(\mathbf{v}_i - \Lambda \mathbf{u}_0\right)'\left(I - \Lambda^2\right)^{-1}\left(\mathbf{v}_i - \Lambda \mathbf{u}_0\right) \tag{4}$$

More details on the CCA approach in RFFA can be obtained in Ouarda et al. (2001).

2.2 Multiple regression

In RFFA, two types of regional models are often considered to predict flood quantiles corresponding to given return periods: the index-flood model and the regression-based model (Ouarda et al., 2008). The index-flood model predicts a target distribution by assuming that all distributions inside a region are proportional to a regional distribution, up to a scale factor called index-flood. The flood quantile of interest at a target location is then calculated from the regional distribution based on the predicted index-flood (e.g., Chebana and Ouarda, 2009; Dalrymple, 1960; Stedinger and Lu, 1995). Conversely, the regression-based model considers directly the at-site estimates of the desired flood quantiles for prediction. Flood quantiles are then predicted at their target locations by the regression equations estimated within the neighborhoods (Pandey and Nguyen, 1999).

Even though they proceed differently, both the index-flood model and the regression-based model may use the same multiple regression techniques to transfer information to an ungauged location. For the sake of simplicity, the term hydrological variables is used to designate the corresponding output variables z_i of these models at location $i = 1, \ldots, n$. Consequently, for the index-flood model, z_i is the index flood, while for a regression-based model the hydrological variable z_i is the flood quantile of interest.

Multiple regression models assume linear interrelation between the hydrological variable z_i and the site characteristics $\mathbf{x}_i$. Consequently, in several cases, transformations are necessary to meet this assumption. For instance, the power law form is frequently used to model flood quantiles:

$$z_i = e^{\beta_0} \times x_{i,1}{}^{\beta_1} \times \ldots \times x_{i,p}{}^{\beta_p} \times \varepsilon_i \tag{5}$$

where $\beta' = \left(\beta_0, \beta_1, \ldots, \beta_p\right)$ are parameters and ε_i is an error term. Applying a logarithmic transformation is sufficient to cast Eq. (5) into a linear model. In general, a proper transformation is assumed for the hydrological variables $y_i = g\left(z_i\right)$ being linearly related to the sites characteristics.

According to previous notations, let $\mathbf{y} = \left(y_1, \ldots, y_n\right)$ be the hydrological variables, $\mathbf{X}$ be the design matrix of the site characteristics $x_{i,j}$ with intercept, and $\varepsilon = \left(\varepsilon_1, \ldots, \varepsilon_n\right)$ be the error term. Hence in matrix notation, a multiple regression model has the form:

$$\mathbf{y} = \mathbf{X}\beta + \varepsilon \tag{6}$$

and according to the least-squares theory, the estimates of the parameters are:

$$\hat{\beta} = \left(\mathbf{X}'\mathbf{X}\right)^{-1} \mathbf{X}'\mathbf{y} \tag{7}$$

2.3 Projection pursuit regression

Some methods predict hydrological variables without the formation of regions, such as physiographical kriging (Castiglioni et al., 2009; Chokmani and Ouarda, 2004), generalized additive models (Chebana et al., 2014) and artificial neural networks (Dawson et al., 2006; Ouarda and Shu, 2009). More recently, Projection Pursuit Regression (PPR) was introduced to provide a flexible nonparametric regression approach to describe the nonlinearity that is present in the relationship between hydrological variables and site characteristics. PPR was used in the RFFA context by Durocher et al. (2015) to directly predict flood quantiles without delineation.

The basic elements of a PPR model are $k = 1, \ldots, m$ functions f_k called terms and defined as:

$$f_k\left(\mathbf{X}\right) = g_k\left(\alpha_k'\mathbf{X}\right) \tag{8}$$

where directions α_k are vectors of coefficients and g_k are smooth functions. The directions α_k are coefficients that respect $\left|\alpha\right| = 1$ and determine a predictor $\alpha_k'\mathbf{X}$ as relevant linear combinations of the site characteristics $\mathbf{X}$. The terms are then combined into a regression model:

$$\mathbf{y} = \mu + \sum_{k=1}^{m} f_k(\mathbf{X}) + \varepsilon \quad (9)$$

where μ is the global mean and ε is a term of error. Notice that the orthogonality between directions α_k is not imposed, hence the predictors $\alpha_k'\mathbf{X}$ and $\alpha_l'\mathbf{X}$ for $k \neq l$ may be correlated. Consequently, PPR allows for interaction between site characteristics, which leads to a large variety of regression models (Hastie et al., 2009).

The components α_k and g_k of the model in (9) are estimated by the least-squares approach (Friedman et al., 1983). For a unique direction ($m=1$), PPR can be estimated by standard nonlinear algorithms (Yu and Ruppert, 2002), but in general a stagewise algorithm is adopted to find a proper solution (Friedman and Tukey, 1974). Comparative studies show that PPR has a similar predictive performance to artificial neural networks (Bishop, 1995; Hwang et al., 1994). However, Durocher et al. (2015) indicated that in RFFA, PPR reduces to more parsimonious models than artificial neural networks, which provides an explicit expression of the regression equations.

3 Methodology

This study deals with neighborhood delineation and more precisely it focuses on the identification of reliable estimates of the hydrological centers of these neighborhoods. For simplicity, the variables forming these centers will be referred to as reference variables, because they represent the reference to evaluate the similarity between a target location and the gauged sites. Reference variables can take different forms, such as site characteristics, hydrological variables or a combination of both. Their nature is important, because it determines the properties that are deemed to be important between close sites. The particularity of the present method is that PPR can be used to predict these neighborhood centers (prior to the RFFA modeling step) when some of the reference variables are unknown hydrological variables. Accordingly, the proposed method will be referred to as RVN for Reference Variable Neighborhoods.

3.1 Estimation of the reference variables

The general procedure can be described by the steps below:

1. Select the reference variables
2. If necessary, predict the reference variables that are not available at the target site
3. Calculate the distance between the reference variables
4. Form the neighborhood based on the previous distance
5. Fit a regional model on the neighborhood
6. Predict the target site

In step 1, the selection of a set of the reference variables can be subjective and depends on the problem at hand. In the present study, backward stepwise selection procedure is considered to remove from an initial set of references variables those that are not contributing to the prediction power of the model. This selection procedure is more objective and depends on performance criteria that will be described in section 3.2.

Step 2 is required only if some reference variables are unknown at the target sites, otherwise, if a target location designated by $i = 0$, the radius of the neighborhood used in step 3 can be computed as $h_i = d\left(\mathbf{t}_i, \mathbf{t}_0\right)$ where d is a metric and $\mathbf{t}_i' = \left(t_{i,1}, \ldots, t_{i,q}\right)$ are the reference variables of the *i*th site. For simplicity, the Euclidian metric d is considered throughout the present study, but other metrics or dissimilarity measures could be employed as well. In particular, the Mahalanobis distance, the weighted distance and the depth function could be considered (Chebana and Ouarda, 2008; Cunderlik and Burn, 2006; Ouarda et al., 2000).

If some hydrological information is unavailable at the target location, the estimation of the hydrological reference variables is necessary to produce an estimate $\mathbf{t}_0 = f(\mathbf{x}_0)$ in step 2 from site characteristics $\mathbf{x}_0$ at the target location. This substitution leads in step 3 to the distance $h_{(i)} = d\left[\mathbf{t}_i, f(\mathbf{x}_0)\right]$, which may be seen as an approximation of the true distance h_i. This study considers PPR models in order to fit every hydrological reference variable as described in section 2.3. The motivations for adopting PPR are that it does not require a prior delineation of regions, it accounts for nonlinear relationships, it has good predictive performances and it leads to a straightforward interpretation of the reference variables when a few directions α_k are necessary (Durocher et al., 2015).

If the hydrological variables $\mathbf{t}_0$ were known at the target location, the distance h_i would be available and the neighborhood that truly regroups the most hydrologically similar sites to the target location can be identified. However, in practice this true neighborhood is unknown. Using instead the estimate $f(\mathbf{x}_0)$ has the effect that some sites are falsely suggested as more hydrologically similar than other sites. Figure 1 illustrates a region with several sites where two neighborhoods are resulting from the RVN method with different predicted centers. The target site is illustrated as a green filled circle and neighborhood is formed of the 10 nearest sites indicated by small empty circles. The other sites are designated by crosses. The red and blue neighborhoods are delineated by circles where the radius is selected to include the 10 nearest sites. The predicted center of the red neighborhood is closer to the target site. Consequently, it can be seen that except from one site, the same sites as the target neighborhood are included (empty circles). On the other hand, the blue neighborhood has a predicted center further to the target site and hence a lower proportion of the sites truly closer to the target are found. It shows the importance of correctly predicting the neighborhood centers in order to identify sites that are truly similar to the target site.

The errors related to prediction of the hydrological reference variables suggest that the RVN method may include an additional source of uncertainty, which is not accurate. Indeed, the same source of uncertainty is present among the sites of a neighborhood delineated on the basis of the site characteristics (*i.e* that the average of the hydrological variables in the neighborhood is not a perfect predictor). This could be seen as an advantage of the RVN method since it directly assesses this source of uncertainty and tries to reduce it.

Step (1-3) are the particularity of the RVN method, while the other steps are common in RFFA and are explained in section 2. In the remainder of this study, step (4) uses a specific type of neighborhoods that is composed of a fixed number of the nearest sites (Eng et al., 2005; Tasker et al., 1996), but could also be constrained to the degree of the homogeneity of the neighborhoods (Ouarda et al., 2001). Consequently, the selected gauged sites can be obtained by sorting $h_{(i)}$ and keeping the desired number of sites. Notice that even though $h_{(i)}$ does not exactly approximate h_i, both distances will lead to the same neighborhoods if they preserve the ranks. Finally, step (5) consists in the estimation of the flood quantiles using either the index-flood or the regression-based model.

Notice that the RVN method may be seen as a generalization of the ROI and the CCA methods in RFFA. Indeed, the ROI method corresponds to the RVN method for which all the reference variables are site characteristics. In that case, $\mathbf{t}_0 = f(\mathbf{x}_0)$ is known and PPR is not necessary in step (2). Similarly, the CCA approach may be seen as the special case for which the reference variables are the canonical pairs in Eq. (4) and CCA is used, instead of PPR, to predict them in step (2).

3.2 Evaluation criteria

For the RVN method presented above, the neighborhood sizes must be calibrated according to an objective criterion. In this regards, the leave-one-out cross-validation approach is a general strategy to assess the performance of the predicted hydrological variables z_i at site $i = 1,\ldots,n$. In turn, each gauged site i is considered as an ungauged target location. From the remaining gauged sites, predicted values $z_{(i)}$ can be obtained without using the hydrological information at the target location. Discrepancies between the sampled and the predicted values are used to define evaluation criteria. Notice that the hydrological variables are transformed $y_i = g\left(z_i\right)$. Hence, if $\bar{y}$ is the sample mean of the y_i, then an appropriate global performance measure is the Nash-Sutcliffe criterion:

$$\text{NHS} = 1 - \frac{\sum_{i=1}^{n}\left[y_i - y_{(i)}\right]^2}{\sum_{i=1}^{n}\left[y_i - \bar{y}\right]^2} \tag{10}$$

Additionally, the predictive performance is examined at the original scale by the relative root mean square error:

$$\text{RRMSE} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}\left(1 - \frac{z_{(i)}}{z_i}\right)^2} \tag{11}$$

The choice of the reference variables is an important aspect and a set of reference variables should be chosen in order to enforce the desired properties. For instance, with the index-flood model the assumption of a regional distribution suggests that, apart from the index-flood, the at-site distributions must be proportional to a regional distribution. A heterogeneity measure based on the dispersion of the L-coefficient of variation (LCV) is shown to be a proper way to ensure that the LCV is relatively constant (Viglione et al., 2007). Accordingly, let I_j be the set of indices for the N nearest gauged sites to the target location j during the cross-validation process. The regional LCV $\hat{\theta}_{(j)}$ is calculated as the average:

$$\hat{\theta}_{(j)} = \frac{1}{N} \sum_{i \in I_j} \theta_i \tag{12}$$

of the at-site LCV θ_i inside the jth region. The heterogeneity measure is defined as:

$$H_{(j)} = \sum_{i \in I_j} \left(\theta_i - \hat{\theta}_{(j)} \right)^2 \tag{13}$$

In their procedure, Hosking and Wallis (1997) used this heterogeneity measure to test the homogeneity of a region, which implies that the regional LCV can be considered constant. Hence, the result of this test allows deciding if a region must be divided into smaller and more homogenous sub-regions. In the present study the size of the neighborhoods is the same for every neighborhood. Hence, if a homogeneity test is performed with a given neighborhood size, some of the neighborhoods will be considered homogenous, while the others will be considered heterogeneous (Das and Cunnane, 2010). However, the heterogeneity measure in Eq. (13) remains a useful indicator of dispersion for the regional LCV $\hat{\theta}_{(j)}$ inside a neighborhood. Consequently, a smaller $H_{(j)}$ suggests that the regional LCV $\hat{\theta}_{(j)}$ is measured with less uncertainty.

To facilitate the interpretation of the results and to ensure the comparability between neighborhoods, the heterogeneity measure $H_{(j)}/N$ is considered instead. The measure represents the sample variance of the LCV for the jth target location. This heterogeneity measure is standardized by H/n, where H is the heterogeneity measure in Eq. (13) calculated on all n available gauged sites. The resulting ratio corresponds to a scale-free heterogeneity measure, where a value under one provides evidence of a less heterogeneous neighborhood in comparison to the whole dataset. Therefore, the Average Heterogeneity Measure (AHM) criterion below is defined as the average of every neighborhood considered in the cross-validation process:

$$\mathrm{AHM} = \frac{1}{N \cdot H} \sum_{j=1}^{n} H_{(j)} \tag{14}$$

This criterion is not specific to a given target location, but represents the global level of heterogeneity resulting from a given delineation method, such as ROI, CCA or RVN. In particular, a delineation method with a smaller AHM suggests that on average a more precise regional LCV is used to predict flood quantiles.

Another desired property for a neighborhood is to lead to estimation models with less uncertainty. For the index-flood model, this implies in particular less uncertainty in the prediction of the index-flood, while for regression-based models, it implies less uncertainty in the prediction of flood quantiles. For a multiple regression model, the uncertainty can be quantified by the residual variance:

$$s_{(j)}^2 = \frac{1}{N} \sum_{i \in I_j} \left(e_{i,(j)} \right)^2 \qquad (15)$$

where $e_{i,(j)}$ is the residual at the ith gauged site, when predicting the jth target location in the cross-validation process. Notice that a regression model fitted on two different neighborhoods (for the same target location) can obtain an identical values, but lead to different levels of uncertainty. In this study, a neighborhood with a smaller residual variance than another one is said to be relatively more efficient.

During the cross-validation process, the sample variance of the regression models can be calculated for every site, which leads to the Average Relative Efficiency (ARE) criterion defined by:

$$\mathrm{ARE} = \frac{1}{ns^2} \sum_{j=1}^{n} s_{(j)}^2 \qquad (16)$$

where the residual variance s^2 is calculated from the multiple regression model on the whole dataset. This criterion is similar to the AHM criterion as it is standardized to a scale-free measure. This criterion can be used to identify the delineation method which achieved on average the smallest residual variances for each neighborhood. The ARE and the AHM criteria are used in the present study, along with the NHS and RRMSE to access the performances of the various models.

4. Applications

4.1 Data

To validate the RVN method on a practical situation, RFFA is carried out in a real-world case study using both the index-flood model and the regression-based model. The hydrological variables of interest are the flood quantiles corresponding to a return period of 100 years, denoted Q100. The analysis is performed on 151 sites located in Southern Quebec, Canada, which are presented in Figure 2. Each site has at least 15 years of data available, with an average length of 31 years. Furthermore, the usual hypotheses of stationarity, homogeneity and independence are verified. Only a brief description of the data and the at-site frequency analysis is provided since the elements were already presented in details in previous studies (e.g., Chokmani and Ouarda, 2004).

The at-site distributions are selected among several families including: generalized extreme values (GEV), Pearson type III (P3), generalized logistic (GLO) and log-normal with 3 parameters (LN3). In general, the estimation of the at-site distribution was achieved by maximum likelihood and the final choices of distributions are based on the Akaike information criterion. Recent studies on the same dataset have identified 4 relevant site characteristics (Chebana et al., 2014; Durocher et al., 2015), which are used in the present analysis: the drainage area or BV (km^2), the fraction of the basin area occupied by lakes or PLAC (%), the annual mean liquid precipitation or PLMA (mm) and the longitude or LON. Proper transformations are applied on these site characteristics in order to obtain approximately standard normal distributions (Chokmani and Ouarda, 2004).

4.2 Determination of the neighborhood centers

The step 1-2 of the RVN method is the selection of the reference variables and, if necessary, the estimation of the hydrological reference variables at the target locations. Two initial groups of reference variables are considered. The first group is based on L-moments only and the second is based on the combination of L-moments and site-characteristics. The acronym LM for L-moment and HYB for Hybrid are used to identify the two groups. More precisely, the L-moments considered for both groups are the sample average (L1), the LCV, the L-coefficient of skewness (LSK) and the L-coefficient of kurtosis (LKT). These reference variables are transformed and standardized to obtain zero mean and unit variance. More precisely, the transformation for L1 and LCV is the logarithm and for LSK and LKT, the transformation is $\log(x-m_x+1)$, where m_x is the minimum of the reference variables. Moreover, a specific implementation of PPR is assumed, which considers the smooth functions g_k in Eq. (8) as cubic spline polynomials with 5 equally spaced knots. The number of knots is validated by cross-validation using the NHS criterion. Notice that for the fitting of LSK, one site has a very low standardized residual of approximately -6. Consequently, this site is considered as an outlier and removed from the estimation of the reference variables. In previous studies (e.g., Chokmani and Ouarda, 2004), this site was identified as one of a few problematic sites that are difficult to predict due to an underestimated drainage area or overevaluated percentage of area covered by lakes. Nevertheless, in the present study, this site is only removed only during the prediction of the reference variables and all sites are included in the rest of the analysis.

Figure 3 shows the fitting of the four reference variables by the PPR models. Cross-validation has selected PPR models with a unique direction α for all reference variables. The PPR equations that describe the relation between the reference variables and the site characteristics are explicit, for instance, the regression equation for the LCV has the form:

$$\begin{aligned}\log(\mathrm{LCV}) = -1.80 + 0.26 \times f\big[&-0.67\times\log(\mathrm{BV}) - 0.09\times\sqrt{PLAC}\\ &+1.27\times\log(\mathrm{PLMA}) + 0.06\times\mathrm{LON} - 1.32\big]\end{aligned} \tag{17}$$

Notice the constant term -1.32 and the norm of direction $|\alpha| \neq 1$ inside the function f in Eq. (17). The difference in (17) in comparison to the general form of the PPR model in Eq. (9) is the consequence of transformations on the explanatory variables. Indeed, during the optimization procedure of a PPR model, it is suggested to scale the explanatory variables in order to avoid the scale effect in the coefficients of the direction α (Hastie et al., 2009). Nevertheless, notice that the formula inside the function f corresponds to a linear model.

Figure 3a shows a strong linear relationship between L1 and the predictor $\alpha'\mathbf{X}$. Conversely, Figures 3b,c,d show mild nonlinearity and hence indicate the need for more flexible models, such as PPR. The predictive performances of the reference variables are evaluated by the NHS criterion with values 91.5%, 33.3%, 6.7% and 55.7% respectively for L1, LCV, LSK and LKT. These results show that L1 is accurately predicted by the site characteristics, while a poor fit is associated to LSK. Indeed, Figure 3c suggests that apart from a few sites on the right of the curve, LSK appears not highly related to the predictor $\alpha'\mathbf{X}$. In comparison, linear models applied on the same reference variables lead to NHS criterion: 90.9%, 28.2%, 7.8% and 48.1% respectively. Remark that NHS criterion is calculated by cross-validation, consequently even though the improved performances by the PPR method appear moderate this represent true fitting improvements.

Due to its poor fit, LSK may not be a proper reference variable for the delineation step. To validate this assumption, the neighborhoods are formed with and without using LSK and the rest of the analysis is carried out for both scenarios. Based on the RRMSE criterion, LSK must be maintained as it is associated to better predictive performances. Similarly, the same procedure is applied to validate in turn the usefulness of each reference variable, which leads to discarding LKT and to maintaining L1, LCV and LSK.

The second group of reference variables contains both the L-moments and the site characteristics. As with the first group, the complete analysis is performed with and without each of the reference variables. The final reference variables that are kept are: BV, PLAC, LCV and LSK. In order to distinguish the two groups of reference variables, RVN-LM will designate the first group with the L-moments only and RVN-HYB will designate the second group with both the L-moments and the site characteristics.

4.3 Results of the index-flood model

At this point, the steps 1-4 of the RVN methodology are performed and the neighborhoods are identified. Notice that for the RVN-LM method, the reference variables include the first three L-moments, which could be used as a moment estimator to deduce the target distribution. This approach is, however, not generally applicable to the present methodology as the reference variables are selected by a stepwise procedure. Moreover, it is necessary to identify a proper family of distributions from regional information, which is achieved here by analyzing the distribution of the gauged sites inside the neighborhoods. The index-flood model and the L-moments algorithm were proven to lead to a reliable procedure to identify a regional distribution and to estimate its parameter (Hosking and Wallis, 1997). In this model, the regional quantile $Q_i(r) = \mu_i Q(r)$ corresponding to a return period r at a target location i, where μ_i is the index-flood. In the present study, the index-flood is taken to be the means of the at-site distributions and is predicted at the target location by multiple regression.

The index-flood model is fitted inside the neighborhoods obtained by each one of the four methods: ROI, CCA, RVN-LM and RVN-HYB. For CCA, two canonical pairs are calculated as described in section 2.1 using flood quantiles corresponding to the 10- and 100-year return periods as hydrological variables. The choice of the regional distribution is made between the four common families of distributions that were mentioned earlier: GEV, GLO, LN3 and P3. The parameters of the regional quantile function $Q(r)$ are calculated from the regional LCV and the regional LSK as the respective averages (see Eq. (12)). Figure 4a shows the L-moment ratio diagram for the regional LSK and LKT with RVN-LM. For each neighborhood, the distribution family is selected as the one having the nearest regional LKT to the theoretical value, given the regional LSK. RVN-HYB is omitted in Figure 4 for the clarity of the graphics, but has similar behaviour to RVN_LM.

Figures 4b,c,d present the L-moment ratio diagrams of the at-site LCV and LSK for three given target locations as an illustration of the gauged sites found in the respective neighborhoods. In these diagrams, the nearest gauged sites selected for RVN-LM, CCA and ROI are highlighted. Figure 4b shows that RVN_LM has a denser cluster of gauged sites in terms of LCV and is approximately centered on the true target. Conversely, Figures 4c and 4d show situations where the true targets do not correspond to the predicted target. Although, all the reference variables are known at the target location for the ROI method, Figures 4b and 4c show that the selected sites are also not located around the true target. This finding is coherent with the results of (GREHYS, 1996a, 1996b) which indicates that delineation according to physiographical similarity can lead to substantially different regions than according to hydrological similarity.

Results of cross-validation are presented in Figure 5. The evaluation criteria are calculated for every neighborhood with size superior to 15 in order to calibrate the model. The tendency illustrated in this figure helps to visualize the evolution of these criteria with better perspective. The comparison of Figures 5a and 5b indicates that the optimal neighborhood sizes for RRMSE and NHS are not always in agreement. In particular, the best RRMSE for the RVN-HYB method is with 24 sites, while the best NHS is with nearly 80 sites. Nevertheless, the optimal values for the three other methods are obtained with approximately 30 sites for both criteria. Figure 5b indicates that all methods have relatively stable NHS between 86% and 87%, but the best NHS is obtained by RVN-LM. Conversely, Figure 5a shows clearer improvements of the calibration in terms of the RRMSE criterion. Hence, the calibrated models are set according to the RRMSE criterion and are represented by circles in Figure 5 and are summarized in Table 1. RVN-HYB, with a RRMSE of 40.1% outperforms the other methods. In particular, a difference of 6.1% and 5.3% is observed respectively with the traditional ROI and CCA methods.

Figures 5c,d present respectively the AHM and the ARE criteria obtained from the considered methods. The AHM criterion indicates that the ROI and the CCA methods have in general lower heterogeneity than the whole dataset, but are outperformed by the RVN-LM and RVN-HYB methods especially for smaller neighborhoods. This quantifies the intuitive assumption that the regional LCV is calculated with less uncertainty when the L-moments are directly considered instead of other reference variables. In particular, the AHM of the ROI method is 72.8% with the optimal neighborhood size of 30. In comparison, the AHM of the RVN-LM method is 14.5% with the optimal neighborhood size of 29 sites, which is considerably lower. Figure 5c shows that the AHM criterion of the RVM-LM method does not reach a similar level to the ROI method until using as much as 120 sites. These results indicate that even for relatively small neighborhoods, the ROI method identifies regions that are only slightly less hydrologically heterogeneous than all sites pooled together. This suggests that, in the present case study, the ROI method has difficulties identifying sites that are similar to the target site in terms of LCV.

As mentioned in section 4.2, previous studies have identified few problematic stations in the considered dataset. Figure 6 presents the residuals between different methods. As it may be difficult to see small improvements by uniquely observing points around the $y = x$ lines, the visualization of Figure 6 is helped by adding a flexible fit of the point cloud, using a standard smoothing spline approach. The resulting red lines indicate if close to x the residuals are lower in average for one of the two methods. In general, the points associated to the largest relative discrepancies are close to the $y = x$ line, which indicates that the sites that are difficult to predict are essentially the same for all methods. However, Figures 6a,b show that the RVN-HYB specifically improves the prediction of the sites with the lowest and largest relative discrepancies as the red line is clearly located under the $y = x$ lines, which explains the improved RRMSE in Table 1. On the other hand, Figures 6c,d demonstrate that at the logarithmic scale, the RVN-LM method achieved predicted values that are mostly similar to the ROI and CCA methods, which explains the similarity of the NHS criteria for all the compared methods.

The present case study is an example of a region where some sites are problematic for likely any methods. In practice, the residuals are not known, consequently we do not know if the target sites of interest will be “problematic” or not. Globally, what Figure 6a indicates is that the RVN-HYB model is more robust (in a certain way), because for the sites that are well predicted by simpler models, such ROI, RVN-HYB will perform in average similarly. However, if the target site is predicted less accurately, the RVN-HYB model will (in average) be better in terms of RRMSE. Consequently, the overall gain may seem of moderate magnitude, but for some problematic stations the gain could be more substantial. In particular, the red lines in the left part of Figure 6a appears mostly influenced by two points, but the two improvements are of 77.2% and 68.5%, which is considerable.

4.4 Results of the regression-based model

Prediction of Q100 at the target location is also performed by the regression-based model using the same delineation methods as with the index-flood model, but with potentially different calibration values for the neighborhood sizes. Consequently, the description of steps 1-4 (in section 3.1) are identical to those of the index-flood approach and are not repeated here.

The fit of the regression-based model is graphically assessed in Figure 7 by Quantile-Quantile plots. It is showed that for all delineation approach the regression-based models correctly predict the flood quantile Q100 at target. Cross-validation criteria for the regression-based model are presented in Figures 8 and summarized in Table 1. As with the index-flood model, Table 1 reveals that the RVN-HYB method leads to the best performance in terms of the RRMSE. Although all methods differ by less than 2% in terms of NHS, results indicate that NHS values corresponding to CCA and RVN-HYB are inferior to those corresponding to the regression model applied on all gauged sites, which corresponds to $n = 150$ in Figure 8b. However, CCA leads to the best relative efficiency as indicated by the ARE criterion in Table 1. Hence, CCA corresponds to the regression models with, on average, the lowest uncertainties. This indicates that flood quantiles may be better reference variables for the regression-based model than for the index-flood model and suggests that in general different reference variables may be more appropriate for different situations. Nevertheless, the two close lines in Figure 8d reveal that for the same neighborhood size the RVN-LM has similar ARE values to CCA. In terms of AHM, Figure 8c is identical to Figure 5c except that new neighborhood sizes are indicated in circles.

Table 1 provides also a comparison between the performance of the index-flood and the regression-based model. In terms of RRMSE and NHS criterion, the two approaches lead to very similar results, which is coherent with what it is reported in other studies (GREHYS, 1996a, 1996b; Haddad and Rahman, 2012). Therefore, similar conclusions can be draw from the two approaches. For instance, in both cases, the RVN-HYB leads to the best results in terms of RRMSE.

5. Conclusions

A general methodology was investigated to improve homogenous properties of neighborhoods in RFFA. A procedure to calculate relevant reference variables at a target location prior to the RFFA was proposed to improve neighborhood properties and to reduce uncertainties. The predicted values of reference variables represent the unknown centers of neighborhoods delineated according to a distance of gauged sites with respect to the centers. The proposed method represents a generalization of both ROI and CCA methods in RFFA. The proposed RVN method has the advantages of accepting various groups of reference variables, of considering nonlinear interrelations and of being more objective since L-moments are used instead of estimated flood quantiles from at-site analysis.

In this study, the reference variables correspond to transformed L-moments. The resulting RVN-LM and RVN-HYB methods were applied on sites located in Southern Quebec, Canada, to predict flood quantiles corresponding to the 100 year return period by both index-flood and regression-based models. The prediction of the reference variables at target locations showed that after proper transformations, L1 can be linearly related to the site characteristics, but no proper transformations are found for the other L-moments. This justifies the consideration of the PPR method to account for the nonlinearity in the prediction of the reference variables. In general, other models, such as generalized additive models or artificial neural networks, could be considered instead of PPR to account for the nonlinearity. Nevertheless, the PPR approach unveils direction vectors that provide explicit, parsimonious and meaningful regression equations.

Although none of the methods performed best for all criteria, cross-validation showed that the proposed RVN method performs well in comparison to the traditional ROI and CCA methods. In both the index-flood and the regression-based model the best RRMSE is obtained by RVN_HYB and the best NHS is obtained by RVN_LM. In particular, the favorable RRMSE values obtained by RVN-HYB are due to more robust estimation of problematic sites. However, RVN_LM has the best balance, because it achieves the best or the second best values for all criteria. Most importantly, the utilization of hydrological reference variables with the CCA and RVN methods has reduced the uncertainty on the regional LCV, the index-flood and the predicted flood quantiles, in comparison to ROI. Consequently, prior modeling of hydrological reference variables was shown to be advantageous to the delineation of neighborhoods in RFFA.

The present study has made specific assumptions in order to investigate the RVN method in well-defined conditions. Nevertheless, the rational of predicting hydrological reference variables in *a priori* analysis remains a valid approach when other choices of regression models, neighborhood forms and metrics are considered. Hence, more comparative studies should be carried out to evaluate alternatives to fixed size neighborhoods and Euclidian distances in the specific context of the RVN framework.

The L-coefficient of skewness is commonly used in RFFA to describe the shape of a distribution. Consequently, to improve the result of the RVN method, further research efforts could focus on improving the prediction of this crucial reference variable. One way to improve the prior analysis of the hydrological reference variables is the consideration of the unequal sampling error. This aspect is often considered in the estimation of flood quantiles in RFFA, but may also play an important role in the prior analysis of the RVN method.

Acknowledgement

Financial support for this study was graciously provided by the Natural Sciences and Engineering Research Council (NSERC) of Canada.

References

Bishop, C.M., 1995. Neural networks for pattern recognition. Oxford university press.

Burn, D.H., 1990. An appraisal of the “region of influence” approach to flood frequency analysis. Hydrol. Sci. J. 35, 149–166. doi:10.1080/02626669009492415

Castiglioni, S., Castellarin, A., Montanari, A., 2009. Prediction of low-flow indices in ungauged basins through physiographical space-based interpolation. J. Hydrol. 378, 272–280. doi:10.1016/j.jhydrol.2009.09.032

Chebana, F., Charron, C., Ouarda, T.B.M.J., Martel, B., 2014. Regional frequency analysis at ungauged sites with the generalized additive model. J. Hydrometeorol. 15, 2418–2428. doi:10.1175/JHM-D-14-0060.1

Chebana, F., Ouarda, T.B.M.J., 2009. Index flood–based multivariate regional frequency analysis. Water Resour. Res. 45. doi:10.1029/2008WR007490

Chebana, F., Ouarda, T.B.M.J., 2008. Depth and homogeneity in regional flood frequency analysis. Water Resour. Res. 44. doi:10.1029/2007WR006771

Chebana, F., Ouarda, T.B.M.J., 2007. Multivariate L-moment homogeneity test. Water Resour. Res. 43. doi:10.1029/2006WR005639

Chokmani, K., Ouarda, T.B.M.J., 2004. Physiographical space-based kriging for regional flood frequency estimation at ungauged sites. Water Resour. Res. 40. doi:10.1029/2003WR002983

Cunderlik, J.M., Burn, D.H., 2006. Switching the pooling similarity distances: Mahalanobis for Euclidean. Water Resour. Res. 42. doi:10.1029/2005WR004245

Cunnane, C., 1988. Methods and merits of regional flood frequency analysis. J. Hydrol. 100, 269–290.

Dalrymple, T., 1960. Flood-frequency analysis. Surv. Water-Supply Pap. 1543.

Das, S., Cunnane, C., 2010. Examination of homogeneity of selected Irish pooling groups. Hydrol. Earth Syst. Sci. Discuss. 7, 5099–5130. doi:10.5194/hessd-7-5099-2010

Dawson, C.W., Abrahart, R.J., Shamseldin, A.Y., Wilby, R.L., 2006. Flood estimation at ungauged sites using artificial neural networks. J. Hydrol. 319, 391–409. doi:10.1016/j.jhydrol.2005.07.032

Durocher, M., Chebana, F., Ouarda, T.B.M.J., 2015. A Nonlinear Approach to Regional Flood Frequency Analysis Using Projection Pursuit Regression. J. Hydrometeorol. doi:10.1175/JHM-D-14-0227.1

Eng, K., Tasker, G.D., Milly, P., 2005. An Analysis of Region-Of-Influence methods for flood regionalization in the Gulf-Atlantic rolling plain. J. Am. Water Resour. Assoc. 41, 135–143. doi:10.1111/j.1752-1688.2005.tb03723.x

Friedman, J., Grosse, E., Stuetzle, W., 1983. Multidimensional Additive Spline Approximation. SIAM J. Sci. Stat. Comput.

4, 291–301. doi:10.1137/0904023

Friedman, J.H., Tukey, J.W., 1974. A projection pursuit algorithm for exploratory data analysis. Comput. IEEE Trans. On 100, 881–890.

GREHYS, 1996a. Presentation and review of some methods for regional flood frequency analysis. J. Hydrol. 186, 63–84.

GREHYS, 1996b. Inter-comparison of regional flood frequency procedures for canadian rivers. J. Hydrol. 186, 85–103.

Griffis, V., Stedinger, J., 2007. The use of GLS regression in regional hydrologic analyses. J. Hydrol. 344, 82–95.

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. J. Hydrol. 430–431, 142–161. doi:10.1016/j.jhydrol.2012.02.012

Hastie, T., Tibshirani, R., Friedman, J.H., 2009. The elements of statistical learning: data mining, inference, and prediction, Springer series in statistics. Springer.

He, Y., Bárdossy, A., Zehe, E., 2011. A review of regionalisation for continuous streamflow simulation. Hydrol. Earth Syst. Sci. 15, 3539–3553. doi:10.5194/hess-15-3539-2011

Hosking, J.R.M., Wallis, J.R., 1997. Regional frequency analysis: an approach based on L-moments. Cambridge Univ Pr.

Hwang, J.-N., Lay, S.-R., Maechler, M., Martin, R.D., Schimert, J., 1994. Regression modeling in back-propagation and projection pursuit learning. Neural Netw. IEEE Trans. On 5, 342–353. doi:10.1109/72.286906

Kjeldsen, T.R., Jones, D.A., 2009. An exploratory analysis of error components in hydrological regression modeling. Water Resour. Res. 45, n/a–n/a. doi:10.1029/2007WR006283

Laio, F., Ganora, D., Claps, P., Galeati, G., 2011. Spatially smooth regional estimation of the flood frequency curve (with uncertainty). J. Hydrol. 408, 67–77. doi:http://dx.doi.org/10.1016/j.jhydrol.2011.07.022

Ouali, D., Chebana, F., Ouarda, T.B.M.J., 2015. Non-linear canonical correlation analysis in regional frequency analysis. Stoch. Environ. Res. Risk Assess. 1–14. doi:10.1007/s00477-015-1092-7

Ouarda, T., Haché, M., Bruneau, P., Bobée, B., 2000. Regional Flood Peak and Volume Estimation in Northern Canadian Basin. J. Cold Reg. Eng. 14, 176–191. doi:10.1061/(ASCE)0887-381X(2000)14:4(176)

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. J. Hydrol. 348, 40–58. doi:10.1016/j.jhydrol.2007.09.031

Ouarda, T.B.M.J., Girard, C., Cavadias, G.S., Bobée, B., 2001. Regional flood frequency estimation with canonical correlation analysis. J. Hydrol. 254, 157–173. doi:10.1016/S0022-1694(01)00488-7

Ouarda, T.B.M.J., Shu, C., 2009. Regional low-flow frequency analysis using single and ensemble artificial neural networks. Water Resour. Res. 45. doi:10.1029/2008WR007196

Pandey, G.R., 1998. Assessment of scaling behavior of regional floods. J. Hydrol. Eng. 3, 169–173. doi:10.1061/(ASCE)1084-0699(1998)3:3(169)

Pandey, G.R., Nguyen, V.-T.-V., 1999. A comparative study of regression based methods in regional flood frequency analysis. J. Hydrol. 225, 92–101. doi:10.1016/S0022-1694(99)00135-3

Reis, D., Stedinger, J., Martins, E., 2005. Bayesian generalized least squares regression with application to log Pearson type 3 regional skew estimation. Water Resour. Res. 41.

Stedinger, J., Lu, L.-H., 1995. Appraisal of regional and index flood quantile estimators. Stoch. Hydrol. Hydraul. 9, 49–75.

Stedinger, J., Tasker, G., 1985. Regional hydrologic analysis: 1. Ordinary, weighted, and generalized least squares compared. Water Resour. Res. 21, 1421–1432. doi:10.1029/WR021i009p01421

Tasker, G., Hodge, S., Bark, S., 1996. Region of Influence regression for estimating the 50-years flood at ungaged sites. Water Resour. Bull. doi:10.1111/j.1752-1688.1996.tb03444.x

Viglione, A., Laio, F., Claps, P., 2007. A comparison of homogeneity tests for regional frequency analysis. Water Resour. Res. 43, n/a–n/a. doi:10.1029/2006WR005095

Yu, Y., Ruppert, D., 2002. Penalized spline estimation for partially linear single-index models. J. Am. Stat. Assoc. 97, 1042–1054. doi:10.1198/016214502388618861

Table 1: Evaluation criteria for the RVN method for optimal neighborhood sizes.

	Model	Size	RRMSE	NHS	AHM	ARE
Index-Flood						
	ROI	30	46.2	86.5	72.8	57.3
	CCA	28	45.4	86.2	41.7	42.9
	RVN-LM	29	45.0	**87.1**	**14.5**	**36.9**
	RVN-HYB	24	**40.1**	86.2	16.5	43.1
Regression-based						
	ROI	30	44.9	86.9	72.8	64.7
	CCA	28	43.5	86.1	41.7	**30.6**
	RVN-LM	39	41.7	**87.6**	17.9	39.8
	RVN-HYB	24	**39.5**	86.2	**16.5**	42.5

Best criteria in bold

List of Figures

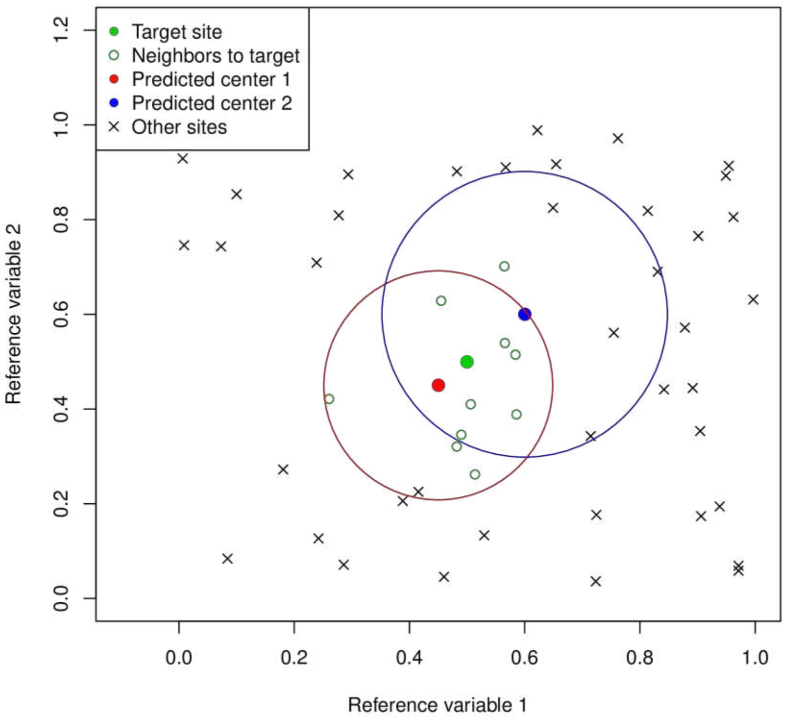

Figure 1: Illustration of the neighborhoods obtained by the RVN method.

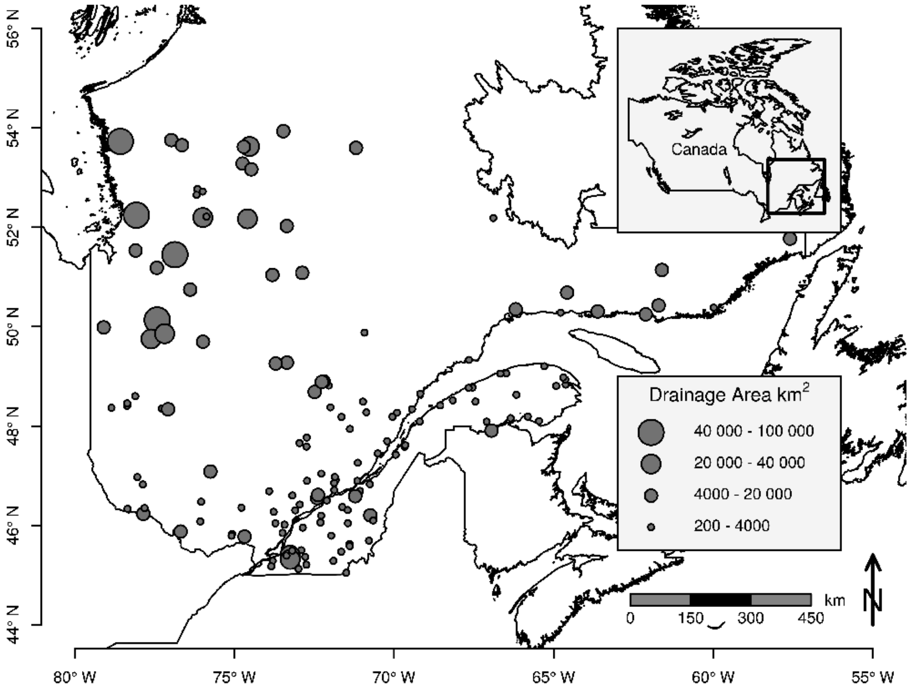

Figure 2: Location of the 151 hydrometric stations in Southern Quebec, Canada.

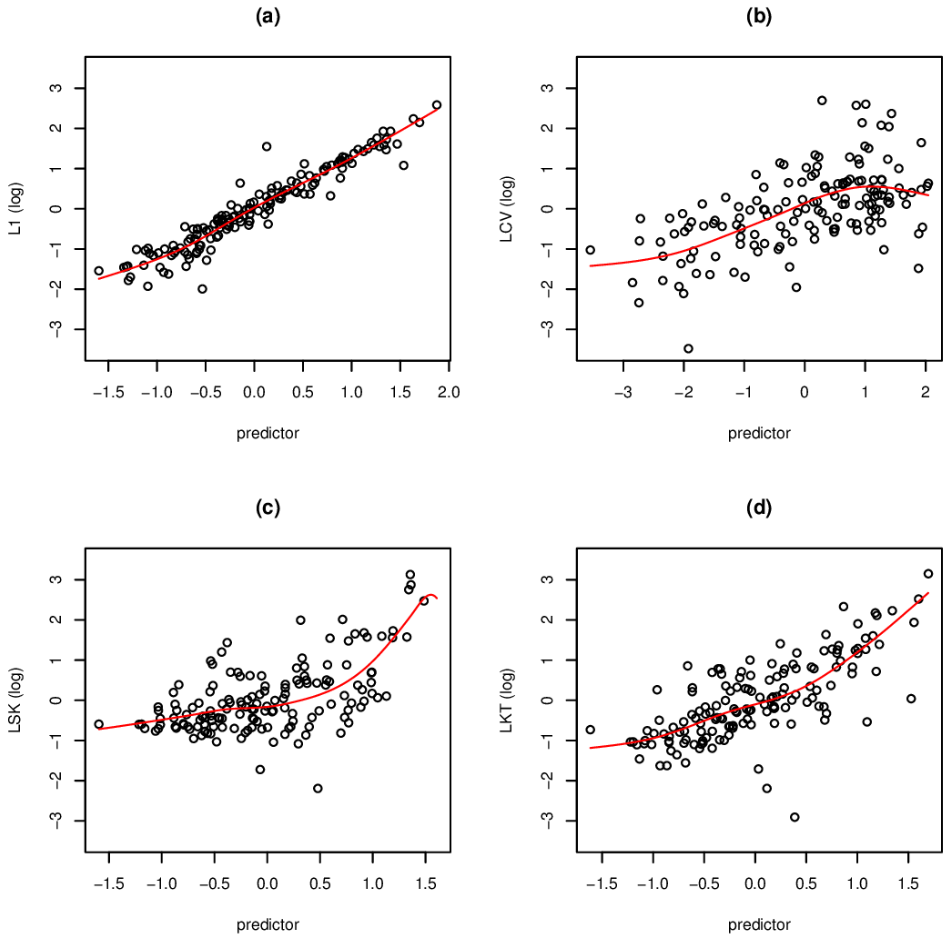

Figure 3: Residuals of the reference variables by PPR methods.

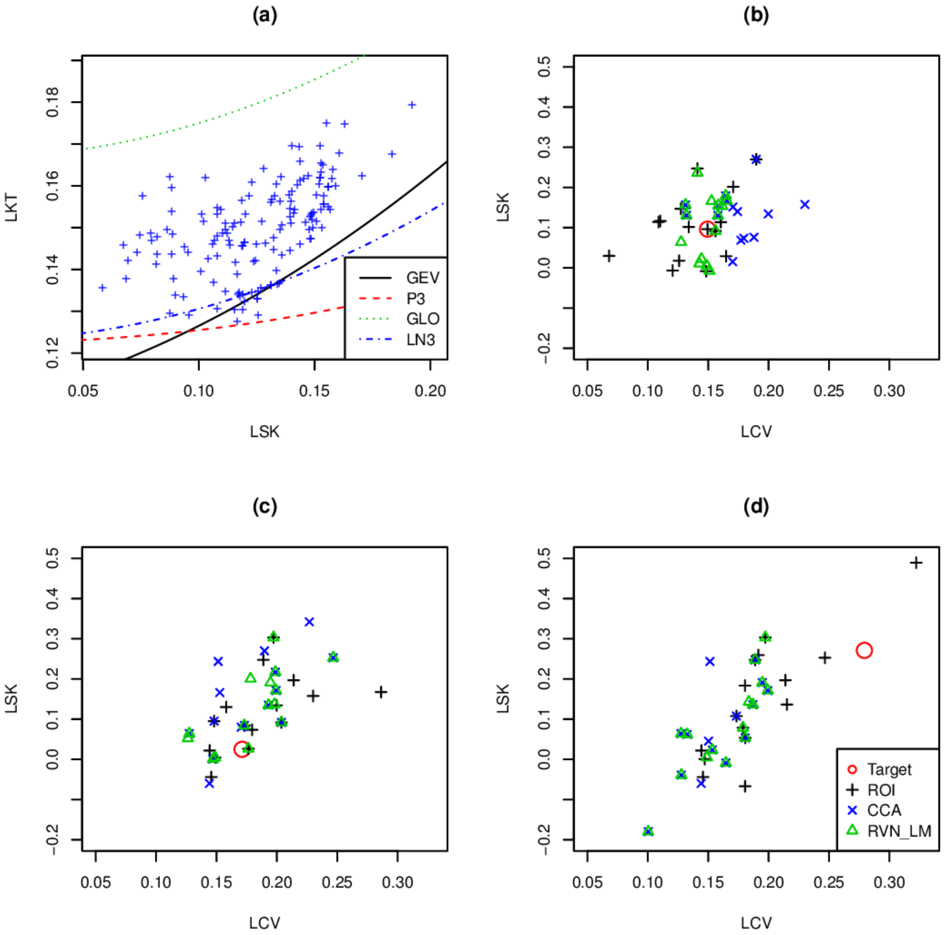

Figure 4: L-moments ratio diagram for index-flood model. (a) Regional L-moments for RVN-LM with 29 gauged sites. (b),(c) and (d) Regional L-moments based on the 15 nearest gauged sites for 3 selected target locations.

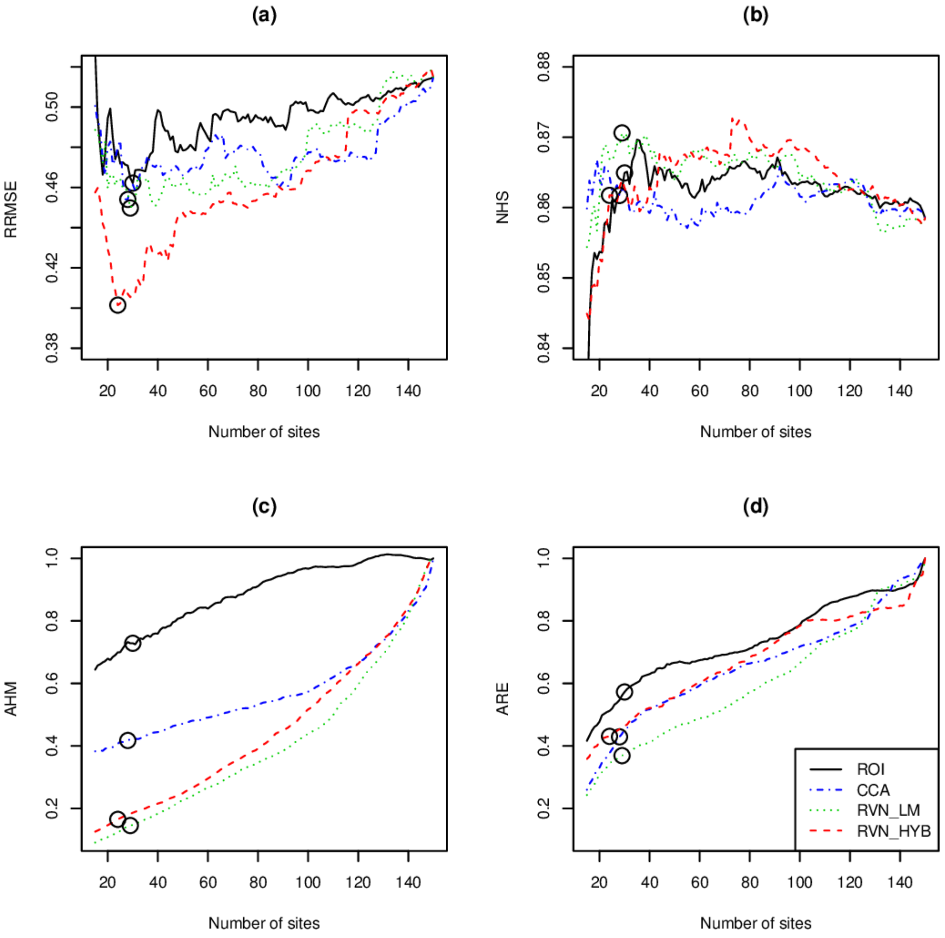

Figure 5: Evaluation criteria for the index-flood model. Calibrated models are represented by circles.

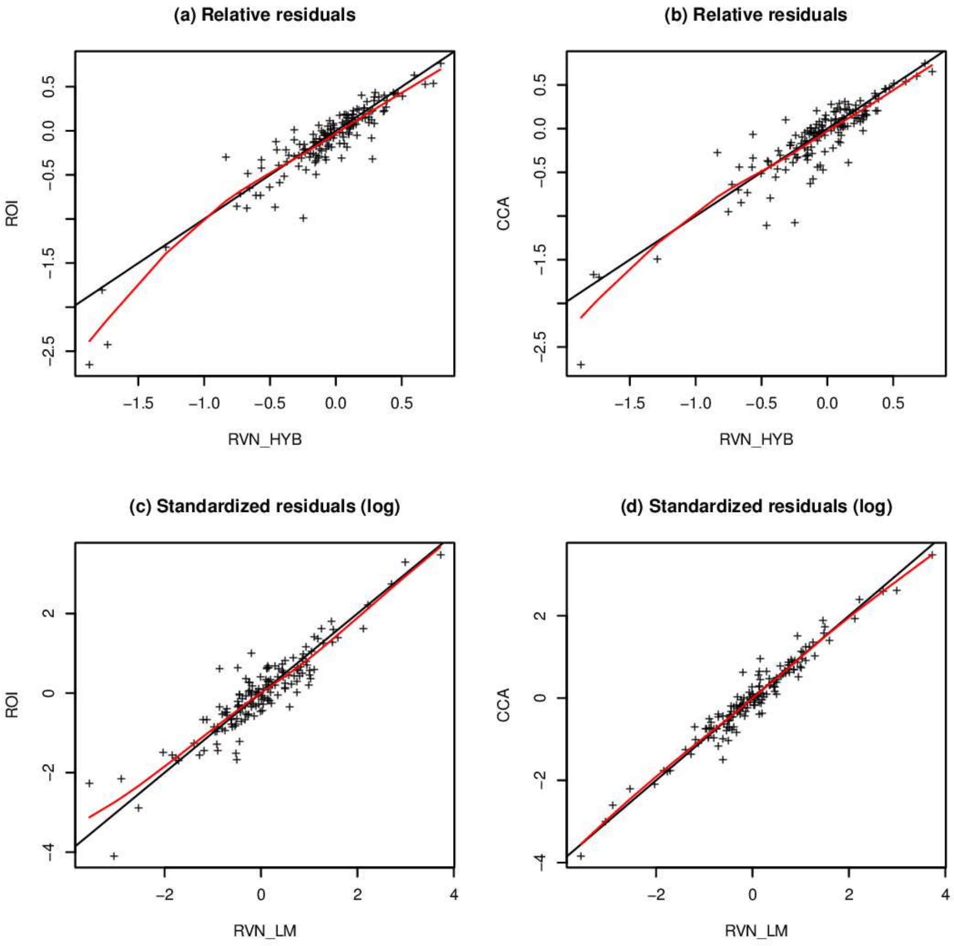

Figure 6: Comparison of the cross-validation residuals for Q100 between different methods. The black line is the unitary slope and the red line is a smooth fitting of the residuals.

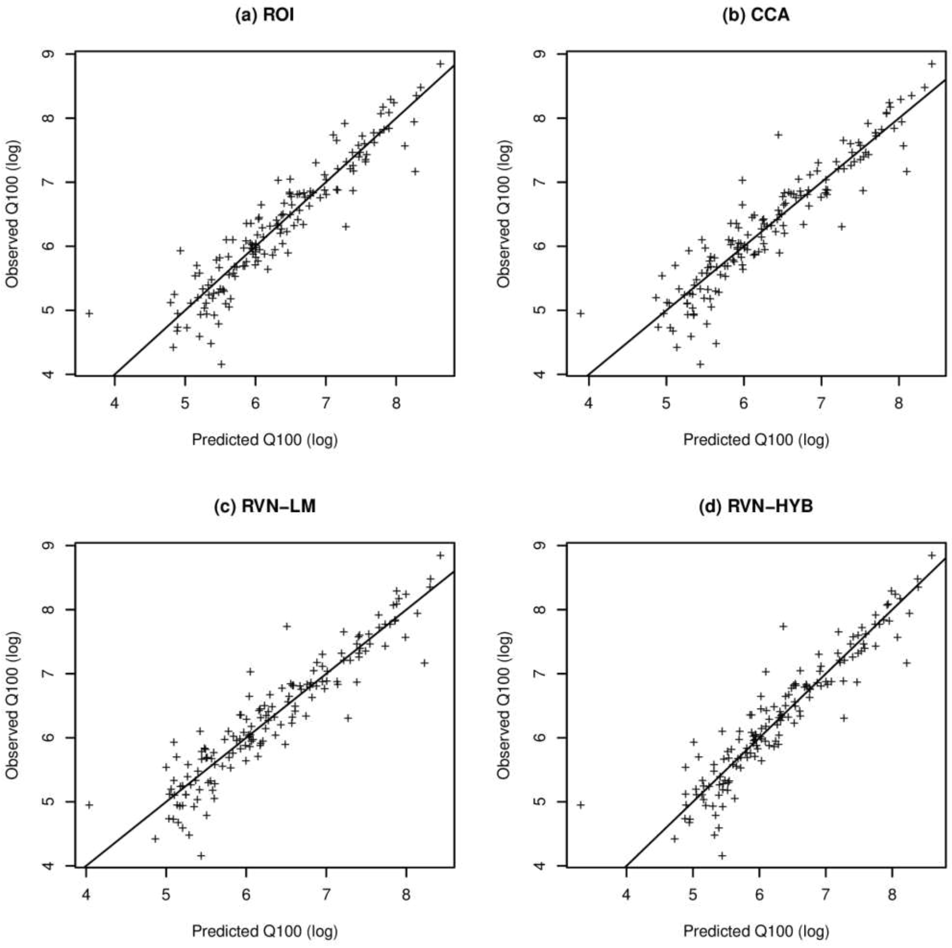

Figure 7: Quantile-Quantile plot of Q100 for the RVN method with regression-based model.

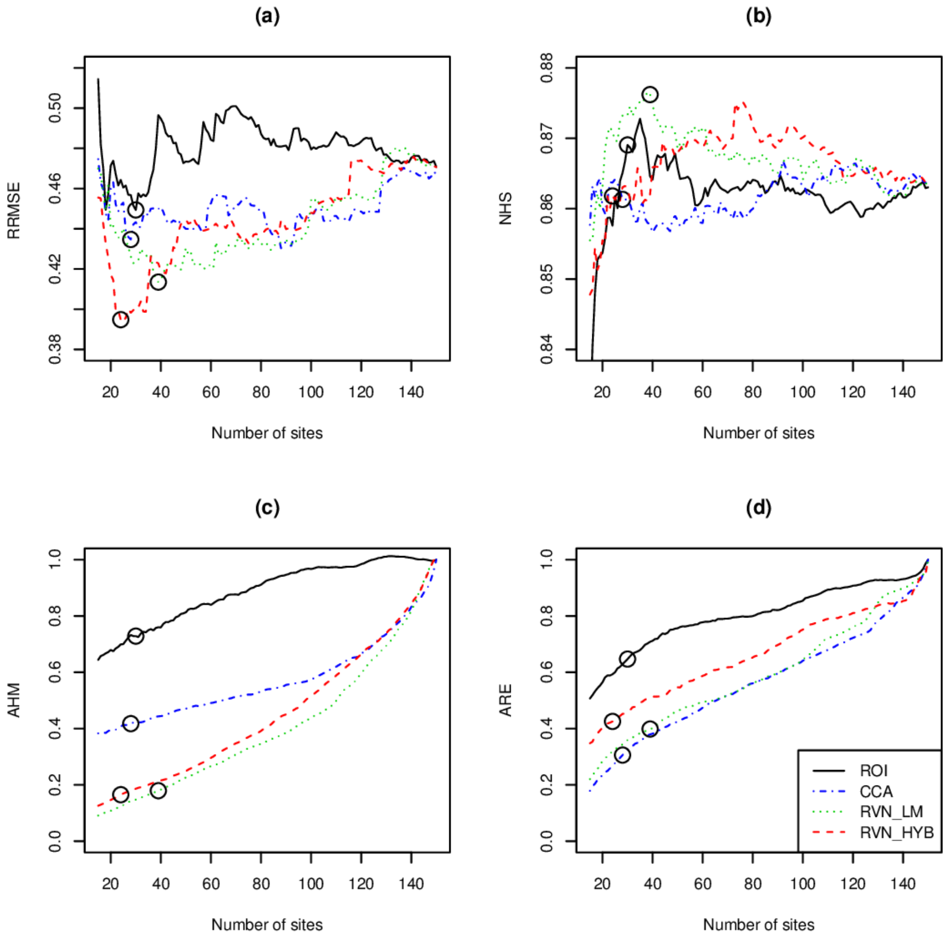

Figure 8: Evaluation criteria for the regression-based model. Calibrated models are represented by circles.