Hydrol. Earth Syst. Sci. Discuss., doi:10.5194/hess-2016-123, 2016
Manuscript under review for journal Hydrol. Earth Syst. Sci.
Published: 31 March 2016

Hydrology and Earth System Sciences
Discussions

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 Frequency Analysis (RFA) to enforce 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 RFA is that hydrological information is not available at target locations. 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.

Hydrol. Earth Syst. Sci. Discuss., doi:10.5194/hess-2016-123, 2016
Manuscript under review for journal Hydrol. Earth Syst. Sci.
Published: 31 March 2016

Hydrology and Earth System Sciences
Discussions

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 Frequency Analysis (RFA) 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).

RFA 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, RFA 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, a notion of distance is needed to evaluate the proximity, or relevance, of each gauged site to the target location and identify the most similar gauged sites. However, when the target location is ungauged, the distance between hydrological variables 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).

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 RFA 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

Hydrol. Earth Syst. Sci. Discuss., doi:10.5194/hess-2016-123, 2016
Manuscript under review for journal Hydrol. Earth Syst. Sci.
Published: 31 March 2016

Hydrology and Earth System Sciences
Discussions

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 estimation model in RFA (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 RFA 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 RFA. 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 RFA, 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:

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

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:

Hydrol. Earth Syst. Sci. Discuss., doi:10.5194/hess-2016-123, 2016
Manuscript under review for journal Hydrol. Earth Syst. Sci.
Published: 31 March 2016

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

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

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

To delineate neighborhoods, the CCA approach considers the canonical scores $\mathbf{u}_i = (a_1, \ldots, a_r)'\mathbf{x}_i$ and $\mathbf{v}_i = (b_1, \ldots, b_r)'\mathbf{y}_i$ 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:

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

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

2.2 Multiple regression

In RFA, 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

Hydrol. Earth Syst. Sci. Discuss., doi:10.5194/hess-2016-123, 2016
Manuscript under review for journal Hydrol. Earth Syst. Sci.
Published: 31 March 2016

Hydrology and Earth System Sciences
Open Access
Discussions

power law form is frequently used to model flood quantiles:

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

where $\beta' = (\beta_0, \beta_1, \ldots, \beta_p)$ 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 $y_i = g(z_i)$ being linearly related to the sites characteristics.

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

$$\mathbf{y} = \mathbf{X}\beta + \varepsilon \quad (6)$$

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

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

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 RFA 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:

$$f_k(\mathbf{X}) = g_k(\alpha_k'\mathbf{X}) \quad (8)$$

where directions α_k are vectors of coefficients and g_k are smooth functions. The directions α_k are coefficients that respect $|\alpha| = 1$ 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,

Hydrol. Earth Syst. Sci. Discuss., doi:10.5194/hess-2016-123, 2016
Manuscript under review for journal Hydrol. Earth Syst. Sci.
Published: 31 March 2016

Hydrology and Earth System Sciences
Discussions

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 RFA, 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 RFA 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 of the RVN method can be described by the main steps below:

(i) Estimation of the hydrological reference variables at the target centers;
(ii) Delineation of the neighborhoods;
(iii) Estimation of the flood quantiles at the target locations.

Step (i) is the particularity of the RVN. If a target location is designated by $i=0$, the radius of the neighborhood can be computed as $h_i = d(\mathbf{t}_i, \mathbf{t}_0)$ where d is a metric and $\mathbf{t}_i' = (t_{i,1}, \ldots, t_{i,q})$ 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).

As hydrological information is unavailable at the target location, the estimation of the hydrological reference variables is necessary to produce an estimate $\mathbf{t}_0 = f(\mathbf{x}_0)$ from site characteristics $\mathbf{x}_0$ at the target location. This substitution leads to the distance $h_{(i)} = d[\mathbf{t}_i, f(\mathbf{x}_0)]$, 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

Hydrol. Earth Syst. Sci. Discuss., doi:10.5194/hess-2016-123, 2016
Manuscript under review for journal Hydrol. Earth Syst. Sci.
Published: 31 March 2016

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).

Figure 1 illustrates two neighborhoods resulting from the RVN method. It shows the importance of correctly predicting the reference variables in order to be representative of the true center of the target location and hence the appropriate sites to be included in the neighborhood. Indeed, in green, the true neighborhood designates the neighborhood that would be delineated if all hydrological variables were known at the target location. Alternatively, the red and the blue neighborhoods are identified from different estimates of the reference variables. Figure 1 indicates that if the reference variables are well predicted, then the corresponding RVN neighborhoods will most likely include the same gauged sites as the true neighborhood.

Steps (ii-iii) are common in RFA and are explained in sections 2.1 and 2.2. In the remainder of this study, step (ii) 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 (iii) 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 RFA. 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 (i). 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 (i).

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 $y_i = g(z_i)$. Hence, if $\bar{y}$ is the sample mean of the y_i, then an appropriate global performance measure is the Nash-Sutcliffe criterion:

Hydrol. Earth Syst. Sci. Discuss., doi:10.5194/hess-2016-123, 2016
Manuscript under review for journal Hydrol. Earth Syst. Sci.
Published: 31 March 2016

Hydrology and Earth System Sciences
Discussions

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

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

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

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

Hydrol. Earth Syst. Sci. Discuss., doi:10.5194/hess-2016-123, 2016
Manuscript under review for journal Hydrol. Earth Syst. Sci.
Published: 31 March 2016

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:

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

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:

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

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 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:

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

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, RFA 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

Hydrol. Earth Syst. Sci. Discuss., doi:10.5194/hess-2016-123, 2016
Manuscript under review for journal Hydrol. Earth Syst. Sci.
Published: 31 March 2016

Hydrology and Earth System Sciences Discussions

corresponding to a return period of 100 years, denoted Q100. The analysis is performed on 151 sites located in Southern Quebec, Canada, for which at least 15 years of data are available and 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 detail 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 first step of the RVN method is the estimation of the hydrological reference variables at the target locations. Two 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. 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 logarithmically and standardized to obtain zero mean and unit variance. For LSK and LKT, an additional translation is necessary to avoid numerical difficulties due to negative values. 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 2 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. Figure 2a shows a strong linear relationship between L1 and the predictor $\alpha'\mathbf{X}$. Conversely, Figures 2b,c,d show nonlinearity and hence indicate the need for a nonlinear model such as PPR. 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:

$$\log(\mathrm{LCV}) = -1.80 + 0.26 \times f\left[-0.67 \times \log(\mathrm{BV}) - 0.09 \times \sqrt{PLAC} + 1.27 \times \log(\mathrm{PLMA}) + 0.06 \times \mathrm{LON} - 1.32\right] \quad (17)$$

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)

Hydrol. Earth Syst. Sci. Discuss., doi:10.5194/hess-2016-123, 2016
Manuscript under review for journal Hydrol. Earth Syst. Sci.
Published: 31 March 2016

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.

The predictive performances of the reference variables are evaluated by the NHS criterion with values 91%, 33%, 7% and 56% 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 2c suggests that apart from a few sites on the right of the curve, LSK appears not highly related to the predictor $\alpha'\mathbf{X}$. 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 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

One of the objectives of RFA is to identify a proper family of distributions from regional information, which is achieved here by analysing the distribution of the gauged sites inside a neighborhood. 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 quantile $Q_i(r)$ corresponding to a return period r at a target location i is of the form $Q_i(r) = \mu_i Q(r)$, 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 3a 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 3 for the clarity of the graphics, but has similar behaviour to RVN_LM.

Figures 3b,c,d present the L-moment ratio diagrams of the at-site LCV and LSK for three given target locations as

Hydrol. Earth Syst. Sci. Discuss., doi:10.5194/hess-2016-123, 2016
Manuscript under review for journal Hydrol. Earth Syst. Sci.
Published: 31 March 2016

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 3b 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 3c and 3d 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 3b and 3c 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 4. 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 4a and 4b 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 4b indicates that all methods have relatively stable NHS between 86% and 87%, but the best NHS is obtained by RVN-LM. Conversely, Figure 4a 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 4. RVN-HYB, with a RRMSE of 40% outperforms the methods RVN-LM and CCA, with a RRMSE of 45% and ROI, with a RRMSE of 46%.

Figures 4c,d present respectively the AHM and the ARE criteria. The AHM criterion indicates that the ROI and the CCA methods have in general lower heterogeneity than the whole dataset, but are outperformed by RVN-LM and RVN-HYB methods especially for smaller neighborhoods. This validates and 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. Moreover, the ARE criterion reveals that RVN-LM leads to neighborhoods with the most relatively efficient regression models for predicting the index-flood. In particular, the calibrated model of the RVN-LM method has an ARE of 36% and the ROI method has an ARE of 57%. Overall, this indicates that a strategic pooling of the gauged sites using hydrological reference variables reduces the uncertainty of both the regional LCV and the index-flood.

As mentioned in section 4.2, previous studies have identified few problematic stations in the considered dataset. Figure 5 presents the relative residuals between different methods. In general, the points associated to the largest 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 5a,b show that the RVN-HYB specifically improves the prediction of the sites with the largest discrepancies as their points are mostly located under the $y = x$ lines, which explains that this method leads to the best RRMSE. On the other hand, Figures 5c,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.

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 different calibration values for the neighborhood sizes. Cross-

Hydrol. Earth Syst. Sci. Discuss., doi:10.5194/hess-2016-123, 2016
Manuscript under review for journal Hydrol. Earth Syst. Sci.
Published: 31 March 2016

Hydrology and Earth System Sciences
Open Access
Discussions

validation results for the regression-based model are presented in Figures 6. As with the index-flood model, Figure 6a 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 6b. However, CCA leads to the best relative efficiency as indicated by the ARE criterion in Figure 6d. 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 6d reveal that for the same neighborhood size the RVN-LM has similar ARE values to CCA. In terms of AHM, Figure 6c is identical to Figure 4c except that new neighborhood sizes are indicated in circles.

5. Conclusions

A general methodology was investigated to improve homogenous properties of neighborhoods in RFA. A procedure to calculate relevant reference variables at a target location prior to the RFA 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 RFA. 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 RFA.

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 RFA 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 RFA, 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.

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

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.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

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)

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.

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

Hydrol. Earth Syst. Sci. Discuss., doi:10.5194/hess-2016-123, 2016
Manuscript under review for journal Hydrol. Earth Syst. Sci.
Published: 31 March 2016

Hydrology and Earth System Sciences
Open Access
Discussions

List of Figures

Hydrol. Earth Syst. Sci. Discuss., doi:10.5194/hess-2016-123, 2016
Manuscript under review for journal Hydrol. Earth Syst. Sci.
Published: 31 March 2016

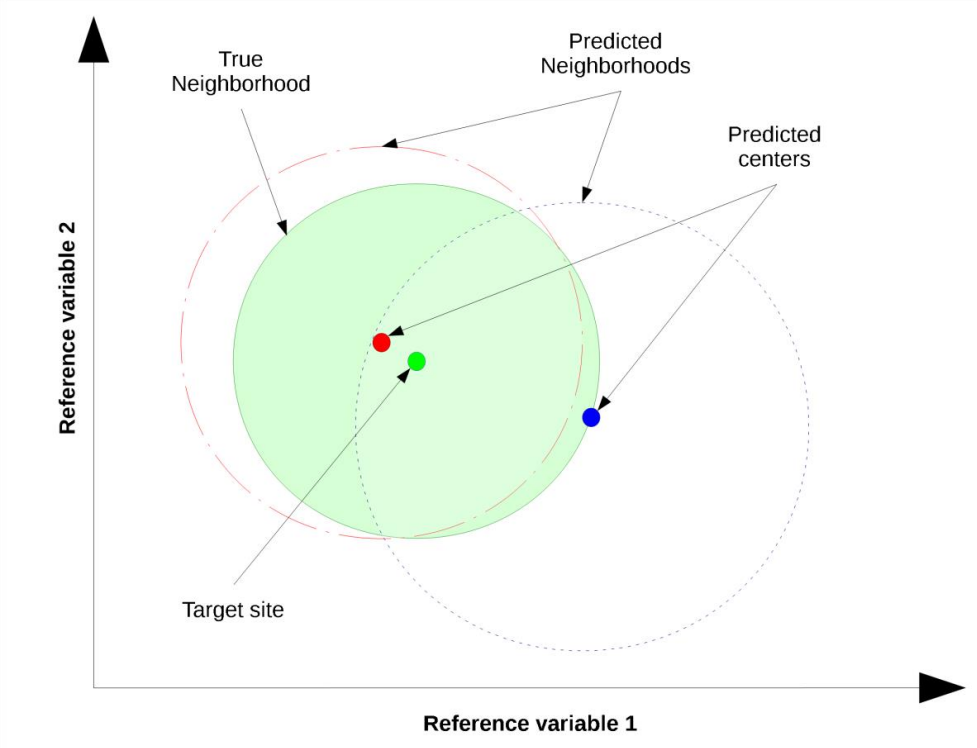

Figure 1: Illustration of the neighborhoods obtained by the RVN method.

Hydrol. Earth Syst. Sci. Discuss., doi:10.5194/hess-2016-123, 2016
Manuscript under review for journal Hydrol. Earth Syst. Sci.
Published: 31 March 2016

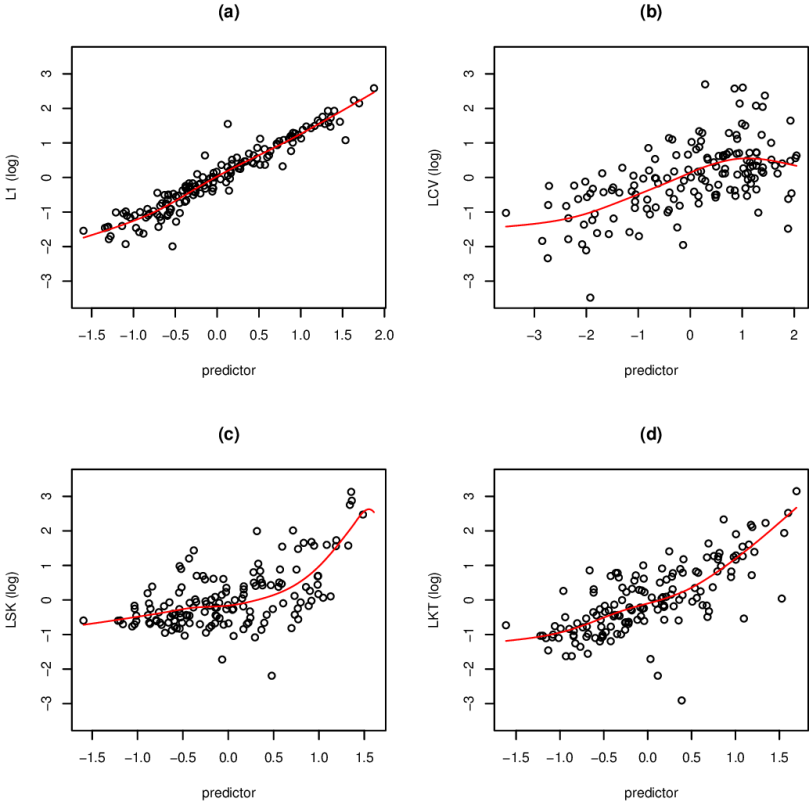

Figure 2: Residuals of the reference variables by PPR methods.

Hydrol. Earth Syst. Sci. Discuss., doi:10.5194/hess-2016-123, 2016
Manuscript under review for journal Hydrol. Earth Syst. Sci.
Published: 31 March 2016

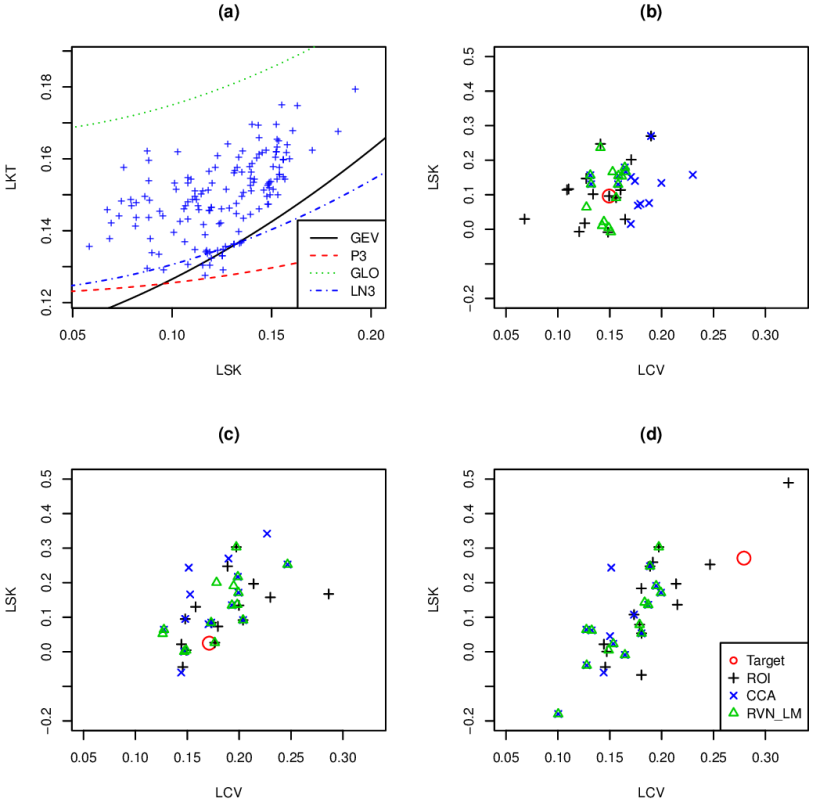

Figure 3: L-moments ratio diagram for index-flood model. (a) Regional L-moments for RVN_LM with 29 gauged sites. (b) Regional L-moments based on the 15 nearest gauged sites for 3 selected target locations.

Hydrol. Earth Syst. Sci. Discuss., doi:10.5194/hess-2016-123, 2016
Manuscript under review for journal Hydrol. Earth Syst. Sci.
Published: 31 March 2016

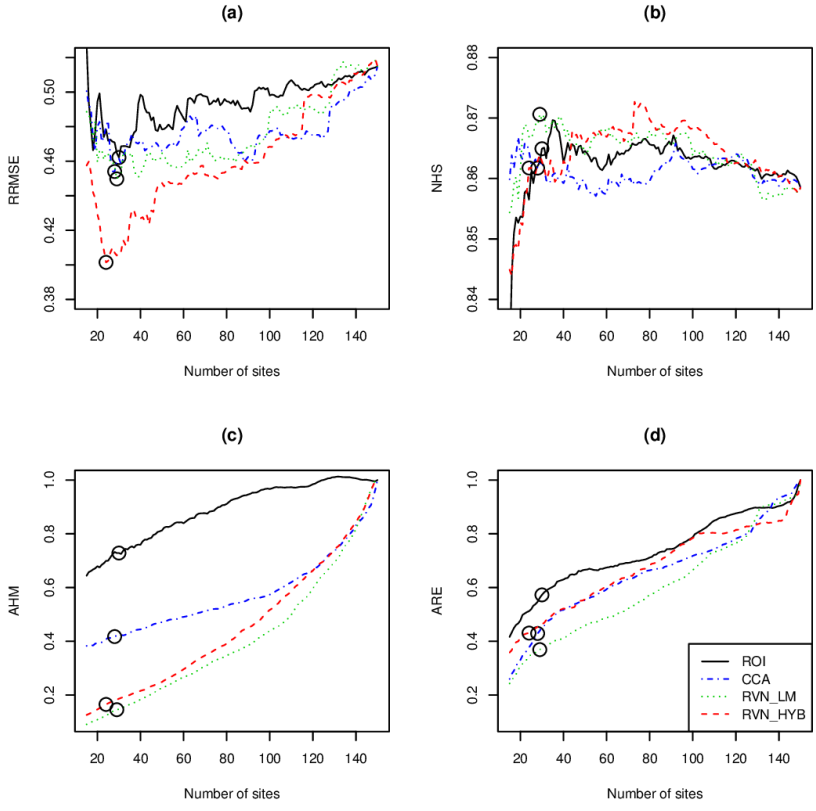

Figure 4: Evaluation criteria for the index-flood model. Calibrated models are represented by circles.

Hydrol. Earth Syst. Sci. Discuss., doi:10.5194/hess-2016-123, 2016
Manuscript under review for journal Hydrol. Earth Syst. Sci.
Published: 31 March 2016

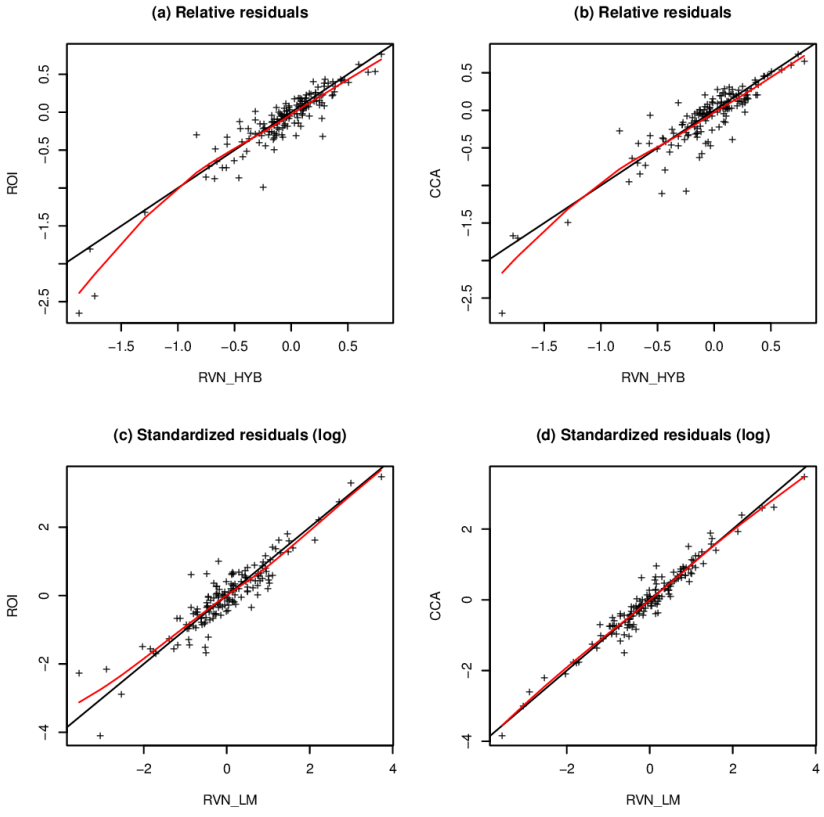

Figure 5: Comparison of the cross-validation residuals for Q100 between different methods. The black line is the unitary slope and the red line is a smooth fitting of the residuals.

Hydrol. Earth Syst. Sci. Discuss., doi:10.5194/hess-2016-123, 2016
Manuscript under review for journal Hydrol. Earth Syst. Sci.
Published: 31 March 2016

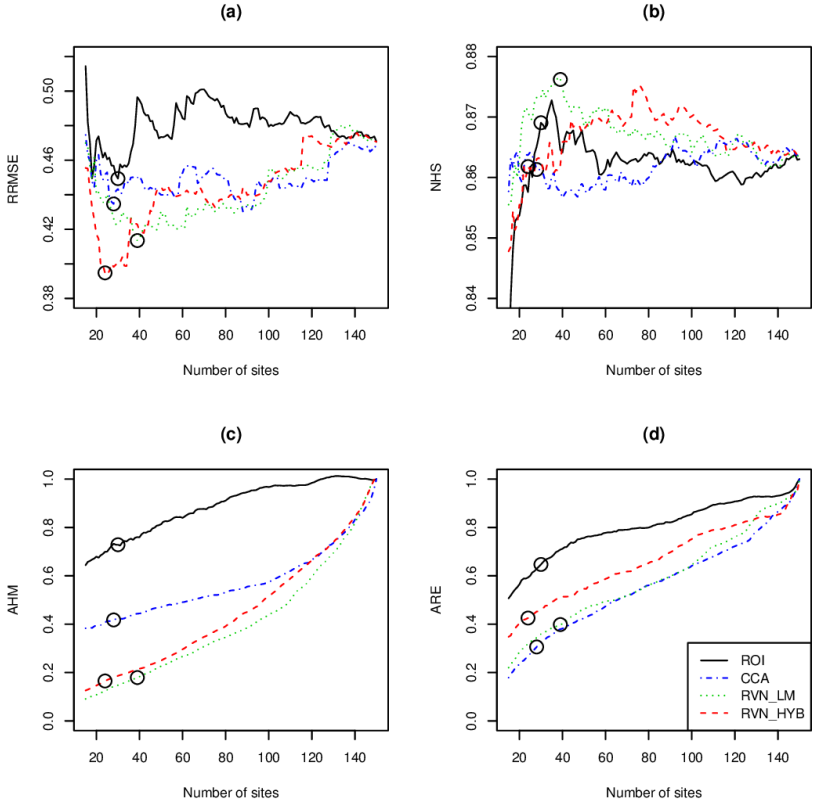

Figure 6: Evaluation criteria for the regression-based model. Calibrated models are represented by circles.