Predicting Future Ranked Statistics and Recorded Values for Some Statistical Distributions

Nidhal Khaleel Ajeel
Technical Instructors Training Institute, Middle Technical University, Iraq.

Received March 12, 2021; Accepted June 28, 2021
ISSN: 1735-188X
DOI: 10.14704/WEB/V18SI04/WEB18135

Abstract

Regional frequency analysis (AFR) brings together a variety of statistical methods aimed at predicting the behavior of extreme hydrological variables at ungauged sites. Regression techniques, geostatistical methods and classification are among the statistical tools frequently encountered in the literature. Methodologies based on these tools lead to regional models that offer a simple, but very useful description of the relationship between extreme hydrological variables and physiometeorological characteristics of a site. These regional models then make it possible to predict the behavior of variables of interest at places where no hydrological information is available. These methods are generally based on restrictive theoretical assumptions, including linearity and normality. These do not reflect the reality of natural phenomena. The general objectives of this paper are to identify the methods affected by these hypotheses, evaluate their impacts and propose improvements aimed at obtaining more realistic and fairer representations.

Projection pursuit regression is a non-parametric method similar to generalized additive models and artificial neural networks that are considered in AFR to take into account the non-linearity of hydrological processes. In a comparative study, this paper shows that regression with revealing directions makes it possible to obtain more parsimonious models while preserving the same predictive power as the other nonparametric methods.

Canonical Correlation Analysis (ACC) is used to create neighborhoods within which a model (e.g. multiple regression) is used to predict hydrologic variables at ungagged sites on the other hand, ACC strongly depends on the assumptions of normality and linearity. A new methodology for delineating neighborhoods is proposed in this paper and uses revealing direction regression to predict a reference point representing hydrological and physiometeorological information that is relevant to these groupings. The results show that the new methodology generalizes that of ACC, improves the homogeneity of neighborhoods and leads to better performance.

In AFR, kriging techniques on transformed spaces are suggested in order to predict extreme hydrological variables. However, a transformation is required so that the hydrological variables of interest derive approximately from a multidimensional normal distribution. This

transformation introduces a bias and leads to suboptimal predictions. Solutions have been proposed, but have not been tested in AFR. This paper proposes the approach of spatial copulas and shows that this approach provides satisfactory solutions to the problems encountered with kriging techniques.

Max-stable processes are a theoretical formalization of spatial extremes and correspond to a more faithful representation of hydrological processes on the other hand; their characterization of extreme dependence poses technical problems which slow down their adoption. In this paper, the approximate Bayesian calculus is examined as a solution. The results of a simulation study show that the approximate Bayesian computation is superior to the standard approach of compound likelihood. In addition, this approach is more appropriate in order to take into account specification errors.

Keywords

Ranked Statistics, Statistical Distributions, Canonical Correlation Analysis.

Probabilistic Models

A stochastic process Z (x) represents a set of random variables indexed by a random vector x of dimension p. to describe the evolution of a random variable, we can then use a stochastic process as a function of explanatory variables x. The law of Z (x) for a given x is called the marginal law and the law of two or more explanatory variables xi is called joint law. In particular, in the geostatistical context, a stochastic process whose joint law is a multidimensional normal law is called a Gaussian random field.

The traditional methods used in AFR require a preliminary analysis which provides the hydrological variables to be regionalized on the other hand, the need for a preliminary step can be avoided by adopting probabilistic models which determine a stochastic process directly on the extreme values (Azzarri, (2006), In this paper, these models are said to be probabilistic since they offer a complete formulation of the probabilities of extreme values. In AFR, the marginal law of a stochastic process Z (x) represents the law of a hydrological variable at a site associated with physiometeorological characteristics x. The analytical formulation of the relation between x and the parameters of the marginal laws of Z (x) is called the response surface.

This has the same role as a regional model for parameter prediction techniques in the traditional approach. This link is illustrated in Figure:1 by a hatched line. These values extremes are generally annual maximums, but can also be threshold exceedances (Berzel, (2016).

Using a hierarchical structure as shown in Figure 2, we observe that a probabilistic model is closer to the physical mechanism used to generate the extreme values. In addition, we see in Figure: 2 four levels which represent the hierarchy of information: daily data, annual maximums, parameters of marginal distributions and quantiles. This hierarchy is useful in order to represent the structure of dependencies between sites at the different levels. The dependence that arises from the existence of a large-scale meteorological phenomenon that creates a spatial dependence for daily data. However, two annual maximums may not be caused by the same event, which shows that the spatial dependence between extreme values is different (step h). Step g also indicates another form of dependence on the parameters of marginal distributions (Biewen, Martin, and Stephen P. Jenkins (2005). This form of dependence, called residual, indicates that part of the relationship between parameters and physiometeorological characteristics is not completely specified by the response surface. Note also that Figure:1 does not directly show any dependence between the hydrological variables, since this relationship is induced by the structures present at the different levels.

The use of hierarchical structures is common in Bayesian approaches. The Bayesian framework was initially introduced in AFR with the aim of improving the frequency analysis of partially gauged sites by making it possible to combine regional information with that of a site (Birkin, (2009)).

The main difficulty with probabilistic models is the modeling of extreme spatial dependence. To avoid this problem, latent variable models have been proposed in the context extreme precipitation (Chaudhuri, (2002), and in the case of river flows (Hentschel, (2008). For these models, each gauged site is associated with a stochastic process whose marginal laws are of the family of extreme values, and whose annual maximums are independent. The term latent variable here emphasizes the fact that the response surface parameters are not directly observed. Latent variable models can be considered for practical purposes when the objective is strictly to model marginal laws, but are unrealistic in their interpretation of extreme spatial dependence. Figure 2 a shows a theoretical example of simulating a latent variable model which highlights the lack of continuity.

In order to obtain more realistic simulations, Horton, (2001) proposed a Bayesian hierarchical model in which the dependence between sites is described by a Gaussian copula, i.e. the marginal distributions follow a law of extreme values and the dependency structure is that of a Gaussian random field. A theoretical example of simulations using a Gaussian copula is presented in Figure:1, on the other hand, the generalization of the

theory of extreme values to stochastic processes indicates that Gaussian copulas do not represent an appropriate dependency structure (Juhn (2013). The generalization of the theory of extreme values to larger dimensions leads to the definition of extreme copulas (Ravallion, Martin (2011). Although there are a number of extreme two-dimensional copulas (Rigby, (2015), for spatial extremes it is necessary to use a copula with a dimension equal to the number of sites. In this case, the extreme t-copula, constructed as the transformation of Student's laws, is more appropriate in the context of probabilistic models in AFR (Rigby, R. A., and Stasinopoulos, D. M. (2015).

In principle, the generalization of the extreme value theory for stochastic processes leads to models based on max-stable processes (Rothe, Christoph (2010). The use of max-stable processes in practice is relatively recent and still presents several challenges. The main difficulty arises from the fact that the plausibility of the data does not have a form analytical practice and prevents the use of usual statistical tools. However, practical representations of max-stable processes allow them to be defined as the superposition of underlying stochastic processes (Stock, sc James H. (2009). These representations thus make it possible to describe the dependency structure of the max-stable processes thanks to the nature of the underlying processes and to perform simulations of these max-stable processes within reasonable time frames (Su, Yu-Sung, Andrew Gelman, Jennifer Hill, and Masanao Yajima (2011),to this end, Figure 3 shows the realization of a Smith model where the underlying processes are represented by elliptical storm cells described as the superimpositions of normal laws. Also, Figure 3 represents a Brown-Resnick process (Van Kerm, Philippe (2013), which represents a max-stable process like the superposition of Weiner processes. As shown in Figure, the Brown-Resnick models lead to a less simplistic structure than the Smith model. Note also that the approach using the extreme t-copula represents a max-stable process which includes the special case where the underlying processes are Gaussian random fields to date, max-stable processes have been used in hydrology only to model extreme precipitation. In particular, a generalization of the flood index model was presented by Wang et al. (2014). This latest study shows how to estimate the regional model taking into account the dependence between sites in the form of max-stable processes. This significantly reduced the uncertainty compared to the traditional L-moments approach. This gain, however, requires the additional specification of a max-stable dependency structure for which the authors noted that a poor specification of the dependency can lead to significant biases.

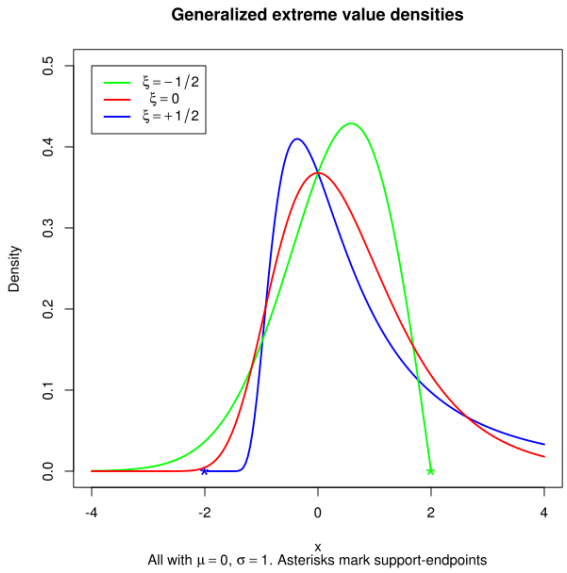

Figure 1 Generalized extreme value densities

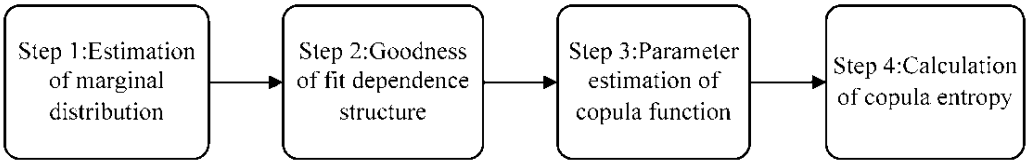

Figure 2 Simulation of stochastic processes with different dependency structures. Laws marginal are unitary Gumbel laws

Proposed Methods and Main Results

In this section we discuss the proposed methods and the main results obtained within the articles presented in this paper. This section is divided into three subsections: choice of data, nonparametric methods and modeling of inter-site dependency. First, a presentation summarizes the data used in the articles. Subsequently, the following two subsections discuss the methods targeted in this paper and underline the particularities of the methods proposed in this paper in order to take into account non-linearity and non-normality. Finally, the important results of this paper are put into context and discussed in a general perspective.

In this paper we have 12,685 residents of the district in Erbil for the years 2019 and 2020. These data include age, gender, and the number of citizens' visits to doctors.

Here there are two typical prediction aunts. Most of the time practitioners only have access to sample data through which we can know the distribution of the number of visits

to the population. On the other hand, the distribution must be predicted for virtual citizens. They are usually considered a very important case.

In order to clarify the application of the method in these two cases, we drew a random sample, for example +, from 310 observations from the year 2019. The age and gender of all individuals in the district were taken, and we imposed the number of visits known to only 310 individuals in the statistical sample presented in Table 1. The goal is to predict the distribution of the number of visits in 2021.

Table 1 Table of statistics, standard deviation

Inhabitance	**male**	**female**		**Male**	**Specimen female**		
No. of individual	12684	13496		132		77	
averages ages in 2019	37 (21)	41	(25)	38	(27)	41	(25)
averages ages in 2020	38 (23)	42	(26)	–		–	
averages No. of visits in 2019	5.2 (6.3)	6.9	(7.2)	6.1	(6.4)	7.9	(7.2)
averages No. of visits in 2020	5.7 (7.5)	8.3	(8.4)	–		–	

The graph shows the number of visits versus age, separately for males and females, in the next figure. Simple local linear regression estimates (lines) indicate a non-linear relationship between mean number of visits with age and gender. The lines differ between males and females and there is evidence of excessive scattering

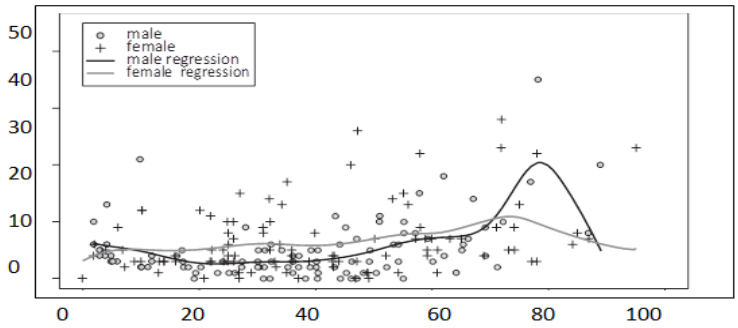

Figure 3 Number of visits to the doctor versus age

Notes: Number of visits to a GP (left, 2019; right, 2020) is a plot versus age for a simple random sample of 310 citizens. Local linear regression estimation with cross validation range.

Choice of Data

When possible, data sets in which preliminary analysis of the gauged sites has been carried out by previous studies have been preferred. This made it possible to concentrate

solely on the prediction part of the ungauged sites. This approach also assures us that the choice of sites is validated and that the usual hypotheses are tested: homogeneity, stationarity and independence. As such, Kouider et al. (2002) built a database of (151) sites located in southern Quebec, Canada, for which an estimate of the flood quantiles of several return periods is available.

Data from Quebec were used in three articles and make it possible to compare the proposed methods with each other. This dataset also has the advantage of having been used for several previous studies for which the flood quantiles were also the hydrological variables of interest. The notation Q100 will be used hereinafter to designate, for example, a quantile of return periods of 100 years.

By previous studies, different selections of physiometeorological characteristics have been considered to explain the behaviors of hydrological variables at sites not. Table: (1) lists the full list of available features. A first selection is proposed by Chokmani et al. (2004) who retained 4 characteristics on the basis of a study of correlations with hydrological variables. The characteristics retained were: the average slope of the watershed (PMBV), the fraction of the watershed occupied by lakes (PLAC), the average annual total precipitation (PTMA) and the number of degree-days below zero degrees Celsius (DJBZ). Note then the decision to exclude the area of the basin (BV) because of the desire to work with flood quantiles standardized by the area of the basin.

Most of the subsequent work kept these 4 variables, but added the air from the watershed (BV), presents a summary of the different characteristics used in the different studies addressed in this paper. The initial selection, however, has been questioned by Chebana et al. (2014) who compared this choice with that obtained by a step-by-step method. This approach made it possible to identify new characteristics on a more objective basis. The same approach was adopted by Durocher et al. (2015a). Note that the combination of the step-by-step method with the non-parametric methods takes into account the non-linearity of the relationship with the variables. There are similarities between the characteristics selected by these last two studies. Indeed, both identified longitude as an important feature and preferred the use of liquid precipitation (PLMA) to total precipitation (PTMA).

Despite the questioning of the choice of physiometeorological characteristics, the methods proposed within this paper must be calibrated on the same characteristics as the previous studies in order to ensure a fair comparison between the methods the proposed method is not compared to methods from previous studies such as physiographic kriging and other

non-parametric methods. A selection based on methods step by step of previous studies was preferred since it leads to better results. In addition, the ACC-based homogeneous region formation approach has been reproduced with these new physiometeorological features.

The data used by Durocher et al. (2015) were the annual precipitation maxima for stations in California. For this study on maxstable processes, it was not possible to directly use the return periods provided by previous analyzes, since the max-stable processes directly use annual maxima as observations. A set of 39 sites were selected. The limited size of this sample was chosen voluntarily because of the complexity of the calculations associated with the method used.

Note that a priori, the dataset of annual floods in Quebec, used within the three other articles of this paper, does not represent an appropriate case study for max-stable processes. Indeed, these represent an asymptotic result based on the maximum of an infinite number of underlying stochastic processes. However, a large part of the annual floods in Quebec are the result of spring melt. Therefore, the annual maximums are essentially the result of one and the same event. The asymptotic arguments of the extreme value theory therefore do not necessarily apply in this case. This reasoning also applies to the choice of marginal laws, i.e. the extreme value theory does not impose a generalized extreme value law. In fact, the frequency analysis of the sites carried out by Kouider et al. (2002) revealed that for a majority of sites, the law selected was of the log-normal or gamma family. Nevertheless, further studies are necessary in order to precisely confirm the real relevance of max-stable processes in the description of annual floods in Quebec.

Table 2 List of basin characteristics available for the Quebec dataset

Variables
Basin area (km2)
Main canal length (km)
Main canal slope (m / km)
Slope of the watershed (°)
Fraction of basin occupied by forests (%)
Fraction of basin occupied by lakes (%)
Average annual total precipitation (mm)
Average annual liquid precipitation (mm)
Mean annual solid precipitation (cm)
Jul-Dec mean liquid precipitation (mm)
Snow level as of March 30 (cm)
Degree-days below 0 Celsius (° / day)
Basin area (km2)
Main canal length (km)

Non-Parametric Methods

1) Existing Nonparametric Methods

Methods considering the non-linearity between hydrological variables and physiometeorological characteristics have already been proposed in AFR. Support for nonlinearity in a regional model is generally done through nonparametric methods for which no specific shape is determined a priori. Note that multiple regression methods within neighborhoods indirectly allow to take into account non-linearity. Indeed, moving from one point of reference to another, a large part of the sites used are the same. Consequently, the parameters of the multiple regression model vary gradually and can be considered as localized approximations of a regional model applied to all the sites studied. A general discussion of this type of statistical approach can be found in Hastie et al. (2009) and neighborhood-type approaches that were considered for the Quebec data are ACC.

For Quebec data, ANNs are also used as quintile prediction techniques by Shu et al. (2007). Their results showed that direct adjustment of a single ANN performs less well than multiple regression within neighborhoods. The improvements they have made are the use of ACC as a pre-processing of physiometeorological characteristics and the use of a bootstrap, or bagging, aggregation method. With these modifications, the approach using the ANNs led to performance superior to the traditional ACC approach, but similar to that of the depth functions.

The regression equations offered by the methods using neighborhoods are straightforward, but are only valid for a target site. In order to obtain nonparametric models with a better understanding of the role of physiometeorological characteristics for all the sites, GAM models were applied to data from Quebec by Chebana et al. These are more practical since they explicitly describe the role of each physiometeorological characteristic for the entire study region. However, these methods performed slightly less well than the ANNs for the data from Quebec.

2) Proposed Nonparametric Methods

Regression with revealing directions, or projection pursuit regression (PPR), was initially introduced by Friedman et al. (2015). To illustrate these models, let x be a vector of several physiometeorological characteristics and y a hydrological variable. A PPR model is then written:

$$y = f_i + \sum_{i=1} \phi_t (\mu_i' \xi) + \varepsilon \text{.....(1)}$$

where μ is the global mean, the f_i are nonlinear functions with zero mean $\sum \phi_t = 0$ et and the α_i are vectors of unit coefficients $\alpha = 1$, or revealing directions, used to define, intermediate predictors $\eta_t = \alpha_t' \mathbf{x}$.. In general, PPR models can use multiple f_i $(\alpha_t' \mathbf{x})$, characteristics and summation of p terms f_i $(\alpha_t' \mathbf{x})$ In these general situations, the use of PPRs is comparable to that of RNAs since it shares several similarities. Indeed, comparative studies have shown that these two methods have similar predictive powers and allow a regression model to be approximated with the desired, on the other hand, PPRs also share certain disadvantages of RNAs which lead to overparametrized models and without a single solution. In addition, when the number of revealing directions is imposing, this number masks the role of the characteristics physiometeorological within each term f_i $(\alpha_t' \mathbf{x})$.

However, the use made of PPRs in this paper is very different from the general case. Durocher et al. (2015a) used PPRs as quintile prediction techniques for annual floods in Quebec. The objective was to test if these models could be useful in AFR when limited to only a few revealing directions. Indeed, the particular case using a single direction can be treated as a distinct approach. In this precise case, the predictor f_i $(\alpha_t' \xi)$ plays the role of a linear model for which f determines an unknown transformation before estimation. The one-way PPR model therefore offers an explicit interpretation of the regional model, similar to the classical regression model. Note that a neuron in an RNA model has the same form a term f_i $(\alpha_t' \mathbf{x})$ in a PPR model. However, a neuron uses nonlinear functions f_i specific, with a limited number of parameters, while for a PPR model f_i are non-parametric and therefore more flexible. This distinction is important since it indicates that an RNA model reduced to a single neuron has limited applications.

PPR models have also been used by Durocher et al. (2015b) to form neighborhoods. This method will be referred to hereafter as RVN, for Reference Variables Neighborhoods. Approaches in AFR using neighborhoods should include three stages: (i) identification of neighborhood centers (ii) formation of neighborhoods and (iii) estimation of regional models. The fundamental point of the RVN method concerns step (i). In this step, the PPRs are used in order to predict reference variables which represent the centers of the neighborhoods. These reference variables may be hydrological variables known at the gauged sites, but unknown at the target sites. Thus, the prediction of the reference

variables at the ungauged sites represents an important preliminary step in order to determine the unknown center of a neighborhood. In contrast, the ROI method considers that all the reference variables are physiometeorological characteristics which are already known. Consequently, step (i) for the ROI method is straightforward and does not require analysis preliminary to the formation of neighborhoods (ii).

The performance of regional models is a central point in the evaluation of the methods proposed in AFR. By relaxing the linearity assumptions, we want to obtain more flexible approaches that will better reflect reality and improve the efficiency of the models. The leave-one-out cross-validation method is adopted as a cross-validation technique in three articles of this paper. This evaluation method considers each target site in turn as an ungauged site. This approach leads to a procedure that is easily reproducible. In addition, the exclusion method was chosen by previous studies serving as a comparison in this paper.

Moreover, for nonparametric methods like PPR, there is not always an adequate measure of complexity (e.g. degree of freedom) as in the case of multiple regression. This prevents the consideration of evaluation criteria such as the Akaike or Schwartz information criterion, which penalizes models on the basis of this measure of complexity. The exclusion method provides a measure of the actual quality of a model's predictions without the use of a complexity measure.

The hydrological variables of interest in Durocher et al. (2015a) are the quantiles Q10 and Q100 which are standardized by the area of the basin. The results of cross-validation using the characteristics of the previous studies, as shown in Table 2b, showed that two methods considered in the previous studies stood out according to the root of the mean relative error (RERM) with a value of approximately 45 % for Q100. These methods are depth functions. For their part, the PPR models and the additive models obtained RERMs of around 48% and the method using neighborhoods constructed from the ACC obtained a RERM of 51%.

However, one aspect was not taken into account in the comparisons of the results. Chokmani et al. (2004) have shown the existence of six problematic sites which have a significant impact on the RERM criterion. Indeed, for the physiographic kriging technique, the RERM goes from 70% to 41% when these sites are excluded, ie a difference of 29%. Likewise, Durocher et al. (2015a) showed that for the PPR method the removal of these six sites reduced the RERM by 14%. From the results available in the previous studies, it was not possible to determine whether the difference in performance

between the PPRs and the other previous methods is mainly due to a better overall estimate or better management of these six problematic sites. The latter nevertheless represent a reality in AFR and the results obtained indicate how the methods used perform under these conditions.

Furthermore, the use of the PPR method in combination with the step-by-step methods reduced the RERM by 8%, to reach 40%. By way of comparison, with the combination of the GAM models with the step-by-step method of Chebana et al. (2014), we find an RERM of 42% which is slightly higher than that of PPR. In addition, note that the PPR method uses a single nonlinear function unlike GAM models which have five nonlinear functions, one for each physiometeorological characteristic. This demonstrates that the telltale direction found in the PPR models is characteristic of the mechanism of flood formation and has led to a more parsimonious representation of the nonlinear aspect than the GAM models, without sacrificing performance.

In the study conducted by Durocher et al. (2015b) comparisons specifically targeted methods with neighborhoods. The choices made for this study mean that the results are not directly comparable with those of previous studies, since Durocher et al. (2015b) considers a quantile that is not standardized by the basin air and the study only uses other physiometeorological characteristics than those of the previous studies. However, the results of the RVN method for the same configuration as that of the previous studies are presented here.

In this case study, the neighborhood centers are formed from a combination of physiometeorological characteristics (BV, PLAC) and L-moments (L-coefficient of variation and L-coefficient of asymmetry). Neighborhoods are made up of a fixed number of gauged sites, determined according to the Euclidean distance between the sites and the center of the neighborhood. Additionally, this RVN method uses a multiple regression model to predict Q100. The results of the RVN method under these conditions lead to an RERM of 42% which is more efficient than the methods considered in the previous studies. Durocher et al. (2015a) also assessed the performance of PPR models using a Nash-Sutcliffe (NSH) criterion and found a value of 71%. For the method of Durocher et al. (2015b) under the same conditions, we obtain an NSH of 72%, which is slightly higher. Overall, these results show that the use of PPRs within the RVN method was more efficient than their direct use as a quantile prediction technique.

Durocher et al. (2015b) have shown the need to consider non-linearity in the RVN method. Indeed, Durocher et al. (2015b) reported that linearity-based ACC performed less well than RVN in terms of REQM and NSH. However, the main advantage of neighborhoods formed by the RVN method is homogeneity. Durocher et al. (2015b) showed that the dispersion of the coefficient of variation was on average smaller for the neighborhoods delimited by

the RVN approach than for the ROI and ACC method. These results show that by choosing the right reference variables, it makes it possible to group together similar sites according to the desired properties and that this choice results in better predictions.

The interpretation of results of the PPR method on revealing directions is also richer. Indeed, Durocher et al. (2015b) have shown that using the usual transformations (logarithms, square roots), the relationship between physiometeorological characteristics and the average annual flood becomes linear. However, nonparametric methods such as PPR are necessary in order to take into account the non-linearity of other aspects, such as L-coefficients of variation and L-asymmetry coefficients, for which no usual transformation is adequate (no linearizable).

Methods for Estimating Interstice Dependence

1) Estimation of the Dependence between Quantiles

Persistence within data can be viewed as a deterministic trend or as a form of dependence between the variables studied. Nonparametric methods describe this persistence as a nonlinear trend, while geostatistical methods take the opposite approach and represent persistence as a dependency. However, in practice, these two modeling approaches can be combined in order to include a trend and a dependency structure in the same model. On the other hand, these two parts must be clearly distinct to avoid any identification problems.

For the data from Quebec, the dependence between the flood quantiles was modeled by Chokmani et al. (2004) and by Nezhad et al. (2010). Chokmani et al. (2004) used ordinary kriging where persistence is treated only as a covariance determined by a semi-variogram. The latter describes the evolution of the covariance as a function of the distance separating two site (s. The nugget effect is a property of the semivariogram which introduces a non-negligible error for the hydrological variables at the sites. In particular, the nugget effect is necessary in AFR since the flood quantiles are not measured, but calculated from daily data. In addition, Chokmani et al. (2004) have shown that the space

constructed from the ACC is more adequate than that of the PCA for data from Quebec. Nezhad et al. (2010) analyzed the same data, but used the residue kriging method in order to introduce a deterministic trend. Specifically, a quadratic trend was fitted to the hydrologic variables, and then the ordinary kriging method was applied to the tailings. These two parts are estimated separately using least squares to fit the trend and the semivariogram.

The main contribution of the spatial copulas framework in the study of data from Quebec is the consideration of a nonlinear predictor and the addition of a linear trend for the standard deviation. This trend shows that the variability tends to decrease in the direction of the predictor provided by the first canonical pair. Durocher et al. (2015c) also considered the use of pairwise likelihood as an estimation method. This method is a special case of the theory of compound likelihood adapted to the context of spatial data (Heagerty et al., 1998, Varin, 2008). Unlike the least squares method used by Nezhad et al. (2010), this method makes it possible to jointly estimate the trend and the dependence in a single step.

An important advantage of the method proposed by Durocher et al. (2015c) on kriging techniques is that it made it possible to calculate the full predictive law of hydrological variables at ungauged sites. Knowledge of this law can be used to calculate the mean of the forecasts which is the best unbiased predictor (Bárdossy et al., 2008). On the other hand, Durocher et al. (2015c) also considered the median of the predictive law which led to better performances in terms of RERM. Indeed, Durocher et al. (2015c) showed that using the median, the spatial copula approach was the best performing geostatistical method with a REQM of 41% for Q100, compared to 70% for Chokmani et al. (2004) and 58% for Nezhad et al. (2010). In fact, the spatial copulas approach has surpassed the best nonparametric methods on the same data.

2) Estimation of the Dependence between the Annual Maximums

As shown in Figure 2, hierarchical models make it possible to model the dependency structure on several levels, which makes it possible to take into account an extreme type of dependence between the annual maximums. California's extreme precipitation was studied using max-stable processes by Shang et al. (2011). The main objective of this study was to study the influence of the Southern Oscillation on extreme precipitation. The consideration of the max-stable processes makes it possible to study the influence of the southern oscillation on all the sites simultaneously. This made it possible to improve the results obtained by analyzing the sites individually one of the limitations of the study by

Shang et al. (2011) was to limit itself to a Smith model which offers a simplistic representation of meteorological phenomena, as shown in Figure 3c.

Faced with the impossibility of explicitly calculating the joint probabilities of a max-stable process, the estimation of the latter is generally carried out according to the theory of compound likelihood. Note that in certain situations, optimizing the compound likelihood can be difficult. In addition, a poor specification of the extreme dependency structure can lead to significant bias (Wang et al., 2014). These difficulties have motivated the search for alternative estimation methods. E. L. Smith et al. (2009), M. Ribatet et al. (2009) and Erhardt et al. (2012) have proposed Bayesian approaches which use algorithms allowing the sampling of the posterior distribution of the parameters. However, these approaches are approximate in the sense that they do not correspond to the true posterior distribution calculated by Bayesian theory. Durocher et al. (2015d) used the Bayesian calculation approximate, or ABC for approximate bayesien computing, as proposed by Erhardt et al. (2012).

This simple approach then leads to an approximation of the posterior distribution of the traditional Bayesian framework and the quality of this approximation is controlled by the thre hold $\hat{\theta}\hat{}$. Small values of $\hat{\theta}$ lead to better approximations, but also lead to rejecting several candidates $\hat{\theta}\hat{}$. Consequently, the choice of $\hat{\theta}$ offers a compromise between the quality of the approximation and the calculation time necessary to draw the number of parameters requested. In practice, faster ABC algorithms have been proposed in order to obtain better approximations than the rejection algorithm, by limiting the number of candidates $\hat{\theta}\hat{}$ rejected.

Erhardt et al. (2012) showed that under certain conditions, the ABC estimator was more efficient than that of the compound likelihood. Note that the ABC procedure strongly depends on the choice of T statistics. Thus Erhardt et al. (2012) considered several groups of statistics and found that statistics based on probability triples were more efficient. By against, Durocher et al. (2015d) chose to use an empirical madogram which is the analogue of a semivariogram in the context of max-stable processes (Cooley et al., 2006). This choice leads to a group of summary statistics of reasonable size which is relatively quick to calculate. This choice was validated using a simulation study by Durocher et al. (2015d) who confirmed that the performance of the ABC approach with the empirical madogram was superior to that obtained with the pairwise likelihood.

Note that the results of the simulation study by Durocher et al. (2015d) conflict to some extent with those of Erhardt et al. (2012). They found better performance from the pairwise likelihood estimator than the ABC approach using the empirical mandogram. Two factors can explain these differences. First, the theoretical madogram considered by Durocher et al. (2015d) was based on an exponential correlation function with a single parameter. That of Erhardt et al. (2012) was based on a Whittle-Matérn correlation function with two parameters. These results suggest that the ABC approach may lose its effectiveness when the dependency structure becomes more complex and the role of each parameter becomes more difficult to identify. Furthermore, Durocher et al. (2015d) used larger ABC samples in addition to post-processing (Blum et al., 2010). This post-processing aims to correct the bias and the variance of the a posteriori distribution resulting from an ABC algorithm in presence of a non-negligible threshold $\hat{\theta}$. These technical improvements may have led to a sufficient improvement of the posterior distribution approximation in order to overcome the pairwise likelihood approach.

In addition, the study by Erhardt et al. (2012) estimated only the dependency structure of a max-stable process. An innovation from Durocher et al. (2015d) is to jointly consider the estimation of marginal distributions and the dependency structure. Durocher et al. (2015d) thus added the L-moments to the empirical mandogram in the group of summary statistics T (y). This choice was also validated by a second simulation study which, however, led to less efficient estimates than those of the paired likelihood. On the other hand, the results obtained are acceptable since they are comparable to those obtained by a model with latent variables, which assumes independence between the sites.

Durocher et al. (2015d) compared the ABC estimator with the pairwise likelihood estimator on extreme precipitation data from California. The response surface retained has marginal laws of the family of extreme values with a constant shape parameter and a dispersion parameter proportional to the location parameters. Note that these assumptions are similar to those of a flood index model. Examination of the empirical madogram showed that the pairwise maximum likelihood estimator failed to adequately estimate the spatial dependence parameter. In addition, the performance obtained from a validation set indicated lower performance than that of the ABC estimator. These results are consistent with the conclusions of Wang et al. (2014), who advised of the importance of choosing well and estimating the dependence of a max-stable process.

Note that Durocher et al. (2015d) and Erhardt et al. (2012) limited their studies to Schlater models, due to the time required to perform a simulation, compared to Brown-Resnick models (Oesting et al., 2012). However, the latter could be more appropriate in the case of extreme precipitation in California. Indeed, Schlater's model is a max-stable process whose underlying processes are Gaussian random fields. This model is more realistic than Smith's model, but has an important limitation since it does not allow independence between two sites. This behavior can be justified for small regions as in the case of Durocher et al. (2015d), but is unrealistic for larger areas where two distant sites are not always affected by the same extreme event. The Brown-Resnick model does not have this limitation and consideration of these models could improve the results for extreme precipitation in California. However, this would make the ABC procedure more difficult to use. The improvement of simulation methods for max-stable processes is currently the subject of study (Dombry et al., 2013, Oesting et al., 2014) and future improvements could make the ABC approach more attractive and more generally applicable in AFR.

Result

The results shown, by using the cross-validation procedure for all the combinations studied, are in Fig. 1. Therefore, the best overall performance should be those that we obtain from the NLCCA-GAM nonlinear model - first when the explanatory variables are those that have been identified. In then the case where the variables are selected in a stepwise fashion (mainly in terms of RRMSE). In what follows, we refer by NLCCA-GAM to the model using the BV, MBS, AMP, FAL, and AMD variants. According to high EC values (> 0.8) and lowest RRMSE values (28.35% for QS100), NLCCA-GAM provides estimates that are accurate compared to all other methods. Based on the RBIAS, the results show that although all models underestimate flood quantities, CCA-GAM is the model with the least bias (-4.8% for QS100). In addition, compared to the NLCCA-GAM approach, the difference is not significant (30.2% difference for QS100).

The results also indicate that the NLCCA-GAM approach produces more accurate estimates when we compare it to the same approach using gradually determined variables, although the difference is not very large. This can be explained by the fact that the criteria used to select the variables (GCV, stepwise-based combination, the used NLCCA solution is the same as in the NLCCA-GAM. Hence, through a more advanced NLCCA parameterization, better results could be achieved by using the stepwise approach.

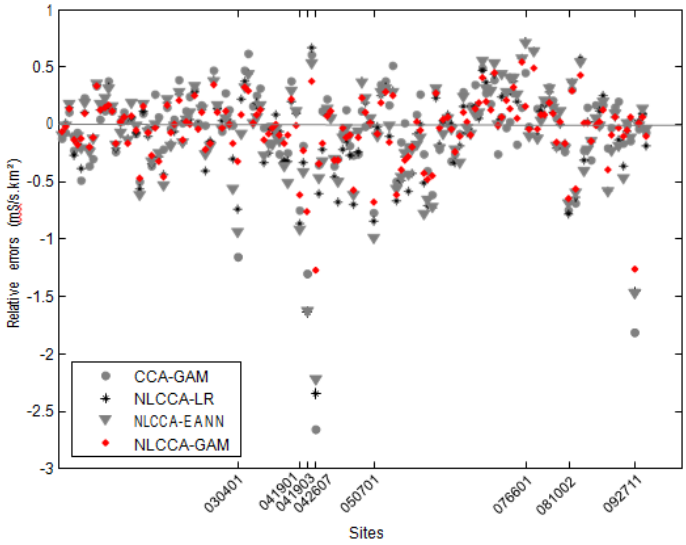

Figure 4 Relative error associate with QS100 calculate at each sites using CCA-GAM, NLCCA-LR, NLCCA-EANN, and NLCCA-GAM

Conclusions

In AFR, many statistical tools are used to model the relationship between hydrological variables and physiometeorological characteristics in order to predict the risks of occurrence of extreme events at ungauged sites. This paper tackled the validity of the assumptions of normality and linearity resulting from the use of these statistical approaches in AFR. The main tools targeted were non-parametric methods, ACC, krigeages techniques and max-stable processes. Although performance is a fundamental aspect of AFR, this paper seeks to highlight the quality of the representation of the models proposed in contexts where extreme hydrological phenomena are nonlinear and non-normal.

Nonparametric methods, such as ANNs, offer limited interpretation of results and require a large amount of information in order to be effective. The PPR method was proposed in order to take into account the non-linearity in the techniques for predicting flood quantiles in Quebec. This method made it possible to adequately adjust the flood quantiles estimated at the sites and to highlight important predictors (or revealing directions). These predictors make it possible to obtain regression equations that are explicit and easy to interpret. Consequently, the proposed method led to more parsimonious models than additive models and ANNs, without sacrificing predictive performance.

The use of the PPR method has also been proposed in order to predict reference variables at ungauged sites. These reference variables were used to construct neighborhoods

making it possible to identify the regional law of ungauged sites and to predict the desired flood quantiles. This method generalizes the ACC approach, considering the non-linearity in the relationship between hydrological variables and physiometeorological characteristics. In the case of the AFR of floods in Quebec, this paper shows that the use of summary statistics, such as L-moments, leads to the formation of more homogeneous neighborhoods. This improvement resulted in a reduction in uncertainties for the models estimated within neighborhoods, which led to improved predictive performance compared to traditional ACC and ROI methods.

In AFR, geostatistical methods are used to predict hydrological variables inside transformed spaces at ungauged sites. This paper proposed spatial copulas as a framework to describe the dependence between flood quantiles by copulas. The spatial copula method has been shown to be superior to kriging techniques which do not make it possible to adequately take into account the non-normality of the regional distribution of hydrological variables and to include a structure taking into account heteroskedasticity. The results obtained by the use of spatial copulas on flood data from Quebec have shown significant reductions in bias in addition to offering the most efficient forecasts of all the methods considered in this paper.

Max-stable processes are a generalization of extreme value theory. These models lead to a more faithful representation of the true mechanism of generation of hydrological extremes and therefore of the dependence between the sites. However, several numerical difficulties result from the absence of an analytical form of joint probabilities. This problem can be solved by using a compound likelihood or an ABC procedure. This paper showed from a simulation study that the ABC approach was in certain circumstances the best method to estimate the dependence structure of max-stable processes.

Research Prospects

This paper proposed several methods which make it possible to improve the current methods encountered in AFR. The results found in this paper lead in turn to new questions that could serve as avenues of research for future work.

The use of max-stable processes has been exclusively considered in AFR for extreme precipitation. However, flood flows can be the consequence of this extreme precipitation. Therefore, it would be interesting to check whether the shape of the dependence between the gauged basins is that of max-stable processes. Several questions on how to specify this structure remain since river flows are not continuous in geographic space. Answers to

these questions could come from geostatistical methods in AFR. In particular, max-stable versions of physiographic and topographic kriging techniques are possible.

The study by Durocher et al. (2015c) considered the issue of heteroskedasticity and this contribution led to good performance. In general, parameter prediction techniques and probabilistic models will take this aspect into account since a dispersion parameter is directly estimated. In traditional methods, the use of homogeneous regions also makes it possible to take into account heteroskedasticity. However, the nonparametric methods considered on the flood data in Quebec assume a constant variance at the logarithmic scale. Future work should be considered to investigate the need to specifically model the variance for nonparametric methods in AFR. In particular, modeling approaches such as proposed by Fan et al. (1998) where the variance is estimated from the residuals should be validated in these situations.

The study by Durocher et al. (2015b) showed that the proximity between reference variables leads to a notion of distance which is beneficial in the formation of neighborhoods and surpasses the ROI method, based on the proximity between the physiometeorological characteristics. This notion distance between reference variables could also be integrated into geostatistical methods in AFR. Indeed, the use of this approach could lead to new spaces within which geostatistical methods would be used to predict hydrological variables. In general, several other proximity concepts could be combined with geostatistical methods in AFR. Other examples would be the dissimilarity resulting from depth functions or the notion of distance in space generated by revealing directions (PPR). These approaches could represent interesting alternatives to ACC and PCA.

The validation of the methods proposed in this paper was carried out on case studies. This approach shows that the proposed methods can be beneficial in certain situations. In future research, simulation studies should be carried out in order to more precisely determine the conditions under which the methods proposed in this paper stand out.

References

Azzarri, C. (2006). Monitoring poverty without consumption data: an application using the Albania panel survey. *Eastern European Economics*, *44*(1), 59-82.

Berzel, A. (2016). Estimating the number of visits to the doctor. *Australian & New Zealand Journal of Statistics*, *48*(2), 213-224.

Biewen, M., & Jenkins, S.P. (2005). A framework for the decomposition of poverty differences with an application to poverty differences between countries. *Empirical Economics*, *30*(2), 331-358.

Birkin, M., & Clarke, M. (2009). The generation of individual and household incomes at the small area level using synthesis. *Regional Studies*, *23*(6), 535-548.

Chaudhuri, (2002). *Assessing Household Vulnerability to Poverty from Cross-Sectional Data: A Methodology and Estimates from Indonesia*.

Hentschel, (2008). Combining census and survey data to trace the spatial dimensions of poverty: A case study of Ecuador. *The World Bank Economic Review*, *14*(1), 147-165.

Horton, (2001). Multiple imputation in practice: comparison of software packages for regression models with missing variables. *The American Statistician*, *55*(3), 244-254.

Juhn, (2013). Wage inequality and the rise in returns to skill. *Journal of political Economy*, *101*(3), 410-442.

Ravallion, M. (2001). Growth, inequality and poverty: looking beyond averages. *World development*, *29*(11), 1803-1815.

Rigby, (2015). Generalized additive models for location, scale and shape. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, *54*(3), 507-554.

Rothe, C. (2010). Nonparametric estimation of distributional policy effects. *Journal of Econometrics*, *155*(1), 56-70.

Stock, J.H. (2009). Nonparametric policy analysis. *Journal of the American Statistical Association*, *84*(406), 567-575.

Su, Y.S.G. (2011). Multiple imputation with diagnostics (mi) in R: Opening windows into the black box. *Journal of Statistical Software*, *45*(2), 1–31.

Van Kerm, P. (2013). Generalized measures of wage differentials. *Empirical Economics*, *45*(1), 465-482.

Sohrabi, B., Vanani, I.R., & Namavar, M. (2019). Investigation of trends and analysis of hidden new patterns in prominent news agencies of Iran using data mining and text mining algorithms. *Webology*, *16*(1), 114-137.