Exploratory functional flood frequency analysis and outlier detection

Fateh Chebana[§], Sophie Dabo-Niang[♯] and Taha B.M.J Ouarda[§]

§ Canada Research Chair on the Estimation of Hydrometeorological Variables, INRS-ETE,
490 rue de la Couronne, Quebec (QC), Canada G1K 9A9

♯ Laboratoire EQUIPPE, Université Charles De Gaulle, Lille 3, maison de la recherche,
domaine du pont de bois, BP 60149, 59653 Villeneuve d'Ascq Cedex, France.

January, 20th 2012

[§] Corresponding author:

Tel: +1 (418) 654-2542

Fax: +1 (418) 654-2600

Email: fateh.chebana@ete.inrs.ca

Abstract: The prevention of flood risks and the effective planning and management of water resources require river flows to be continuously measured and analyzed at a number of stations. For a given station, a hydrograph can be obtained as a graphical representation of the temporal variation of flow over a period of time. The information provided by the hydrograph is essential to determine the severity of extreme events and their frequencies. A flood hydrograph is commonly characterized by its peak, volume and duration. Traditional hydrological frequency analysis (FA) approaches focused separately on each of these features in a univariate context. Recent multivariate approaches considered these features jointly in order to take into account their dependence structure. However, all these approaches are based on the analysis of a number of characteristics, and do not make use of the full information content of the hydrograph. The objective of the present work is to propose a new framework for frequency analysis using the hydrographs as curves: functional data. In this context, the whole hydrograph is considered as one infinite dimensional observation. This context allows to provide more effective and efficient estimates of the risk associated with extreme events. The proposed approach contributes to addressing the problem of lack of data commonly encountered in hydrology by fully employing all the information contained in the hydrographs. A number of functional data analysis tools are introduced and adapted to flood FA with a focus on exploratory analysis as a first stage towards a complete functional flood FA. These methods, including data visualization, location and scale measures, principal component analysis as well as outlier detection, are illustrated in a real-world flood analysis case study from the province of Quebec, Canada.

Key Words: Functional data, frequency analysis, hydrology, flood, outliers, exploratory analysis, principal component analysis.

1 Introduction

Extreme hydrological events such as floods, droughts and rain storms may have significant economic and social consequences. Hydrological frequency analysis (FA) procedures are essential and commonly used for the analysis and prediction of such extreme events, which have a direct impact on reservoir management and dam design. Flood FA is based on the estimation of the probability $P(X > x_T)$ of exceedence of the event x_T corresponding to a quantile of a given return period T e.g. T = 10, 50 or 100 years. The random variable X is commonly taken to be the peak of the flood which is the maximum of the daily streamflow series during a hydrological year or season. Relating the magnitude of extreme events to their frequency of occurrence, through the use of probability distributions, is the principal aim of FA (Chow et al., 1988).

The accurate estimation of the risk associated with the design and operation of water infrastructures requires a good knowledge of flood characteristics. Indeed, an overestimation of the design flood leads to an over-sizing of hydraulic structures and, would therefore involve additional costs, while underestimation of design floods leads to material damages and loss of human lives. Flood FA is commonly employed to study this risk. It has been traditionally carried out for the analysis of flood peaks in a univariate context. The reader is referred, e.g. to Cunnane (1987) and Rao and Hamed (2000).

In general, a flood is described through a number of correlated characteristics, e.g. peak, volume and duration. The univariate treatment of each flood characteristic ignores their dependence structure. Consequently, the univariate framework is less representative of the phenomenon and reduces the risk estimation accuracy. Thereafter, several authors focused on the joint treatment of flood characteristics through the use of a number of multivariate techniques such as multivariate distributions and copulas (e.g. Yue et al., 1999; Shiau, 2003; Zhang and Singh, 2006; Chebana and Ouarda, 2011a). Multivariate studies contributed to the improvement of the estimation accuracy and provide information concerning the dependence structure between flood characteristics. The multivariate framework is applied in several hydrological events, such as floods, droughts and storms. For instance in floods, it is used for hydraulic structure design and extreme event prediction purposes (see Chebana and Ouarda, 2011a for recent references).

Despite their usefulness, univariate and multivariate FA approaches have a number of limitations and drawbacks. The separate or joint use of hydrograph characteristics constitutes a major simplification of the real phenomenon. Furthermore, the way these characteristics can be determined is neither unique nor objective (in particular, flood starting and ending dates). In addition, each flood characteristic can be seen as a real-valued transformation of the hydrograph, e.g. the peak is the maximum. For hydrological applications, the bivariate setting is largely employed to treat two hydrological variables. A limited number of studies deals with the trivariate one, e.g. Serinaldi and Grimaldi (2007) and Zhang and Singh (2007). The trivariate models generally suffer from less representativity and formulation complexity. Note that, in general, the number of associated parameters grows up rapidly with the dimension of the model and therefore the generated uncertainty increases. In addition, higher dimensions are not considered in hydrological practice. Finally, given the lack of data in hydrology, working with a limited number of extracted characteristics represents a loss of information in comparison to the overall available series.

The main data source in FA is daily streamflow series, which during a year constitutes a hydrograph, from which the univariate and multivariate variables are extracted. The total information that is available in a hydrograph is necessary for the effective planning of water resources and for the design and management of hydraulic structures. The entire hydrograph, as a curve with respect to time, can be considered as a single observation within the functional context. In the univariate and the multivariate settings an observation is respectively a real value and a vector. Therefore, the functional framework which treats the whole hydrograph as a functional observation (function or curve) is more representative of the real phenomena and makes better use of available data. Figure 1 illustrates and summarizes the three frameworks.

In the hydrological literature, there were some efforts towards a representation of the hydrograph as a function, such as in the study of the design flood hydrograph, e.g. Yue et al. (2002), and in the flow duration curve study e.g. by Castellarin et al. (2004) where the mean, median and variation are presented as curves. These studies underlined the importance to consider the shape of the hydrograph which is necessary, for instance, for water resources planning, design and management. The shape of flood hydrographs for a given river may change according to the observed storm or snowmelt events. More practical issues and examples related to the hydrograph can be found for instance in Yue et al. (2002) or Chow et al. (1988). Note that the main flood characteristics, i.e. peak, volume and duration, can not completely capture the shape of the hydrograph. The study of the hydrographs in Yue et al. (2002), and similar studies, are simplistic and limited, as they approximated the flood hydrograph using a two-parameter beta density and considered only single-peak hydrographs. On the other hand, the flow duration curve approach (Castellarin et al., 2004) is in the univariate setting and the presented functional elements (e.g. mean and median curves) are important but remain limited. The previous studies show the need to introduce a statistical framework to study the whole hydrograph and to perform further statistical analysis. The functional framework is more general and more flexible and can represent a large variety of hydrographs.

Functional data are becoming increasingly common in a variety of fields. This has sparked a growing attention in the development of adapted statistical tools that allow to analyze such kind of data. For instance, Ramsay and Silverman (2005), Ferraty and Vieu (2006) and Dabo-Niang and Ferraty (2008) provided detailed surveys of a number of parametric and nonparametric techniques for the analysis of functional data. In practice, the use of functional data analysis (FDA) has benefited from the availability of the appropriate statistical tools and high performance computers. Furthermore, the use of FDA allows to make the most of the information contained in the functional data. The aims of FDA are mainly the same as in the classical statistical analysis, e.g. representing and visualizing the data, studying variability and trends, comparing different data sets, as well as modeling and predicting. The majority of classical statistical techniques, such as principal component, linear models, confidence interval estimation and outlier detection, were extended to the functional context (e.g. Ramsay and Silverman, 2005). The application of FDA has been successfully carried out, for instance, in the case of the El Niño climatic phenomenon (Ferraty et al., 2005) and radar wave curve classification (Dabo-Niang et al., 2007). Dabo-Niang et al. (2010) proposed a spatial heterogeneity index to compare the effects of bioturbation on oxygen distribution. Delicado et al. (2008) and Monestiez and Nerini (2008) considered spatial functional kriging methods to model different temperature series. Sea ice data are treated in the FDA context by Koulis et al. (2009).

The functional methodology constitutes a natural extension of univariate and multivariate hydrological FA approaches (see Figure 1). This new approach uses all available data by employing the whole hydrograph as a functional observation. In other words, FDA permits to exhaustively analyze hydrological data by conducting one analysis on the whole data instead of several univariate or multivariate analysis. In addition, the approach proposed by Yue et al. (2002) can be generalized in the FDA context where it becomes more flexible and includes hydrographs with different shapes such as multi-peak ones.

Given the above arguments, for hydrological applications, the functional context could be seen as an alternative framework to the univariate and multivariate ones, or it can also be employed as a parallel complement to bring additional insight to those obtained by the two other frameworks. The main objective of the present paper is to attract attention to the functional nature of data that can be used in all statistical techniques for hydrological applications through the FDA framework. A second objective is to introduce some of the FDA techniques, point out their advantages and illustrate their applicability in the hydrological framework. In the present paper, we focus on hydrological FA.

Four main steps are required in order to carry out a comprehensive hydrological FA: i) descriptive and exploratory analysis and outlier detection, ii) verification of FA assumptions, i.e. stationarity, homogeneity and independence, iii) modeling and estimation and iv) evaluation and analysis of the risk. The first step (i) is commonly carried out in univariate hydrological FA as pointed out, e.g. by Rao and Hamed (2000), Kite (1988) and Stedinger et al. (1993) whereas in the multivariate framework it was investigated recently by Chebana and Ouarda (2011b). Contrary to the univariate setting, exploratory analysis in the multivariate and functional settings is not straightforward and requires more efforts. Table 1 summarizes the four FA steps and their status in each one of the three frameworks. It is indicated that the specific aim of the present paper is to treat step (i) which deals with data visualization, location and scale measures as well as outlier detection. A new non-graphical method to detect functional outliers is also proposed in the present paper. The presented techniques are applied to floods based on daily streamflow series from a basin in the province of Quebec, Canada.

Exploratory data analysis as a preliminary step of FA is useful for the comparison of hydrological samples and for the selection of the appropriate model for hydrological variables. It consists in a close inspection of the data to quantify and summarize the properties of the samples, for instance, through location and scale measures. Outliers can have negative impacts on the selection of the appropriate model as well as on the estimation of the associated parameters. In order to base the inference on the right data set, detection and treatment of outliers are also important elements of FA (Barnett and Lewis, 1998). Therefore, it is essential to start with the basic analysis (step i) in order to perform a complete functional FA.

This paper is organized as follows. The theoretical background of functional statistical methods is presented in Section 2 in its general form. In Section 3, the functional framework is adapted to floods. The functional FA methods are applied, in Section 4, to a real-wold case study representing daily streamflows from the province of Quebec, Canada. A discussion as well as a comparison with multivariate FA are also reported in Section 4. Conclusions and perspectives are presented in the last section.

2 Functional data analysis background

This section presents the general functional techniques. It is composed of four parts representing FDA phases: first, data smoothing is discussed, second location and scale parameters are introduced, then functional principal component analysis (FPCA) is described and finally data visualization and outlier detection methods are presented.

Data are generally measured in discrete time steps such as hours or days. Therefore, the first phase in FDA consists in the conversion of observed discrete data to functional data. Once the discrete data are transformed to curves, they can be analyzed in the functional framework. In a descriptive statistical study, it is of interest to obtain estimates of the location and scale parameters within FDA. The next phase in the considered FDA is to extract information from functional data using FPCA where the corresponding scores to these components are the basis for visualization and outlier detection.

2.1 Data smoothing

The objective of this step is to prepare data to be used in the FDA context. As a preparation step of the data to be employed, it is analogous to the step of extracting peaks in univariate FA or peak and volume series in the multivariate FA. Note that the statistical object of FDA is a function (curve) as shown in Figure 1. However, the curves are not observed, instead, only discrete measurements of the curves are available. In the case where data series are of good quality and long enough records, one can simply interpolate the measurements to obtain the curves, e.g. for rainfall series. Otherwise, smoothing can be required. This is typically the case for diffusive processes like in the present study of floods. However, even in the first case, smoothing could be necessary depending on the goal of the study (e.g. Ramsay and Silverman, 2005).

Let $\mathbf{Y}_i = (y_i(t_1), ..., y_i(t_T))$, $i = 1, ..., n$ be a set of n discrete observations where each $t_j \in$

$\mathcal{C} \subset \mathbb{R}^+$, $j = 1, ..., T$ is the jth record time point from a given time subset $\mathcal{C}$. For a fixed observation i, each set of measurements $(y_i(t_1), ..., y_i(t_T))$ is converted to be a functional data (curve) denoted $y_i(t)$ by using a smoothing technique where the index t covers the continuous subset $\mathcal{C}$. To this end, we suppose that the discrete observation $(y_i(t_j))_{j=1,...,T}$ is fitted using the regression model:

$$y_i(t_j) = x_i(t_j) + \epsilon_{ij} \quad i = 1, ..., n \text{ and } j = 1, ..., T \tag{1}$$

where ϵ_{ij} are the errors and the functions $x_i(.)$ are linear combinations of basis functions $\phi_k(.)$, that permit to explain most of the variation contained in the functional observations:

$$x_i(t) = \sum_{k=0}^{p} c_k^i \phi_k(t) \qquad \text{for } t \in \mathcal{C}. \tag{2}$$

The functional data set $(y_i(t))_{i=1,...,n}$ are then given by:

$$y_i(t) = \hat{x}_i(t) = \sum_{k=0}^{p} \hat{c}_k^i \phi_k(t), \quad t \in \mathcal{C} \tag{3}$$

where the estimated coefficients $\hat{c}_k^i$ are obtained by minimizing the following sum:

$$SSE(i) = \sum_{j=1}^{T} (y_i(t_j) - x_i(t_j))^2, \text{ for } i = 1, ..., n \tag{4}$$

For more details, the reader is referred, for instance, to Ramsay and Silverman (2005). A number of possible types of basis $\phi_k(.)$ have been presented in the literature. Most of the practical situations are treated with the well-known basis, such as, polynomial, wavelet, Fourier and the various Spline versions. Fourier and $B-$Spline basis are widely employed in the FDA context. The functional representation uses Fourier series for periodic or near periodic data. When the data are far from being periodic, spline approximations are commonly used in FDA for most problems involving non-periodic data (Ramsay and Silverman, 2005). Splines are more flexible but more complicated than Fourier series. The latter allows capturing the seasonal variability while the Spline series captures high and low values of the data (Ramsay and Silverman, 2005, Koulis et al., 2009). In general, the basis functions or the smoothing method to use should be based on objective considerations depending mainly on the nature of the data to be studied. Fourier basis functions $(\phi_k(.))_{k=0,\ldots,p}$ are defined by:

$$\phi_0(t) = 1,\ \phi_{2j-1}(t) = \sin(jwt),\ \phi_{2j}(t) = \cos(jwt), w = 2\pi/T. \tag{5}$$

Splines are piecewise polynomials defined on subintervals of the range of the observations. In each subinterval, the Spline is a polynomial function with a fixed degree but could be with different shapes. For instance, when the polynomial degree is three, we talk about cubic splines. For a comprehensive review about splines, the reader is referred to De Boor (2001).

Note that the aim of using the above expansion (3) is to obtain smooth functions to be employed as observations in FDA. In this case, the expansion series need not to be interpreted since the interest is not to extract a signal from the whole series. However, the number of basis functions to be selected is important where a large number leads to over-fitting of the data while a small number leads to under-fitting. Hence, the smoothing degree of the obtained functions to be employed as observations depends on the aim of the analysis, e.g. in principal component analysis, the aim is to capture a large variability rather than to reach the peaks. For more flexibility, a penalty term can be added to (4) to ensure the regularity of the smoothed functions. More details can be found for instance in Ramsay and Silvermann (2005) and Wahba (1990).

2.2 Location and scale parameters for functional variables

In a descriptive statistical study, we generally begin by looking for centrality and dispersion properties of a given sample. A location parameter summarizes the data and indicates where most of the data are located. Scale parameters are useful to measure the dispersion of a sample and also to compare different samples. These notions are useful in hydrology since they appear in almost all commonly employed probability distributions and models. In hydrology, location curves can also be used to characterize a given basin and to proceed to comparison or grouping of a set of basins. The scale measures can be used in a similar way but at a second level. In the setting of real or multivariate random variables, this is usually done through the mean, median, mode, variance, covariance and correlation. To avoid the possibility of missing important information, it is generally recommended to employ more than one measure for each feature. For instance, by looking only at the mean of the sample one might miss a possible heterogeneity in the population which would be captured by the mode. Obviously, these same problems will also appear when one studies a sample composed of curves $\{y_i(t),\ t \in \mathcal{C}\}, i = 1, ..., n$. In this setting, it is straightforward to define the mean curve $\bar{y}(.)$ of the sample as:

$$\bar{y}(t) = \frac{1}{n} \sum_{i=1}^{n} y_i(t), \quad t \in \mathcal{C}. \tag{6}$$

One has to use this mean curve carefully according to the shape of the data. For instance, if the data exhibit a high roughness degree the mean curve could be less informative.

Robust and efficient alternatives to the sample mean are the median and the trimmed mean (e.g., Ouarda and Ashkar, 1998). In the functional context, both measures are based on the statistical notion of depth function which is initially introduced in the multivariate context. The aim of depth functions is to extend the notion of ranking for a multivariate sample. These functions are introduced by Tukey (1975) and are introduced and applied to water sciences by Chebana and Ouarda (2008). Recently, the notion of depth has been extended to functional data (e.g. Fraiman and Muniz, 2001 and Febrero et al., 2008). The reader is referred to Chebana and Ouarda (2011b) for hydrological applications and a brief review and to Zuo and Serfling (2000) for a general and detailed description.

Fraiman and Muniz (2001) presented the definition of trimmed means in the functional context which are based on the empirical α-trimmed functional region. It is defined by $TR_\alpha := \{x,\ D_n(x) \geq \alpha\}$ for $0 < \alpha < 1$ where $D_n(.)$ is an empirical functional depth function, as the various ones defined, e.g. in Fraiman and Muniz (2001) and Febrero et al. (2008) where the corresponding formulations are explicitly given. A depth-based functional trimmed mean can be defined as the average over the $y_i(t),\ i = 1, ..., n$ that belong to the empirical trimmed region TR_α:

$$\bar{y}_\alpha(t) = \frac{1}{|TR_\alpha|} \sum_{y_i \in TR_\alpha} y_i(t), \quad t \in \mathcal{C} \tag{7}$$

where $|A|$ is the cardinal of the set A. For functional observations, the median curve is the deepest function in the sample $\{y_1, ..., y_n\}$. It maximizes the depth function $D_n(.)$:

$$Y_{median} = \operatorname{argmax}_{x \in \{y_1, ..., y_n\}} D_n(x) \tag{8}$$

where $\operatorname{argmax}_{z \in A} g(z)$ stands for the element in the set A that maximizes the function g.

From a theoretical point of view, the mode as a location measure, when it exists, is the value that locally maximizes the probability density f of the underlying variable. Developments and applications related to nonparametric density estimation in this context can be found in Dabo-Niang et al. (2007). An estimator of the modal curve can be obtained by:

$$Y_{mode} = \text{argmax}_{x \in \{y_1,\dots,y_n\}} f_n(x) \tag{9}$$

where f_n is an estimate of the density f.

The median and mode given respectively in (8) and (9) are natural extensions of their multivariate counterparts. However, they are rarely used in practice because of their complex computations. Alternatively, they are commonly defined on the basis of the bivariate scores of a functional principal component analysis of the curves observations as described in Section 2.4 below.

Variability is one of the important quantities to be evaluated and analyzed in statistics. For multivariate data, the reader is referred to Liu et al. (1999) and Chebana and Ouarda (2011b). The simplest way to define a variance function in the functional context is by:

$$var_y(t) = \frac{1}{n-1}\sum_{i=1}^{n}\left(y_i(t) - \bar{y}(t)\right)^2, \qquad t \in \mathcal{C} \tag{10}$$

The covariance function summarizes the dependence structure across different argument values:

$$cov_y(s,t) = \frac{1}{n-1}\sum_{i=1}^{n}\left(y_i(s) - \bar{y}(s)\right)\left(y_i(t) - \bar{y}(t)\right), \qquad s,t \in \mathcal{C} \tag{11}$$

The variability of the functional sample is analyzed by plotting the surface $cov_y(s,t)$ as a function of s and t as well as the corresponding contour map.

Note that, for functional observations, several types of variability can occur such as the variability within the same observation or between the different observations. In addition, functional principal component analysis is also employed to explore the variability between observations. The reader is referred to Ramsay and Silverman (2005) and the following sections for a presentation of the functional principal component analysis.

2.3 Functional principal component analysis (FPCA)

Principal component analysis (PCA), as a multivariate procedure, is usually employed to reduce the dimensionality by defining new variables as linear combinations of the original ones and which capture the maximum of the data variance. After converting the data into functions, functional PCA (FPCA) allows to find new functions that reveal the most important type of variation in the curve data. Note that these new functions cannot be in the Fourier or Spline basis since their aim is not to smooth but rather to produce a reasonable summary of the data by maximizing the capture of the variability. The FPCA method maximizes the sample variance of the scores (defined below) subject to orthonormal constraints. It decomposes the centered functional data in terms of an orthogonal basis as described in the following.

Let $y_i(t), i = 1, ..., n$ be the functional observations obtained by smoothing the observed discrete observations $(y_i(t_1), ..., y_i(t_T)),\ \ i = 1, ..., n.$

By definition, the mean curve is a way of variation common to most curves that can be isolated by centering. Let $(y_i^*(t) = y_i(t) - \bar{y}(t))_{i=1,...,n}$ be the centered functional observations where $\bar{y}(t)$ is the mean function of $(y_1(t), ..., y_n(t))$ given by (6). A FPCA is then applied to $(y_i^*(t))_{i=1,...,n}$ to create a small set of functions, called also harmonics, that reveals the most important type of variation in the data.

The first principal component of $(y_i^*(t))_{i=1,...,n}$ denoted by $w_1(t)$ is a function such that the variance of the corresponding real-valued scores $z_{i,1}$ written as:

$$z_{i,1} = \int_{\mathcal{C}} w_1(s) y_i^*(s) ds,\ i = 1, ..., n \tag{12}$$

is maximized under the constraint $\int_{\mathcal{C}} w_1(s)^2 ds = 1$. The next principal components $w_k(t)$ are obtained by maximizing the variance of the corresponding scores $z_{i,k}$:

$$z_{i,k} = \int_{\mathcal{C}} w_k(s) y_i^*(s) ds, \ i = 1, ..., n \tag{13}$$

under the constraints $\int_{\mathcal{C}} w_k(s) w_j(s) ds = 0, \ k \geq 2, \ k \neq j$. As in the multivariate setting, the interpretation of the principal component function w_k is slightly difficult as it depends on the type of data being used and may require nonstatistical considerations. A useful way consists in examining the plots of the overall mean function and perturbations around the mean based on w_k's. The perturbation functions are obtained as suitable multiples of the considered w_k, namely:

$$\bar{y} \pm 2\sigma_{\omega_k} * \omega_k, \ k = 1, ..., K \tag{14}$$

where σ_{ω_k} is the square root of the variance (eigenvalue) of the corresponding kth principal component. This presentation allows to isolate the perturbations about the mean across time and then assess the variability of the observations. Note that the principal components w_k are optimal, according to the maximization in (12) or (13), but are not unique. Therefore, any rotation with an orthogonal matrix of the w_k is also optimal and orthonormal. A well-known choice of such matrices is the VARIMAX. These rotated components can be useful for the interpretation. More technical details can be found, for instance, in Ramsay and Silverman (2005). On the other hand, the regularity of the harmonics $w_k(.)$ can be controlled. Rice and Silverman (1991) and Silverman (1996) extended this traditional functional PCA to the regularized FPCA (RFPCA) that maximizes the sample variance of the scores subject to penalized constraints.

2.4 Functional data visualization and outlier detection methods

In general, outliers represent gross errors, inconsistencies or unusual observations and should be detected and treated (Barnett and Lewis, 1998). Univariate outliers are well defined and their detection is straightforward (e.g. Hosking and Wallis, 1997; Rao and Hamed, 2000). This topic is also relatively well developed in the multivariate setting (e.g. Dang and Serfling, 2010). The identification and the treatment of outliers constitute an important component of the data analysis before modeling. For hydrologic data, outlier detection is a common problem which has received considerable attention in the univariate framework. In the multivariate setting, the problem is well established in statistics. However, in the hydrologic field the concepts are much less established. A pioneering work in this direction was recently presented by Chebana and Ouarda (2011b). As it is the case in the univariate and multivariate settings, outliers may have a serious effect on the modeling of functional data.

In this section, we focus on visualization methods that help to explore and examine certain features, such as outliers, that might not have been apparent with summary statistics. Different outlier detection methods exist in the functional context literature(e.g. Hardin and Rocke, 2005; Febrero et al., 2007; Filzmoser et al., 2008). However, Hyndman and Shang (2010) showed, on the basis of real data, that their methods are more able to detect outliers and computationally faster. The methods proposed by Hyndman and Shang (2010) are graphical and consist first in visualizing functional data through the rainbow plot, and then in identifying functional outliers using the functional Bagplot and the functional highest density region (HDR) boxplot. The latter two methods can detect outlier curves that may lie outside the range of the majority of the data, or may be within the range of the data but have a very different shape. These methods can also exhibit curves having a combination of these features. In practice, depending on the nature of the data, the two outlier detection methods can give different results.

As pointed out by Jones and Rice (1992) and Sood et al. (2009), the considerable amount of information contained in the original functional data is captured by the first few principal components and scores. The outlier identification methods of Hyndman and Shang (2010) considered here are based on these first two score vectors. Let $y_i(t), w_k(t)$ and $z_{i,k}$ be respectively the smoothed observed curves, the principal component curves and the corresponding scores obtained from the FPCA decomposition (Section 2.3). Let $(z_{1,1}, ..., z_{n,1})$ and $(z_{1,2}, ..., z_{n,2})$ be the first two vector scores and $z_i = (z_{i,1}, z_{i,2})$ the bivariate score points. At the end of this section, a non-graphical outlier detection method is proposed on the basis of $z_i = (z_{i,1}, z_{i,2})$.

2.4.1 Rainbow plot

The rainbow plot, proposed by Hyndman and Shang (2010), is a simple presentation of all the data, with the only added feature being a color palette based on an ordering. In the functional context, this ordering is either based on functional depth or data density indices. These indices are evaluated from the bivariate score depths and kernel density. The bivariate score depth is given by:

$$OT_i = d(z_i, Z), \quad Z = \{z_j \in \mathbb{R}^2; \, j = 1, ..., n\} \tag{15}$$

where $d(.,.)$ is the halfspace depth function introduced by Tukey (1975). Tukey's depth function at z_i is defined as the smallest number of data points included in a closed half-space containing z_i on its boundary. The observations are decreasingly ordered according to their depth values OT_i. The first ordered curve represents the *median curve*, while the last curve can be considered as the *outermost curve* in a sample of curves. As indicated in Section 2.2, this median curve based on

Tukey depth function of the bivariate principal scores z_i will be used in the following adaptation to floods. Let θ be this Tukey bivariate depth median defined as $\theta = \arg\max_z d(z, Z)$. If there are several maximizers, the Tukey bivariate depth median can be taken as their center of gravity. The second way of ordering functional observations is based on the kernel density estimate (e.g. Scott, 1992) at the bivariate principal component scores :

$$OD_i = \widehat{f}(z_i) = \frac{1}{n}\sum_{j=1}^{n}\frac{1}{h_j}K\left(\frac{z_i - z_j}{h_j}\right), \quad i \neq j,\, i = 1, ..., n \tag{16}$$

where $K(.)$ is the kernel function and h_j is the bandwidth for the j^{th} bivariate score points $\{z_j\}$. The functional data $\{y_i(t)\}$ are then ordered in a decreasing order with respect to OD_i. Hence, the first curve with the highest OD is considered as the *modal curve* while the last curve with the lowest OD can be considered as the most *unusual curve*. This modal curve will also be used in the following application.

The smoothed observations are presented with colors according to the values of OT and OD. The curves close to the center are red while the most outlying curves are violet.

2.4.2 Functional Bagplot

The bivariate Bagplot is introduced by Rousseeuw et al. (1999) and is based on the halfspace depth function. It is employed by Chebana and Ouarda (2011b) for multivariate hydrological data. The functional Bagplot version is obtained from the bivariate Bagplot based on the first two principal scores $z_i = (z_{i,1}, z_{i,2})$ given in Section 2.3. Each curve in the functional Bagplot is associated with a point in the bivariate Bagplot. Similar to the bivariate Bagplot, the functional Bagplot is composed by three elements: the Tukey median curve, the functional inner region and the functional outer region. The inner region includes 50% of the observations whereas the outer region covers either 95% or 99% of the observations. The outer region is obtained by inflating the inner region by a factor ϱ. Hyndman and Shang (2010) suggested that the factor ϱ could take the values 1.96 or 2.58 in order to include respectively 95% or 99% of the curves in the outer region. These values of ϱ correspond to the case where the bivariate scores follow the standard normal distribution. Finally, points outside the outer region are considered as outliers.

2.4.3 Functional HDR boxplot

The functional HDR boxplot corresponds to the bivariate HDR boxplot of Hyndman (1996) applied to the first two principal component scores $z_i \in \mathbb{R}^2$. The bivariate HDR boxplot is constructed using the bivariate kernel density estimate $\widehat{f}(z)$. An HDR with order $\alpha \in (0,1)$ is defined as:

$$R_{1-\alpha} = \{z \in \mathbb{R}^2 : \widehat{f}(z) \geq f_{1-\alpha}, \} \tag{17}$$

where $f_{1-\alpha}$ is such that $\int_{R_{1-\alpha}} \widehat{f}(t)dt = 1-\alpha$ and $\widehat{f}$ is defined by (16). An HDR can be seen as a density contour with expanding coverage decreasing with α. The associated bandwidth h_j in $\widehat{f}$ is selected by a smooth cross validation procedure (Duong and Hazelton, 2005).

The functional HDR boxplot is composed of the mode defined as $\arg\sup_z \widehat{f}(z)$, the 50% inner region ($R_{50\%}$) and the 99% outer highest density region ($R_{1\%}$). For an HDR with 95% outer region, one can take $R_{5\%}$ instead of $R_{1\%}$. Curves excluded from the outer functional HDR are considered as outliers.

The difference between detecting outliers by the Bagplot and by the HDR boxplot lies mainly in the way the inner and outer regions are established. The Bagplot uses a depth function (Tukey) and the estimated median curve (based on the Tukey depth function of the first bivariate scores z_i) while the HDR uses the density estimate of the z_i and its mode. Hence, the most outlier curves from HDR are unusual compared to the mode curve whereas those detected by the Bagplot are unusual with respect to the median curve.

In connection with the multivariate setting, as indicated in Chebana and Ouarda (2011b), the points outside the fence of the Bagplot are considered as extremes rather than outliers. Chebana and Ouarda (2011b) considered the approach proposed by Dang and Serfling (2010) to detect outliers. This approach is based on the evaluation of the outlyingness of each observation, the determination of a threshold and then the identification of the observations that exceed this threshold are considered as outliers. The outlyingness values are simple decreasing transformations of depth functions. In the present study, we propose to consider this approach based on the first two scores. A brief presentation of the approach is given in Chebana and Ouarda (2011b), section 2.6.

The above graphical approaches should be considered as preliminary indications for suspected observations. The latter could be seen as extremes rather than outliers (see e.g. Chebana and Ouarda, 2011b). In addition, the approach by Dang and Serfling (2010) is based on the outliyngness criteria and the corresponding threshold is empirical and not necessarily normally-based (instead of the values of the inflating central region 1.96 or 2.58).

3 Adaptation to floods

The first and most important adaptation for floods lies in the nature of hydrological data. The main data source in hydrology is daily flow from a given station. Flows can also sometimes be available on an hourly, instantaneously or any other time scale. In the following, we focus on daily data and we assume it is recorded during a number n of years of measurements, $\mathbf{Y}_i =$

$(y_i(t_1), ..., y_i(t_T))'$, $i = 1, ..., n$, $j = 1, ..., T$, with $T = 365$ days and $y_i(t_j)$ is the flow measured at the day t_j of the ith year. The time subset index $\mathcal{C}$ is then the interval $[1, 365]$. According to this kind of data, we have n discrete observations $\{y_i(t_j),\ j = 1, ..., 365\},\ i = 1, ..., n$. The observation $\{y_i(t_j),\ j = 1, ..., 365\}$ denotes the daily flow for the ith year. A functional observation constitutes a year starting from January 1st to December 31st. However, it can be cut out in different ways according to the seasonal characteristics of the geographical area of interest. For instance, for most parts of Canada, it is possible to define the March-June season for spring floods and the July-October season for fall floods.

The discrete observed data $(y_i(t_j))_{j=1,...,T}$ are to be converted to smooth curves $y_i(t)$ as temporal functions with a base period of $T = 365$ days and with $p = 52$ weeks non-constant basis functions as in (2). This smoothing can be obtained through the two well-known Fourier and $B-$Spline basis. Usually, the flow data of the whole series present some seasonal variability and periodicity over the annual cycle. Therefore, Fourier basis are preferred. Although the two smoothing methods do not give identical results, the differences between them in this adaptation are generally insignificant to affect interpretations. The choice of $p = 52$ can be justified to capture the flow variation within a week. Since in flood studies, the peak value is important, in order to ensure that the smooth curves reach the associated peaks, it may be recommended to consider values of p greater than 52. Nevertheless, this could lead to irregular curves which could not reasonably capture the entire flow variation.

The nonparametric approach presented in Section 2.4.3, using the kernel density estimate of z_i's, is employed for curve ordering and outlier detection and not for estimation purposes. Note that even though nonparametric approaches have been employed in hydrological FA in the univariate context (see e.g. Adamowski and Feluch, 1990; Ouarda et al., 2001), they are still of limited use for the hydraulic design of major structures (Singh and Strupczewski, 2002). In addition, the mode as a location measure is useful to detect the presence of inhomogeneity in the data. In hydrological FA, the mode is not commonly used since, generally, data should pass a homogeneity test. Therefore, the fitted models should be unimodal.

Generally, in hydrology, there are two main sources of outliers. The data may be incorrect and/or the circumstances around the measurement may have changed over time (Hosking and Wallis, 1997). However, a detected outlier can also represent true but unusual observed data. In the present functional context, outlier curves have different magnitudes and shapes compared to the rest of the observed curves.

4 Case study

The methods described in Section 2 are applied to hydrological data series by using the adaptation presented in Section 3. In the following, the data are described, and functional as well as analogous multivariate results are presented and discussed. More precisely, the conversion of the discrete data to be employed as continuous functions is the first preliminary step. Then, the different location functions are obtained and the variability of the sample is studied directly as well as using the FPCA. The latter are also used for data visualization and as a preliminary tool to identify outliers. These outliers are checked by the previously presented approaches and interpreted on the basis of meteorological data. The last subsection provides results using multivariate approaches for comparison purposes.

4.1 Data description and smoothing

The data series is a daily flow (m^3/s) from the Magpie station with reference number 073503. It is located at the outflow of the Magpie lake in the Côte-Nord region in the province of Quebec, Canada. The area of the drainage basin is 7 230 km^2 and the flow regime is natural. Data are available from 1979 to 2004. Figure 2 indicates the geographical location of the Magpie station.

According to the present dataset, we have $n = 26$ discrete observations $y_i(t_j),\ t_j \in \mathcal{C} = [1, 365],\ i = 1, ..., n$. The ith discrete observation $\{y_i(t_j),\ j = 1, ..., 365\}$ denotes the daily flow measurements for the ith year which is converted to a smooth curve $\{y_i(t), t \in \mathcal{C}\}$. This is done through the technique based on Fourier series expansion. This smooth representation of flow data is done with a 365-day base period ($T = 365$ days) and 52-week non-constant basis functions ($p = 52$). The obtained functional observations are given in Table 2 with the corresponding univariate and bivariate samples. This table allows to have an overall view of the data within the three frameworks.

Figure 3a illustrates the whole daily flow series. It shows that the data are nearly periodic. As indicated above, this periodicity can justify the use of Fourier basis. A number of observed hydrographs with the corresponding Fourier and B-splines smoothing curves are presented in Figure 3b. They show that the Fourier and B-Splines smoothing are similar and indicate also that the peaks are generally reached. Figure 4 displays the standard deviation of the residuals $\widehat{\epsilon}_{ij} = y_i(t_j) - \widehat{x}_i(t_j)$ over j after smoothing the flow data. It gives the residual variations across days, within each year. We observe that these errors are generally very small and do not exceed 32 m^3/s. The highest errors are associated with the years 1981, 1999 and 2002.

Note that, other values of p, both smaller and larger than 52, e.g 4, 12, 90, 122, 182, 300, were also tested. Even though, large values of p, e.g. close to the number of observations per year (here 365), allow to reach almost all the daily flow points including the peaks, the obtained curves are not smooth or regular enough and also do not allow to capture enough of the variance by the first few principal components. Small values of p, e.g. 4,12 give a very bad quality of smoothing, where a large amount of daily flow points are not reached, particularly the high and low values. Therefore, it is reasonable to choose a number p which combines the quality of smoothing (related to (4)) and a high percentage of explained variance by PCA analysis. In the present application, the choice $p = 52$ fits reasonably the discrete data except for some extreme points corresponding to a number of years (e.g. 1980, 1989 and 1993) where the resulting differences between the real peaks and the smooth ones are less than 150 m^3/s, see Table 2.

4.2 Functional results

Figure 5 presents the smooth location curves (mean, median and mode). It shows that generally the maximum flow occurs in late April and early May followed by a recession during May and June. This phenomenon is common in Canada where floods are mainly caused by snow melt during the Spring season. On the right tail of the curves, we observe a small flood which occurs in the autumn and which is caused generally by liquid precipitations. This kind of flood is exhibited by the mode. In both floods, spring or autumn, we observe that the mode is always higher than the mean and the median. The mean seems to be more regular and can not reach high flow values. Therefore, it is useful to consider all these location measures. These location curves lead to different basin characterization through the whole event rather than just some of its parts or summaries and therefore allow for comprehensive basin comparisons.

The bivariate (temporal) variance-covariance surface obtained from (11) as well as the corresponding contour are presented in Figure 6. We observe that the main part of the variability occurs in the middle of the year and it is negligible elsewhere. That is, the highest variability occurs approximately between April and late June. This period corresponds approximately to the highest flows. This measure has the advantage of providing information concerning the variance structure and also when it occurs.

The principal components are obtained by FPCA on the centered observations $y^*(.)$. The variance rates accounted for by each one of the first four principal components are respectively 39.5%, 24.0%, 14.4% and 5.4%. These components account for 83.3% of the total variance of the flow. The centered principal components are presented in Figure 7a. The perturbations of these first four principal components about the mean, as given in (14), are presented in Figure 7b.

From Figure 7, where the first two principal components accumulate 63.5% of the total variance, one can observe that the station flow is most variable between April and July. This variation dominates the variation occurring between July and the end of the year and which is associated with the third and fourth components, and represents 19.9% of the total variance. This finding is, for all practical purposes, consistent with the one obtained from the variance-covariance surface (Figure 6). More precisely, the first two principal components w_1 and w_2 are representative of the spring floods whereas w_3 and w_4 are more likely to represent autumn floods.

The scores corresponding to the first four principal components are given in Table 3. Given the high variation rate captured by the first principal components, the corresponding variation indicates that the years for which the first or the second principal score is higher (resp. lower) have higher (resp. lower) flow during April to July. Therefore, the year 1981 represents the highest variability during this period followed by the year 1999. On the other hand, the smallest variability of the flow during April-July is associated with the year 1987. Other years could be considered also with low flow variability, such as 1986 and 2002. The flow variability associated with the years 1981, 1986, 1987, 1999 and 2002 is unusual where some of the corresponding curves (1981,1987, 1999) are displayed with the location curves in Figure 8.

In order to check the above unusual years, the outlier detection methods described in Section 2 are employed. Other functional methods are also tested, such as the functional depth method of Febrero et al. (2007) and the Integrated squared error method of Hyndman and Ullah (2007). However, these two methods gave either too many or no outliers. Hence, the corresponding results are omitted.

Figure 9 presents the rainbow plots based on the bivariate depth ordering and the density ordering indices (15) and (16) respectively. The colors indicate the ordering of the curves where the blue curves are the closest to the center. The red and black outlier curves correspond to 1981 and 1999 respectively. Results show that both methods lead to a similar ordering especially for the years associated with high or low ordering.

The bivariate Bagplot associated with the first two principal scores as well as the corresponding functional Bagplot for both 95% and 99% of probability coverage are presented in Figure 10. We observe that the curve corresponding to the year 1981 is outside the outer bivariate Bagplot region for both 95% and 99% cases. It corresponds to the red curve in the associated functional Bagplot (Figure 10c,d). Hence, this year is considered as an outlier according to Tukey depth, as described in Section 2. However, when considering the 95% Bagplot, the additional outlier curve that is detected is the one corresponding to 1987 as shown in Figure 10b. Note that generally when outliers are relatively near the median, the functional Bagplot is not a good way to detect them (Hyndman and Shang, 2010). Even though it is not the case here, it is also more appropriate to use the functional HDR boxplot.

The bivariate HDR and the associated functional HDR boxplots of the smooth flow curves are presented in Figure 11 for both 95% and 99% of probability coverage. The only detected outlier with 99% coverage probability is 1981 which is outside the bivariate HDR outer region. In the present case, we can deduce that the flow corresponding to the year 1981 is the most outlier, has a different magnitude and shape compared to the other curves and is not near the median. Hence, we can conclude that 1981 is an effective outlier according to the HDR Boxplot. When the probability coverage is 95%, another outlier is detected and corresponds to the year 1999 as shown in Figure 11b. This curve is closer to the median than the curve corresponding to 1981 (Figure 8), that is why the functional HDR boxplot is more able to detect it as outlier than the functional Bagplot.

As discussed in Section 2.4, the HDR boxplot and the Bagplot are graphical outlier detection methods and their thresholds are based on normality. Therefore, the above detected years can be considered as extreme curves and could be outliers. The approach developed by Dang and Serfling (2010) is applied on the first two functional principal component scores Z of the dataset. We evaluated Spatial, Mahalanobis and Tukey outlyingness functions for the bivariate score series. The corresponding thresholds are obtained by selecting the ratio of false outliers to 15% and the true number of outliers as 5 (the same choices as in Chebana and Ouarda, 2011b and Section 4.3 below). Hence, the threshold corresponds to the 0.97-quantile of the outlyingness values. Figure 12 presents the detected outliers. We observe that the Tukey outlyingness function detects several years as outliers (including 1981, 1987, 1999 and 2002) whereas the year 1981 is detected by the three outlyingness functions. In addition, the year 1987 corresponds to the second largest Spatial and Mahalanobis outlyingness values and its value is very close to 1999 with the Mahalanobis function. Note that Tukey outlyingness is not recommended by Dang and Serfling (2010). Therefore, the year 1981 can be considered as an effective outlier to be checked. The years 1987 and 1999 could be detected by Spatial and Mahalanobis outlyingness and considering a larger true number of outliers than 5 (with values of 5%, 10% and 20% of the ratio of false outliers, the results remain the same). Note that the above suspected years of 1986 and 2002 can be considered as extremes and not outliers.

Even though the curve of 1981 is the only effective outlier, in the following we examine also the years 1987 and 1999 since they are close to the thresholds. We observe from Figure 8 that the curves of 1981, 1987 and 1999 are clearly different from the location curves and from the general shape of curves. Indeed, based on the corresponding hydrographs, the curve of 1981 is characterized by very high peak and volume whereas 1987 seems to correspond to a dry year since the flow was the lowest during the Spring season. The flood corresponding to the year 1999 has also a high peak, although lower than the one corresponding to 1981.

The detected outliers can be explained on the basis of meteorological data. The following interpretations are drawn on the basis of the data available in Environment Canada's Web site http://www.climat.meteo.gc.ca/climateData/canada_f.html. For 1981, which corresponds to the most important flood for this basin, an important amount of snow was accumulated in early Winter (October-November to January) followed by thaw and rain during February-March. For the outlier corresponding to 1987, the comparison with the preceding and following years reveals that during the fall of 1987 there was much less rain and the temperatures were very cold, whereas the end of Winter was hot. Hence, all the snow melted earlier compared to other years. The curve of 1999 is relatively higher than the location curves and corresponds to an important quantity of snow on the ground with high temperatures in March. In conclusion, the above detected years seem to be actually observed and do not correspond to incorrect measurements or circumstance changes over time. Hence, these observations should be kept and employed for further analysis. However, it is recommended to use robust statistical methods to avoid sensitivity of the obtained results (e.g. modeling and risk evaluation) to outliers.

4.3 Multivariate results

For comparison purposes, a multivariate study based on Chebana and Ouarda (2011b) is carried out on the present dataset. We focus here on the flood peak Q and the flood volume V as they are among the most important and studied flood characteristics (e.g. Yue et al., 1999 and Shiau, 2003). The bivariate series (Q,V), given in the first three columns of Table 4, are obtained from the daily flow series using an automated version of the algorithm of Pacher (2006). Note that the multivariate approaches presented in Chebana and Ouarda (2011b) are mainly based on depth functions. The Tukey depth function is considered in the present section. The corresponding depth values of each bivariate observation are reported in the fourth column of Table 4. The location and scale results are presented in Table 5. Results with other measures (such as the trimmed mean and dispersion) are obtained but not presented due to space limitations and in order to maintain the coherence with the FDA approach.

We observe that Q and V of the bivariate median correspond to those of the median curve obtained in the previous section. Indeed, in both multivariate and functional frameworks, the median corresponds to the year 1980. However, the Q and V of the bivariate mean vector are quite different from those resulting from the mean curve. The mean vector is (Q = 859.15, V = 2138.70) whereas, when using Pacher's (2006) algorithm, the Q and V of mean curve are respectively 673.09 and 2230.46. We observe also that the difference is larger for the peak than for the volume. This result could be explained by the effect of the detected outliers on the mean which is not the case for the median. Note that the outliers do not necessarily have the same impact in the multivariate and the functional frameworks.

Figure 13a presents the bivariate (Q, V)-Bagplot where the median, the central and the outer regions are indicated as well as some particular observations (corresponding to years suspected as outliers from Section 4.2). Note that the outer region is obtained by inflating the central region by a factor of 3 instead of 1.96 or 2.58 as in the functional Bagplot (Figures 10a,b). We observe that the shape of the bivariate (Q, V)-Bagplot is not similar to the functional Bagplot and to the HDR boxplot based on the first two functional principal component scores $z_i = (z_{i,1}, z_{i,2})$. The values in Tables 3 and 4 as well as the corresponding figures (Figures 10a,b, 13a) indicate that the first two functional principal component scores z_i capture the information from the hydrograph in a different way than do (Q, V). The former are based on an optimization procedure whereas the latter have physical significance. Nevertheless, both ways are useful to understand flood dynamics and should be used in a complementary manner. This finding should be studied more thoroughly in future research by considering a number of case studies.

The bivariate (Q, V)-variability is evaluated both in a matrix form (Table 5) and by using scalar curve (Figure 13b). Note that the variability is particularly useful when comparing at least two data sets for the same kind of series (e.g. same variable or same vector). It is appropriate to compare the univariate peak scale with the functional one since the flood peak has the same unit and scale as the daily flow which is not the case for the volume. Hence, we observe that the peak variance has the same magnitude as in the functional context as it can be seen from Table 5 and Figure 6. One can also appropriately standardize the Q and V variables in order to compare the variances of the vector (Q, V) and the functional context.

The procedure employed in Chebana and Ouarda (2011b) for outlier detection is based on depth outlyingness measures and the corresponding thresholds. The reader is referred to Chebana and Ouarda (2011b) or Dang and Serfling (2010) for more details about the outlyingness expressions and threshold determination. In the present section, three outlyingness measures are evaluated on the (Q, V) series, i.e. Tukey (TO), Mahalanobis (MO) and Spatial (SO). Their values are presented in the last three columns of Table 4. To obtain the threshold that the outlyingness of an outlier should exceed, we considered a ratio of false outliers of 15% among the allowed ones and we also allowed 5 true outliers (the same choices as in Chebana and Ouarda, 2011b). Hence, the threshold corresponds to the empirical 97%-quantile of the outlyingness values. The obtained threshold values are 0.9231, 0.8676 and 0.9462 respectively for TO, MO and SO. Consequently, 1981 is detected by all measures, 1987 is detected only by TO and it has also the second highest outlying value by MO and SO but smaller than the corresponding thresholds. The measure TO detects several other outliers, including 1999 and 2002, which all have the same TO value (equal to the threshold). However, if a quantile of order higher than 97% is considered, by modifying the parameters related to the threshold, the TO does not detect any outliers. These results are consistent with those of the functional framework in the sense that the most unusual observations are detected in both frameworks. However, the proposed approach that consists in applying the Dang and Serfling (2010) technique on the first two score series z_i seems to be justified and more reliable.

5 Summary and concluding remarks

The first aim of the present paper is to introduce the functional framework to hydrological applications based on the curve nature of the data to be employed and analyzed. The FDA framework can be seen as a natural extension of the multivariate FA where the latter is gaining popularity and usefulness in meteorological and hydrological studies. In the present study we introduced a number of FDA notions and techniques and adapted them to the hydrological context, and more specifically to floods. The techniques within the first functional FA step deal with visualization, location estimation, variability quantification, principal component analysis and outlier detection. A new non-graphical (numerical) outlier detection method is proposed which combines multivariate and functional techniques.

An application is carried out to demonstrate the potential of employing FDA techniques in hydrology. The application deals with the natural streamflow series of the Magpie station in the province of Quebec, Canada. Results regarding location measures such as mean, median and modal curves, are obtained. The variability is studied as a simple bivariate function surface and also by using principal component analysis. Outlier curves are identified by the most efficient methods and interpretations are given based on meteorological data. For comparison purposes, a brief bivariate study of flood peak and volume is carried out. Even though FDA is an extension of multivariate analysis, it is recommended to perform both approaches to obtain a complete understanding of floods and to make the appropriate decisions.

From the elements discussed in the introduction and the results obtained in the case study, the following concluding remarks can be drawn and a number of limitations and perspectives are given:

I) Drawbacks of previous approaches: The following drawbacks represent the motivation and the need to introduce the functional framework in hydrological applications:

1. The separate or joint use of hydrograph characteristics constitutes a major simplification of the real phenomenon;
2. Given the lack of data in hydrology, working with a limited number of extracted characteristics represents a loss of a part of the available information;
3. The way these characteristics are determined is neither unique nor objective;
4. The multivariate analysis is a simplification of the hydrological phenomena since it is based on flood characteristics which are simple transformations of the hydrograph;
5. In the multivariate setting, the complexity of the models, the fitting and estimation difficulty, the number of parameters and the associated uncertainty increase with the dimension;
6. The importance of the hydrograph shape is shown in studies such as Yue et al. (2002) where the approaches approximating the flood hydrograph using probability densities are limited for instance to single-peak hydrographs;
7. The main flood characteristics, peak, volume and duration, can not completely capture the shape of the hydrograph;

8. Even though, in the flow duration curve studies, e.g. Castellarin et al. (2004), a number of functional elements, such as mean and median curves, are presented, they are limited and do not have a functional statistical foundation;

II) Conceptual advantages of the functional framework: The functional framework presents some general advantages which contribute to overcome the previous drawbacks at different levels:

1. The functional framework treats the whole hydrograph as a functional observation (function or curve) which is more representative of the real phenomena;
2. It employs the maximum of the available information in the data where the impact of the lack of data in hydrology can be reduced;
3. The functional framework is more general and more flexible and can represent a large variety of hydrographs;
4. The functional methodology constitutes a natural extension of univariate and multivariate hydrological FA approaches;
5. The location curves and functional scale measures can be used to characterize a given basin and to proceed to comparison or grouping of a set of basins;
6. FDA allows to perform a single analysis on the whole data instead of several univariate or multivariate analysis;
7. The approaches dealing with hydrograph shape, e.g. the one proposed by Yue et al. (2002), can be generalized in the FDA context using smoothing techniques;

8. The functional setting avoids the definition and the evaluation of flood characteristics. Therefore, it does not require specific algorithms and avoids subjective evaluations; and the associated uncertainty can be reduced in the analysis;

III) Concluding remarks from the application: The following points are drawn as specific results of the FDA application to the case study:

1. The location curves (mean, median and mode) give more information concerning the hydrological regime in the basin than the univariate and multivariate approaches by adding temporal aspects. These curves allow to summarize the information contained in the data for a given basin, and hence make comparisons between basins and group basins with similar regime;
2. The bivariate (temporal) variance-covariance surface as well as the corresponding contour give an additional insight to the hydrological regime variability than the real-value or matrix in the univariate and multivariate contexts;
3. In addition to quantifying the variability, functional scale measures indicate when it occurs;
4. The case study results show that the mode is useful to characterize high flood values, the variability is very high during spring season and the principal components are shown to describe the variability in spring floods and autumn floods. The detected outliers are checked to be real observations and therefore it is suggested to use robust methods in any further analysis;
5. The first two functional principal components capture the information from the hydro-

graph in a different way than do (Q, V). Nevertheless, both ways are useful to understand flood dynamics and should be used in a complementary manner;

6. The FPCAs represent a new way to distinguish the different flood events in a given year. Indeed, the few first principal components can be used to identify where in the hydrograph the variation dominates and can be used to characterize flood events, e.g the first two principal components are representative of the spring floods whereas the two others represent autumn floods;

7. In the functional context, outlier curves have different magnitudes and shapes compared to the rest of the observed curves. In the univariate and multivariate settings, the shape is not considered and can not be captured even by using several variables;

8. The functional results obtained in this study are generally coherent with those of the multivariate analysis but give more insight to the hydrological phenomena such as in terms of location measures, variability and principal components;

IV) Limitations and perspectives of the functional framework: The present study presented exploratory functional tools that are important on their own and it constitutes also a basis for the next steps for a reliable FDA-based hydrological FA, especially in terms of model selection and risk evaluation. Several perspectives are promising and can be carried out in future research:

1. Although the study focused on floods, the presented FDA methodology can be adapted and applied to treat other hydro-meteorological variables such as droughts, precipitations, storms and heat waves;

2. FDA relies on the smoothing step. Therefore, a careful inspection of the resulting curves is recommended, for instance, to ensure the regularity of the smoothed functions, to reach a majority or special points such as peaks or to capture enough of the variance by the first few principal components. Even though a number of elements in this direction are given in the present study, it could be of interest to develop general criteria and objective choices depending on the objective of the analysis;

3. The classification of the curves of a given site as well as the clustering of sites in a region on the basis of the full hydrograph are also topics of interest;

4. Inferential aspects, such as modeling for prediction purposes, represent also important issues for future research efforts;

5. Future investigations should also deal with hypothesis testing as well as regression modeling.

Acknowledgement Financial support for this study was graciously provided by the Natural Sciences and Engineering Research Council (NSERC) of Canada, and the Canada Research Chair Program. The authors would also like to thank the authors of the *FDA R* and *Rainbow R* packages. The authors wish to thank the Editor, Associate Editor and three anonymous reviewers for their useful comments which led to considerable improvements in the paper.

References

Adamowski, K. and Feluch, W. (1990), Nonparametric flood frequency analysis with historical information. *ASCE Journal of Hydraulic Engineering*, 116, 1035-1047.

Barnett, V., and T. Lewis (1998), *Outliers in statistical data*, 3rd ed., Wiley, Chichester.

Barnett, V. (2004), *Environmental statistics : methods and applications*, Wiley, Chichester, England ; Hoboken, N.J.

Castellarin, A., Vogel, R. and Brath, A. (2004), A stochastic index flow model of flow duration curves. *Water Resources Research*. 40, W03104.

Chebana, F. and Ouarda, T.B.M.J. (2007), Multivariate L-moment homogeneity test. *Water Resources Research*, 43, W08406. doi:10.1029/2006WR005639.

Chebana, F. and Ouarda, T.B.M.J. (2008), Depth and homogeneity in regional flood frequency analysis. *Water Resources Research*, 44, W11422. doi : 10.1029/2007WR006771

Chebana, F., Ouarda, T.B.M.J. and Duong, T.C. (2010), Testing for stationarity, homogeneity and trend in multivariate hydrologic time series: a review, *INRS-ETE, report I-271*: Quebec. p. 73.

Chebana, F. and Ouarda T.B.M.J. (2011a), Multivariate quantiles in hydrological frequency analysis. *Environmetrics*, 22, 63-78, DOI: 10.1002/env.1027

Chebana, F. and Ouarda, T.B.M.J. (2011b), Depth-based multivariate descriptive statistics with hydrological applications. *J. Geophys. Res.,* 116, D10120, doi:10.1029/2010JD015338.

Chow, V. T., Maidment, D. R. and L. R. Mays (1988), *Applied Hydrology*, McGraw-Hill, New York.

Cuevas, A. and Febrero, M. and Fraiman, R. (2006), On the use of bootstrap for estimating functions with functional data, *Computational Statistics and Data Analysis*, 51, 1063-1074.

Cunnane C. (1987), Review of statistical models for flood frequency estiamtion. In Hydrologic Frequency Modeling, Singh V.P. (ed.). Reidel, Dordrecht: The Netherlands; 49-95.

Dabo-Niang, S., F. Ferraty and P. Vieu (2007), On the using of modal curves for radar waveforms classification. *Computational statistics and data analysis*, 51, 4878-4890.

Dabo-Niang, S. and F. Ferraty (2008), *Functional and Operatorial Statistics*, Springer Verlag, New-York.

Dabo-Niang, S.; Yao, A.F., Pischedda, L.; Cuny, P. and Gilbert, F. (2010), Spatial mode estimation for functional random fields with application to bioturbation problem. *Stochastic Environmental Research and Risk Assessment*, 24 (4), 487-497.

Dang, X. and Serfling, R. (2010), Nonparametric depth-based multivariate outlier identifiers, and masking robustness properties. *Journal of Statistical Planning and Inference*, 140, 198-213.

de Boor, C. (2001). *A Practical Guide to Splines*, Revised Edition. New York, Springer.

Delaigle, A. and Hall, P. (2010), Defining probability density for a distribution of random functions. *Annals of Statistics*, 38, 1171-1193.

Delicado, P., Giraldo, R. and Mateu, J. (2008), Point-wise Kriging for spatial prediction of functional data. *In Functional and Operatorial Statistics*. Editors S. Dabo-Niang and F. Ferraty, Physica-Verlag, 135-141.

Duong, T. and Hazelton, M. L. (2005), Cross-validation bandwidth matrices for multivariate kernel density estimation. *Scandinavian Journal of Statistics*, 32, 485-506.

Febrero, M., Galeano, P. and Gonzalez-Manteiga, W. (2007), A functional analysis of NOx levels: location and scale estimation and outlier detection, *Computational Statistics*, 22, 411-427.

Febrero, M., Galeano, P. and Gonzalez-Manteiga, W. (2008), Outlier detection in functional data by depth measures, with application to identify abnormal NOx levels, *Environmetrics*, 19, 331- 345.

Ferraty, F. and Vieu, P. (2006), *Nonparametric functional data analysis*. Springer-Verlag, New York.

Ferraty, F.; Rabhi, A. and Vieu, P. (2005), Conditional quantiles for functional dependent data with application to the climatic El-NINO phenomenon. *Sankhya: The Indian Journal of Statistics,* 67, 378-398.

Filzmoser, P., Maronna, R. and Werner, M. (2008), Outlier identification in high dimensions. *Computational Statistics and Data Analysis*, 52, 1694-1711.

Fraiman, R. and Muniz, G. (2001), Trimmed means for functional data. *Test*, 10, 419-440.

Hardin, J. and Rocke, D. M. (2005), The distribution of robust distances. *Journal of Computational and Graphical Statistics*, 14(4), 928-946.

Hosking, J. R. M. and Wallis, J. R. (1997), *Regional Frequency Analysis: An Approach Based on L- Moments*. Cambridge University Press.

Hyndman, R. J. (1996), Computing and graphing highest density regions. *The American Statistician*, 50, 120-126.

Hyndman, R. J. and Ullah, M. S. (2007), Robust forecasting of mortality and fertility rates: A functional data approach. *Computational Statistics and Data Analysis*, 51, 4942-4956.

Hyndman, R. J. and Shang, H.L. (2010), Rainbow plots, bagplots and boxplots for functional data. *Journal of Computational and Graphical Statistics*, 19, 29-45..

Jones, M. C. and Rice, J. A. (1992), Displaying the important features of large collections of similar curves. *The American Statistician*, 46, 140-145.

Khaliq, M.N., Ouarda, T.B.M.J., Gachon, P., Sushama, L. and St-Hilaire, A. (2009), Identification of hydrological trends in the presence of serial and cross correlations: A review of selected methods and their application to annual flow regimes of Canadian rivers. *Journal of Hydrology*, 368, 117-130.

Kite, G.W. (1988), *Frequency and Risk Analysis in Hydrology*. Water Resources Publications, Littleton, Colorado

Koulis, T.; Thompson, M.E. and LeDrew, E. (2009), A spatio-temporal model for Antarctic sea ice formation. *Environmetrics.*, 20, 68-85.

Kundzewicz, Z.W., Graczyk, D., Maurer, T., Pinskwar, I., Radziejewsky, M., Svensson, C., and Szwed, M. (2005), Trend detection in river flow series: 1. Annual maximum flow. *Hydrol. Sci. J.* 50, 797-810.

Liu, R. Y., Parelius, J. M. and Singh, K. (1999), Multivariate analysis by data depth: Descriptive statistics, graphics and inference. *The Annals of Statistic*, 27, 783-858.

Mizuta, M. (2006), Discrete functional data analysis. In: *Proccedings in Computatiobal Statistics. Physica-Verlag, Springer, Berlin*, 361-369.

Monestiez, P. and Nerini (2008), A Cokriging method for spatial functional data with applications in oceanology. *In Functional and Operatorial Statistics*. Editors S. Dabo-Niang and F. Ferraty, Physica-Verlag, 237-242.

Ouarda, T. B. M. J. and Ashkar, F. (1998), Effect of trimming on LP III flood quantile estimates. *Journal of Hydrologic Engineering*, 3, 33-42.

Ouarda, T. B. M. J.; Girard, C.; Cavadias, G. S. and Bobée, B. (2001), Regional flood frequency estimation with canonical correlation analysis. *Journal of Hydrology*, 254, 157-173.

Pacher, G. W. (2006), Détermination objective des paramètres des hydrogrammes [In French]. Personal note.

Ramsay, J. O. and Silverman, B. W. (2002), *Applied Functional Data Analysis: methods and case studies*. Springer, New York, London.

Ramsay, J. O. and Silverman, B. W. (2005), *Functional Data Analysis*, 2nd edn, Springer, New York.

Rao, A. R. and Hamed, K.H. (2000), *Flood Frequency Analysis*. CRC Press: Boca Raton.

Rice, J.A. and Silvermann, B.W. (1991), Estimating the mean and covariance structure nonparametrically when data are curves. *Journal of the Royal Statistical Society B*, 53, 233-243.

Rousseeuw, P. J.; Ruts, I. and Tukey, J. W. (1999), The bagplot: A bivariate boxplot. *The American Statistician,* 53, 382-387.

Salvadori, G., De Michele, C., Kottegoda, N.T. and Rosso, R. (2007), *Extremes in Nature: An Approach Using Copulas*. Springer.

Scott, D.W. (1992), *Multivariate density estimation: theory, practice, and visualization*, Wiley, New York.

Serinaldi, F. and Grimaldi, S. (2007), Fully nested 3-copula: Procedure and application on hydrological data. *Journal of Hydrologic Engineering*, 12, 420-430.

Shiau, J. T. (2003), Return period of bivariate distributed extreme hydrological events. *Stochastic Environmental Research and Risk Assessment*, 17, 42-57.

Silvermann, B.W. (1996), Smoothed functional principal components analysis by choice of norm. *The Annals of Statistics*, 24, 1-24.

Singh, V.P. and Strupczewski, W.G. (2002), On the status of flood frequency analysis. *Hydrological Processes*, 16, 3737-3740.

Sood, A., James, G. M. and Tellis, G. J. (2009), Functional regression: a new model for predicting market penetration of new products, *Marketing Science*, 28, 36-51.

Stedinger, J.R., Vogel, R.M. and Foufoula-Georgiou, E. (1993), *Frequency Analysis of Extreme Events*. In: Handbook of Hydrology, D.R. Maidment (Editor). McGraw-Hill Inc., New York, New York.

Tukey, J. W. (1975), Mathematics and the picturing of data, in Proceedings of the International Congress of Mathematicians (Vancouver, B. C., 1974), 2, 523-531.

Wahba, G. (1990), *Splines Models for Observational data*, Philadelphia, SIAM.

Warner, R. M. (2008), *Applied statistics : from bivariate through multivariate techniques*, SAGE Publications, Thousand Oaks, Calif.

Yue, S.; Ouarda, T. B. M. J.; Bobée, B.; Legendre, P. and Bruneau, P. (1999), The Gumbel mixed model for flood frequency analysis. *Journal of Hydrology*, 226, 88-100.

Yue, S., Ouarda, T. B. M. J., Bobée, B., Legendre, P. and Bruneau, P. (2002), Approach for describing statistical properties of flood hydrograph, *Journal of Hydrologic Engineering*, 7, 147.

Zhang, L., and Singh V.P. (2006), Bivariate flood frequency analysis using the copula method. *Journal of Hydrologic Engineering*, 11, 150-164.

Zhang, L. and Singh, V.P. (2007), Trivariate flood frequency analysis using the Gumbel-Hougaard copula. *Journal of Hydrologic Engineering*, 12, 431-439.

Zuo, Y. and Serfling, R. (2000), General notions of statistical depth function. *The Annals of Statistics*, 28, 461-482.

	Framework		
FA Steps	**Univariate**	**Multivariate**	**Functional**
i) Exploratory analysis & outlier detection	Large literature : Cunnane 1987 Kite 1988 Stedinger et al 1993 Rao & Hamed 2000	Very sparse literature : Chebana & Ouarda 2011b	The specific aim of the present paper
ii) Checking the FA assumptions: stationarity homogeneity independence	Large literature: Yue et al 2002 Kundzewicz et al 2005 Khaliq et al 2009	Very sparse literature : Chebana et al 2010	To be developed
iii) Modeling & estimation	Large literature : Cunnane 1987 Bobée & Ashkar 1991	Large recent literature: Shiau 2003 Zhang & Singh 2006 Salvadori et al 2007	To be developed
iv) Risk evaluation & analysis	Large literature : Chow et al. 1988	Little but growing literature : Shiau 2003 Chebana & Ouarda 2011a	To be developed

Table 1: FA steps in the three frameworks.

Note: in the univariate framework, step (i) is straitforward and is generally not treated separately;

The references are given only as examples from the literature for space limitation.

Year	z_1	z_2	z_3	z_4
1979	-1180.34	1174.70	1457.11	373.28
1980	42.34	329.59	-115.61	121.06
1981	**2613.26**	**2046.54**	-448.91	225.00
1982	**1947.70**	-704.01	500.88	-600.22
1983	-1673.67	1095.67	1936.68	-133.87
1984	843.83	822.51	-10.20	-238.71
1985	903.82	-1171.34	0.07	-419.37
1986	**-1737.16**	-185.44	466.40	36.39
1987	**-1671.02**	**-1615.92**	285.93	10.93
1988	-130.59	393.79	-732.23	18.73
1989	-633.73	-176.40	-663.42	87.96
1990	-669.66	-519.85	-375.08	-324.09
1991	529.43	-604.15	-281.41	-52.87
1992	-465.75	-725.79	-342.08	944.50
1993	-374.77	-753.53	-315.64	19.32
1994	1058.79	68.18	451.04	1281.09
1995	-268.43	463.72	-933.12	-268.28
1996	748.05	235.06	560.82	-575.86
1997	**1085.56**	-516.08	447.61	482.64
1998	-1557.15	428.41	-504.03	-165.38
1999	-1306.02	1809.15	-621.20	-641.45
2000	879.38	-20.22	145.23	-178.76
2001	-1173.90	-702.48	-837.66	133.55
2002	1134.07	**-1692.79**	796.87	-433.89
2003	-67.68	120.028	-1087.29	315.00
2004	1123.65	400.67	219.25	-16.70

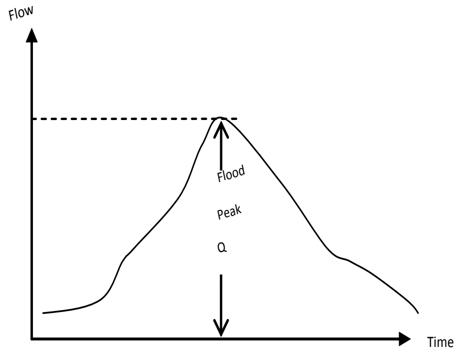
Table 2: First four principal component scores. The bold characters indicate the largest and the smallest values for the first and the second component.

Year	Peak	Volume	TD	MO	SO	TO
1979	886.67	2088.92	0.2692	0.0571	0.1361	0.4615
1980	849.67	2357.02	0.3846	0.1971	0.1567	0.2308
1981	1456.67	3909.14	0.0385	**0.8851**	**0.9563**	**0.9231**
1982	1270.00	2443.15	0.0385	0.8032	0.6246	**0.9231**
1983	974.67	3012.18	0.0769	0.6700	0.8500	0.8462
1984	1056.67	2751.69	0.1154	0.4713	0.6857	0.7692
1985	787.00	1574.21	0.1538	0.4623	0.4815	0.6923
1986	610.33	1536.34	0.1154	0.5306	0.6026	0.7692
1987	344.33	1069.86	0.0385	0.8225	0.9204	**0.9231**
1988	843.33	2374.49	0.3077	0.2390	0.2455	0.3846
1989	678.67	1534.53	0.1923	0.4534	0.5395	0.6154
1990	506.33	1752.06	0.0769	0.7223	0.5603	0.8462
1991	740.00	2260.57	0.1538	0.4461	0.3003	0.6923
1992	710.80	1128.71	0.0385	0.7223	0.8923	**0.9231**
1993	666.80	1407.32	0.1538	0.5400	0.6964	0.6923
1994	932.90	2722.55	0.1538	0.4802	0.6113	0.6923
1995	868.77	2192.44	0.3462	0.0068	0.0324	0.3077
1996	886.90	2476.36	0.3077	0.2644	0.3562	0.3846
1997	697.30	2665.87	0.0385	0.7817	0.6607	**0.9231**
1998	825.00	1843.60	0.3077	0.1963	0.2717	0.3846
1999	1306.67	2652.26	0.0385	0.8042	0.7450	**0.9231**
2000	858.90	2492.65	0.2308	0.3526	0.4095	0.5385
2001	732.50	1188.92	0.0769	0.7053	0.8076	0.8462
2002	999.60	1485.36	0.0385	0.8045	0.6758	**0.9231**
2003	1004.93	1883.80	0.1538	0.6236	0.4102	0.6923
2004	842.57	2802.32	0.0769	0.6783	0.7252	0.8462

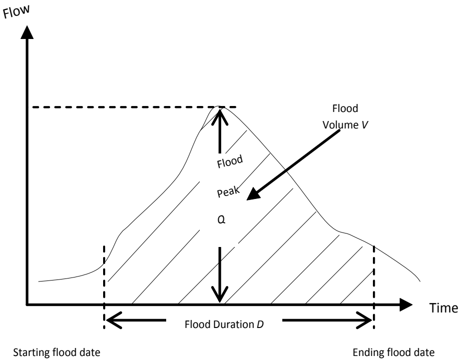
Table 3: Multivariate results for flood peak and volume. TD: Tukey Depth, MO: Mahalanobis Outlyingness, SO: Spatial Outlyingness, and TO: Tukey Outlyingness. Bold characters indicate the values of the outlying measure corresponding to the detected outlier.

	Peak	Volume
Mean (vector)	859.15	2138.70
Tukey median (vector)	847.72	2216.22
Dispersion (matrix)	57316.61	113915.10
	113915.10	457040.80

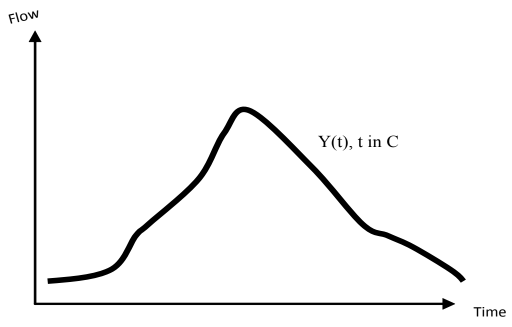
Table 4: Multivariate results for flood peak and volume: location and scale parameters.

(a) 

Framework: univariate
Variable : peak Q
Nature : real value
Series : $q_1, ..., q_n$
Duration: n years
References: Rao and Hamad (2000)
Cunnane (1987)

(b) 

Framework: multivariate
Variables: peak Q, volume V and duration D
(the most important and most studied)
Nature: vector
Series : e.g. $(q_1, v_1, d_1), ..., (q_n, v_n, d_n)$
Duration: n years
References: Yue et al. (1999)
Shiau (2003)
Zhang and Singh (2006)
Chebana and Ouarda (2011a)

(c) 

Framework: functional
Variable: $Y(t), t \in \mathcal{C}$,
the whole hydrograph
Nature: function
Series: $(Y_1(t), ..., Y_n(t)), t \in \mathcal{C}$
Duration: n years
References: The object of the present paper

Figure 1: Illustration of the different approaches (a) univariate (b) multivariate and (c) functional with the corresponding types of variables, series and a number of references.

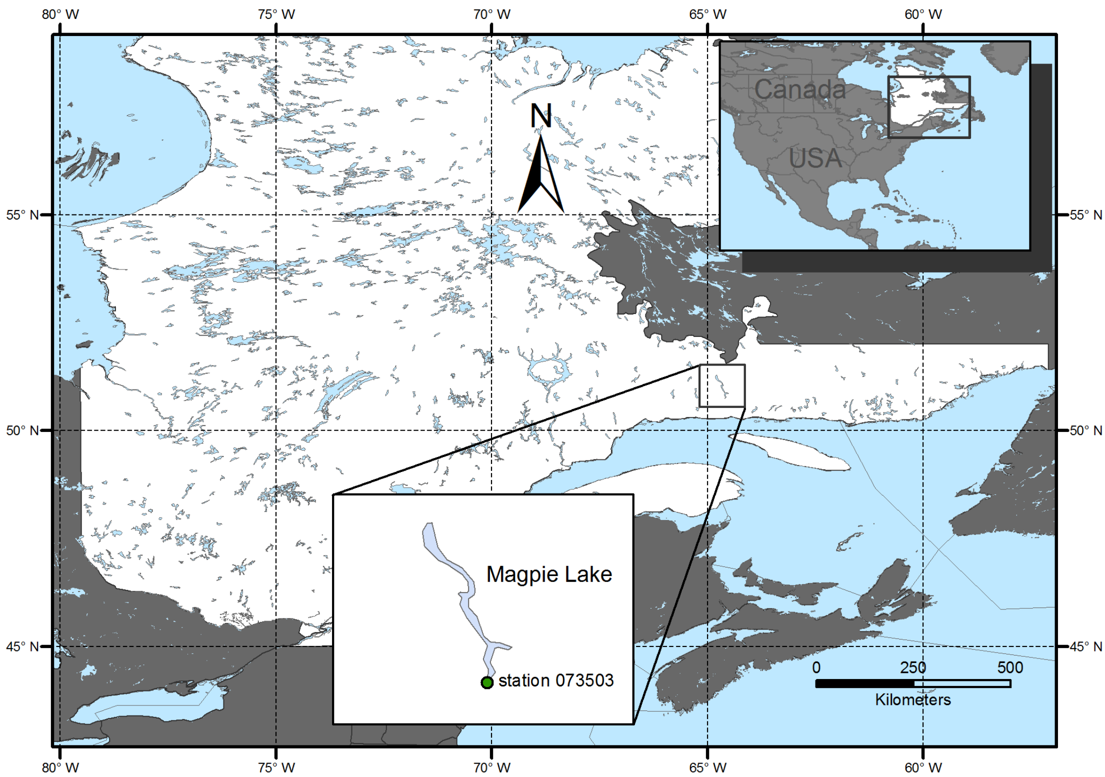


Figure 2: Geographical location of the Magpie station.

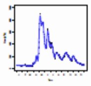
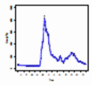
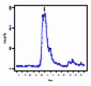
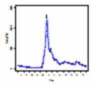
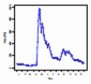
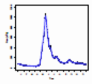
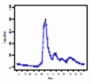
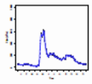
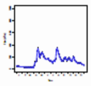
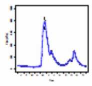
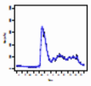
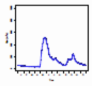
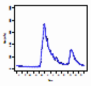
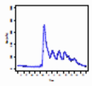
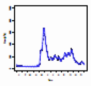
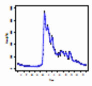
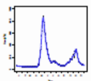
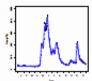
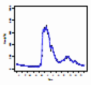
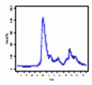
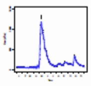
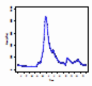
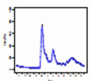
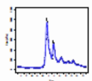
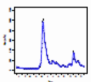
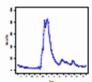
Year	Univ. Q	Biv. $\begin{pmatrix} Q \\ V \end{pmatrix}$	Func. Hydrograph	Year	Univ. Q	Biv. $\begin{pmatrix} Q \\ V \end{pmatrix}$	Func. Hydrograph
1979	886.7	$\begin{pmatrix} 886.7 \\ 2088.9 \end{pmatrix}$		1992	710.8	$\begin{pmatrix} 710.8 \\ 1128.7 \end{pmatrix}$	
1980	849.7	$\begin{pmatrix} 849.7 \\ 2357.0 \end{pmatrix}$		1993	666.8	$\begin{pmatrix} 666.8 \\ 1407.3 \end{pmatrix}$	
1981	1456.7	$\begin{pmatrix} 1456.7 \\ 3909.1 \end{pmatrix}$		1994	932.9	$\begin{pmatrix} 932.9 \\ 2722.5 \end{pmatrix}$	
1982	1270.0	$\begin{pmatrix} 1270.0 \\ 2443.1 \end{pmatrix}$		1995	868.8	$\begin{pmatrix} 868.8 \\ 2192.4 \end{pmatrix}$	
1983	974.7	$\begin{pmatrix} 974.7 \\ 3012.2 \end{pmatrix}$		1996	886.9	$\begin{pmatrix} 886.9 \\ 2476.4 \end{pmatrix}$	
1984	1056.7	$\begin{pmatrix} 1056.7 \\ 2751.7 \end{pmatrix}$		1997	697.3	$\begin{pmatrix} 697.3 \\ 2665.9 \end{pmatrix}$	
1985	787.0	$\begin{pmatrix} 787.0 \\ 1574.2 \end{pmatrix}$		1998	825.0	$\begin{pmatrix} 825.0 \\ 1843.6 \end{pmatrix}$	
1986	610.3	$\begin{pmatrix} 610.3 \\ 1536.3 \end{pmatrix}$		1999	1306.7	$\begin{pmatrix} 1306.7 \\ 2652.3 \end{pmatrix}$	
1987	344.3	$\begin{pmatrix} 344.3 \\ 1069.9 \end{pmatrix}$		2000	858.9	$\begin{pmatrix} 858.9 \\ 2492.6 \end{pmatrix}$	
1988	843.3	$\begin{pmatrix} 843.3 \\ 2374.5 \end{pmatrix}$		2001	732.5	$\begin{pmatrix} 732.5 \\ 1188.9 \end{pmatrix}$	
1989	678.7	$\begin{pmatrix} 678.7 \\ 1534.5 \end{pmatrix}$		2002	999.6	$\begin{pmatrix} 999.6 \\ 1485.4 \end{pmatrix}$	
1990	506.3	$\begin{pmatrix} 506.3 \\ 1752.1 \end{pmatrix}$		2003	1004.9	$\begin{pmatrix} 1004.9 \\ 1883.8 \end{pmatrix}$	
1991	740.0	$\begin{pmatrix} 740.0 \\ 2260.6 \end{pmatrix}$		2004	842.6	$\begin{pmatrix} 842.6 \\ 2802.3 \end{pmatrix}$	

Figure 3: Data for each one of the three frameworks: univariate, bivariate and functional.

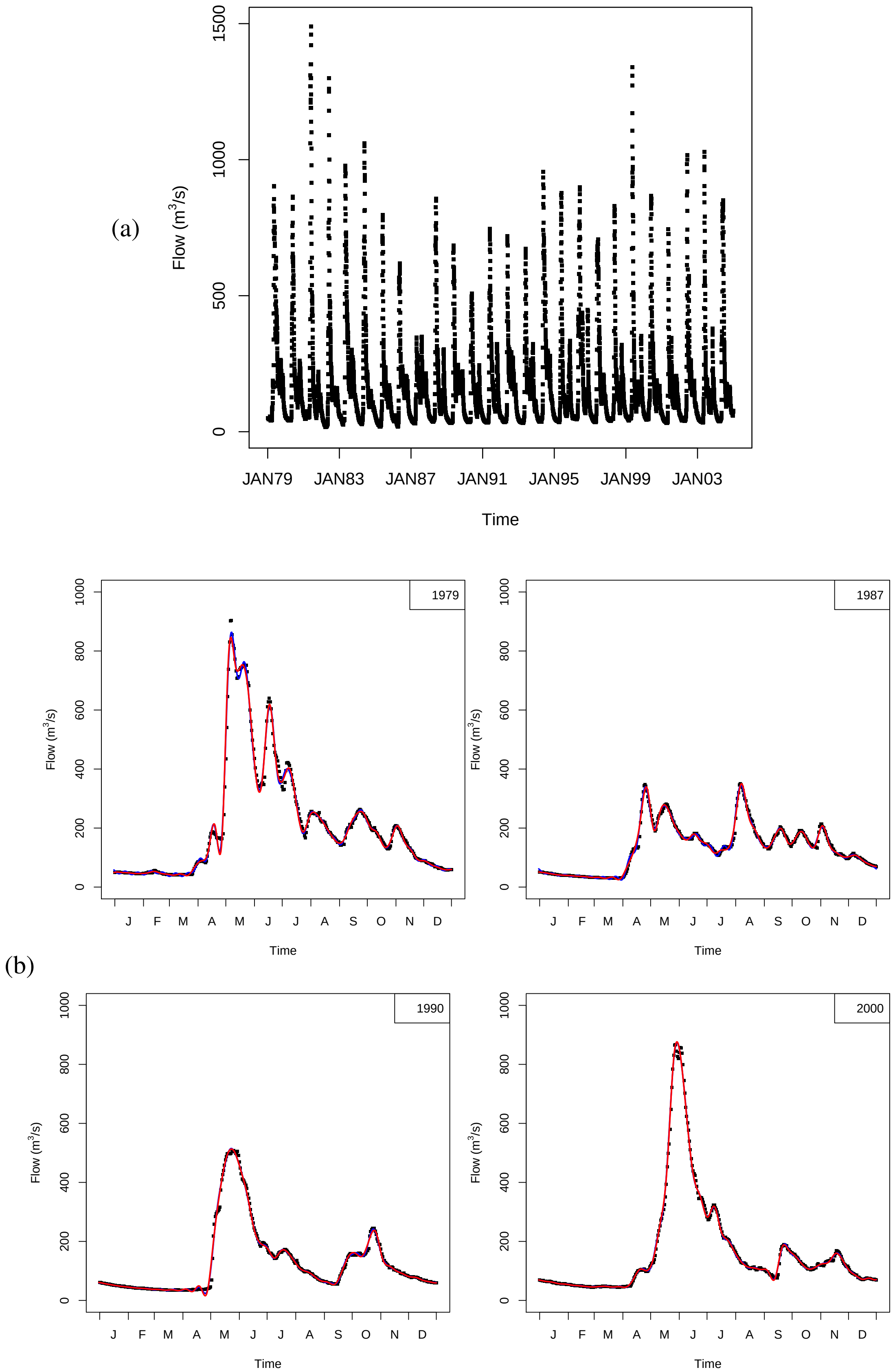


Figure 4: The representation of all the data (in (a)) and illustration of discrete hydrographs and the corresponding smoothing curves (Fourier in blue and B-Splines in red) for some selected years (in (b))

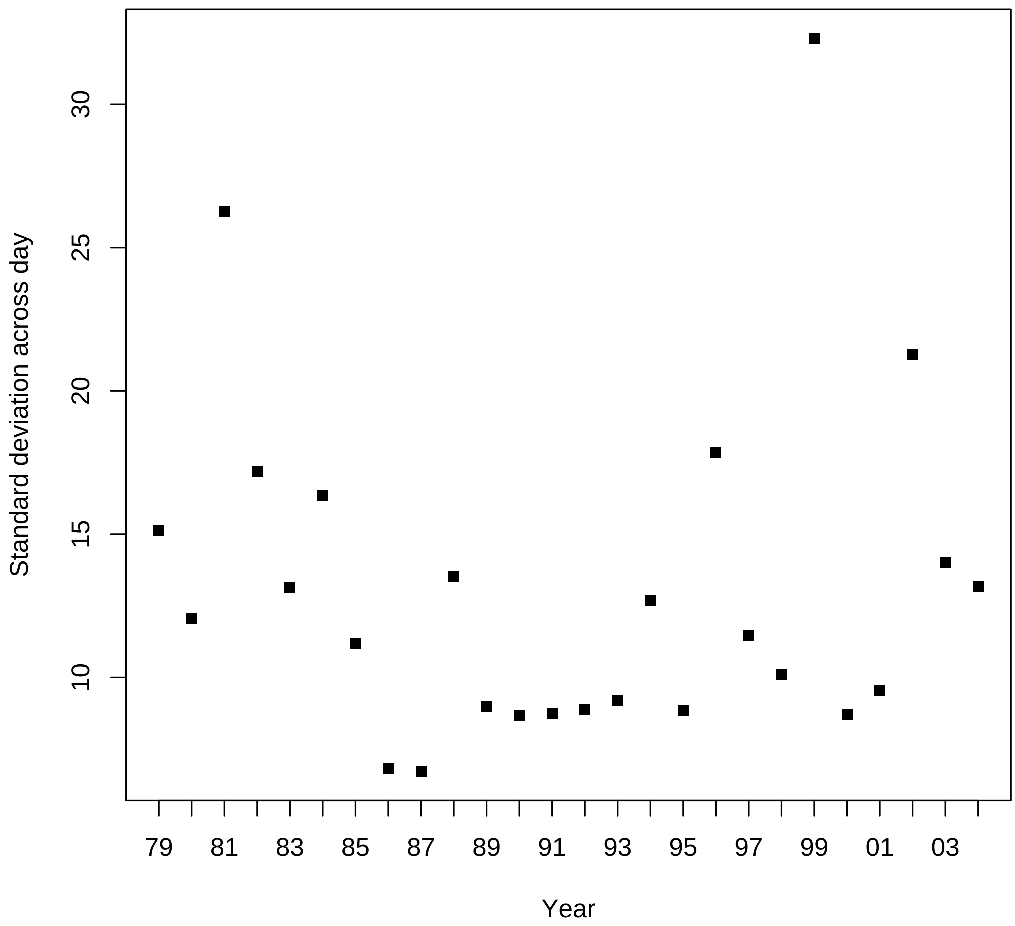


Figure 5: Standard deviations of the residuals from the smooth flows across day

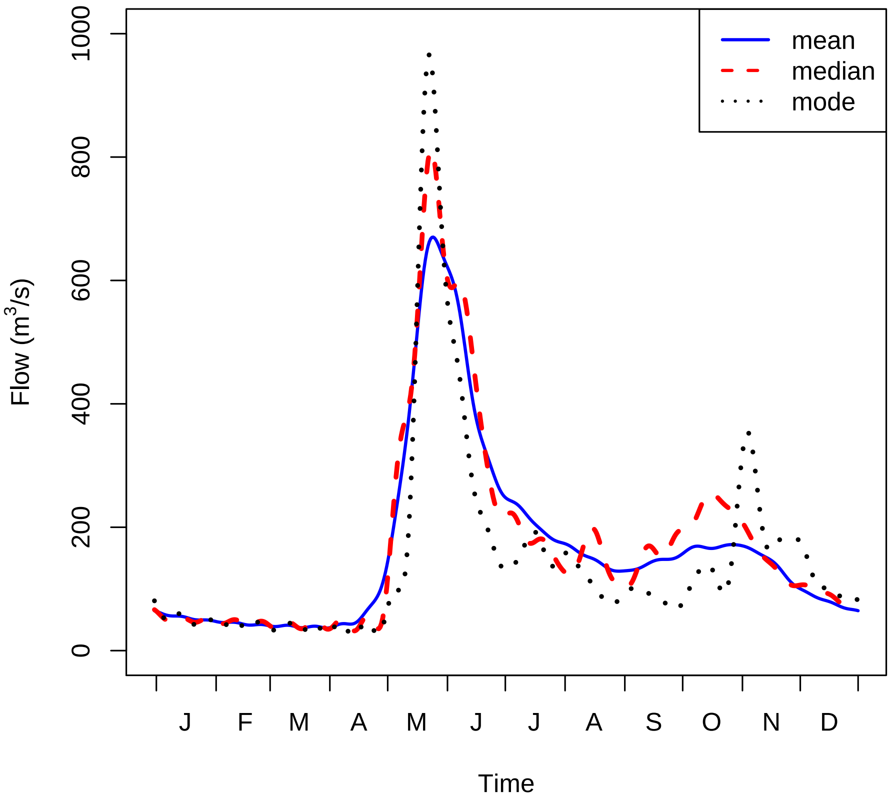
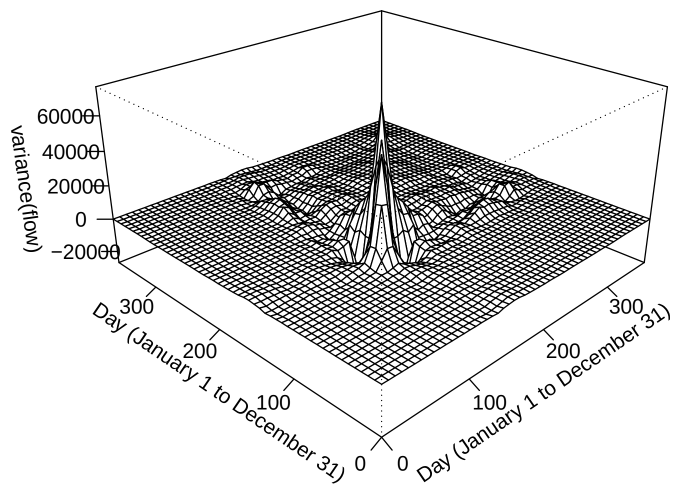
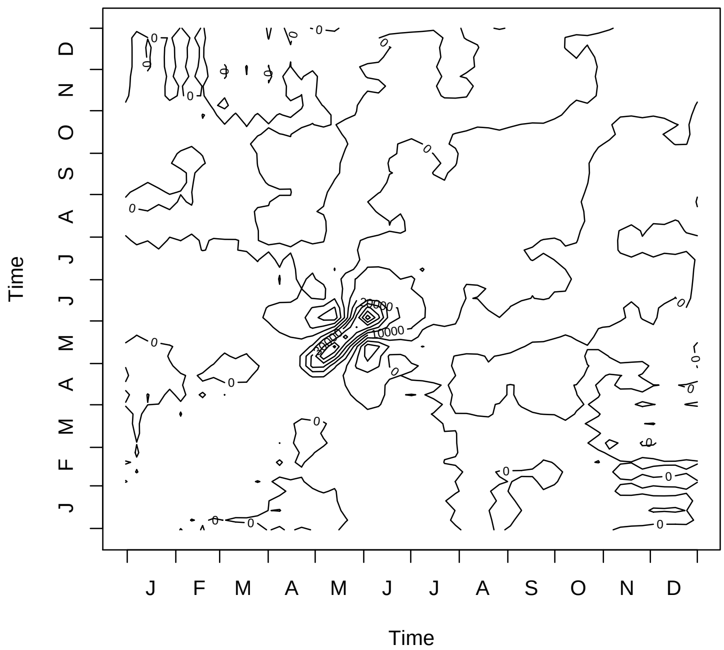


Figure 6: Fourier smoothed location curves : the mean, the median and the mode



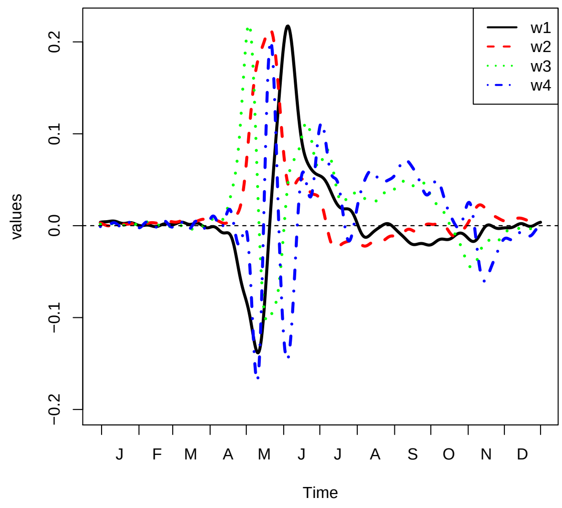
(a)



(b)

Figure 7: Estimated variance-covariance (a) surface of the flow curves for years 1979 to 2004 and (b) the corresponding contour map

(a)



(b)

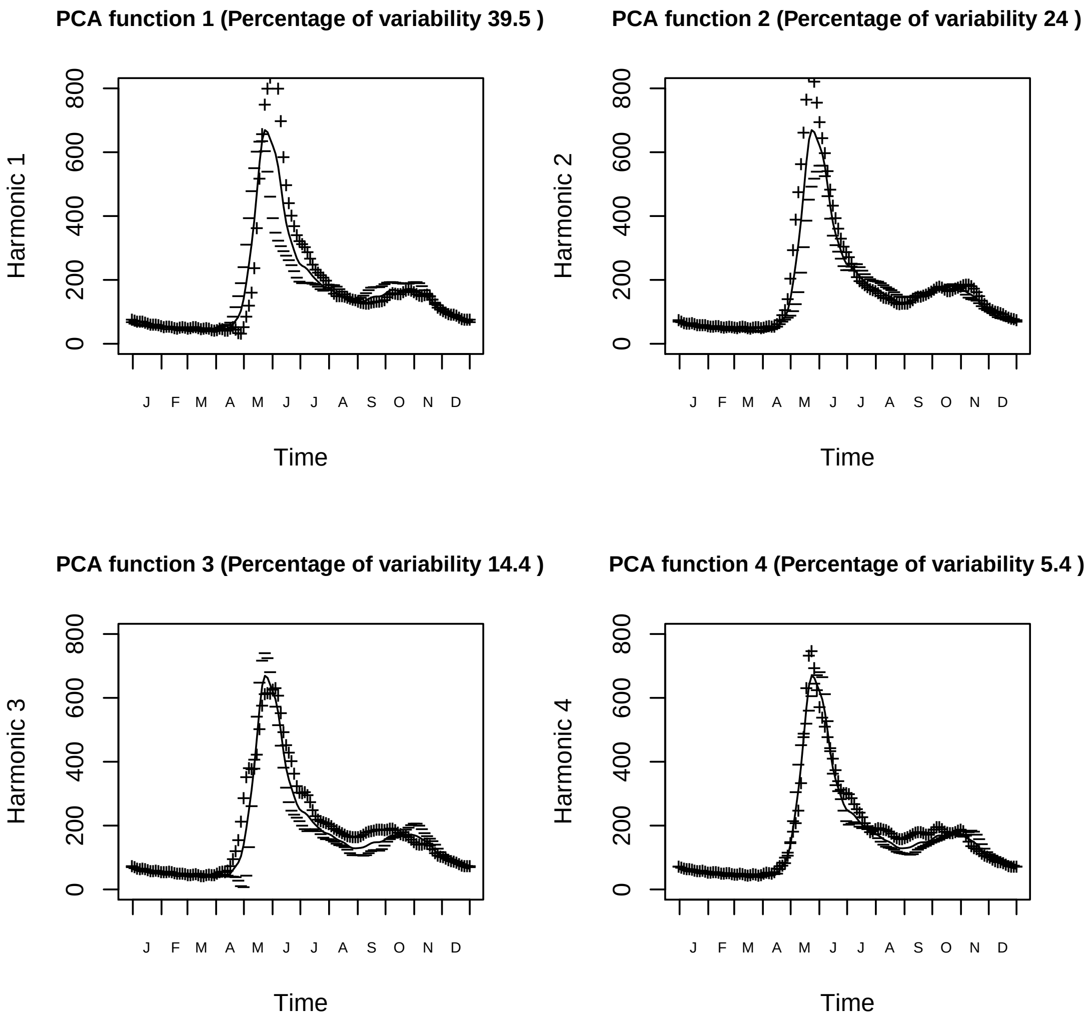


Figure 8: First four smoothed principal components: (a) centered components; (b) components with variation about the mean $\bar{y}$. Negative and positive perturbations are indicated respectively by the minus (-) and plus (+) symbols.

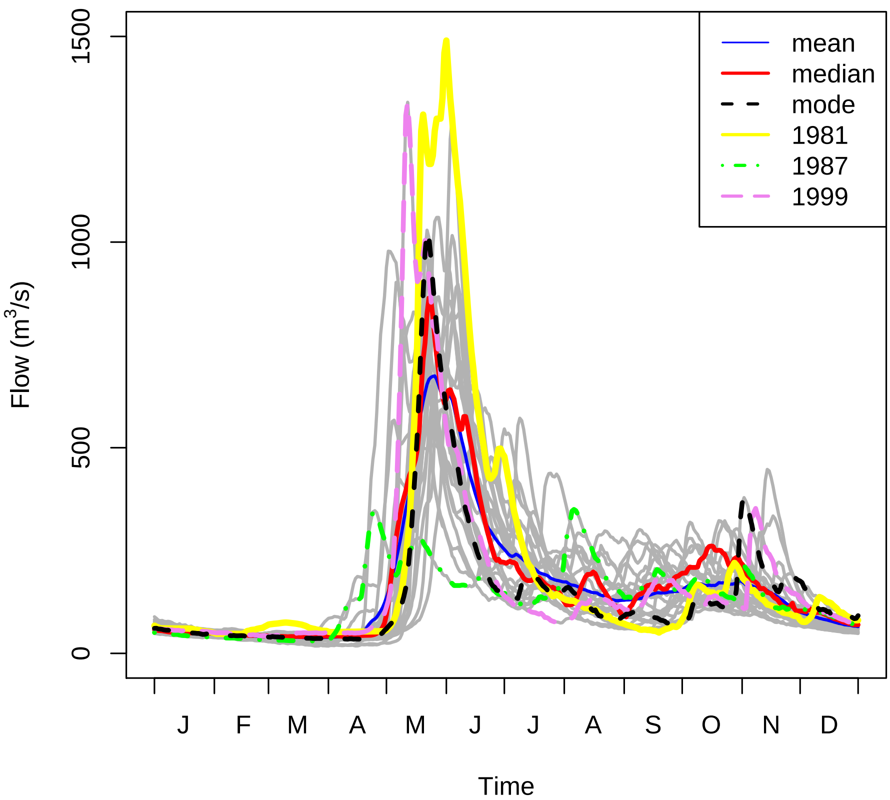
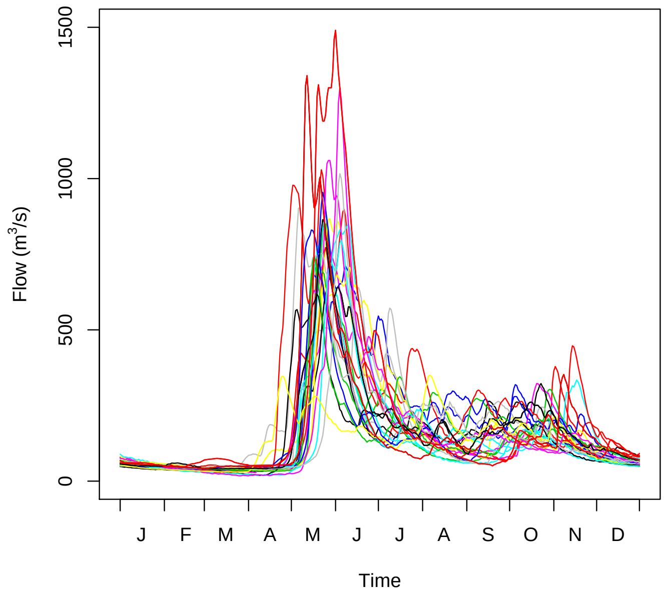
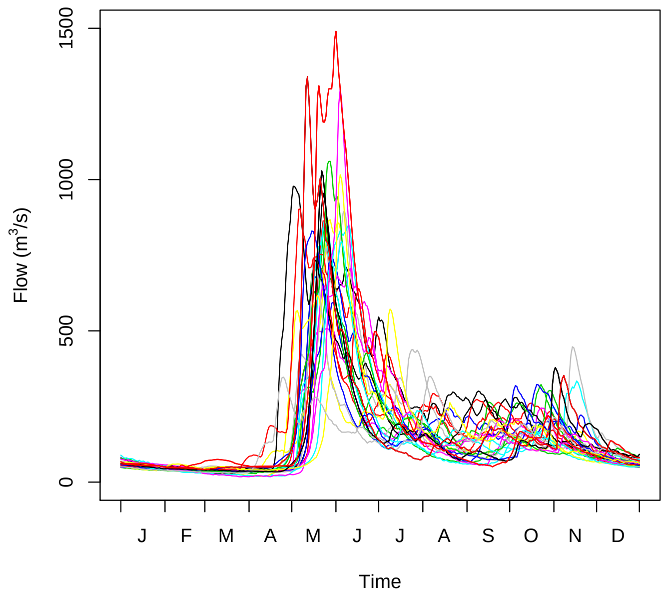


Figure 9: Curves corresponding to the suspected years (based on principal component scores) with the mean, median and mode curves



(a)



(b)

Figure 10: Rainbow plots of the flow curves for years 1979 to 2004 using (a) the bivariate score depth and (b) the kernel density estimate.

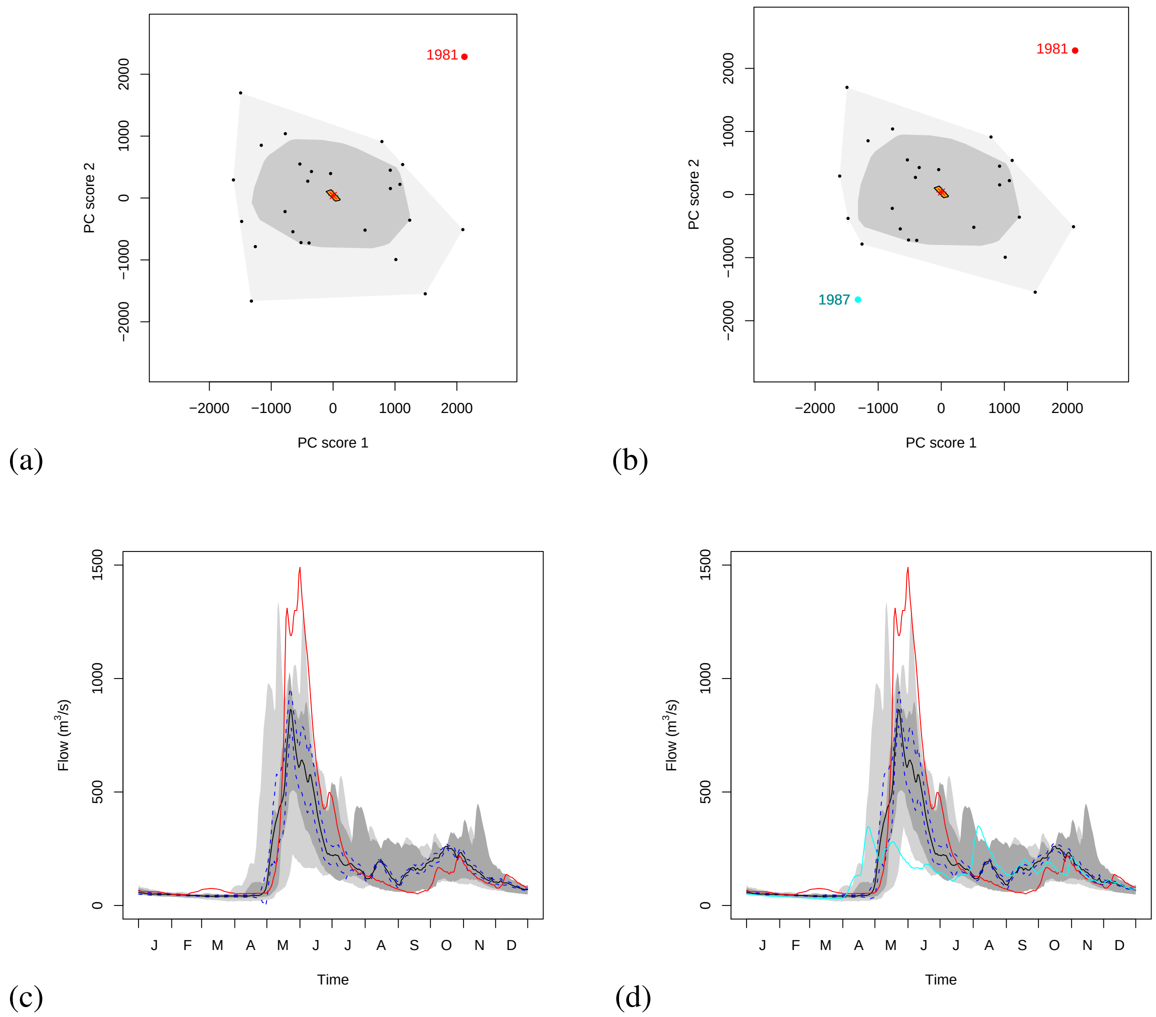


Figure 11: Bivariate score Bagplot with (a) 99% and (b) 95% of probability coverage and the corresponding functional Bagplot with (c) 99% and (d) 95% of probability coverage. The solid black curve shows the median curve and in blue are presented its 95% or 99% point-wise confidence intervals while in (a) and (b) the red asterisk is the Tukey median of the bivariate principal scores

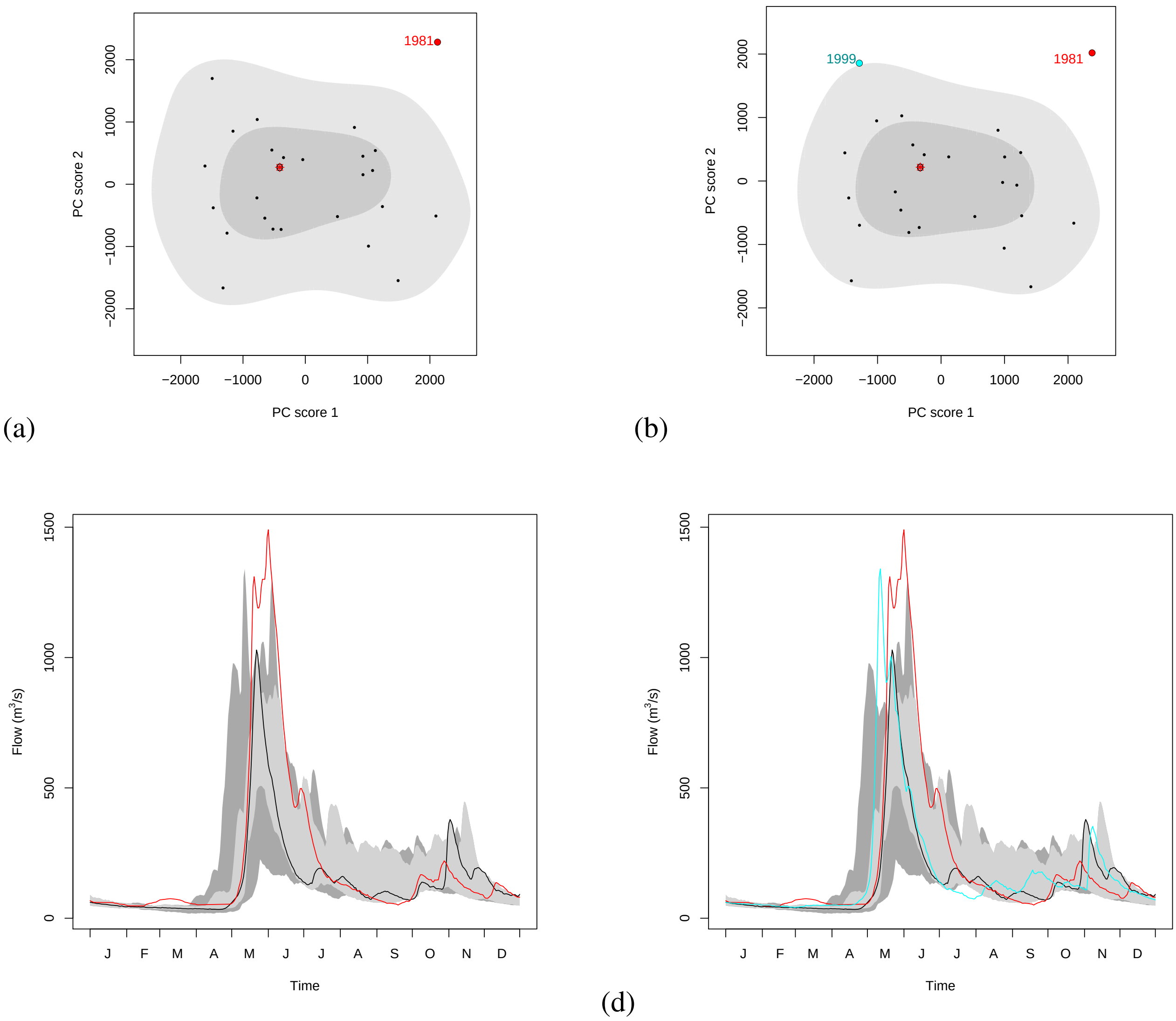


Figure 12: Bivariate score HDR boxplot with (a) 99% and (b) 95% of probability coverage and the corresponding functional HDR boxplot with (c) 99% and (d) 95% of probability coverage. The solid black curve shows the modal curve and in blue are presented its 95% or 99% pointwise confidence intervals while in (a) and (b) the red asterisk is the mode

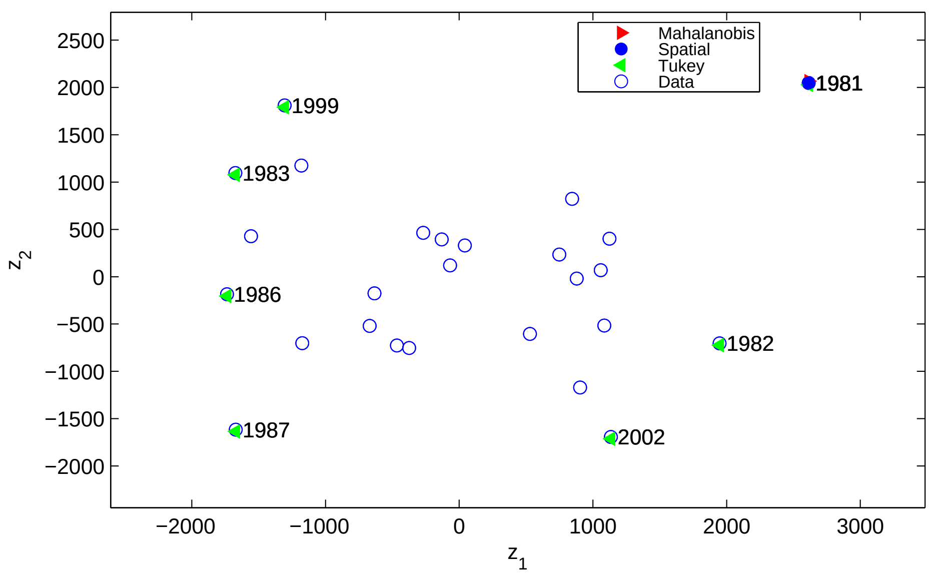
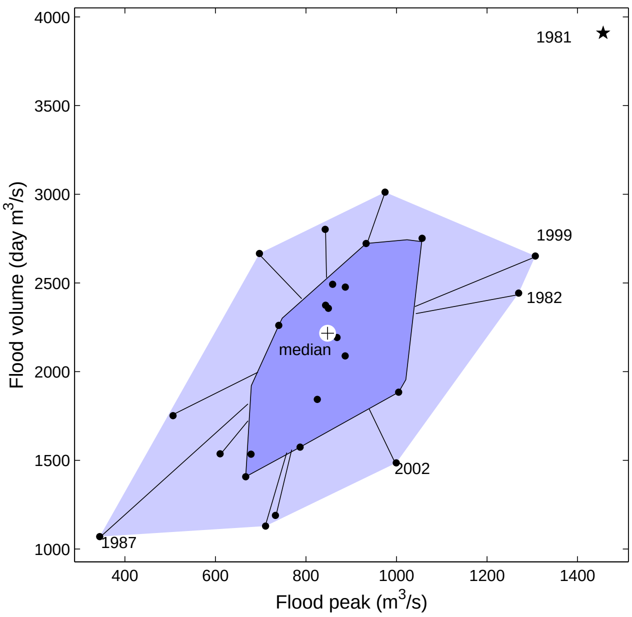
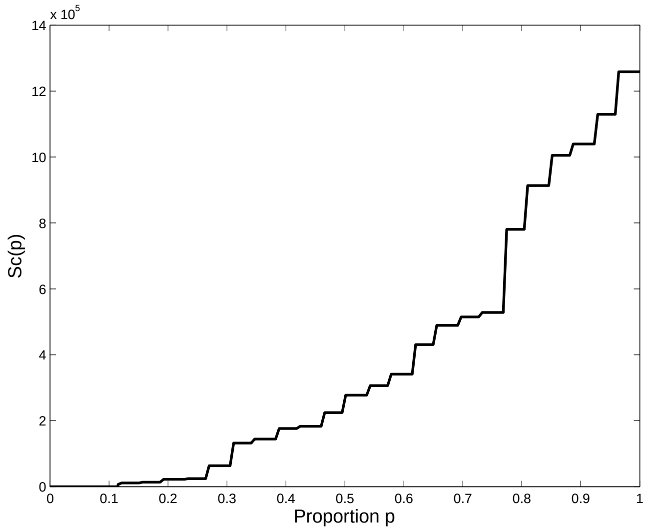


Figure 13: Outlier detection using the Outlyingness approach applied on the first two scores



(a)



(b)

Figure 14: Bivariate results : (a) Bagplot with the median and some particular years and (b) Scalar scale function