Journal of
Flood Risk Management

CIWEM

River flow modelling using nonparametric functional data analysis

A. Quintela-del-Río and M. Francisco-Fernández

University of A Coruña, A Coruña, Spain

Correspondence
Alejandro Quintela-del-Río, Universidade da Coruña, Departamento de Matemáticas, Facultad de Informática, Elviña, 15071 A Coruña, Spain
Email: aquintela@udc.es

DOI: 10.1111/jfr3.12282

Key words
Extreme values; forecasting; functional data; nonparametric estimation; river flow.

Abstract

Time series and extreme value analyses are two statistical approaches usually applied to study hydrological data. Classical techniques, such as autoregressive integrated moving-average models (in the case of mean flow predictions), and parametric generalised extreme value fits and nonparametric extreme value methods (in the case of extreme value theory) have been usually employed in this context. In this article, nonparametric functional data methods are used to perform mean monthly flow predictions and extreme value analysis, which are important for flood risk management. These are powerful tools that take advantage of both, the functional nature of the data under consideration and the flexibility of nonparametric methods, providing more reliable results. Therefore, they can be useful to prevent damage caused by floods and to reduce the likelihood and/or the impact of floods in a specific location. The nonparametric functional approaches are applied to flow samples of two rivers in the United States. In this way, monthly mean flow is predicted and flow quantiles in the extreme value framework are estimated using the proposed methods. Results show that the nonparametric functional techniques work satisfactorily, generally outperforming the behaviour of classical parametric and nonparametric estimators in both settings.

Introduction

Prediction of future values is essential for the design of water systems, and control measures will be more effective if the process is reliable. Likewise, management and scheduling of areas exposed to flood risk rely heavily on tools for frequency analysis of hydrological extremes.

Numerous studies have been carried out on hydrological problems using statistical methods. Among them, time series prediction is topical in this field (Toth *et al.*, 2000; Tamea *et al.*, 2005; Wu *et al.*, 2009). Research studies on time series also include linear models for forecasting river flows (see Wang *et al.*, 2009, and references therein). Among the several techniques to model time series, autoregressive integrated moving-average (ARIMA) models described well the data analysed in the present research and, therefore, they were employed to fit the hydrological time series studied. Moreover, with this choice, a similar comparison (between ARIMA models and nonparametric functional methods) to that performed in some works widely cited in the literature (Ferraty *et al.*, 2005) can be carried out. Basically, ARIMA models are preferred for time series of short-memory type (the autocorrelation structure decreases quickly), while, in other cases, hydrological processes are of long-memory type. Other possible alternatives not considered in the present research would be, for example, fractional Gaussian noise and broken line models (Koutsoyiannis, 2000).

Statistics of extremes (Coles, 2001) is also one of the most significant techniques in frequency analysis (Katz *et al.*, 2002; Singh *et al.*, 2005; Saf, 2009). Daily, monthly, or annual maximum time series of river flow recordings are typically represented by the generalised extreme value (GEV) distribution.

ARIMA and GEV fitting are typical examples of parametric modelling. A different type of statistical model applied to hydrological data involves using nonparametric curve estimation methods, which does not require restrictive assumptions on the distribution of the population of interest. Several papers have applied nonparametric estimation methods to hydrological time series to carry out predictions as well as to perform extreme value analysis (Lall *et al.*, 1993; Guo *et al.*, 1996; Sharma *et al.*, 1997; Kim and Heo, 2002; Wang *et al.*, 2009; Quintela-Del-Río, 2011).

Further details regarding nonparametric techniques (including theoretical motivations, practical applications to several scientific fields and references) may be found in, for instance, the books of Ruppert *et al.* (2003) or Wasserman (2005).

Time series were recently analysed by nonparametric functional data analysis (NFDA) (Ramsay and Silverman, 2005; Ferraty and Vieu, 2006). NFDA works with data consisting of curves or multidimensional variables. Different procedures using these techniques have been applied to several complex real problems (Besse *et al.*, 2000; Hall *et al.*, 2001; Fernández De Castro *et al.*, 2005; Castellano Méndez *et al.*, 2009).

This article focuses on applying NFDA techniques in prediction problems and extreme value analysis in the setting of hydrology. The organisation of the article is as follows. *Statistical methods* presents the statistical methods used in this article. These methods correspond to ARIMA models for time series prediction (*Time series analysis, ARIMA models*), and GEV parametric estimators and nonparametric methods in extreme value analysis (*The GEV distribution* and *Nonparametric estimators*, respectively). Next, the new proposals using NFDA to study both problems (time series prediction and extreme value analysis) are presented (*NFDA applied to time series analysis*). In *Hydrological data*, these tools are applied to river flow data from two sites in the United States. Finally, in *Discussion*, a general discussion of the results is included.

Statistical methods

Time series analysis, ARIMA models

Let $\{Z_t\}_{t\in\mathbb{R}}$ be a stochastic process or time series observed until a time T. Usually, the process is observed at N discretised times, and the observations are denoted by $\{Z_1, \ldots, Z_N\}$. To predict a future value Z_{N+s}, the simplest way consists of taking into account one single past value. This is done by constructing a two-dimensional statistical sample of size $n = N - s$, by setting $X_i = Z_i$ and $Y_i = Z_{i+s}$, with $i = 1,\ldots, N - s$. Therefore, the problem is converted into a standard prediction problem of a response Y, given an explanatory variable X. This can be generalised by considering the following autoregressive process of order p:

$$Z_{i+s} = m\left(Z_i,\ldots,Z_{i-p+1}\right) + \varepsilon_i, i = p,\ldots,N-s, \tag{1}$$

where ε_i is the error process, assumed to be independent of Z_i, and the aim is to estimate the function $m(\cdot)$.

A first approximation consists in assuming that $m(\cdot)$ belongs to a particular class of functions, only depending on a finite number of parameters to be estimated, such as the ARIMA(p, d, q) models (Singh *et al.*, 2005). If d is a non-negative integer, then $\{Z_t\}$ is said an ARIMA(p, d, q) process if $Y_t = (1 - B)^d Z_t$ is a causal ARMA(p, q) process, where B is the backward shift operator defined by $B^j Z_t = Z_{t-j}$, $j = 0, \pm 1, \pm 2, \ldots$ (Brockwell and Davis, 1991). Note that the process $\{Z_t, t = 0, \pm 1 \pm 2, \ldots\}$ is said an ARMA(p, q) process if $\{Z_t\}$ is stationary and if for every t, $Z_t - \varphi_1 Z_{t-1} - \ldots - \varphi_p Z_{t-p} = \varepsilon_t + \theta_1\varepsilon_{t-1} + \ldots + \theta_q\varepsilon_{t-q}$ where $\{\varepsilon_t\}$ is a process of error terms, generally assumed to be uncorrelated random variables with mean 0 and variance σ^2. In a ARIMA(p,d,q), p is the order (number of time lags) of the autoregressive model, d is the degree of differencing (the number of times the data have had past values subtracted), and q is the order of the moving-average model. The good practical properties of ARIMA models have led to regularly use them to study hydrological problems. Some relevant papers on this topic are, for example, Montanari *et al.* (1997), Toth *et al.* (2000), or Tamea *et al.* (2005).

The prediction problem can be tackled using nonparametric methods. To apply these methods only some mild regularity conditions on the function $m(\cdot)$ have to be assumed. The 'curse of dimensionality problem' (Wand and Jones, 1995, p. 90) is particularly troublesome in this nonparametric framework. It has to do with the selection of the number of past values to consider in the model. This is indeed an important question. The lower the number of past predictors, the less flexible the model is, but when the lag increases, a large number of observations are needed to obtain good estimates of the model parameters. This number increases exponentially as the dimension becomes larger.

Extreme value analysis

The GEV distribution

Suppose $X_1, \ldots, X_n$ is a sequence of extreme values with a common distribution function F. In the context of this article, these variables can be the maximum river flows measured in a specific period of time (24 h, a month, a year, etc.). Classical parametric extreme value theory uses the idea that, under certain regularity conditions (Fisher and Tippett, 1928), the limit of the distribution function F of the maximum is the GEV distribution. Its cumulative distribution function is:

$$F_\theta(x) = \begin{cases} exp\left\{-\left[1+\gamma(x-\mu)/\sigma\right]^{-1/\gamma}\right\} & \text{if}\gamma \neq 0 \\ exp\left\{-exp\left[-(x-\mu)/\sigma\right]\right\} & \text{if}\gamma = 0 \end{cases} \tag{2}$$

with $\theta = (\mu, \sigma, \gamma)$. Here, μ is the location parameter, $\sigma > 0$ is the scale parameter, and γ is the shape parameter. Mean and standard deviation are obtained as functions of these parameters (Coles, 2001). The range of definition of the GEV distribution depends on γ. If $\gamma \neq 0$, $F_\theta(x)$ is defined for x such that $1 + \gamma(x - \mu)/\sigma > 0$, while if $\gamma = 0$, it is defined for $-\infty < x < \infty$. Various values of the shape parameter yield the extreme value type I, II, and III distributions. Specifically, the three cases $\gamma = 0$, $\gamma > 0$, and $\gamma < 0$ correspond to the Gumbel, Fréchet, and 'reversed' Weibull distributions, respectively. Using the random sample of extreme values, an estimator $\hat{\theta}$ for θ can be obtained. Then, substituting F by $F_{\hat{\theta}}$, estimators of some important functions in this framework can be defined. For instance:

1. The function providing the probabilities of exceedance. In the context of this article, it corresponds to the function that, for a river flow c, gives the probability of obtaining a flow larger than c (per unit of time). It is defined as

$$R(c) = P(X > c) = 1 - F(c). \tag{3}$$

2. The flow quantile, defined as the value of the flow that can be expected to be once exceeded during a T period of time. For each value of T, it is given by

$$FQ(T) = F^{-1}\left(1 - \frac{1}{T}\right) \tag{4}$$

3. The mean return period or recurrence interval of a particular river flow c, defined as an estimator of the interval of time between events of this flow. It can be expressed as the inverse of the probability that a flow c will be exceeded in one period of time:

$$RT(c) = \frac{1}{P(X > c)} = \frac{1}{1 - F(c)}. \tag{5}$$

An application of these expressions is given in *Extreme value analysis.*

Nonparametric estimators

The main advantage of working with nonparametric methods is that they are model-free, that is, no specific functional form is required for the parameters or curves to be estimated. Several nonparametric estimators for different functions of interest have been developed in the last decades. In this work, kernel estimators of the density function and the distribution function are used.

Let X be a continuous random variable, with density function f and distribution function F. Given a random sample $X_1, \ldots, X_n$, each X_i having the same distribution as X, the Parzen-Rosenblatt nonparametric kernel estimator (Parzen, 1962) of f is defined by:

$$f_h(x) = \frac{1}{nh}\sum_{i=1}^{n} K\left(\frac{x - X_i}{h}\right) \tag{6}$$

where K is a kernel function (normally, K is a density function with some regularity conditions) and $h = h(n) \in \mathbb{R}^+$ is the smoothing parameter (or bandwidth) that regulates the amount of smoothing to be used. From the relation between a density function and a distribution function, a nonparametric kernel estimator of the distribution function can be directly constructed:

$$F_h(x) = \int_{-\infty}^{x} f_h(t)\mathrm{d}t = \frac{1}{n}\sum_{i=1}^{n} H\left(\frac{x - X_i}{h}\right) \tag{7}$$

where $H(u) = \int_{-\infty}^{u} K(t)\mathrm{d}t$ is the distribution function of the kernel K.

Using Eqn (7), nonparametric estimators of the probabilities of exceedance, the flow quantiles, and the recurrence intervals defined in Eqns (3)–(5), respectively, can be obtained:

$$R_h(c) = 1 - F_h(c) \tag{8}$$

$$FQ_h(T) = F_h^{-1}\left(1 - \frac{1}{T}\right) \tag{9}$$

and

$$RT_h(c) = \frac{1}{1 - F_h(c)} \tag{10}$$

An important first step to compute (8), (9) and (10) is the selection of the smoothing parameter h. Popular techniques to select the bandwidth are the modified cross-validation (Bowman et al., 1998; Quintela-Del Río, 2011) and plug-in methods (Lall et al., 1993; Quintela-Del-Río, 2011). In the examples presented in this work, a cross-validation bandwidth selection method is used.

In an extreme value framework, it can be of interest to estimate the flow quantiles or the return periods for extremely large events. In a hydrological context, Lall *et al.* (1993) found that the previous nonparametric estimators can suffer from boundary problems. Some authors have addressed extrapolation issues using nonparametric estimators like those given in Eqn (9) or (10). They basically focused on studying the influence of the kernel function and the bandwidth parameter in the final results. Regarding the kernel, while Guo *et al.* (1996) proposed to use a Gumbel kernel and Lall *et al.* (1993) discussed the use of a variable kernel distribution function estimator to tackle this problem, Adamowski and Feluch (1990) tested Gaussian, Gumbel, and Epanechnikov kernels in flood frequency analysis and found that the choice of the kernel is not important, and the shape of the kernel does not affect extrapolation accuracy. As for the smoothing parameter, the use of variable or local bandwidths to address the extrapolation problem was discussed in Adamowski (1989) or Guo *et al.* (1996). Note that the variance of the (parametric or nonparametric) estimators can increase significantly when the interest is to estimate extremely large flow quantiles. For this, in that case, the results obtained should be considered carefully. In this article, the methods will always be applied for values inside the range of the observed data.

Nonparametric kernel quantile function estimators based on smoothing the empirical quantile function are proposed and studied by Moon and Lall (1994) and Apipattanavis *et al.* (2010). They follow similar ideas, but while in Moon and Lall (1994) the Gasser–Müller estimator (Gasser and Müller, 1984) with higher order kernel is used, in Apipattanavis *et al.* (2010) the smoothing process is carried out employing the local polynomial estimator (Fan and Gijbels, 1996) with a local bandwidth.

Functional data, NFDA techniques

Let $\{(\chi_i, Y_i), i = 1, \ldots, n\}$ be a sample of n random pairs, each distributed as $(\mathcal{X}, Y)$, where the variable $\mathcal{X}$ is of functional nature (a curve), and Y is scalar. Formally, $\mathcal{X}$ is a random variable valued in some semi-metric functional space E, and $d(\cdot,\cdot)$ denotes the associated semi-metric, according to the definition (Ferraty and Vieu, 2006):

1. $\forall\, \mathbf{x} \in E,\ d(\mathbf{x}, \mathbf{x}) = 0$.
2. $\forall\, \mathbf{x}, \mathbf{y}, \mathbf{z} \in E,\ d(\mathbf{x}, \mathbf{y}) \le d(\mathbf{x}, \mathbf{z}) + d(\mathbf{z}, \mathbf{y})$.

The conditional cumulative distribution of Y given $\mathcal{X}$ is defined for any $y \in \mathbb{R}$ and any $\chi \in E$ by:

$$F(y|\chi) = P(Y \le y | \mathcal{X} = \chi) \tag{11}$$

A functional variable can be considered a generalisation of a multidimensional variable, assuming that the variable χ is p-dimensional, with p an integer (for example, $p = 12$ for the monthly mean flow in the 12 months of a year). In this case, the functional space would be $E = \mathbb{R}^p$ and the semi-metric could be the classical Euclidean distance or some equivalent measure (Ramsay and Silverman, 2005).

Both parametric and nonparametric methods can be used in functional data applications. The monograph of Ferraty and Vieu (2006) provides a benchmark of nonparametric curve estimation for functional data. As shown in this book, the conditional distribution $F(\cdot|\chi)$ given in Eqn (11) can be nonparametrically estimated by:

$$\hat{F}_n(y|\chi) = \frac{\sum_{i=1}^{n} K\left(\frac{d(\chi,\chi_i)}{h}\right) H\left(\frac{y - Y_i}{g}\right)}{\sum_{i=1}^{n} K\left(\frac{d(\chi,\chi_i)}{h}\right)} \tag{12}$$

where K is a kernel function and H is defined as the distribution of another kernel density function K_0, that is, $H(x) = \int_{-\infty}^{x} K_0(u)\mathrm{d}u$. Parameters g and h are smoothing parameters or bandwidths (they could take the same value).

Equation (12) is a direct extension of the nonparametric estimator of a conditional distribution function [$F(y|X = x)$, for (X, Y) real random variables (Hall *et al.*, 1999)]. The main difference between functional and non-functional estimators lies in the use of a semi-metric $d(\chi, \chi_i)$ instead of the Euclidean distance $\|\chi - \chi_i\|$. Several types of kernel functions and semi-metrics can be considered (see Ferraty and Vieu, 2006, Sections 3.2–3.4), depending, essentially, on the data at hand. Theoretical optimality properties of the estimator (Eqn (12)) can be found in Quintela-Del-Río (2008).

An important advantage of NFDA techniques is that the framework model reduces to a bivariate setting and, therefore, the curse of dimensionality problem is basically avoided. Additionally, the boundary problems of nonparametric estimators, previously described, can be partially avoided in functional data estimation. This fact requires a proper choice of the semi-metric (Ferraty and Vieu, 2006).

NFDA applied to time series analysis

As it is well known, ARIMA models are constrained by their particular structure and the number of past values used in the statistical model for prediction purposes. NFDA methods overcome these two restrictions, because of the nonparametric nature of the approaches and dividing the observed seasonal time series into a sample of curves. In *Monthly mean flow prediction*, NFDA methods are applied to predict monthly mean flows in practical situations, and the performance of these approaches is compared with that obtained when ARIMA models are employed.

To analyse a monthly mean series as a set of functional data, the original time series is converted into annual curves. Note that if there were some months in which the corresponding measures were not available, the curves would not have the same number of components (this is known as an unbalanced data setting), and more complex specific preprocessing would be required (see Section 3.6 of Ferraty and Vieu, 2006). Let $\{Z_k\}_{k=1}^{N}$ be the complete time series. For $i = 1, \ldots, n$, the annual curves, $\chi_i = (\chi_i(1), \ldots, \chi_i(12))$, are constructed, where

$$\forall t \in \{1,2,\ldots,12\}, \qquad \chi_i(t) = Z_{12\cdot(i-1)+t} \tag{13}$$

corresponds to the monthly mean flows of the ith year. Each annual curve is considered as a continuous path (i.e. $\chi_i = \{Z_{12\,\cdot\,(i\,-\,1)\,+\,t},\ t \in [0, 12)\}$), but observed only at

12 discretised points. Thus, the time series consists of a sample of n dependent functional data $\chi_1, \ldots, \chi_n$.

In this way, much information from the past of the time series can be taken into account, but still using for the past a single continuous object (exactly 1 year). For more insight on this issue, let us suppose, for instance, that the time series could be measured p-times each year with $p > 12$. In this case, the functional data analysis will consider the whole continuous past year and the same asymptotic behaviour remains, independent of p.

In order to predict the monthly mean flow in the year $n + 1$, the following process was carried out. For $i = 1, \ldots, n - 1$ and for any fixed δ in $\{1, \ldots, 12\}$, take $Y_i(\delta) = \chi_{i+1}(\delta)$, i.e. $Y_i(\delta)$ denotes the monthly flow in the month δ and the year $i + 1$. Thus, a sample of $n - 1$ pairs, $\{\chi_i, Y_i(\delta)\}_{i=1}^{n-1}$, with $Y_i(\delta)$ a real variable and χ_i a functional one, is available. According to *Functional data, NFDA techniques*, a predictor of $Y_n(\delta)$, knowing χ_n, can be achieved by estimating the median of the conditional distribution:

$$\hat{Y}_n(\delta) = \hat{t}_{0.5} = \hat{F}_n^{-1}(0.5|\chi_n) \tag{14}$$

where $\hat{F}_n(\cdot|\chi_n)$ is the estimated distribution of $Y(\delta)$ given χ_n. Repeating this step for $\delta = 1, \ldots, 12$, the mean values of the flow for the $(n + 1)$th year can be predicted.

In the functional data context of this article, another approximation consists in considering a regression model like Eqn (1) and using a nonparametric kernel functional method to estimate the regression function $m(\cdot)$. Considering the sample data of functional covariates and a scalar response, $\{\chi_i, Y_i(\delta)\}_{i=1}^{n-1}$, the nonparametric functional estimator (Ferraty and Vieu, 2006) has the expression:

$$\hat{m}(\chi) = \frac{\sum_{i=1}^{n-1} Y_i(\delta) K\left(\frac{d(\chi,\chi_i)}{h}\right)}{\sum_{i=1}^{n-1} K\left(\frac{d(\chi,\chi_i)}{h}\right)} \tag{15}$$

Equation (15) constitutes a functional alternative based on regression techniques to the approach previously used based on median estimation. Using Eqn (15), the flows of the $(n + 1)$th year can be predicted calculating

$$\hat{Y}_n(\delta) = \hat{m}(\chi_n) \quad (\delta = 1, \ldots, 12) \tag{16}$$

NFDA applied to extreme value analysis

Denote by t_α the α-order quantile of the distribution of Y given a particular value of χ. From the conditional distribution function, the α-order quantile is defined as:

$$t_\alpha = F^{-1}(\alpha|\chi), \quad \forall \alpha \in (0,1) \tag{17}$$

Using the estimator given in Eqn (12), a nonparametric estimator of t_α in Eqn (17) is readily obtained by

$$\hat{t}_\alpha = \hat{F}_n^{-1}(\alpha|\chi) \tag{18}$$

Several asymptotic properties of this estimator are shown in Ferraty *et al.* (2005). Equation (18) can be immediately used as an estimator of the flow quantiles Eqn (4). *Extreme value analysis* presents an application of this approximation using a time series of a river in the United States.

The problem of extreme quantile estimation using functional data has also been addressed in Gardes *et al.* (2010), where nonparametric estimators of quantiles from heavy-tailed distributions when functional covariate information is available are studied.

Hydrological data

In this section, the functional nonparametric techniques are applied to two time series of river flow (measured in cubic metres per second, m^3/s), in the United States, which were downloaded from the National Water Information System (NWIS) of the United States, http://waterdata.usgs.gov. The free statistical software R (R Development Core Team, 2016) was employed to implement the different procedures. Specific packages used in this process are cited below.

Firstly, flow data of Salt River near Roosevelt, AZ, were selected. The annual peak flow data for this river were considered by Katz *et al.* (2002), where they used a GEV distribution. A study is also available in Anderson and Meerschaert (1998), who found that the monthly mean flow is quite seasonal and possesses a heavy-tailed distribution. These data have been also used in nonparametric studies (Quintela-Del-Río, 2011). In this article, Salt River hydrological data are employed to examine the approaches on flow prediction and extreme value analysis. Additionally, a monthly mean flow time series of Christina River at Coochs Brigde, DE, was also considered (Senior and Koerkle, 2003; Celebioglu, 2006). These data are only used in the time series prediction application, but not to perform extreme value analysis. Lower flow values, compared with those of Salt River, are obtained here (Figures 2 and 4).

These two rivers were selected because they belong to two different climate areas with disparate temperatures and significant differences in rainfall throughout the year (see Figure 1 for a location map). Christina River at Coochs Bridge at Delaware (United States) is influenced by an

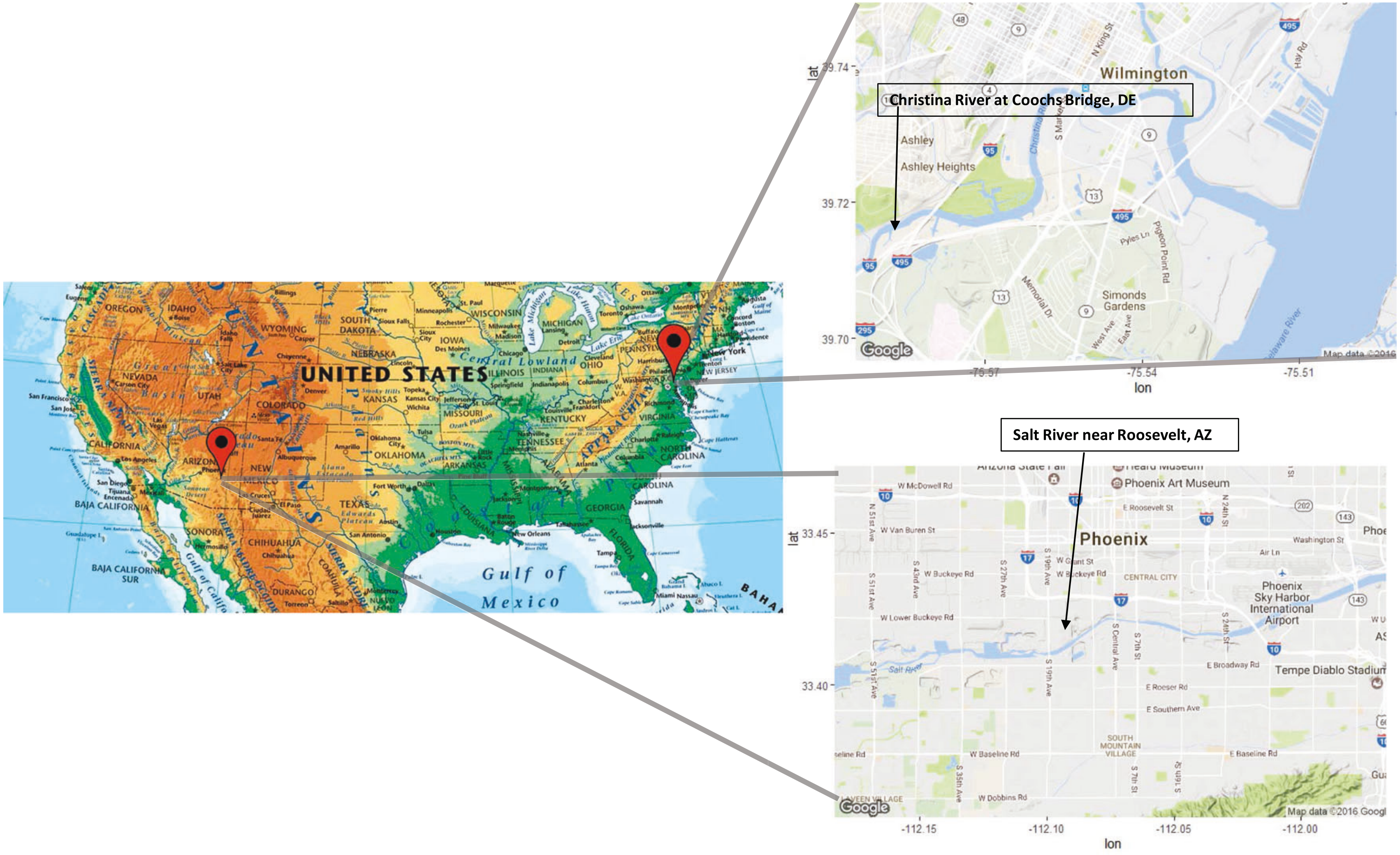

Figure 1 Location map of Salt River near Roosevelt, AZ, and Christina River at Coochs Brigde, DE.

Table 1 Descriptive statistics for the monthly mean flow variable of Salt River and Christina River

Statistic	Salt River	Christina River
Minimum	2.11	0.03
1st quartile	5.88	0.32
Median	9.13	0.62
Mean	23.32	0.82
3rd quartile	24.72	1.11
Maximum	381.47	4.68
Standard deviation	35.93	0.67
Skewness	4.01	1.80
Kurtosis	23.13	7.85

Atlantic climate, with high humidity and stable precipitations. The average annual temperature in this location is about 13 °C, and the average annual precipitation is around 1168 mm. Salt River near Roosevelt, Arizona, belongs to a Continental area, with high average annual temperature (over 21 °C), and an average annual precipitation around 635 mm. Thus, the performance of the NFDA techniques can be compared in different scenarios.

Monthly mean flow prediction

Monthly mean flow data of both rivers, from January 1944 to December 2009, are considered. The number of observations in this time interval is 792. In both cases, no missing values appear, and the quality of the records is guaranteed by the information of the web page of the NWIS.

Firstly, a descriptive statistical analysis of both time series is performed. Table 1 presents the most usual descriptive statistics for the data of the two rivers. In both cases, high values for the kurtosis and the skewness (to the right), and the presence of maximum values far away from the rest of data, according to a heavy-tailed distribution, can be observed.

The mean monthly time series, which does not fit a normal distribution, can be normalised using a log-transformation function in order to remove the periodicity of the original series (Keskin *et al.*, 2006; Wang *et al.*, 2009). In Figure 2, Salt River data, before and after the logarithmic transformation are shown. Figure 3 presents the estimated density functions computed with Eqn (6) using these data. In Figure 4, similar plots to those in Figure 2, but for Christina River, are displayed.

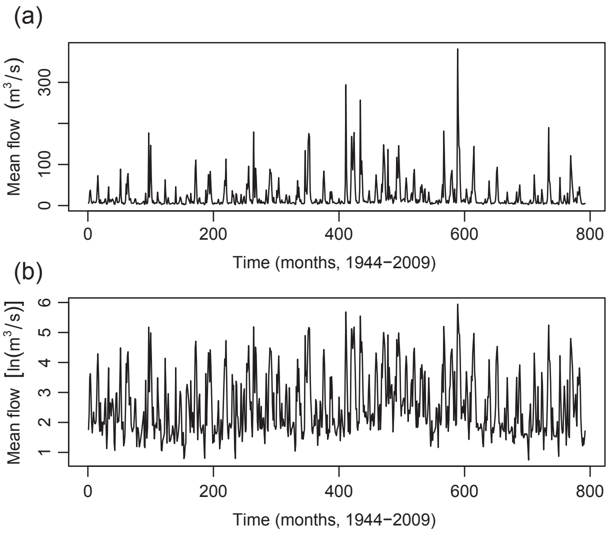

Figure 2 Salt River monthly mean flow data. (a) Original data (measured in m^3/s). (b) Natural logarithm of original data.

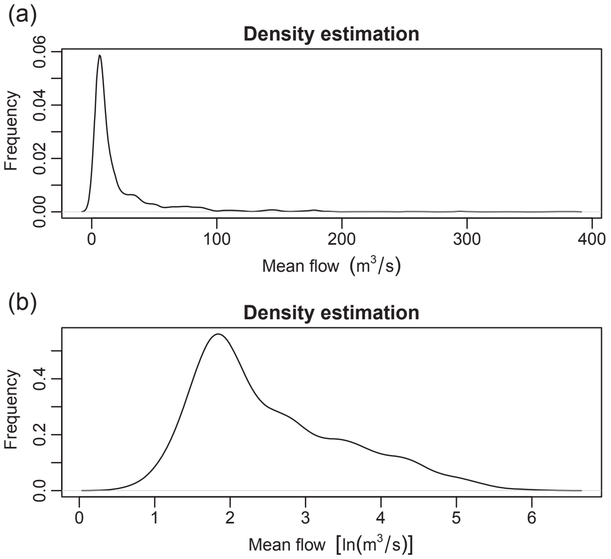

Figure 3 Nonparametric density estimates of Salt River flow data. (a) Original data. (b) Natural logarithm of original data.

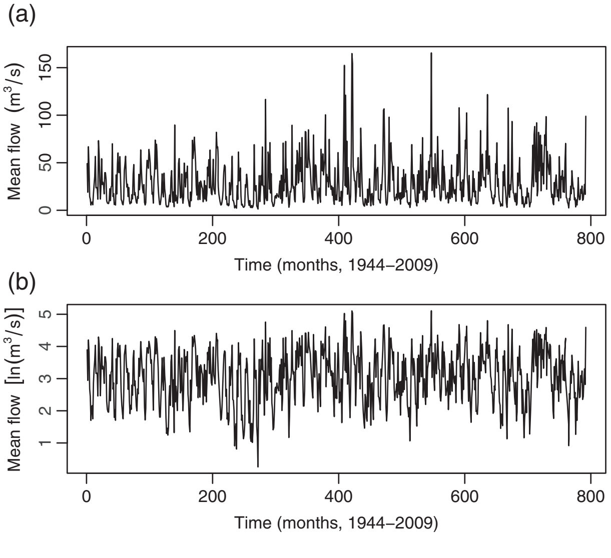

Figure 4 Christina River monthly mean flow data. (a) Original data (measured in m^3/s). (b) Natural logarithm of original data.

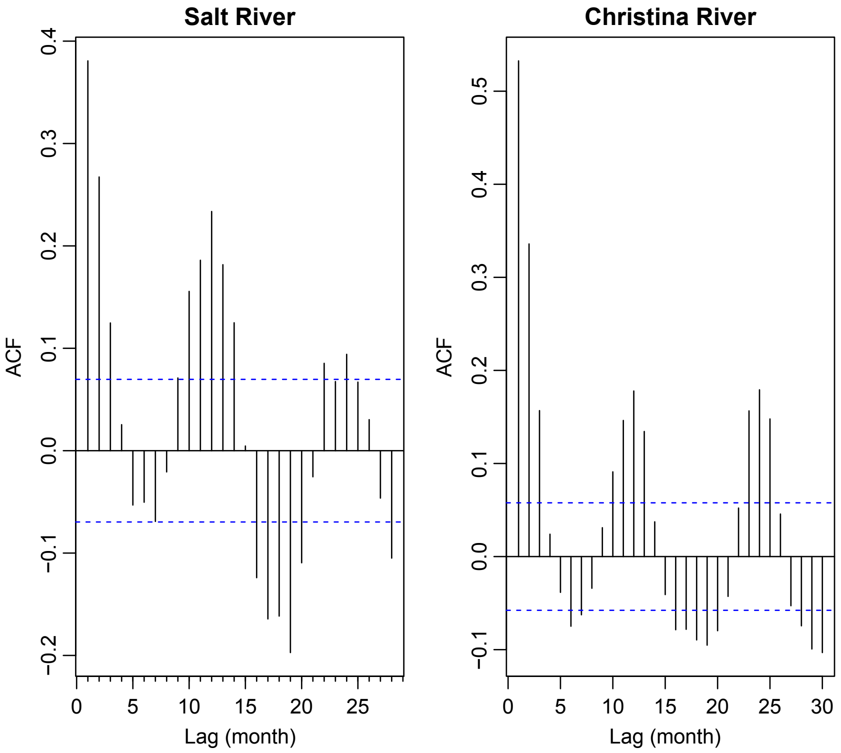

Figure 5 Autocorrelation functions (ACF) of Salt River and Christina River monthly mean flow data before the logarithmic transformation.

In Figures 5 and 6, the autocorrelation functions for the data of both rivers before and after the logarithmic transformation, respectively, are shown. The plots present different dependence structures, and suggest that the ARIMA modelling could be a possible approximation for prediction purposes.

To perform a functional analysis of the series and following *NFDA applied to time series analysis*, the original time series are converted into annual curves. In this case, there are no missing data and measures for all the months are available. Therefore, the number of annual curves is 66. To validate the performance of the approaches, the values in the 66th year (2009) are predicted using the values from the 65 previous years, and these predictions are compared with the real values in that year. To apply the nonparametric functional methods, two bandwidths have to be selected. To do this, in a first step, considering the first 64 years, the 65th is used as a validation step. As explained in *NFDA applied to time series analysis*, given the sample $\{\chi_i, Y_i(\delta)\}_{i=1}^{64}$, a predictor of $Y_{64}(\delta)$, knowing χ_{64}, can be achieved using Eqn (14). Repeating this step for $\delta = 1, \ldots, 12$, the mean values of the flow for the 65th year can be predicted. The NFDA estimators are applied using kernels based on the Epanechnikov density, $G(u) = 0.75(1 - u^2)1_{[-1,1]}(u)$, taking $K(u) = 2G(u)1_{[0,1]}(u)$ and $H(u) = \int_{-\infty}^{u} G(t)\mathrm{d}t$. On the other hand, the bandwidths h and g are selected by minimising the prediction error over the 65th year, that is $\sum_{\delta=1}^{12}\left[\hat{Y}_{64}(\delta) - Y_{64}(\delta)\right]^2$, and the Functional Principal Components Analysis semi-metric is used (for more details, see Ramsay and Silverman, 2005).

Next, in a second step, given $\{\chi_i, Y_i(\delta)\}_{i=1}^{64}$ and the previous selected parameters h and g, $\hat{F}_n(\cdot|\chi_{65})$ is estimated and a predictor of $Y_{65}(\delta)$ for $\delta = 1, \ldots, 12$, using the corresponding estimator of the median of the conditional distribution $F(\cdot|\chi_{65})$ given in Eqn (14), is obtained. Additionally, the nonparametric functional estimator of the mean function (Eqn (15)), based on regression techniques, was also applied. In this case, the monthly mean flows of the 66th year were predicted using Eqn (16) for $n = 65$. The software for computing the NFDA, programmed in R, can be freely obtained at the web http://www.math.univ-toulouse.fr/staph/npfda/.

A parametric ARIMA model is also fitted to the time series, by means of the package forecast of the software R. In this package, automatic methods to select the order of the model and also to estimate the corresponding parameters are implemented. In this case, an ARIMA(1, 0, 2) is

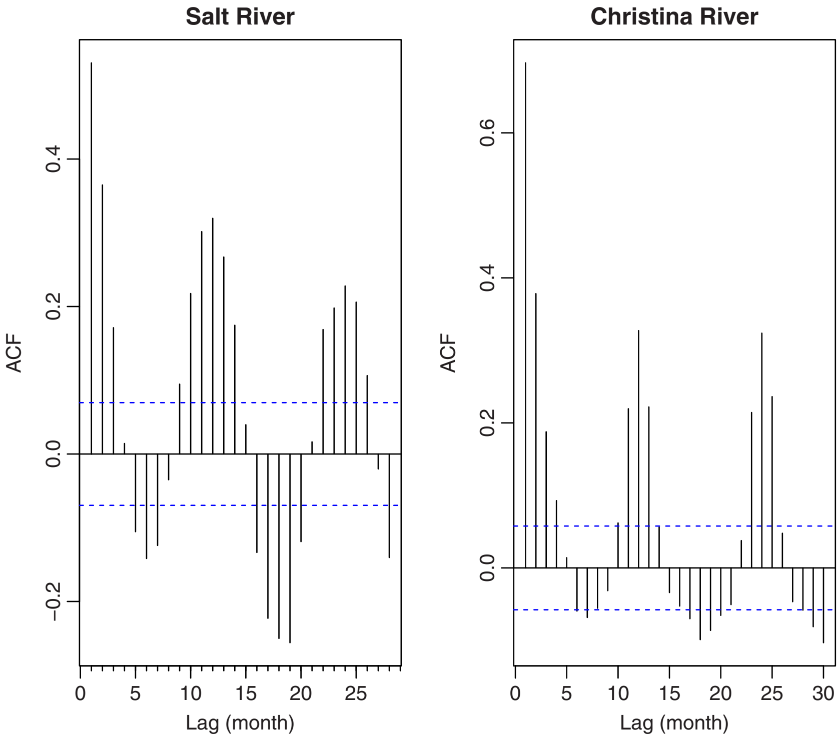

Figure 6 Autocorrelation functions (ACF) of Salt River and Christina River monthly mean flow data after the logarithmic transformation.

fitted for Salt River data and an ARIMA(4, 0, 4) for Christina River data.

Figure 7 shows, for Salt River data, the predicted values for the 66th year (dashed line) together with the real values in the 66th year (solid line). All the data considered are the natural logarithm of the real values. Figure 8 is the equivalent plot for the Christina River data set. In each case, the top panel corresponds to the functional modelling using the predictions based on regression (Eqn (16)), the middle panel shows the functional modelling using the predictions based on the median (Eqn (14)), and the bottom panel presents the ARIMA approach.

A numerical comparison for obtaining the best predictor is made using the mean squared error (MSE) criterion, that is,

$$\mathrm{MSE} = \frac{1}{12}\sum_{\delta=1}^{12}\left[\hat{Y}_{65}(\delta) - Y_{65}(\delta)\right]^2 \qquad (19)$$

The MSE values in both rivers are given in Table 2. In the first row, the results using NFDA based on regression (Eqn (16)) are presented. The results obtained applying NFDA methods based on the median (Eqn (14)) are shown in the second row. Finally, in the third row, MSE values using ARIMA models are given.

As observed in Figures 7 and 8, NFDA predictions methods provide better fits to the real series. The ARIMA predictions are, basically, the mean values. Moreover, it can be observed in Table 2 that the MSE errors are lower using the NFDA techniques, and the best criterion is that using the median as the predicted value in the two time series.

Extreme value analysis

In this section, NFDA techniques are applied for extreme value analysis. Equations described in *NFDA applied to extreme value analysis* are used, and the results obtained are compared with those using the parametric GEV and nonparametric estimators presented in *The GEV distribution* and *Nonparametric estimators*, respectively. In this case, only Salt River data are available. The maximum daily flow data of this river, from 1 January 1987 to 31 December 2009, are used to calculate flow quantile estimates as indicated in Eqn (4). In Figure 9, a boxplot computed with these data is presented. It can be observed the very asymmetric and heavy-tailed data distribution, with a lot of extreme values corresponding to high quantiles of the variable. Similar information can be deduced from Table 3, where the most usual descriptive statistics for the maximum daily flow variable are shown.

The considered values from years 1987 to 2008 (inclusive) are used in the estimation process, and the corresponding estimates are checked with the real values in the year 2009.

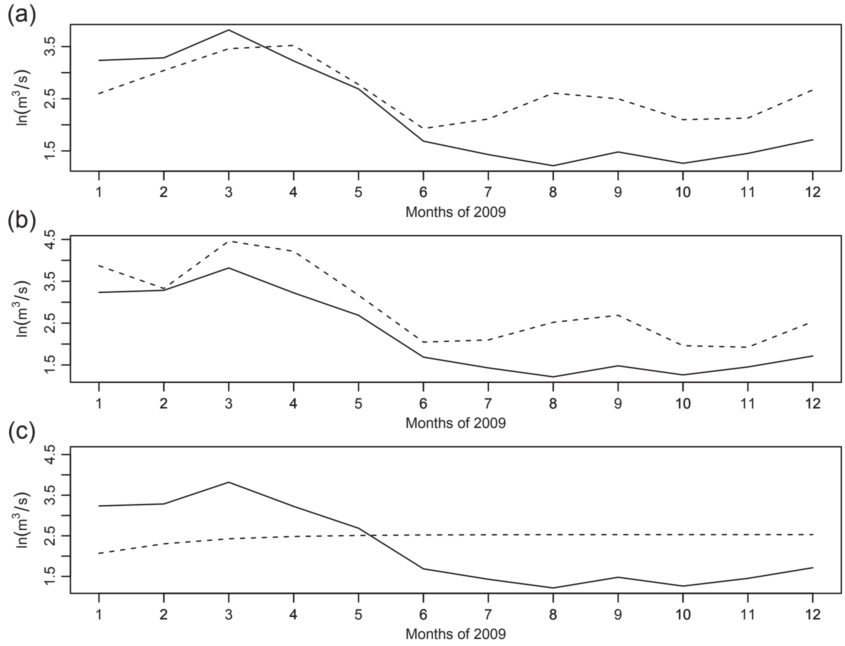

Figure 7 Predicted values for Salt River monthly mean flows in 2009 (dashed line) and real values in that year (solid line). (a) Nonparametric functional data analysis (NFDA) modelling based on regression. (b) NFDA modelling based on the median. (c) Autoregressive integrated moving-average approach.

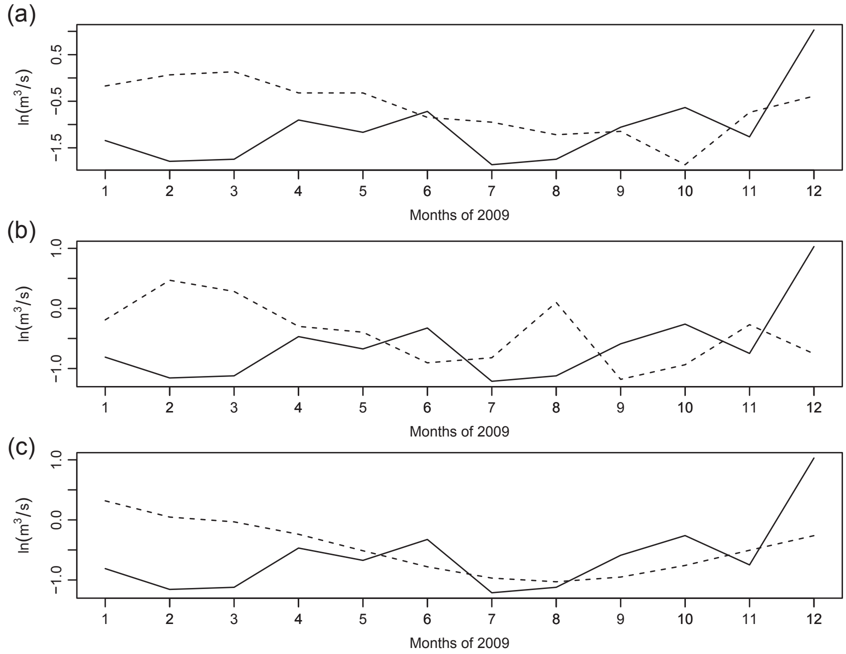

Figure 8 Predicted values for Christina River monthly mean flows in 2009 (dashed line) and real values in that year (solid line). (a) Nonparametric functional data analysis (NFDA) modelling based on regression. (b) NFDA modelling based on the median. (c) Autoregressive integrated moving-average approach.

In the classical parametric GEV estimation (*The GEV distribution*), the data need to be independent, or, at least, the dependence has to decrease suitably fast with increasing time separation (Smith, 1989). However, nonparametric estimators (both of functional and non-functional type) can be correctly applied in this field and have good

Table 2 MSEs of the monthly mean flow predictors in the 66th year (2009) using different methods (NFDA based on regression, in the first row; NFDA based on the median, in the second row; and ARIMA models in the third row), for Salt River and Christina River

	River	
Method	Salt River	Christina River
NFDA based on regression	0.5965	0.7388
NFDA based on the median	0.5208	0.4969
ARIMA models	1.0818	0.9430

MSE, mean squared error; NFDA, nonparametric functional data analysis; ARIMA, autoregressive integrated moving average.

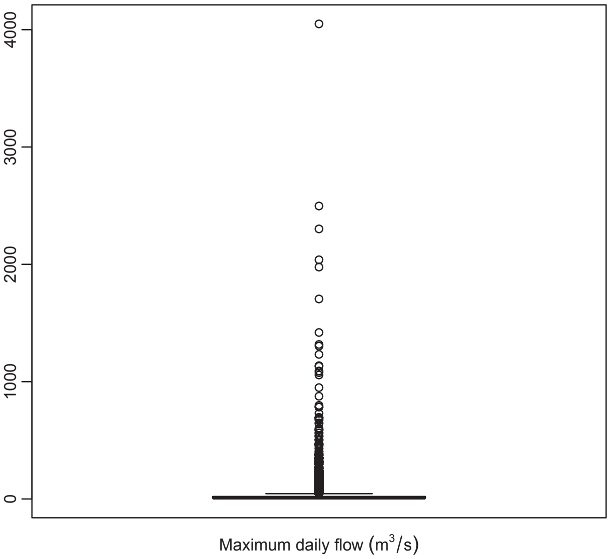

Figure 9 Boxplot of Salt River maximum daily flow data, measured in m³/s.

Table 3 Descriptive statistics for Salt River maximum daily flow variable

Minimum	1.95
1st quartile	5.55
Median	8.26
Mean	27.25
3rd quartile	21.42
Maximum	4050.00
Standard deviation	94.51
Skewness	19.63
Kurtosis	592.33

theoretical properties, although the assumption of independence is not strictly fulfilled (Youndjé and Vieu, 2006; Quintela-Del-Río, 2008).

The first step to apply NFDA (now, to calculate flow quantiles) is to construct functional data from a sample of daily maxima, in the same way as in *Monthly mean flow prediction*. As daily data are available, functional data composed of the corresponding values of each month are constructed. Unlike the situation in *Monthly mean flow prediction*, now the number of components changes from one functional variable χ to another (unbalanced data setting). This happens because the months do not have the same number of days. All months are considered to have 31 measures, interpolating linearly the two closest values for each value that originally does not exist. Therefore, each functional observation consists of 31 data.

Here, the construction of the functional data is analogous to Eqn (13):

$$\forall t \in \{1,2,\ldots,31\}, \qquad \chi_i(t) = Z_{31\cdot(i-1)+t} \tag{20}$$

where $\{Z_k\}_{k=1}^{n}$ denotes the complete time series of daily maxima, and $\chi_i = (\chi_i(1), \ldots, \chi_i(31))$ the daily data of the ith month. Now, our focus is on the estimation of the conditional distribution function of the variable of each daily maximum, conditioned on the values in the previous month.

The comparison between the classical parametric GEV methods, the nonparametric techniques and the NFDA approaches is carried out in the following steps: a set of values for levels c_i from $i = 1, \ldots, 20$ is selected. Specifically, c_1 is chosen as the median of the data (up to the year 2008), and c_{20} as the quantile of order 0.95. The sequence c_i consists of 20 equally spaced points. Using the true measures of the last year 2009, the number of days in which the values c_i were exceeded can be computed. Thus, the recurrence intervals, using the corresponding empirical distribution function, $F_n(c_i) = \frac{\text{number of measures } \le c_i}{365}$, in Eqn (5), can be approximated. These estimators are:

$$\hat{R}T(c_i) = \frac{1}{1 - F_n(c_i)}, \quad i = 1, \ldots, 20 \tag{21}$$

Now, any estimation method of the flow quantiles (Eqn (4)), using the values in Eqn (21), should provide an approximated value of the true values c_i. The flow quantiles are estimated using the classical parametric methods, the nonparametric approaches and also by means of our approximation based on NFDA methods, described below.

Parametric GEV approach

For each $i = 1, \ldots, 20$, the flow quantiles are estimated by

$$\hat{c}_{i\theta} = F_{\hat{\theta}}^{-1}\left(1 - \frac{1}{\hat{R}T(c_i)}\right) \tag{22}$$

For this, the package evir of the software R, that estimates the GEV parameters by maximum likelihood, is used.

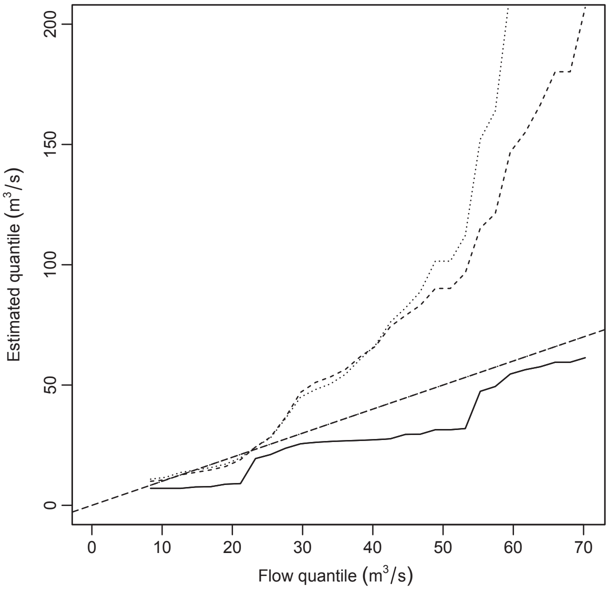

Figure 10 Estimations of the quantiles using the parametric generalised extreme value (GEV) estimator, nonparametric kernel method, and the nonparametric functional data analysis approach for Salt River data (parametric GEV estimations with a dotted line, nonparametric with a dashed line, and nonparametric functional estimations with a solid line. The dashed diagonal line represents the true quantiles to be estimated.

Nonparametric approach

For each $i = 1, \ldots, 20$, nonparametric estimators of the flow quantiles are calculated, as indicated in Eqn (9): $\hat{c}_{ih} = F_h^{-1}\left(1 - \frac{1}{\hat{R}T(c_i)}\right)$, where the bandwidth, obtained by cross-validation, is $h = 13.66$.

NFDA approach

Equation (14) can be adapted to estimate any quantile. Then, for each $i = 1, \ldots, 20$, estimators of the flow quantiles are obtained by estimating the conditional quantile of order $1/\hat{R}T(c_i)$ by the expression:

$$\hat{Y}_{(n/31)-1}(\delta) = \hat{F}_n^{-1}\left(1 - \frac{1}{\hat{R}T(c_i)} | \chi_{(n/31)-1}\right), \quad \delta = 1, 2, \ldots, 31 \tag{23}$$

where, for each δ, $Y_j(\delta) = \chi_{j+1}(\delta)$ and $\chi_{(n/31)-1}$ denotes the functional data composed of the 31 measures of the penultimate month. Then, for each day δ an estimated value is available, and the functional nonparametric estimate of the flow quantile, denoted by $\hat{c}_{iF}$, will be the sample mean of these daily values $\hat{Y}_{(n/31)-1}(\delta)$. The same kernels, bandwidths, and semi-metric as in the example in *Monthly mean flow prediction* are used.

To compare mathematically the three approaches, the relative mean absolute error (RMAE) of $\hat{c}_{i\theta}$, $\hat{c}_{ih}$, and $\hat{c}_{iF}$ is computed, given by:

$$\text{RMAE} = \frac{1}{20}\sum_{i=1}^{20} \frac{|c_i - \hat{c}_{i*}|}{c_i} \tag{24}$$

where $\hat{c}_{i*}$ can be $\hat{c}_{i\theta}$, $\hat{c}_{ih}$, or $\hat{c}_{iF}$. The results obtained are

$$\text{RMAE} = 1.13, 0.71, \text{ and } 0.26 \tag{25}$$

for the parametric GEV, nonparametric, and NFDA estimates, respectively. Therefore, the minimum error is obtained with the NFDA techniques. On the other hand, Figure 10 shows the quantile estimations with the previous proposals for Salt River data (parametric GEV estimations with a dotted line, nonparametric with a dashed line, and NFDA estimations with a solid line. The dashed diagonal line represents the true values to be estimated). The long-tailed distribution observed in Figure 9 clearly reveals the difficulty in the extreme value estimation process (Serinaldi, 2009). However, the NFDA approach, considering each functional datum as the complete set of values for each month, gives more precise estimations than those obtained with the parametric GEV or the simple nonparametric methods. The largest differences between the estimates occur at the highest levels, where the good approximations of the NFDA estimates are observed and it is more important to have reliable prediction techniques. Note that a multivariate approach would be possible in the parametric GEV and the nonparametric settings, but, in this case, a vector composed of 30 predictor variables would be necessary. This high value makes very difficult (if not impossible) this kind of approximation in practice.

Discussion

Statistical techniques are usually applied to address practical problems in hydrology. In this article, two of them, monthly mean river flow prediction and extreme value analysis, are the focus of the research. NFDA approaches, combining nonparametric methods with functional data, are used in this setting.

The main objective of this article was to apply different NFDA techniques to two particular hydrological problems, and to test their behaviour in comparison to more classical approaches. The nonparametric functional methods were applied to real data of two rivers in the United States. The different alternatives were validated using the final year in the database as a testing sample, and the rest of the years as the training sample.

In the prediction setting, two nonparametric functional proposals, based on the median and the mean, respectively, were applied and compared with classical ARIMA models.

The results showed that NFDA approaches, especially those based on the median, had a better performance than the classical ARIMA models.

The previous approaches could be extended including available information like daily precipitation, daily temperature, or any other climatic covariate. Several models similar to those presented incorporating covariates have been proposed and studied previously. For example, the dynamic regression (ARIMAX) combines the Box-Jenkins models with the linear regression, obtaining a more general model for the study of the time series (Shumway and Stoffer, 2011). This kind of models simply adds covariates to the general expression of an ARIMA model, but the covariate coefficients are hard to interpret.

An alternative approach could be the application of regression models with ARMA (or ARIMA) errors. This includes the use of parametric, nonparametric, or semiparametric approaches. In a hydrological context, Castellano-Méndez *et al.* (2004) presents a study of the Xallas River (northwest of Spain), using Box-Jenkins and neural networks methods, incorporating exogenous variables such as rainfall information. Regarding the case of functional methods, covariates could be included in the problem through the use of semi-functional partial linear models (Aneiros-Perez and Vieu, 2006). This approach uses a nonparametric kernel procedure; the output is scalar, and a functional covariate and multivariate non-functional covariate are considered. Functional regression between functional explanatory variables and a scalar response is also possible using the backfitting algorithm (Febrero-Bande and González-Manteiga, 2011). This would allow including functional covariates in the model. The application of these techniques to our data would require the availability of some relevant climatic variables. Unfortunately, these variables are not available in the managed databases. However, a more deep study of this issue could be carried out in a future research.

Regarding the extreme value analysis, the estimation of the flow quantiles has been the focus of the study. These values play an important role in hydrological problems, because they are directly linked with flood analysis. The new NFDA approach performed better than the parametric GEV estimators, producing more close estimations to the true values. On the other hand, it is well known that due to the small number of extreme values in a sample, it is usually difficult to obtain reliable estimations. These estimations could be improved by using more precise bandwidth parameters. The bandwidth parameter selection in NFDA remains, nowadays, as an open problem. The development of data-driven techniques for computing optimal bandwidths will produce directly the improvement of the promising results obtained in the quantile estimation problem.

In general, the approaches proposed in this article yielded accurate estimates of both the functions of interest, such as the cumulative distribution function or the function providing the probabilities of exceedance, and derived parameters, as, for example, the flow quantiles. They also captured more complex patterns in the data providing better future estimations. Therefore, they represent a better alternative to the classical methods regularly used in this framework, being useful tools for environmental agencies to manage hydrological risks including those of floods.

Acknowledgements

The authors have no conflict of interest to declare. This research has been partially supported by the Spanish Ministry of Science and Innovation Grants MTM2011-22392 and MTM2014-52876-R for M.F.-F. We are grateful to three referees and to the editors of the Journal of Flood Risk Management, for constructive and helpful suggestions.

References

Adamowski K. A Monte Carlo comparison of parametric and nonparametric estimation of flood frequencies. *J Hydrol* 1989, **108**, 295–308.

Adamowski K. & Feluch W. Nonparametric flood frequency analysis with historical information. *J Hydraul Eng* 1990, **116**, 1035–1047.

Anderson P.L. & Meerschaert M.M. Modeling river flows with heavy tails. *Water Resour Res* 1998, **34**, 2271–2280.

Aneiros-Perez G. & Vieu P. Semi-functional partial linear regression. *Stat Prob Lett* 2006, **76**, 1102–1110.

Apipattanavis S., Rajagopalan B. & Lall U. Local polynomial–based flood frequency estimator for mixed population. *J Hydraul Eng* 2010, **15**, 680–691.

Besse P., Cardot H. & Stephenson D. Autoregressive forecasting of some functional climatic variations. *Scand J Stat* 2000, **27**, 673–687.

Bowman A., Hall P. & Prvan T. Bandwidth selection for the smoothing of distribution functions. *Biometrika* 1998, **85**, 799–808.

Brockwell P. & Davis R. *Time series: theory and methods*. New York: Springer Verlag, 1991.

Castellano Méndez M., Franco A., Cartelle D., Febrero-Bande M. & Roca E. Identification of NO_x and ozone episodes and estimation of ozone by statistical analysis. *Water Air Soil Pollut* 2009, **198**, 95–110.

Castellano-Méndez M., González-Manteiga W., Febrero-Bande M., Prada-Sánchez J.M. & Lozano-Calderón R. Modelling of the monthly and daily behaviour of the runoff of the Xallas River using Box–Jenkins and neural networks methods. *J Hydrol* 2004, **296**, 38–58.

Celebioglu T. Simulation of hydrodynamics and sediment transport patterns in Delaware Bay. PhD Thesis, Drexel University, 2006.

Coles S. *An introduction to statistical modeling of extreme values*. London: Springer Verlag, 2001.

Fan J. & Gijbels I. *Local polynomial modelling and its applications*. London: Chapman & Hall, 1996.

Febrero-Bande M. & González-Manteiga W. Generalized additive models for functional data. In: F. Ferraty, ed. *Recent advances in functional data analysis and related topics*. Physica-Verlag, Berlin 2011, 91–96.

Fernández De Castro B., Guillas S. & González-Manteiga W. Functional samples and bootstrap for predicting sulfur dioxide levels. *Technometrics* 2005, **47**, 212–222.

Ferraty F. & Vieu P. *Nonparametric functional data analysis*. New York: Springer-Verlag, 2006.

Ferraty F., Rabhi A. & Vieu P. Conditional quantiles for dependent functional data with application to the climate El Niño Phenomenon. *Sankhya* 2005, **67**, 378–398.

Fisher R.A. & Tippett L.H.C. Limiting forms of the frequency distributions of the largest or smallest member of a sample. *Proc Cambridge Philos Soc* 1928, **24**, 180–190.

Gardes L., Girard S. & Lekina A. Functional nonparametric estimation of conditional extreme quantiles. *J Multivar Anal* 2010, **101**, 419–433.

Gasser T. & Müller H.G. Estimating regression functions and their derivatives by the kernel method. *Scand J Stat* 1984, **11**, 171–185.

Guo S.L., Kachroo R.K. & Mngodo R.J. Nonparametric kernel estimation of low flow quantiles. *J Hydrol* 1996, **185**, 335–348.

Hall P., Rodney C.L. & Yao Q. Methods for estimating a conditional distribution function. *J Am Stat Assoc* 1999, **94**, 154–163.

Hall P., Poskitt D.S. & Presnell B. A functional data-analytic approach to signal discrimination. *Technometrics* 2001, **43**, 1–9.

Katz R.W., Parlange M.B. & Naveau P. Statistics of extremes in hydrology. *Adv Water Resour* 2002, **25**, 1287–1304.

Keskin M.E., Taylan D. & Terzi O. Adaptive neural-based fuzzy inference system (ANFIS) approach for modelling hydrological time series. *Hydrol Sci J* 2006, **51**, 588–598.

Kim K. & Heo J. Comparative study of flood quantiles estimation by nonparametric models. *J Hydrol* 2002, **260**, 176–193.

Koutsoyiannis D. A generalized mathematical framework for stochastic simulation and forecast of hydrologic time series. *Water Resour Res* 2000, **36**, 1519–1533.

Lall U., Moon Y.I. & Bosworth K. Kernel flood frequency estimators: bandwidth selection and kernel choice. *Water Resour Res* 1993, **29**, 1003–1015.

Montanari A., Rosso R. & Taqqu M.S. Fractionally differenced ARIMA models applied to hydrologic time series: identification, estimation, and simulation. *Water Resour Res* 1997, **33**, 1035–1044.

Moon Y. & Lall U. Kernel quantile function estimator for flood frequency analysis. *Water Resour Res* 1994, **30**, 3095–3103.

Parzen E. On estimation of a probability density function and mode. *Ann Math Stat* 1962, **32**, 1065–1076.

Quintela-Del-Río A. Hazard function given a functional variable: nonparametric estimation under strong mixing conditions. *J Nonparametr Stat* 2008, **20**, 413–430.

Quintela-Del-Río A. On bandwidth selection for nonparametric estimation in flood frequency analysis. *Hydrol Process* 2011, **25**, 671–678.

R Development Core Team *R: a language and environment for statistical computing*. Vienna: 2016. http://www.R-project.org.

Ramsay J.O. & Silverman B.W. *Functional data analysis*. New York: Springer-Verlag, 2005.

Ruppert D., Wand M.P. & Carroll R.J. *Semiparametric regression*. Cambridge: Cambridge University Press, 2003.

Saf B. Regional flood frequency analysis using L-moments for the West Mediterranean Region of Turkey. *Water Resour Manage* 2009, **23**, 531–551.

Senior L.A., Koerkle E.H. Simulation of streamflow and water quality in the Christina River subbasin and overview of simulations in other subbasins of the Christina River basin, Pennsylvania, Maryland, and Delaware, 1994-98. Number 03-4193 in Water-resources investigations report. US Department of the Interior, US Geological Survey: New Cumberland, Pennsylvania, 2003.

Serinaldi F. Assessing the applicability of fractional order statistics for computing confidence intervals for extreme quantiles. *J Hydrol* 2009, **376**, 528–541.

Sharma A., Tarboton D.G. & Lall U. Streamflow simulation: a nonparametric approach. *Water Resour Res* 1997, **33**, 291–308.

Shumway R.H. & Stoffer D.S. *Time series analysis and its applications*. New York: Springer, 2011.

Singh V.P., Wang S.X. & Zhang L. Frequency analysis of nonidentically distributed hydrologic flood data. *J Hydrol* 2005, **307**, 175–195.

Smith R. Extreme value analysis of environmental time series: an application to trend detection in ground-level ozone. *Stat Sci* 1989, **4**, 367–393.

Tamea S., Laio F. & Ridolfi L. Probabilistic nonlinear prediction of river flows. *Water Resour Res* 2005, **41**, W09421.

Toth E., Brath A. & Montanari A. Comparison of short-term rainfall prediction models for real-time flood forecasting. *J Hydrol* 2000, **239**, 132–147.

Wand M.P. & Jones M.C. *Kernel smoothing*. London: Chapman & Hall, 1995.

Wang W.C., Chau K.W., Cheng C.T. & Qiu L. A comparison of performance of several artificial intelligence methods for forecasting monthly discharge time series. *J Hydrol* 2009, **374**, 294–306.

Wasserman L. *All of nonparametric statistics*. New York: Springer-Verlag, 2005.

Wu C.L., Chau K.W. & Li Y.S. Predicting monthly stream flow using data-driven models coupled with data preprocessing techniques. *Water Resour Res* 2009, **45**, W08432.

Youndj´e E. & Vieu P. A note on quantile estimation for long dependent stochastic processes. *Stat Prob Lett* 2006, **76**, 109–116.