

Streamflow forecasting using functional regression

Pierre Masselot^{a,*}, Sophie Dabo-Niang^b, Fateh Chebana^a, Taha B.M.J. Ouarda^{a,c}

^a*Centre Eau-Terre-Environnement (ETE), Institut national de la recherche scientifique (INRS), 490 de la Couronne, Québec, QC, G1K 9A9, Canada*

^b*Université Charles-de-Gaulle Lille 3, Domaine du pont de bois, Laboratoire EQUIPPE, maison de la recherche, BP 149, 59653 Villeneuve d'Ascq cedex, France*

^c*Institute center for water and environment (iWATER), Masdar institute of science and technology, PO Box 54224, Abu Dhabi, United Arab Emirates*

Abstract

Streamflow, as a natural phenomenon, is continuous in time and so are the meteorological variables which influence its variability. In practice, it can be of interest to forecast the whole flow curve instead of points (daily or hourly). To this end, this paper introduces the functional linear models and adapts it to hydrological forecasting. More precisely, functional linear models are regression models based on curves instead of single values. They allow to consider the whole process instead of a limited number of time points or features. We apply these models to analyse the flow volume and the whole streamflow curve during a given period by using precipitations curves. The functional model is shown to lead to encouraging results. The potential of functional linear models to detect special features that would have been hard to see otherwise is pointed out. The functional model is also compared to the artificial neural network approach and the advantages and disadvantages of both models are discussed. Finally, future research directions involving the functional model in hydrology are presented.

Keywords: Functional data, Streamflow hydrograph, Functional linear models, Regression

1. Introduction

Streamflow forecasting is an important topic in hydrology. Being able to precisely forecast streamflows is crucial for the adequate management of water resources systems. Therefore, there is a wide body of literature concerning streamflow forecasting, developing and applying a wide range of forecasting methods. For such a purpose, two types of models can be identified (Fortin et al., 1997): a) physical models which apply deterministic equations to a set of input variables (such as physiographic features or rainfall) to obtain the desired streamflow values, and b) statistical models which model streamflows in a probabilistic way and which take into account the uncertainty in observed data. The latter are often cheaper to perform. The focus of this paper is on statistical models.

A large number of statistical models have been proposed for streamflow forecasting. Two major classes of such models can be distinguished: time series and regression models. The former is mainly based on the modelling of the streamflow autocorrelation structure while the latter focuses on the correlation between a response (streamflow)

*Corresponding author

Email address: pierre-lucas.masselot@ete.inrs.ca (Pierre Masselot)

and explanatory variables (or covariates) regardless of the time structure. Hence, the forecasts of such explanatory variables are used to forecast future streamflows. Typical explanatory variables are precipitations and temperatures (Sene, 2009). Some models are typically used to long term forecast such as linear regression (*e.g.* Vogel et al., 1999), principal component regression (*e.g.* Garen, 1992; Eldaw et al., 2003), partial least square regression (*e.g.* Tootle et al., 2007) and wavelet regression (*e.g.* Kisi, 2009). However, these models only forecast the streamflow volume at large time scales (usually seasonal or annual). This does not provide information about the shape of the hydrograph within these long periods, lacking accuracy about the moment when flows will be peaking.

There are also a number of models used to forecast short-term streamflows. The main models are artificial neural networks (*e.g.* Zealand et al., 1999; Kisi, 2007; Chokmani et al., 2008; Makkeasorn et al., 2008), wavelet regression (*e.g.* Kisi, 2009; Sahay & Sehgal, 2013) and support vector regression (*e.g.* Yu et al., 2006). However, these models can only model and predict few days (or few data points) at a time which is useful for detecting future extrema but not for forecasting overall trends in the streamflow process, such as the whole spring flood hydrograph for instance (such as illustrated in Figure 1).

In order to combine both short-term and long-term forecasting benefits, it is proposed to forecast streamflows using the whole streamflow process instead of single characteristics. This approach consists in considering the hydrograph as a continuous curve, *i.e.* a *functional data*, as proposed by Chebana et al. (2012). The statistical framework dealing with functional data is called *functional data analysis* (FDA). It was introduced by Ramsay (1982) and widely popularized by Ramsay & Silverman (2005). Such an approach became possible because a number of natural phenomena, such as streamflows, are now measured at a very fine scale or on a real time basis. This paradigm shift caused by FDA implies that the basic datum is now a curve instead of a scalar value.

FDA has recently been introduced in hydrology by Chebana et al. (2012) including the description and analysis framework of functional hydrographs and Ternynck et al. (2016) used FDA to classify hydrographs. The present paper aims to introduce functional linear models (FLM) to hydrograph forecasting. A complete overview of FLM models is available in Ramsay & Silverman (2005) as well as Ferraty & Vieu (2006). Among these models, two are particularly applicable for streamflow forecasting purposes: i) the scalar response FLM (FLM-S) expressing a traditional scalar variable according to functional variables and ii) the fully FLM (FLM-F) which aims at explaining a functional variable according to other functional variables. Table 1 provides an overview of existing FLM models with references.

The FLM-S, which models a scalar response according to functional predictors, is perhaps the most studied model among FLM. It originated from Hastie & Mallows (1993) and has known several theoretical developments since, *e.g.*, Cardot et al. (1999, 2003b). The FLM-S is interesting because of its potential applications for long-term flood forecasting. This model is indeed well suited for modelling a hydrologic feature (such as a flood duration or volume) based on the complete history of the predictors (*e.g.* temperatures or precipitations). This allows the use of more information than, for instance, classical linear regression. When using non functional predictors, a classical

model with such fine information would be difficult to calibrate because it requires a number of lagged variables, resulting in a large number of coefficients to estimate, as well as a high collinearity between predictors. Thus, the uncertainty associated to such a classical model would be dramatically high. Since instead of simple scalar values the covariates of the FLM-S are curves, there is no need for many lagged variables and collinearity is no longer a problem (Cuevas et al., 2002).

The other model used in the present paper is the FLM-F introduced by Ramsay & Dalzell (1991) and studied theoretically by, *e.g.*, Yao et al. (2005). An advantage of this model is its natural applicability to time series. Indeed, time series often contain features such as autocorrelation, trend or seasonality, creating spurious relationship and inducing wrong conclusions in classical linear models (*e.g.* Granger & Newbold, 1974; Hoover, 2003). Even though several statistical methods have been developed to overcome this major drawback (*e.g.* Phillips, 1987), linear models applied to time series always need a careful exploration of the considered time series. The FLM-F does not suffer from this drawback since a long sequence of data points can be considered as a single curve (such as a whole hydrograph). Moreover, its results are easier to interpret than many of the classical models used for day to day forecasting (*e.g.*, ANN or support vector regression). The FLM-F model provides a way to forecast the overall trend of the hydrograph without a complex parametrization (illustrated by Figure 1). Finally, FLM-F allows changing the relationship between the predictor and the response over time, while classical models implicitly assume that the relationship does not change over time. Although it could be possible to fit a regression model without this assumption, this would result in a complicated model while it is a natural characteristic of FLM-F.

Globally, FDA is a topic receiving increasing attention in the theoretical and applied statistics literature. A number of reference textbooks dealing with functional data analysis already exist, such as Bosq (2000), Ramsay & Silverman (2005), Bosq & Blanke (2008), Dabo-Niang & Ferraty (2008), Ramsay et al. (2011) and Horváth & Kokoszka (2012). A large number of fields have already seen successful FDA applications, such as image processing (Cardot et al., 2003a), medicine (Ratcliffe et al., 2002a,b), genetics (Müller et al., 2008), ecology (Bel et al., 2011), marketing (Sood et al., 2009), economy (Goia et al., 2010) and transportation systems (Chiou, 2012). All these applications show the increasing interest in applying FLMs for practical purposes.

It is important to note that the expected improvements from using the functional framework may be mostly in terms of interpretation and statistical justification rather than raw performances. This includes more informations concerning the phenomena and conceptually more relevant results. We insist that functional regression methods are relevant since they are intuitively well suited and theoretically justified for time-related data. Indeed, these models represent a simple and clever way to exhibit many features of a phenomenon or a relationship between several variables.

The paper is organized as follows. After recalling the basics of data smoothing, functional linear models are introduced in section 2. Then, a case study on streamflow forecasting is presented in section 3. Section 4 concludes the study.

2. Functional linear models

This section introduces FLMs starting with the necessary step of data smoothing. Then, the FLM-S and FLM-F are introduced and the procedure for their fitting by using empirical data is presented.

2.1. Data smoothing

The main characteristic of a function $x(\cdot)$ is its infinite dimensionality, meaning that there exists a value of $x(t)$ for each possible value of $t \in \mathbb{R}$ (Ramsay & Silverman, 2005; Ferraty & Vieu, 2006). Discrete measurements are thus not sufficient to represent such a datum, since they do not provide values between $x(t_j)$ and $x(t_{j+1})$. The usual way to represent an *a priori* unknown function, is to express it as a sum of analytically known basis functions. This is actually the same thing as data smoothing, as it can be seen in other frameworks such as generalized additive models (Hastie & Tibshirani, 1986; Chebana et al., 2014). However, in FDA it is a preliminary step aiming to prepare data to be functions while it is an aim itself in classical modelling. To obtain a set of n functions $x_i(\cdot), i = 1, \dots, n$ with the same distribution (such as a collection of n annual streamflow curves), the same basis functions must be employed for each curve. Hence, a functional datum $x_i(\cdot)$ can be expressed as:

$$x_i(t) = \sum_{k=1}^{K_x} c_{ik} \phi_k(t) ; t \in \Omega \quad (1)$$

where $\phi_1(\cdot), \dots, \phi_{K_x}(\cdot)$ is a set of known basis functions. In practice the basis functions used are either Fourier basis when data are periodic, or B-spline when data are not periodic. However, other basis functions are possible such as wavelets which perform well when there are more local features. The relevance of the basis expansion approach lies in the fact that the problem of estimating a function is reduced to the estimation of a set of scalar coefficients c_{ik} . These coefficients are estimated by minimizing the penalized sum of square criterion:

$$PSSE = \sum_{j=1}^T (z_j - x_i(t_j))^2 + \lambda \int D^2 x_i(t) dt \quad (2)$$

where $D^2 x(t)$ represents the second derivative of $x(t)$, $z_{ij}, j = 1, \dots, T$ are the measurements used to fit the function $x_i(\cdot)$ and the t_{js} are the corresponding measurement times. For instance, z_{ij} could be the measured streamflow of a river on day j of the year i and $T = 365$, meaning that $x_i(t)$ seeks to represent the flow over the i^{th} year. Since observed data are usually noisy, in practice, the penalty term $\lambda \int D^2 x(t) dt$ is added to the sum of square criterion (2) in order to penalize the rough $x_i(t)$. The parameter λ controls the severity of the penalization, which means that the larger λ is, the lesser the resulting number of basis K will be. λ is estimated by minimizing a cross-validation (CV) score which is an estimation of the prediction error, *i.e.* the error we make when trying to predict new data (Stone, 1974). These approaches are extensively explained and discussed in chapters 3 to 5 of Ramsay & Silverman (2005).

Finally, note that the formulation of curves as in (1) allows the user to adapt the smoothing to catch special features of the curve. For instance, when special features of the curves need to be correctly estimated (such as peaks in frequency analysis), the use of B-spline basis functions with more breakpoints at locations where these features happen more often can prove to be useful. At these locations, the curve can thus adapt more easily to the targeted features.

2.2. Functional linear models

Now that the definition and construction of a functional data has been presented, this section introduces functional linear models, *i.e.* how to model either a scalar or a functional response using functional predictors.

2.2.1. Functional linear models for scalar response

The functional linear model for scalar response seeks to explore the influence of a set of curves $X^{(j)}(t)$, ($j = 1, \dots, q$), at each time $t \in \Omega$ on a scalar response Y . To give an example, Y can be the flow volume during a given period and the FLM-S gives the evolution of the influence of variables such as precipitations on the final flow volume. For simplicity and clarity purposes, the model with a single covariate $X(t)$ is presented. Nevertheless, we indicate at the end of the section how to generalize the model with several covariates.

The functional linear model for scalar response is, in the case of a single covariate, given by:

$$y_i = \alpha + \int_{\Omega} \beta(t)x_i(t) dt + \epsilon_i \quad i = 1, \dots, n \quad (3)$$

where $(x_i(\cdot), y_i)$ are observed data from $(X(\cdot), Y)$, α is the intercept term, $\beta(\cdot)$ is the coefficient function, ϵ_i is the error term for observation i and n is the number of observations. Here the traditional regression coefficient is replaced by the coefficient function $\beta(t)$ which gives the influence of $x_i(t)$ on y_i at each time t . For instance, in the ninth chapter of Ramsay et al. (2009), the model (3) is used to fit the logarithm of total annual precipitation (y_i) according to temperature curves ($x_i(\cdot)$) over a year (meaning that $\Omega = [0; 365]$). This model can also be seen as a generalization of any linear model for which the curve $x_i(\cdot)$ represents an infinity of scalar covariates.

Since the coefficient $\beta(\cdot)$ is a function, it is infinite dimensional. The functional model in (3) has thus an infinity of degrees of freedom, which means that there is an infinity of solutions. Therefore, in order to reduce the degrees of freedom to a finite value, the coefficient $\beta(\cdot)$ has to be expanded using a set of basis functions $\theta_k(t)$, $k = 1, \dots, K_{\beta}$, in the same manner as equation (1), *i.e.* $\beta(t) = \sum_{k=1}^{K_{\beta}} b_k \theta_k(t)$. The fitting of the FLM-S is then reduced to estimate a finite number of coefficients b_k .

Let $\phi_k(t)$, $k = 1, \dots, K_x$ be the set of basis functions used to represent the functional covariate $X(\cdot)$. An observed curve from $X(\cdot)$ is thus expressed as $x_i(t) = \sum_{k=1}^{K_x} c_{ik} \phi_k(t)$ similarly to equation (1). Note that, since the curves $x_i(\cdot)$ are observed, the coefficients c_{ik} are estimated before fitting the FLM-S.

To explain how the coefficients b_k are estimated, it is convenient to use the matrix notation of the basis expansion (1) for both $\beta(\cdot)$ and the set of observed curves $\mathbf{x}(\cdot) = (x_1(\cdot), \dots, x_n(\cdot))$. They are thus expressed as:

$$\beta(t) = \Theta'(t)B \quad \text{and} \quad \mathbf{x}(t) = C\Phi(t) \quad (4)$$

where $\Theta(t) = (\theta_1(t), \dots, \theta_{K_\beta}(t))'$ and $\Phi(t) = (\phi_1(t), \dots, \phi_{K_x}(t))'$ are the vectors containing the basis functions evaluated at time t . B is the vector containing the K_β coefficients b_k to estimate, and since $\mathbf{x}(t)$ is a set of n curves, C is a $n \times K_x$ matrix containing the coefficient c_{ik} at line i and column k .

The matrix form in (4) is convenient because it allows expressing the functional linear model in (3), in the matrix form:

$$\mathbf{y} = \alpha + \int_{\Omega} C\Phi(t)\Theta'(t)B dt + \epsilon \quad (5)$$

for the vector \mathbf{y} containing all observed scalar responses. Since C and B do not depend on t , it is possible to evaluate the $K_x \times K_\beta$ matrix $J_{\Phi\Theta} = \int_{\Omega} \Phi(t)\Theta'(t) dt$, and thus rewrite (5) as $\mathbf{y} = \alpha + CJ_{\Phi\Theta}B + \epsilon$. Finally, by setting $\chi = \begin{bmatrix} \mathbf{1} & CJ_{\Phi\Theta} \end{bmatrix}$ and $\mathbf{B} = (\alpha, B)'$, the functional linear model (3) is written as :

$$\mathbf{y} = \chi\mathbf{B} + \epsilon \quad (6)$$

which is similar to a traditional linear model with design matrix χ and coefficient vector \mathbf{B} . This model is fit by ordinary least squares.

The shape of the functional coefficient $\beta(\cdot)$, and therefore the accuracy of the response forecasts depends on the value K_β which controls the smoothness of the estimation. Thus, as for the smoothing of data curves, the smoothness of $\beta(\cdot)$ can be controlled by directly choosing a small K_β or by adding a roughness penalty to the SSE criterion. The latter allows to define a basis with a large number of functions, for which several coefficients are shrunk to zero. The criterion to be minimized is thus the penalized SSE:

$$\begin{aligned} PENSSE_\lambda(\alpha, \beta) &= \sum_{i=1}^n \left[y_i - \alpha - \int_{\Omega} \beta(t)x_i(t) dt \right]^2 + \lambda \int_{\Omega} [L\beta(t)]^2 dt \\ &= \|\mathbf{y} - \chi\mathbf{B}\|^2 + \lambda\mathbf{B}'R\mathbf{B} \end{aligned} \quad (7)$$

where L is a linear differential operator applied to $\beta(t)$ and

$$R = \begin{bmatrix} 1 & & \dots \\ \vdots & \int_{\Omega} [L\Theta(t)][L\Theta(t)]' dt & \end{bmatrix}$$

is a $(K_\beta + 1) \times (K_\beta + 1)$ penalization matrix. The first row and column of 1's are here to take into account the intercept α . The model is now similar to a Ridge regression and thus has the same solution (Hoerl & Kennard, 1970).

Traditionally, L is defined as the second derivative (or acceleration) of $\beta(\cdot)$ in order to penalize rough functions, but other differential operators are also possible, such as the harmonic acceleration (Ramsay & Silverman, 2005, p. 93). The coefficient λ controls the severity of the penalization, and thus the smoothness of $\beta(\cdot)$. In practice, λ is chosen by minimizing the CV criterion (Stone, 1974).

The FLM-S is not tied to only one predictor and is generalized for several functional predictors and optional scalar predictors. There is also the possibility of using different time intervals for the covariates if the covariates have different time lags. In hydrology, this allows to consider variables such as snow height only during the months when they are not null (November to April in Canada). This generalization and more topics are covered in a number of reference textbooks (*e.g.* Ramsay & Silverman, 2005; Horváth & Kokoszka, 2012).

2.2.2. Fully functional linear model

This section presents the fully functional linear model (FLM-F) which links the whole response curve to the whole covariate curve. This is a more general model than the “concurrent” model (also known as “point-wise FLM” or “varying coefficient model”) developed for the case when response and covariate are both functional (*e.g.* Hastie & Tibshirani, 1993; Ramsay & Silverman, 2005, chapter 14) but which links only time t of the covariate to the same time t of the response. Although an estimation method allowing the use of multiple functional covariates in the FLM-F is currently appearing (see the working paper Ivanescu et al., 2014), we present here the basic method with only one covariate of Ramsay & Silverman (2005), for simplicity purposes.

The FLM-F is expressed as:

$$y_i(t) = \alpha(t) + \int_{\Omega_1} \beta(s, t)x_i(s) ds + \epsilon_i(t) ; t \in \Omega_2 \quad (8)$$

where $\beta(s, t)$ is now a function of both s and t , which means that it is a surface. Indeed, $\beta(s, t)$ gives the influence of $x_i(\cdot)$ at time s on $y_i(\cdot)$ at time t , allowing also Ω_1 and Ω_2 to be different. This flexibility generalizes the use of lags in a linear model since it allows the significant lags to change over time. For instance, this model is able to take into account the fact that the lag between rainfall and the resulting streamflow can change over time.

Note that the FLM-F includes the time domain $s > t$. When data are time series this means that there is an effect of the predictor on the response backwards in time. Attempts have been made at developing a “historical FLM” (HFLM) taking account only of the times $s < t$, (Malfait & Ramsay, 2003; Kim et al., 2011). However these models result in less smooth $\beta(\cdot, \cdot)$ surfaces. Moreover, in practice the surface is very close to zero where $s > t$ (means no backward influence). Thus the present paper introduces the general case for clarity and simplicity purposes.

For the reason of infinite degrees of freedom exposed in section 2.2, the surface $\beta(\cdot, \cdot)$ has to be expressed using basis functions. Since $\beta(\cdot, \cdot)$ is bi-dimensional, it is expanded using a tensor product expansion (also called outer product) which is a multi-dimensional version of (1). The coefficient surface is thus expressed in terms of K_1

functions $\eta_k(s)$ and K_2 function $\theta_l(t)$, *i.e.*:

$$\beta(s, t) = \sum_{k=1}^{K_1} \sum_{l=1}^{K_2} b_{kl} \eta_k(s) \theta_l(t) = H(s) B \Theta(t) \quad (9)$$

where $H(s)$ and $\Theta(t)$ are respectively the vectors containing the K_1 functions $\eta_k(s)$ and K_2 function $\theta_l(t)$. The coefficients to estimate are stored in the $K_1 \times K_2$ matrix B .

As in the case of FLM-S, the FLM-F is fit through the minimization of a penalized sum of square criterion. Since the response is a function, the sum of squares is integrated, *i.e.* the criterion to minimize is:

$$ISSE(\alpha, \beta) = \int \sum_{i=1}^n \left[y_i(t) - \alpha(t) - \int_{\Omega_1} \beta(s, t) x_i(s) ds \right]^2 dt + PEN_0(\alpha) + PEN_1(\beta) + PEN_2(\beta) \quad (10)$$

where $PEN_0(\alpha) = \lambda_0 \int_{\Omega_1} [L_0 \alpha(t)]^2 dt$ is a penalization on α while $PEN_1(\beta) = \lambda_1 \int_{\Omega_1} \int_{\Omega_2} [L_1 \beta(s, t)]^2 ds dt$ and $PEN_2(\beta) = \lambda_2 \int_{\Omega_1} \int_{\Omega_2} [L_2 \beta(s, t)]^2 ds dt$ are both penalizations on β . L_0 , L_1 and L_2 are differential linear operators. Two distinct penalisation terms are used for $\beta(., .)$ because of the tensor product which makes independent the two dimensions of the surface. This allows the complexity of the surface to be controlled independently on each dimension. As in the case of data smoothing and FLM-S, the parameters λ_0 , λ_1 and λ_2 are chosen by minimizing a cross-validation criterion.

The fully functional linear model in (8) is fitted using the same heuristic as the FLM-S of section 2.2. The ISSE criterion (10) can be expressed in a matrix form and then be derived to obtain an equation to resolve (called the normal equation). Here, the quantity to estimate is a matrix instead of a vector in the FLM-S. The normal equations have to be rearranged using the Kronecker product in order to estimate the vector $vec(B)$ which is the matrix B rearranged column-wise. The computational details are not shown here. They can be found in chapter 16 of Ramsay & Silverman (2005).

An alternative way to provide an estimate for $\beta(., .)$ in (8) was developed by Yao et al. (2005). This estimate uses the functional principal components (FPC, Ramsay, 1982) as basis functions. More precisely, the FPCs of $X(s)$ are used to expand $\beta(s, t)$ on the s dimension and the FPCs of $Y(t)$ on the t dimension. The associated coefficients are proportional to the covariance between the FPC scores of $X(s)$ and $Y(t)$. Details of the procedure can be found in chapter 8 of Horváth & Kokoszka (2012). Although this latter estimation procedure has good theoretical convergence properties, it does not allow to control the smoothness of the estimation better than the procedure explained above. Moreover, the choice of the FPCs to consider is difficult (*e.g.* Goldsmith et al., 2011). This is why the application presented here favors the approach of Ramsay & Silverman (2005).

Finally, as for the FLM-S, one can wonder if there are identifiability constraints on the number of basis for $\beta(s, t)$. In the present case, the inverted matrix is a Kronecker product of two matrices. This means that this matrix can be inverted if and only if the two matrices constituting the Kronecker product are not singular. It can be shown that this is the case, similarly to section 2.2.1, only if K_1 is lower than both n and K_2 .

3. Application to streamflow

Following the work of Chebana et al. (2012) which described how to explore functional hydrographs, this section applies FLMs introduced in section 2 to forecast streamflows according to precipitation curves. In this application we focus on summer/autumn floods (floods caused by liquid precipitations) *i.e.* streamflow data from July to October are considered for each year. To illustrate the FLM-S of section 2.2.1, the flow volume during the four mentioned months is modelled with the curves of precipitation as covariates. Then, the FLM-F of section 2.2.2 is used to predict the whole streamflow curve between July and October using the precipitation curve. Precipitations are considered from June to October in order to have past values for the first days of the streamflows time span. However, the model is applicable to the whole year. In fact, many possibilities exist with the functional linear model (e.g. only spring floods, only a month, etc. . .).

3.1. Data description

The present application considers the Dartmouth station with federal reference number 01BH005. It is located on the Dartmouth river, 1.6 km upstream from the Ruisseau du Pas de Dame in the Gaspésie region of the province of Quebec, Canada. The drainage basin area is 626 km^2 and the flow regime of the river is natural. The geographical location of the station is shown in Figure 2. The data consists in a daily streamflow series (m^3/s) from 1970 to 2012 as well as daily total precipitation series available from 1981 to 2012. The common years between the series are thus 1981 to 2012. This means that, in the following, the sample size is $n = 32$. The streamflow curves domain is the continuous interval $\Omega = [30; 153]$ and the precipitation time domain is $\Omega_p = [0; 153]$.

Chebana et al. (2012) smoothed daily streamflows using 53 Fourier basis functions to obtain annual streamflow curves, which corresponds to a basis per week. Since the present application considers only the period July-October, the curves do not cover an entire year and Fourier basis appears less suited to smooth functional data. All the data used in this application (including precipitation data) are thus smoothed using B-spline basis which are more suited than Fourier basis for non periodic data.

Note that streamflow and precipitation data cannot physically have negative values, which means that there is an extra constraint in the smoothing of this data. Although it is possible to smooth under constraint (see Ramsay & Silverman, 2005, chapter 6), we choose here to smooth the logarithm of both streamflow and precipitation series. This has also the advantage of adding some symmetry around the mean curve for these series. In order to apply the logarithm for days with no precipitations, the values are set to 0.05 mm since the minimum value recorded is 0.2 mm. Note that the results are robust to the value attributed to the days without precipitation. Streamflow and precipitation curves are displayed in their original scale, *i.e.* the inverse transformation is applied.

Figure 3 shows the leave-one-out cross-validation (LOOCV) curve for choosing the parameter λ in the smoothing of streamflows (for the FLM-F only) and precipitation curves (for both FLM-S and FLM-F). The smoothing of streamflow values is straightforward since a clear minimum appears for $\lambda = 10^{-1.25}$. This corresponds to 81 B-spline

basis functions. The process is less straightforward for precipitations since the minimum is for a very high value of λ which corresponds to only 2 basis functions. This translates the difficulty of predicting precipitation processes (Suhaila et al., 2011). Since this would result in straight lines for precipitations, the λ value chosen is the lowest with a CV value inside the standard error of the minimum (in other words, below the horizontal shaded line in Figure 3b). This results in $\lambda = 10^{4.25}$ and 7 basis functions.

Figure 4 shows two examples of the resulting smoothed streamflows (with $\lambda = 10^{-1.25}$). Because of the high number of basis functions, the peaks are well reached by this smoothing. Peaks are probably the most important feature of streamflows and it is important that they are well represented. One look at the mean curve allows a characterisation of streamflows on this period. Indeed, it shows that streamflows are generally low during the middle of summer and begin to increase when the fall season begins.

The same examples for precipitation curves are shown in Figure 5. It is the opposite of streamflows here since the curves are extremely smooth and only show periods when precipitations are more likely to happen. Indeed, a smoothing with more basis functions would have a prediction error that is too high. This means that precipitations are a phenomenon that is extremely difficult to predict. All the curves shown here are then used in the FLM applications of the next section.

3.2. Forecasting of streamflow volume

For water resources management, it is important to know the amount of incoming water. Thus, it is of interest to estimate and forecast the total streamflow during a given period. This is often achieved through the use of regression models using the mean or total of some covariates (*e.g.* Garen, 1992; Eldaw et al., 2003) or through regional frequency analysis (Ouarda et al., 2000). We illustrate in this section how to use the FLM-S for such a purpose.

Using the data described in section 3.1, the method applied is the FLM-S (3) described in section 2.2.1, with the variables

- y_i : the response as the logarithm of the sum of daily streamflow values from the 1st of July to the 31st of October for year i ($i = 1, \dots, n = 31$). The logarithm is used because total streamflows are strongly lognormal (Vogel & Wilson, 1996);
- $x_i(\cdot)$: the explanatory variable is represented by precipitation curves from the 1st of June to the 31st of October for year i ($i = 1, \dots, 31$).

Precipitations are considered up to one month earlier than streamflows in order to consider all precipitations that could influence the streamflow volume.

To fit the FLM-S, the parameter λ controlling the smoothness of the $\beta(\cdot)$ coefficient curve must be chosen by minimizing the PENSSE criterion. Figure 6 shows the LOOCV scores for different values of λ . There is an obvious minimum for $\lambda = 10^{-2}$, which corresponds to 22 basis functions on the B-spline basis and a rough $\hat{\beta}(\cdot)$ curve.

The $\hat{\beta}(\cdot)$ curve obtained by fitting the FLM-S with $\lambda = 10^{-2}$ is shown in Figure 7. The 95% confidence interval is obtained by estimating a standard error curve and multiplying it by the quantiles of the standard gaussian distribution (Ramsay et al., 2011, p.141). The low amplitude of the $\hat{\beta}(\cdot)$ curve is due to the spreading of the influence of precipitation over the entire time span. Moreover, recall that the response of this model is the log volume. The $\hat{\beta}(\cdot)$ curve shown in Figure 7 is not smooth and looks like an oscillation. This is because the data number ($n = 32$) is small and the $\hat{\beta}(\cdot)$ curve is sensitive to the small features in precipitation data. However, it shows two periods where the oscillations have a larger amplitude : the second half of July and the beginning of the fall, indicating two periods where streamflows are more influenced by precipitations. At the end of July, streamflows can be low because the snow melt is over and the river has almost dried out. Therefore, any rainfall has a large influence on streamflows. Moreover, the large amplitude period towards the end of July also quickly follows the rainy period which occurs at the beginning of the summer, as indicated by Figure 5c. The same reason can be evoked for the fall period of the $\hat{\beta}(\cdot)$ curve which aligns with the beginning of autumn rainfalls (Figure 5c).

In order to assess the performances of the FLM-S, the fit with $\hat{\beta}(\cdot)$ is compared to a traditional linear model with the same response variable (denoted “LM” in the following). The explanatory variable of LM is the sum of daily precipitations between the 1st of June to the 31st of October. This model results in a coefficient that is equal to 0.004 which is significantly different from 0 ($p - value = 5e^{-6}$). This coefficient is a kind of integration of the $\hat{\beta}(\cdot)$ curve, the latter can be seen as the detail of the LM coefficient.

The scatterplot of Figure 8 compares the fitted values \hat{y}_i of the FLM-S and the LM. The fit is visually better for the FLM-S since the points are closer to the $y = x$ line (which represents a perfect fit). Table 2 illustrates the scores for different performance indicators (RMSE, R^2 , LOOCV and bias). It also shows better performances for the FLM-S. Indeed, the FLM-S displays a higher R^2 and a lower RMSE than the LM. Note that the R^2 is close to one for the FLM-S (equal to 94%), indicating an excellent fit. Note also that the difference between the two models is smaller with the LOOCV criterion. The complexity of the FLM-S seems somehow to balance its better performances. However, this should vanish with longer data records. Finally, note that the bias of the FLM-S is not null because of the regularization used for the fit. It is however close to zero and very small compared to the scale of the response (it accounts for less than 0.01% of the mean of the response).

3.3. Hydrograph forecasting

3.3.1. FLM-F fitting

In this application, we are not interested in a single feature of the streamflow process but rather in the whole hydrograph. The objective is to forecast the hydrograph using the FLM-F given by (8). Although recent developments of the FLM-F allow its use with several covariates (Ivanescu et al., 2014), only precipitations are used here for simplicity purposes.

A proper estimation of the FLM-F needs the choice of three regularization parameters, *i.e.* one for the intercept curve $\alpha(\cdot)$ (λ_0), and two for the coefficient surface $\beta(\cdot, \cdot)$ (λ_1 for the s dimension and λ_2 for the t dimension). The

resulting 3-dimensional plot is not shown here but the LOOCV (which, in this case, is a "leave-one-year-out" CV) is minimized when the three parameters are $\lambda_0 = \lambda_1 = \lambda_2 = 10^5$.

The estimated coefficients $\alpha(\cdot)$ and $\beta(\cdot, \cdot)$ are shown in Figure 9. The $\alpha(\cdot)$ curve illustrates the shape of the hydrograph without the influence of precipitations. The curve reaches its minimum during the middle of the summer (the beginning of august) and rises afterwards when the fall season starts. The $\beta(\cdot, \cdot)$ surface is a bump on the diagonal which indicates the positive influence of precipitation on the hydrograph, with a slight translation to the left to show delay between rainfall and the increase of streamflow. This influence is stronger during autumn season when more floods occur. This can be explained by the fact that there is less evaporation during this period of the year when temperature is not high. There is also a little extension of the bump to the top of the surface during July which could mean that, since the ground is dry at this time, infiltration is larger which leads to a significant contribution to surface flows two months later.

3.3.2. Comparison with the artificial neural network approach

To understand the strengths and weaknesses of the FLM-F, it has to be compared to other commonly used forecasting methods using exogeneous covariates in hydrology. Among these methods (reviewed in the introduction), ANN models are the most widely used because of their ability to simulate complex relationship. Thus, for comparison purposes, ANN models are considered to forecast the log streamflow as well. Following number of previous references (*e.g.* Anctil & Lauzon, 2004; Chau et al., 2005; Yonaba et al., 2010; Govindaraju & Rao, 2010, chapters 1 and 2), ANN models are used with the 3 previous days of precipitations as covariates. For more generality, the covariate lag could have been selected using a cross validation procedure (*e.g.* Haddad et al., 2013). However, for comparison purposes, the model design is based on what is found in the literature. The ANN estimation is made with one hidden layer containing 5 nodes (the number of nodes is chosen using the leave-one-out cross-validation such as in Wu & Chau, 2010, for instance).

Figure 10 shows the mean prediction error estimated by CV for each year. Performances of ANN models and FLM-F depend on the year, and it is hard to discriminate the two methods using only this figure. Three particular examples of predicted hydrographs $\hat{y}_i(\cdot)$ using the FLM-F and the ANN models are shown in Figure 11 to help understand when the FLM-F performs better than ANN models. It is immediately visible that the FLM-F and the ANN approaches have very different behaviours. The FLM-F actually predicts the global shape of the hydrograph, while the ANN model predicts only the short term patterns such as the peaks, but does not forecast well the streamflow accumulations. Hence, it appears clear that the FLM-F model is suited to match the trend and hence will perform better for years with few peaks. Moreover, FLM-F seems able to match some behaviours such as July droughts (*e.g.* Figure 11b). However, when there are more short-term features such as peaks, the ANN performs better than the FLM-F (*e.g.* Figure 11c). The FLM-F forecast smoothness is mainly due to the smoothness of the $\beta(\cdot, \cdot)$ surface. The consideration of less smooth precipitation curves and $\beta(\cdot, \cdot)$ surface increases the prediction error since precipitations are extremely difficult to predict with precision. Note that the small number of years of record

(32 years) does not allow for a rough $\beta(.,.)$ surface and the FLM-F should improve with more data years.

Table 3 shows the global fitting criteria and indicates that the ANN model leads globally to better performances. Indeed, its RMSE value is lower and its R^2 value is higher than those of the FLM-F. However, the difference between the FLM-F and the ANN model is smaller for the CV score which is the only criterion that considers their performance on new data not used for the calibration of the models. This means that the FLM-F performs almost as well as ANN models for long time forecasting. Figure 12 provides an explanation for this by showing the distribution of the CV along the time domain, for the mean over all years. The CV of ANN models is shown to be increasing with the horizon, meaning that the accuracy of the forecasting decreases with the forecasting horizon. This is due to the fact that it is not able to predict whole events. Conversely, the CV does not increase for the FLM-F and is even smaller during the fall season. Figure 12 also shows the main drawback of the FLM-F which is that it does not predict well the short-term variations. Indeed, during July streamflows are often low and thus are very dependant on day-to-day rainfall. This is when the mean error of the FLM-F is the highest. Summer streamflows are more difficult to forecast than fall streamflows using only precipitations. This is in agreement with the shape of the $\hat{\beta}(.,.)$ surface which has less amplitude for summer streamflows.

4. Conclusions

The purpose of the present work is to introduce and adapt functional linear models to the hydrological framework. This work follows the paper of Chebana et al. (2012) which shows the relevancy of using the functional framework in hydrology. After an introduction to FLM models, they are applied to streamflow forecasting based on precipitation curves. Conceptually, FLM are perfectly suited for time series regression since they provide a solution to the problems caused by autocorrelation and non stationarity in time series. Moreover, section 2 shows that elegant estimation methods have been developed to manage infinite dimensional data.

The application of the FLM-S to forecast streamflow volumes provides interesting insights for the interpretation of the results. Indeed, results suggest that precipitations influence streamflow especially in the middle of summer and at the beginning of the fall season. Moreover, this model outperforms the somewhat simple linear regression, in terms of volume forecasting accuracy. The shape of the influence of precipitations on streamflows is refined with the application of the FLM-F. This model highlights the large influence of July precipitations on streamflows two-months later. The importance of summer precipitations on the beginning of the fall season streamflows is also obvious. If ANN models show slightly better performances than the FLM-F, the latter show an ability to forecast the global shape of the hydrograph. Moreover, FLM-F performances do not decrease with the increase of the horizon like ANN models. Moreover, the better performances of the ANN model are due to the fact that we have used the observed precipitation data. However, in practice, precipitation forecasts are not as accurate and the long term forecasts of ANN models would be less accurate than in this application.

The application presented in this paper is relatively simple and deals with a single case study. The present work

focuses mainly on the method itself, and illustrates it on a single application. However, a large number of other hydrological applications can be considered. For instance, one could orient the application to other hydrologic events such as spring floods or droughts. In such cases, the curves can be streamflows, snowfalls and temperatures during different seasons. Similar applications can also be performed for different regions representing different climates. Setting a FDA application requires a proper definition of the targeted event and time intervals. FDA can also be used for other purposes such as regional estimation at ungauged sites or estimation of missing data (for instance a continuous part of a hydrograph). A large number of applications need to be investigated in the field of hydrology.

There are a few limitations to the application of FLMs. One of the particularities of the hydrological framework is the importance of peak streamflow values. For functional data, reaching the peaks necessitates an important amount of basis functions. Similarly to the multiple regression framework, FLMs are based on the modelling of mean curves which do not always reach the desired peaks. Doing so necessitates models that use complex curves which could decrease the performances of the model. Furthermore, the complexity allowed by FLMs also depends on the number of available curves. When using long curves such as in the present work, fitting complex models requires a large number of data years which are not always available.

The discussion presented above leads to a number of perspectives. First, it can be of interest to apply the historical FLM either from Malfait & Ramsay (2003) or from Kim et al. (2011) and compare the results with the FLM-F. Second, Following the recent development of the R package `refund`, a wide body of literature on FLMs is emerging. Notably, we can cite a new estimation method based on mixed models which allows the use of several functional covariates in the FLM-F (Ivanescu et al., 2014). This method expresses the FLM-F as an additive model in order to be fit efficiently (such as in Wood, 2006, for instance). Such an estimation method also has the advantage of providing well justified confidence intervals (Goldsmith et al., 2011). Also part of `refund` is the recent development of functional generalized additive models (McLean et al., 2014). Third, an important feature of functions as mathematical objects is the possibility to derive them. This can lead to insights on the variation of streamflow processes and can also be a path for the study of curve peaks. A fourth perspective lies in the use of functional autoregressive models (*e.g.* Damon & Guillas, 2002) in order to forecast future streamflow phenomena using past streamflow curves. Finally, it appears important in the future to take advantage of the emerging body of literature on functional geostatistics (*e.g.* Delicado et al., 2010; Caballero et al., 2013; Ignaccolo et al., 2014) to model the spatial dependence between hydrological sites.

Acknowledgements

Financial support for this study was graciously provided by the Natural Sciences and Engineering Research Council (NSERC) of Canada and by the Nord-Pas de Calais Regional Council, France. Moreover, the authors would like to thank the international relations ministry of Quebec (ministère des Relations internationales, de la Francophonie et du Commerce extérieur du Québec) and the french general consulate in Quebec (consulat général de France au Québec) for their financial contribution to this France-Québec cooperation project. Finally, the authors are grateful

to the authors of the R package *fda* (Ramsay et al., 2014) for the very handy tools they have provided.

References

References

- Anctil, F., & Lauzon, N. (2004). Generalisation for neural networks through data sampling and training procedures, with applications to streamflow predictions. *Hydrology and Earth System Sciences Discussions*, *8*, 940–958.
- Bel, L., Bar-Hen, A., Petit, R., & Cheddadi, R. (2011). Spatio-temporal functional regression on paleoecological data. *Journal of Applied Statistics*, *38*, 695–704.
- Bosq, D. (2000). *Linear processes in function spaces: theory and applications* volume 149. Springer.
- Bosq, D., & Blanke, D. (2008). *Inference and prediction in large dimensions* volume 754. John Wiley & Sons.
- Brumback, B. A., & Rice, J. A. (1998). Smoothing spline models for the analysis of nested and crossed samples of curves. *Journal of the American Statistical Association*, *93*, 961–976.
- Caballero, W., Giraldo, R., & Mateu, J. (2013). A universal kriging approach for spatial functional data. *Stochastic Environmental Research and Risk Assessment*, *27*, 1553–1563.
- Cardot, H., Faivre, R., & Goulard, M. (2003a). Functional approaches for predicting land use with the temporal evolution of coarse resolution remote sensing data. *Journal of Applied Statistics*, *30*, 1185–1199.
- Cardot, H., Ferraty, F., & Sarda, P. (1999). Functional linear model. *Statistics & Probability Letters*, *45*, 11–22.
- Cardot, H., Ferraty, F., & Sarda, P. (2003b). Spline estimators for the functional linear model. *Statistica Sinica*, *13*, 571–592.
- Chau, K., Wu, C., & Li, Y. (2005). Comparison of several flood forecasting models in yangtze river. *Journal of Hydrologic Engineering*, *10*, 485–491.
- Chebana, F., Charron, C., Ouarda, T. B. M. J., & Martel, B. (2014). Regional frequency analysis at ungauged sites with the generalized additive model. *Journal of Hydrometeorology*, *15*, 2418–2428.
- Chebana, F., Dabo-Niang, S., & Ouarda, T. B. M. J. (2012). Exploratory functional flood frequency analysis and outlier detection. *Water Resources Research*, *48*, W04514.
- Chiou, J.-M. (2012). Dynamical functional prediction and classification, with application to traffic flow prediction. *The Annals of Applied Statistics*, *6*, 1588–1614.

- Chokmani, K., Ouarda, T. B. M. J., Hamilton, S., Ghedira, M. H., & Gingras, H. (2008). Comparison of ice-affected streamflow estimates computed using artificial neural networks and multiple regression techniques. *Journal of Hydrology*, *349*, 383–396.
- Cuevas, A., Febrero, M., & Fraiman, R. (2002). Linear functional regression: The case of fixed design and functional response. *Canadian Journal of Statistics*, *30*, 285–300.
- Dabo-Niang, S., & Ferraty, F. (2008). *Functional and operatorial statistics*. Springer.
- Damon, J., & Guillas, S. (2002). The inclusion of exogenous variables in functional autoregressive ozone forecasting. *Environmetrics*, *13*, 759–774.
- Delicado, P., Giraldo, R., Comas, C., & Mateu, J. (2010). Statistics for spatial functional data: some recent contributions. *Environmetrics*, *21*, 224–239.
- Eldaw, A. K., Salas, J. D., & Garcia, L. A. (2003). Long-range forecasting of the Nile river flows using climatic forcing. *Journal of Applied Meteorology*, *42*, 890–904.
- Ferraty, F., & Vieu, P. (2006). *Nonparametric functional data analysis: theory and practice*. Springer.
- Fortin, V., Ouarda, T., Rasmussen, P., & Bobée, B. (1997). Revue bibliographique des méthodes de prévision des débits. *Revue des sciences de l'eau/Journal of Water Science*, *10*, 461–487.
- Garen, D. (1992). Improved techniques in regression-based streamflow volume forecasting. *Journal of Water Resources Planning and Management*, *118*, 654–670.
- Goia, A., May, C., & Fusai, G. (2010). Functional clustering and linear regression for peak load forecasting. *International Journal of Forecasting*, *26*, 700–711.
- Goldsmith, J., Bobb, J., Crainiceanu, C. M., Caffo, B., & Reich, D. (2011). Penalized functional regression. *Journal of Computational and Graphical Statistics*, *20*, 830–851.
- Govindaraju, R. S., & Rao, A. R. (2010). *Artificial Neural Networks in Hydrology*. Springer Publishing Company, Incorporated.
- Granger, C. W. J., & Newbold, P. (1974). Spurious regressions in econometrics. *Journal of Econometrics*, *2*, 111–120.
- Haddad, K., Rahman, A., A Zaman, M., & Shrestha, S. (2013). Applicability of monte carlo cross validation technique for model development and validation using generalised least squares regression. *Journal of Hydrology*, *482*, 119–128.

- Hastie, T., & Mallows, C. (1993). [a statistical view of some chemometrics regression tools]: Discussion. *Technometrics*, *35*, 140–143.
- Hastie, T., & Tibshirani, R. (1986). Generalized additive models. *Statistical Science*, *1*, 297–310.
- Hastie, T., & Tibshirani, R. (1993). Varying-coefficient models. *Journal of the Royal Statistical Society. Series B (Methodological)*, *55*, 757–796.
- Hoerl, A. E., & Kennard, R. W. (1970). Ridge regression: Biased estimation for nonorthogonal problems. *Technometrics*, *12*, 55–67.
- Hoover, K. D. (2003). Nonstationary time series, cointegration, and the principle of the common cause. *The British Journal for the Philosophy of Science*, *54*, 527–551.
- Horváth, L., & Kokoszka, P. (2012). *Inference for functional data with applications* volume 200. Springer.
- Ignaccolo, R., Mateu, J., & Giraldo, R. (2014). Kriging with external drift for functional data for air quality monitoring. *Stochastic Environmental Research and Risk Assessment*, *28*, 1171–1186.
- Ivanescu, A. E., Staicu, A.-M., Scheipl, F., & Greven, S. (2014). Penalized function-on-function regression. *Johns Hopkins University, Dept. of Biostatistics Working Papers*, .
- Kim, K., Şentürk, D., & Li, R. (2011). Recent history functional linear models for sparse longitudinal data. *Journal of Statistical Planning and Inference*, *141*, 1554–1566.
- Kisi, O. (2007). Streamflow forecasting using different artificial neural network algorithms. *Journal of Hydrologic Engineering*, *12*, 532–539.
- Kisi, O. (2009). Wavelet regression model as an alternative to neural networks for monthly streamflow forecasting. *Hydrological processes*, *23*, 3583–3597.
- Makkeasorn, A., Chang, N. B., & Zhou, X. (2008). Short-term streamflow forecasting with global climate change implications – a comparative study between genetic programming and neural network models. *Journal of Hydrology*, *352*, 336–354.
- Malfait, N., & Ramsay, J. O. (2003). The historical functional linear model. *Canadian Journal of Statistics*, *31*, 115–128.
- McLean, M. W., Hooker, G., Staicu, A.-M., Scheipl, F., & Ruppert, D. (2014). Functional generalized additive models. *Journal of Computational and Graphical Statistics*, *23*, 249–269.
- Müller, H.-G., Chiou, J.-M., & Leng, X. (2008). Inferring gene expression dynamics via functional regression analysis. *BMC bioinformatics*, *9*, 60.

- Ouarda, T., Haché, M., Bruneau, P., & Bobée, B. (2000). Regional flood peak and volume estimation in northern canadian basin. *Journal of Cold Regions Engineering*, *14*, 176–191.
- Phillips, P. C. B. (1987). Time series regression with a unit root. *Econometrica*, *55*, 277–301.
- Ramsay, J. (1982). When the data are functions. *Psychometrika*, *47*, 379–396.
- Ramsay, J., Wickham, H., Graves, S., & Hooker, G. (2011). fda: Functional data analysis. *R package version*, *2*.
- Ramsay, J. J. O., Hooker, G., & Graves, S. (2009). *Functional data analysis with R and MATLAB*. Springer.
- Ramsay, J. O., & Dalzell, C. (1991). Some tools for functional data analysis. *Journal of the Royal Statistical Society. Series B (Methodological)*, (pp. 539–572).
- Ramsay, J. O., & Silverman, B. (2005). *Functional data analysis*. (2nd ed.). Wiley Online Library.
- Ramsay, J. O., Wickham, H., Graves, S., & Hooker, G. (2014). *fda: Functional Data Analysis*. URL: <http://CRAN.R-project.org/package=fda> r package version 2.4.4.
- Ratcliffe, S. J., Heller, G. Z., & Leader, L. R. (2002a). Functional data analysis with application to periodically stimulated foetal heart rate data. ii: Functional logistic regression. *Statistics in Medicine*, *21*, 1115–1127.
- Ratcliffe, S. J., Leader, L. R., & Heller, G. Z. (2002b). Functional data analysis with application to periodically stimulated foetal heart rate data. i: Functional regression. *Statistics in Medicine*, *21*, 1103–1114.
- Sahay, R. R., & Sehgal, V. (2013). Wavelet regression models for predicting flood stages in rivers: a case study in eastern india. *Journal of Flood Risk Management*, *6*, 146–155.
- Sene, K. (2009). *Hydrometeorology: forecasting and applications*. Springer.
- Sood, A., James, G. M., & Tellis, G. J. (2009). Functional regression: A new model for predicting market penetration of new products. *Marketing Science*, *28*, 36–51.
- Stewart-Koster, B., Olden, J. D., & Gido, K. B. (2014). Quantifying flow-ecology relationships with functional linear models. *Hydrological Sciences Journal*, *59*, 629–644.
- Stone, M. (1974). Cross-validatory choice and assessment of statistical predictions. *Journal of the Royal Statistical Society. Series B (Methodological)*, *36*, 111–147.
- Suhaila, J., Jemain, A. A., Hamdan, M. F., & Wan Zin, W. Z. (2011). Comparing rainfall patterns between regions in peninsular malaysia via a functional data analysis technique. *Journal of Hydrology*, *411*, 197–206.
- Ternynck, C., Ben Alaya, M. A., Chebana, F., Dabo-Niang, S., & Ouarda, T. B. M. J. (2016). Streamflow hydrograph classification using functional data analysis. *Journal of Hydrometeorology*, .

- Tootle, G., Singh, A., Piechota, T., & Farnham, I. (2007). Long lead-time forecasting of u.s. streamflow using partial least squares regression. *Journal of Hydrologic Engineering*, *12*, 442–451.
- Vogel, R., & Wilson, I. (1996). Probability distribution of annual maximum, mean and minimum streamflow values in the united states. *Journal of Hydrologic Engineering*, *1*, 69–76.
- Vogel, R., Wilson, I., & Daly, C. (1999). Regional regression models of annual streamflow for the united states. *Journal of Irrigation and Drainage Engineering*, *125*, 148–157.
- Wood, S. (2006). *Generalized additive models: an introduction with R*.
- Wu, C. L., & Chau, K. W. (2010). Data-driven models for monthly streamflow time series prediction. *Engineering Applications of Artificial Intelligence*, *23*, 1350–1367.
- Yao, F., Muller, H.-G., & Wang, J.-L. (2005). Functional linear regression analysis for longitudinal data, . (pp. 2873–2903).
- Yonaba, H., Anctil, F., & Fortin, V. (2010). Comparing sigmoid transfer functions for neural network multistep ahead streamflow forecasting. *Journal of Hydrologic Engineering*, *15*, 275–283.
- Yu, P.-S., Chen, S.-T., & Chang, I. F. (2006). Support vector regression for real-time flood stage forecasting. *Journal of Hydrology*, *328*, 704–716.
- Zealand, C. M., Burn, D. H., & Simonovic, S. P. (1999). Short term streamflow forecasting using artificial neural networks. *Journal of Hydrology*, *214*, 32–48.

Name	Response Y	Covariates $X^{(j)}$	Reference
Classical regression	Scalar	Scalar	(Vogel et al., 1999)
FANOVA	Functional	Scalar	(Brumback & Rice, 1998)
FLM for scalar response	Scalar	Functional	(Stewart-Koster et al., 2014)
Concurrent	Functional	Functional	(Hastie & Tibshirani, 1993)
Fully functional	Functional	Functional	(Ramsay & Silverman, 2005, chapter 16)

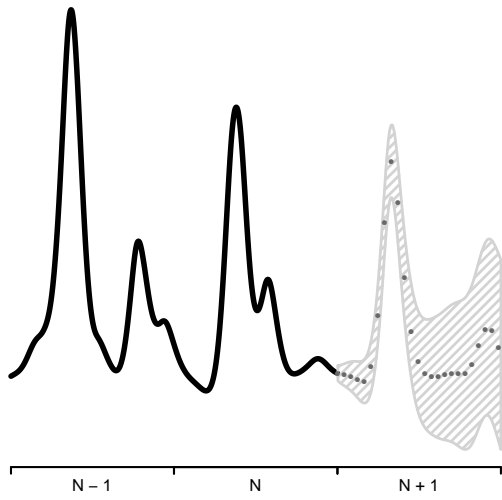
Table 1: Types of functional linear models.

Criterion	FLM-S	LM
Bias	0.001	0.000
RMSE	0.808	2.327
LOOCV	0.183	0.192
R^2	0.940	0.506

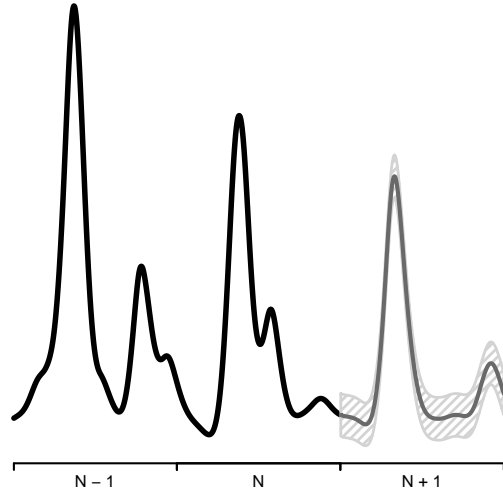
Table 2: Values of several criteria in order to compare the FLM-S to the LM. For the bias, RMSE and LOOCV, the lower the criterion, the better the model is, and inversely for the R^2 .

Criterion	FLM-F	ANN
Bias	0.000	0.000
RMSE	0.671	0.211
LOOCV	0.795	0.610
R^2	0.452	1.000

Table 3: Values of several criteria in order to compare the FLM-F to the ANN. For the bias, RMSE and LOOCV, the lower the criterion, the better the model is, and inversely for the R^2 .



(a) Classical methods such as ANN or Wavelet regression.



(b) Functional regression

Figure 1: Illustration of the difference between pointwise classical methods such as ANN models or wavelet regression and functional regression for forecasting purposes. The dashed area indicates the shape of expected confidence intervals for forecasts.

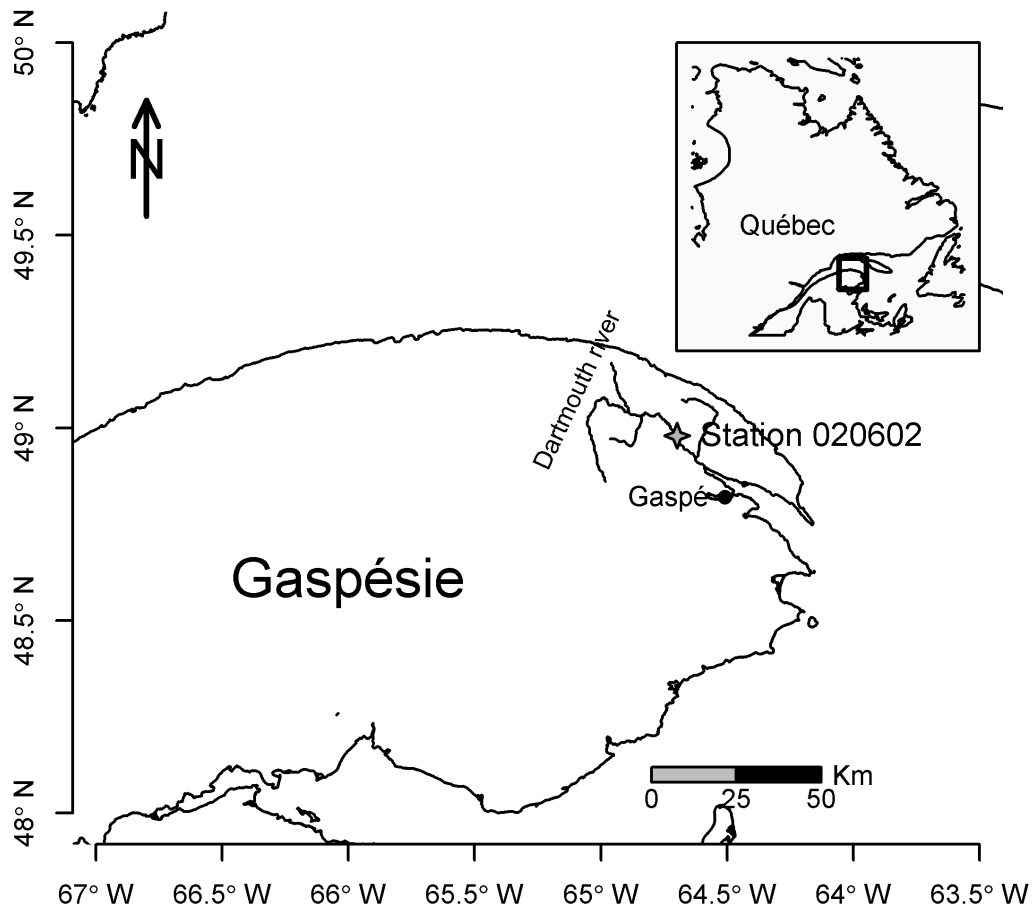


Figure 2: Geographical location of the Dartmouth station.

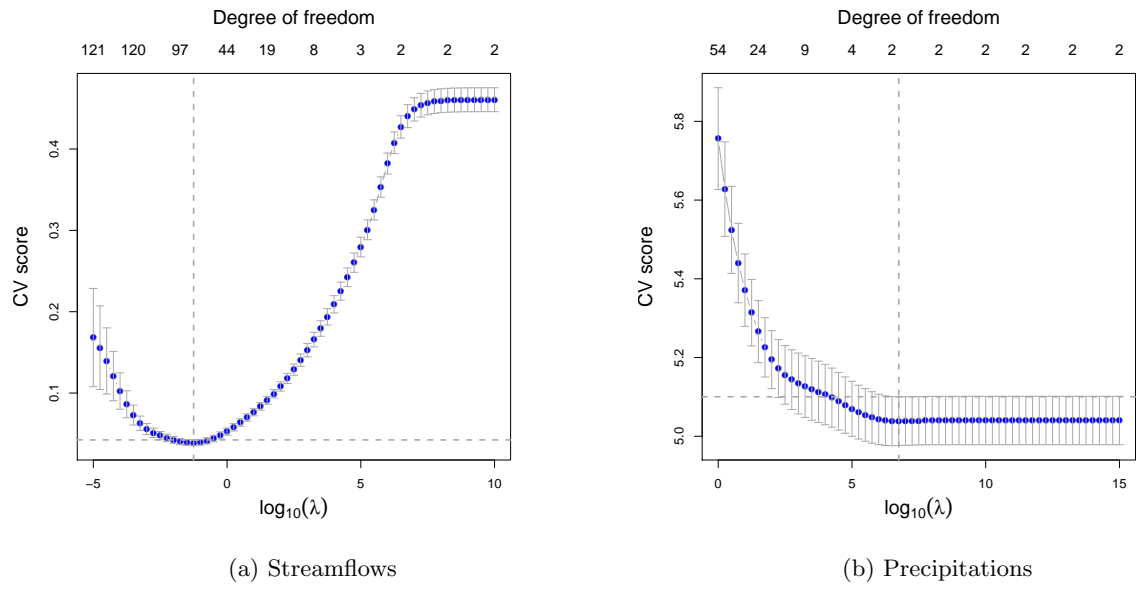


Figure 3: 10-fold CV curves for different values of λ in the smoothing of streamflows and precipitation curves. The bars at each point represent the standard errors of the CV values. The vertical dashed line aims at spotting the minimum CV value and the horizontal dashed line is the minimum CV value plus its standard error.

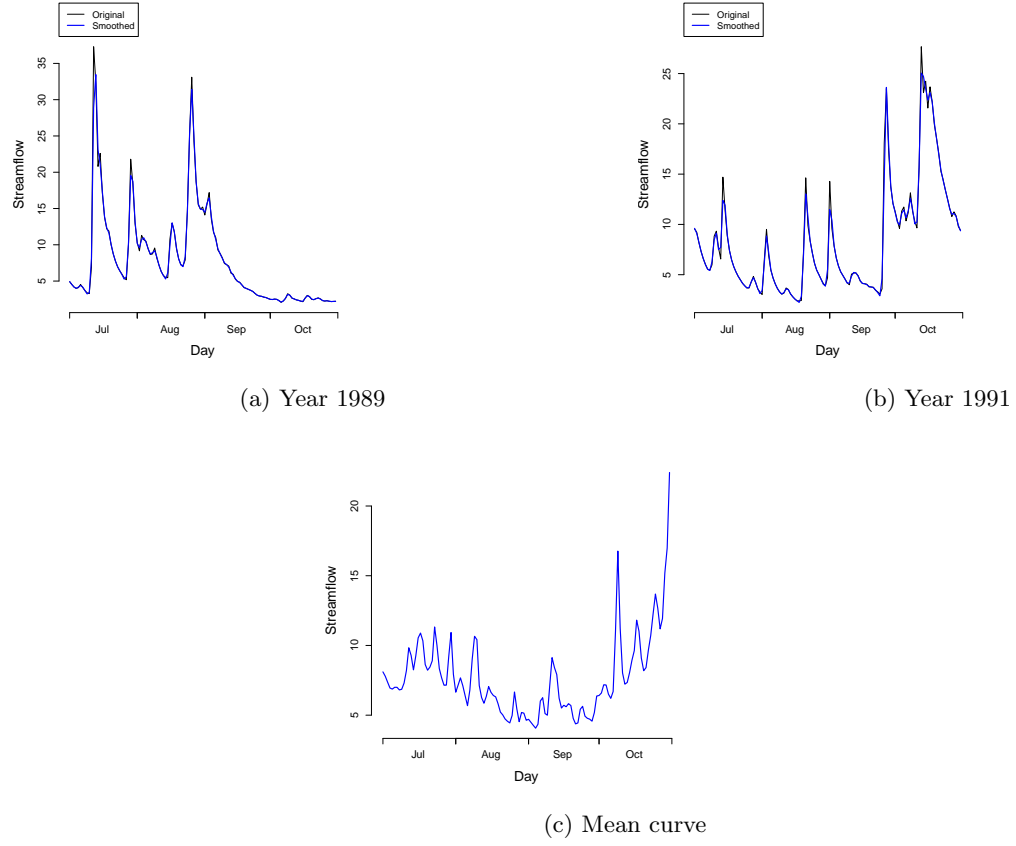
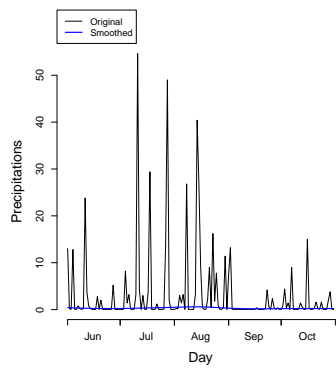
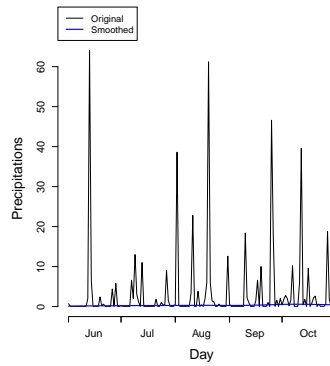


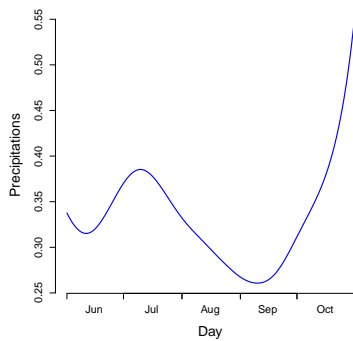
Figure 4: Examples of observed and smoothed streamflows for the years 1989, 1991 and the mean curve for the period 1981-2012. The black lines are observed streamflows and blue lines are the smoothed ones.



(a) Year 1989



(b) Year 1991



(c) Mean curve

Figure 5: Examples of observed and smoothed precipitations for the years 1989, 1991 and the mean curve for the period 1981-2012. The black lines are observed precipitations and blue lines are the smoothed ones.

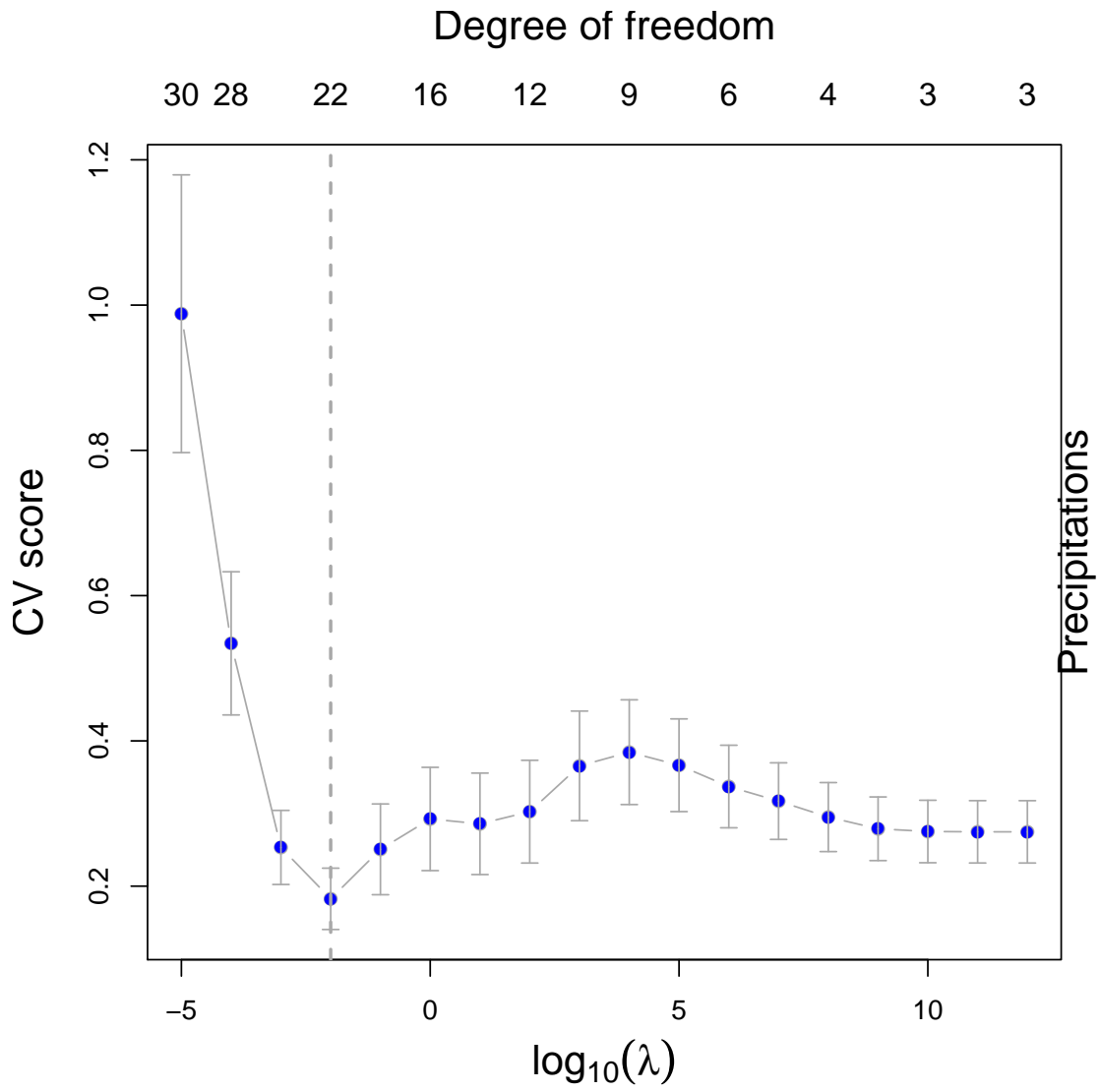


Figure 6: Leave-one-out CV curve for the parameter λ in the FLM-S. Bars indicate the standard error of the LOOCV values.

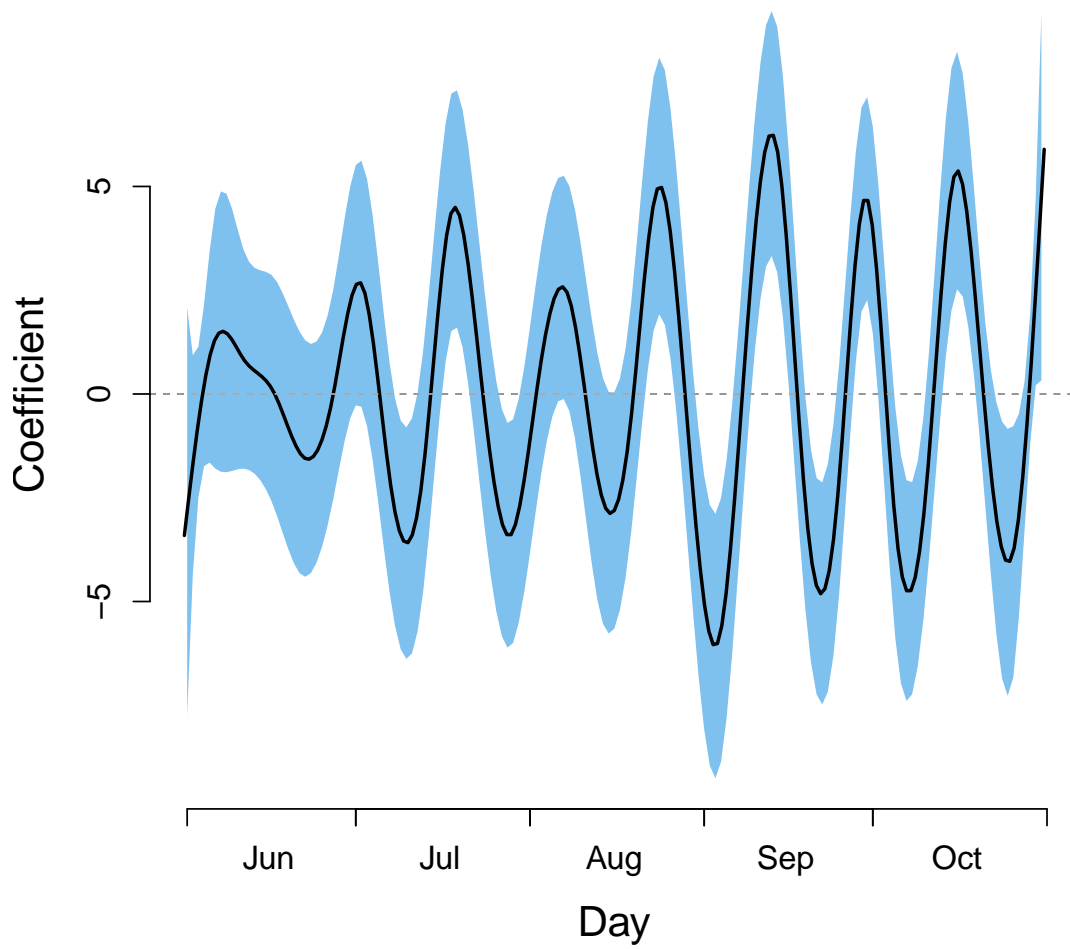


Figure 7: Estimated $\hat{\beta}(\cdot)$ function for precipitations in the FLM-S. The blue area corresponds to the pointwise 95% confidence interval.

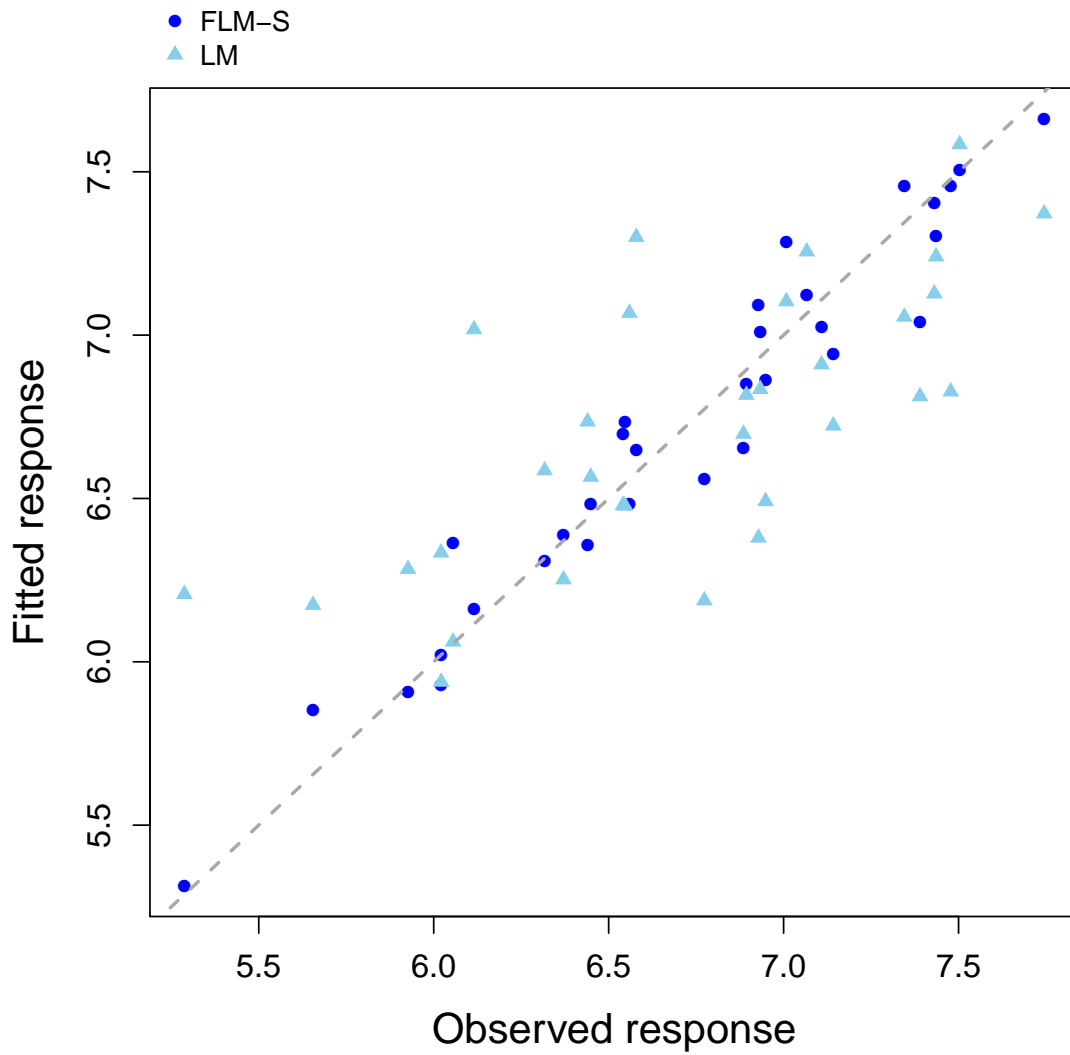
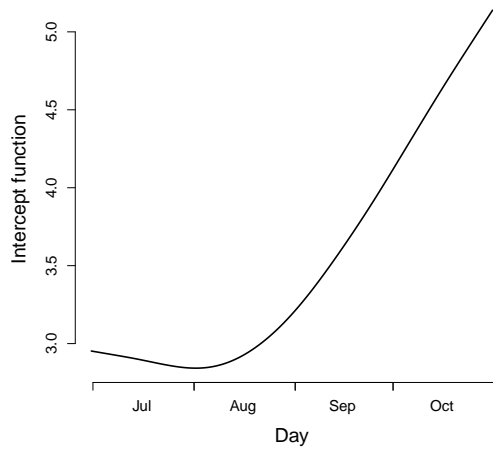
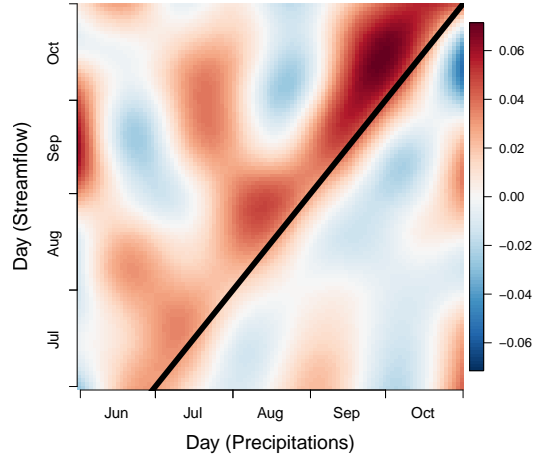


Figure 8: Scatterplot of fitted vs. observed response values for the sum of streamflows. The plot contains the fitted values of the models FLMS (blue circles) and LM (cyan triangles).



(a) Intercept curve $\hat{\alpha}(\cdot)$



(b) Coefficient surface $\hat{\beta}(\cdot, \cdot)$

Figure 9: Estimated functional coefficients of the FLM-F. The $\hat{\alpha}(\cdot)$ curve shows the expected shape of the hydrograph without any influence of precipitations and the $\hat{\beta}(s, t)$ surface shows the influence of precipitations on the hydrograph. For the latter, the dimension s is in the abscissa and the dimension t in the ordinate. The thick black line indicates the times $s = t$

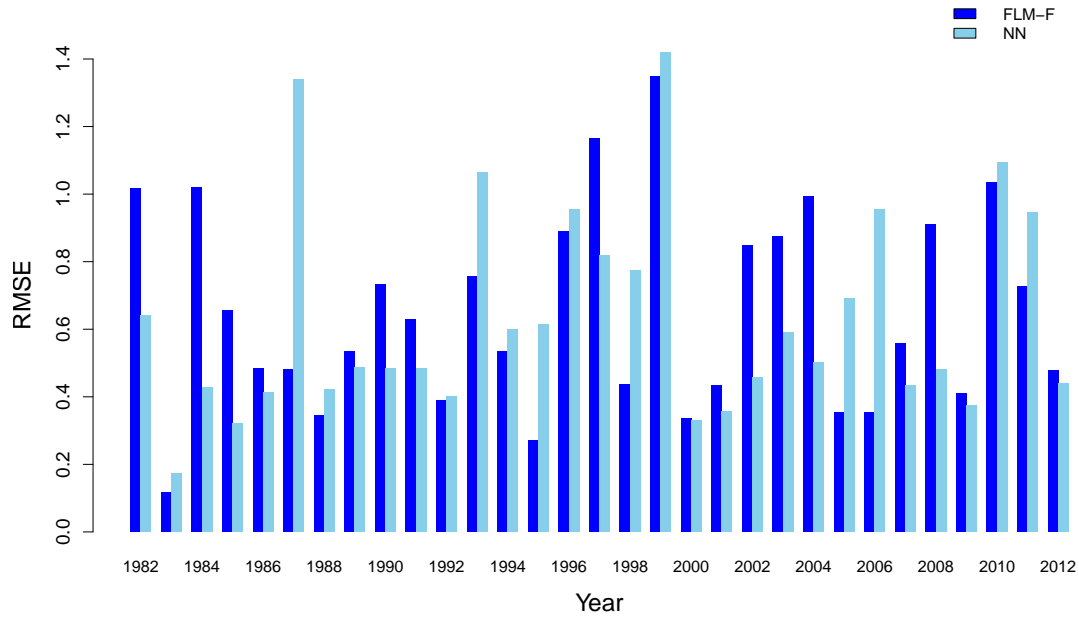
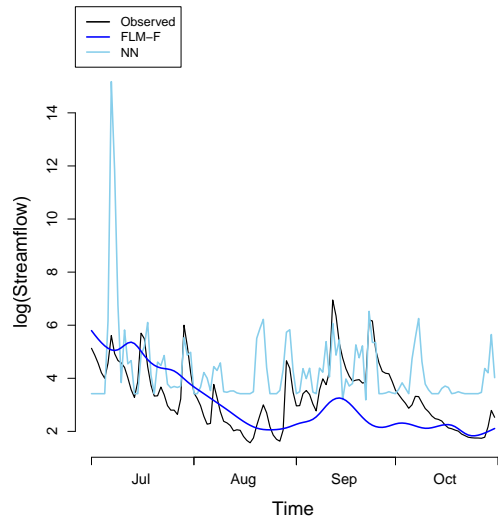
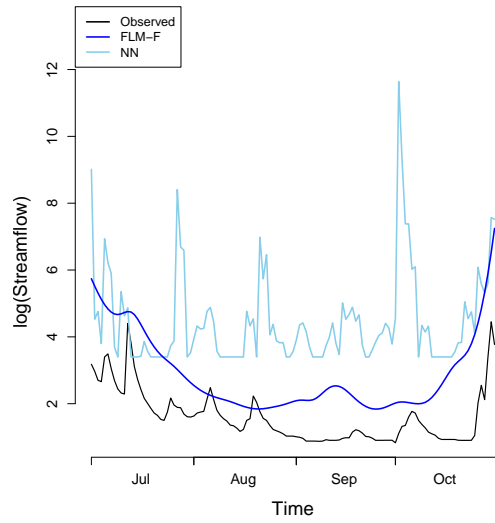


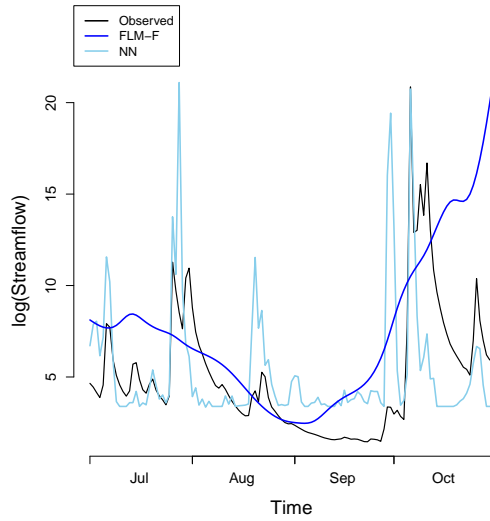
Figure 10: Mean prediction error over all data points for each year.



(a) Year 1983



(b) Year 2009



(c) Year 2009

Figure 11: Predicted hydrograph with the estimated FLMF and the estimated ANN model. For each year, the model is fitted on every other years and the remaining year is predicted.

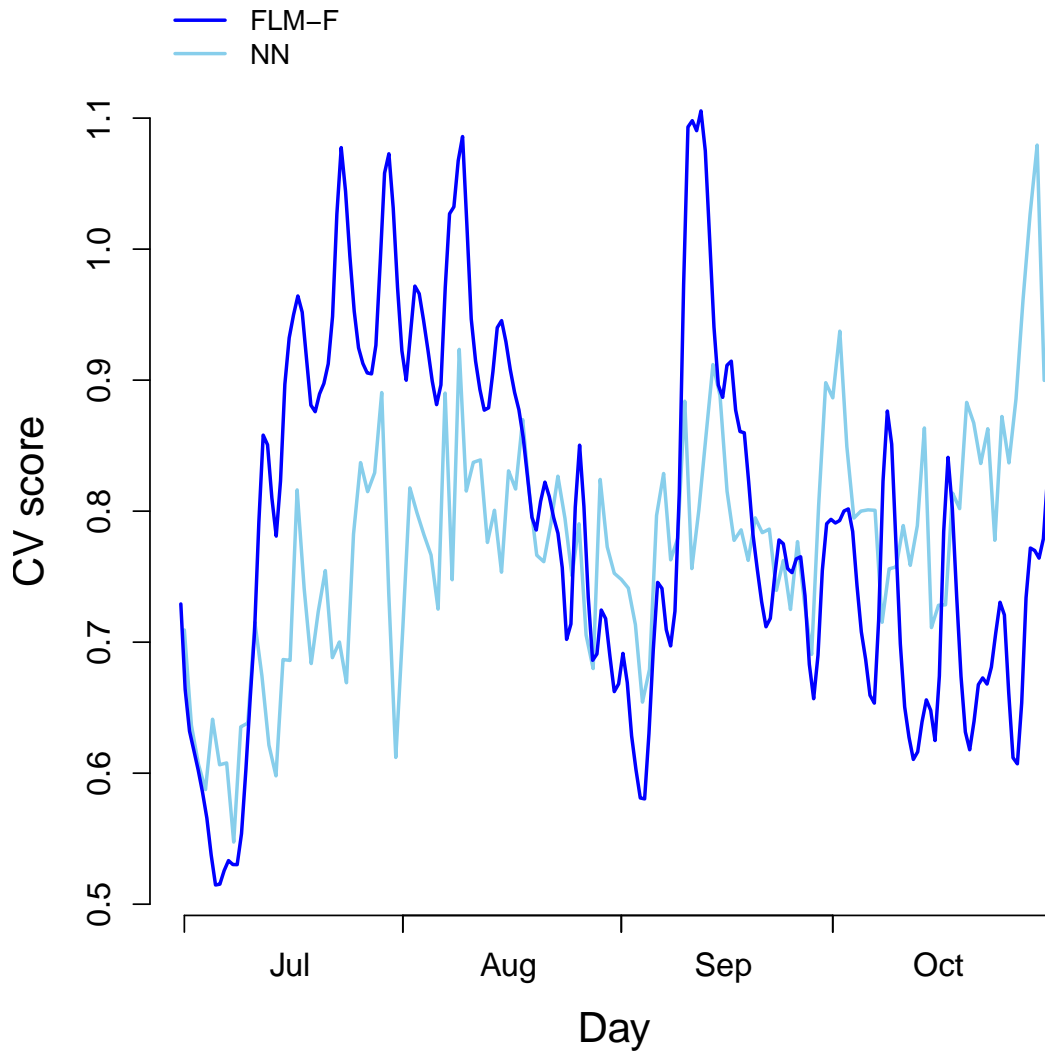


Figure 12: Mean prediction error curve estimated by CV over all years.