Bayesian modeling of hydrographs

Article *in* Water Resources Research · October 2007

DOI: 10.1029/2006WR005376

CITATIONS
8

READS
127

4 authors, including:

L. Perreault
Hydro-Québec Research Institute - IREQ
83 PUBLICATIONS **2,428** CITATIONS
SEE PROFILE

Anne-Catherine Favre
Grenoble Institute of Technology
120 PUBLICATIONS **5,176** CITATIONS
SEE PROFILE

WATER RESOURCES RESEARCH, VOL. 43, W10432, doi:10.1029/2006WR005376, 2007

Bayesian modeling of hydrographs

James Merleau,[1] Luc Perreault,[1] Jean-François Angers,[2] and Anne-Catherine Favre[3]

Received 25 July 2006; revised 27 February 2007; accepted 9 May 2007; published 26 October 2007.

[1] This article presents a new approach to model yearly hydrographs with daily or weekly streamflow measurements. The method considers yearly hydrographs as a sample of functions to be modeled nonparametrically in a Bayesian setting. The functional data analysis framework provides great flexibility to reproduce the features of yearly hydrographs, while the Bayesian probabilistic model ensures statistical coherence between the flood variables and the shapes of flood events. The proposed methodology is applied to two samples of hydrographs from two watersheds in the province of Quebec.

Citation: Merleau, J., L. Perreault, J.-F. Angers, and A.-C. Favre (2007), Bayesian modeling of hydrographs, *Water Resour. Res.*, *43*, W10432, doi:10.1029/2006WR005376.

1. Introduction

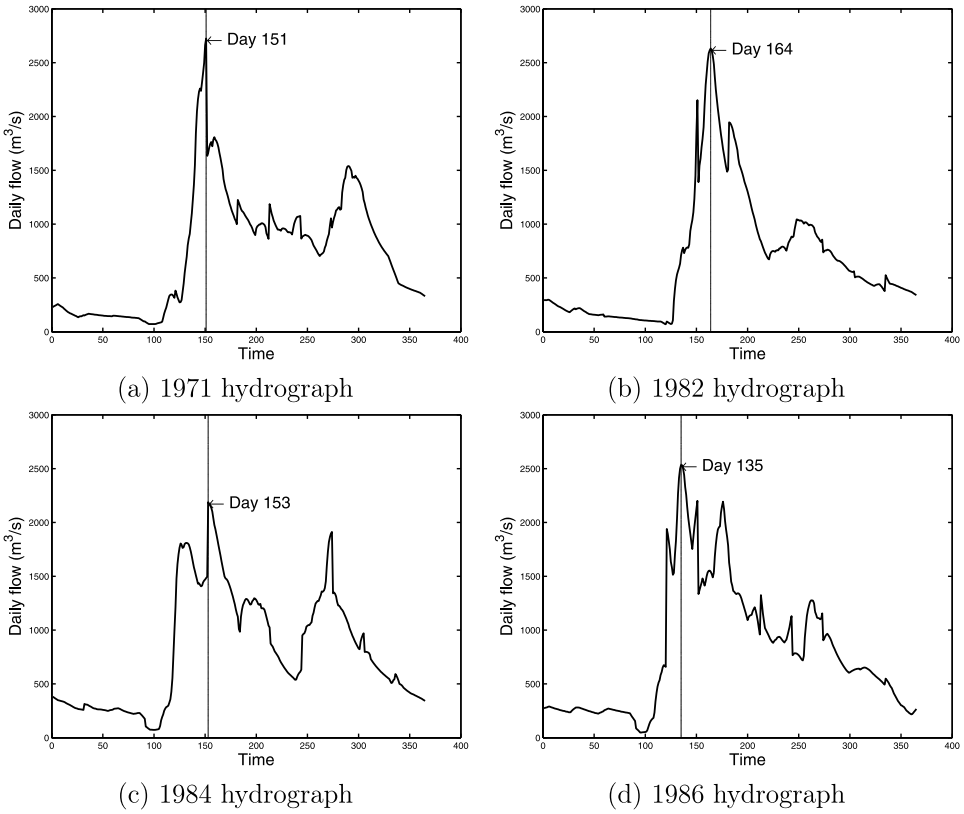
[2] Statistical modeling of hydrographs is important for many engineering purposes, in particular for energy planning and the design of power plants. Hydrographs are studied in these decision-making contexts to ensure good water management and human population safety. For example, modeling of extreme hydrographs is necessary for the construction of dams, which need to contain and evacuate large quantities of water. In this context, synthetic hydrographs, which preserve a realistic shape but simultaneously have extreme flood volumes and/or flood peaks, are of interest for engineering planning. A good model to simulate extreme hydrographs thus needs to reproduce hydrographs with the aforementioned characteristics. In a water management context, hydrograph modeling has to be able to fulfill two main purposes. The first of these is to obtain a reference hydrograph for a given river, while the second consists in generating synthetic hydrographs that can occur with a given probability. It is difficult to construct a reference hydrograph since key features of different yearly hydrographs for a given river will happen at different times of the year and these features will often vary regarding their shapes (see Figures 1 and 2). For the purpose of generating hydrographs, a good model needs to be flexible enough to encompass a large variety of shapes which can be encountered in practice, since water management decisions depend heavily on these shapes. Several techniques have been set forth to model and simulate hydrographs. Some of these focus on flood events while others attempt to capture the stochastic process, which governs water flow. The former methods usually model flood variables statistically, construct a design-flood hydrograph separately and combine the two levels of modeling to simulate hydrographs. The latter methods are based on a time series analysis and are most often used to simulate a diversity of possible hydrographs for a given time horizon.

[3] In the present paper, we propose a novel approach to model yearly hydrographs. Our method considers yearly hydrographs as a sample of functions to be modeled nonparametrically in a Bayesian setting. As will be shown, this functional data analysis framework offers the required flexibility to reproduce the characteristics of yearly hydrographs, but also provides a probabilistic model which ensures coherence between the flood variables and the shapes of flood events. Before exposing our new methodology, we will indicate the difficulties of conducting a statistical analysis of hydrographs and present the solutions that have been put forward in the literature.

[4] Figure 1 illustrates four yearly hydrographs with daily measurements, while the same four yearly hydrographs with weekly measurements are shown in Figure 2. All these hydrographs come from the same basin in northern Quebec. The first observation corresponds to the first measurement taken at the beginning of January, while the last observation corresponds to the last measurement at the end of December. The spring flood, mainly governed by snowmelting, is present on each of the four hydrographs and starts roughly around the 100th day of each year; autumn floods, governed by heavy rainfall, are also present and occur roughly between days 250 and 325. The four spring floods show a wide variety of shapes, intensity and duration; the time at which the flood peak happens, indicated by a vertical line, also varies between the different years. These differences are due to the climatic conditions and the amount of accumulated snow which vary from one year to the next; the presence of late spring liquid precipitations also affects the spring flood events and might cause secondary peaks. It is interesting to contrast the hydrographs of Figures 1 and 2 regarding some of their main characteristics. The hydrographs of Figure 2, with weekly time increments, are obviously smoother than the ones represented in Figure 1 (daily time increments), which causes the flood peaks to be flatter in Figure 2, especially for the hydrograph illustrated in Figure 2a. It is important to note that the main flood structures in Figure 1 can also be seen in Figure 2, although attenuated in certain cases. We thus see that complex structures are present for hydrographs with both daily and

[1] Institut de recherche d'Hydro-Québec, Varennes, Quebec, Canada.
[2] Département de mathématiques et de statistique, Université de Montréal, Montréal, Quebec, Canada.
[3] Chaire en hydrologie statistique, INRS, Eau-Terre et Environnement, Quebec, Quebec, Canada.

0043-1397/07/2006WR005376$09.00



(a) 1971 hydrograph (b) 1982 hydrograph

(c) 1984 hydrograph (d) 1986 hydrograph

Figure 1. Four yearly hydrographs with daily measurements. On each plot, the vertical line indicates the day at which the annual peak occurred. (a) 1971 hydrograph; (b) 1982 hydrograph; (c) 1984 hydrograph; (d) 1986 hydrograph.

weekly time increments. An adequate hydrograph model therefore needs to capture these structures.

[5] Flood peak, flood volume and flood duration are considered to be the main variables that summarize flood events and they are usually studied statistically through univariate or bivariate flood frequency analysis [*Yue et al.*, 1999; *Javelle et al.*, 2002]. It is clear that studying the statistical distributions of these three variables is of major importance, but as pointed out by *Yue et al.* [2002], it is not enough to fully describe flood events because of the impact of their shapes in a water management situation. The approach often adopted in practice is to do a flood frequency analysis and proceed to calculate return periods for the different flood variables. Separately, one of the following construction methods is used to create a reference flood hydrograph and it is then adjusted to have the properties with the desired return periods.

[6] A comprehensive review of the different methods to construct a design-flood hydrograph is given by *Yue et al.* [2002] and the interested reader is referred to the article for further details. Adopting the four categories listed by Yue et al., the construction methods are the traditional unit hydrograph (TUH) methods, the synthetic unit hydrograph (SUH) methods, the typical hydrograph (TH) methods and the statistical (S) methods. The TUH and SUH methods are based on hydrological principles. The TUH methods assume that the runoff response to rainfall is time invariant and that this response is linear as a function of rainfall. The SUH methods are based on empirical relationships that appear to exist between the parameters of a unit hydrograph and the physical characteristics of a drainage basin. A substantial number of articles have been devoted to the TUH methods [*Sherman*, 1932; *Doodge*, 1959; *Chow*, 1964; *Chow et al.*, 1988; *Pilgrim and Cordery*, 1993; *Yue and Hashino*, 2000] and will not be discussed further here; the same applies to the SUH methods [*Snyder*, 1938; *U.S. Soil Conservation Service US-SCS*, 1985]. In fact, these approaches are not designed to produce realistic synthetic hydrographs for basins in northern regions like Quebec where major floods are not the result of rainfall but mainly come from snowmelting at the onset of spring. It is precisely for this reason that engineers in northern countries have relied on the TH methods.

[7] The TH methods [*Nezhikhovsky*, 1971; *Sokolov et al.*, 1976] are widely used by practitioners. In this approach, a typical flood hydrograph, usually the one with the highest peak or the largest volume, is chosen from a river's sample of flood hydrographs. Each water flow value of the chosen flood hydrograph is then multiplied by a constant in order to get a flood peak and/or a flood volume corresponding to a given return period. This method considers the flood of the hydrograph as a function but it relies on a single historical realization. Therefore it does not use all the information available in the sample of historical flood hydrographs.

[8] The S methods, which include the approach put forward by *Yue et al.* [2002], consist of modeling the shape of each flood event by a probability density function, usually a gamma or a beta distribution. Yue et al. pursue

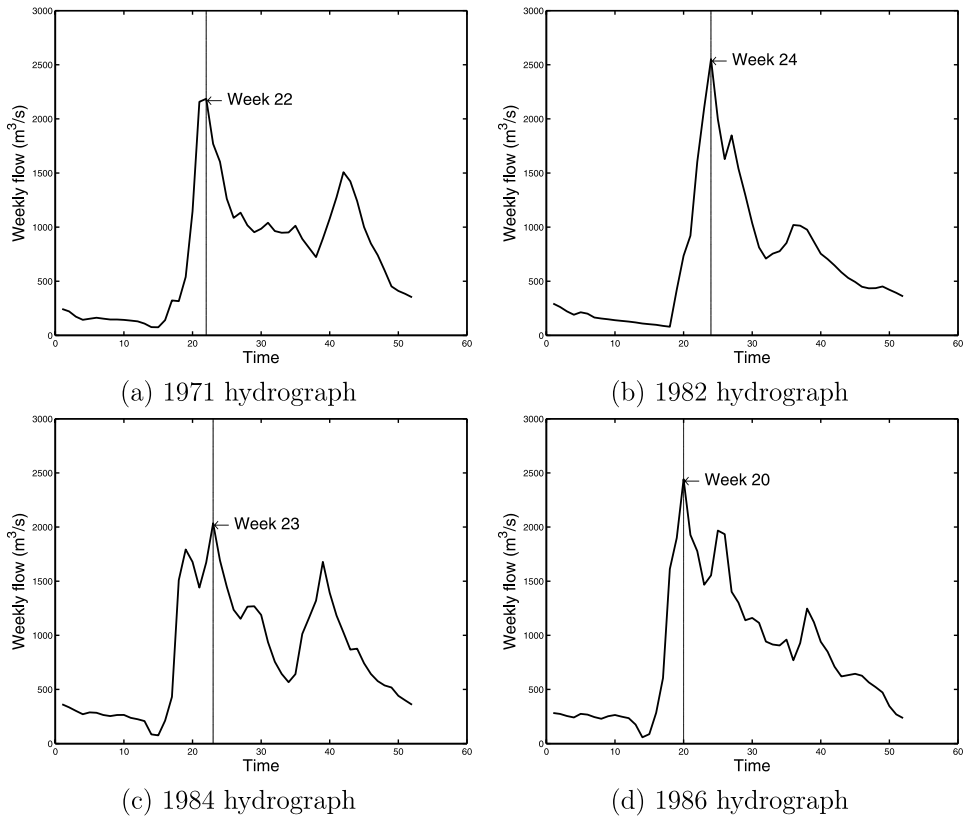


Figure 2. Four yearly hydrographs with weekly measurements. On each plot, the vertical line indicates the week at which the annual peak occurred. (a) 1971 hydrograph; (b) 1982 hydrograph; (c) 1984 hydrograph; (d) 1986 hydrograph.

this methodology further by studying shape variables of the adjusted beta distributions to the flood hydrographs. The shape variables, namely the shape mean and the shape standard deviation, are then considered as independent random variables and are each statistically modeled by a lognormal distribution. This enables the authors to consider return periods for the two shape variables. While incorporating a better probabilistic component to the problem by modeling the shape of flood events by two variables which are analyzed in a probabilistic framework, it seems to us that it is necessary to go further by considering hydrographs and their flood events as complex functions, and not restrict the shape of a flood to a model containing only two parameters.

[9] Finally, modeling techniques based on time series are mostly used to generate a wide range of hydrographs which are considered to be statistically probable scenarios. Because of the complexity of the underlying processes, the time series models often need to include numerous parameters to capture the observed statistical properties of hydrographs. Periodic autoregressive moving average (PARMA) models [*Salas et al.*, 1980; *Salas et al.*, 1982; *Vecchia et al.*, 1983; *Rasmussen et al.*, 1996] or PARMAX models [*Perreault and Latraverse*, 2001; *Ouhib*, 2005], which include explanatory variables, seem to be able to reproduce observed properties of hydrographs. However, the period of these models is usually taken to be the time increment of the series, therefore leading to an excessively large number of parameters for daily or weekly data. Furthermore, these methods cannot simulate hydrographs with fixed flood volumes and/or flood peaks because of their stochastic nature.

[10] Statistical modeling of hydrographs is a complex multivariate problem since the objective is to reproduce the characteristics of a sample of functions. Yearly hydrographs, and their flood events, constitute complex functional data and should therefore be analyzed statistically in a functional data analysis framework [*Ramsay and Silverman*, 2005]. For instance, it should be clear that the flood events illustrated in Figure 1 could not be reproduced by only one beta or gamma distribution since these distributions are unimodal functions. One could complexify the S methods by using a mixture of probability distribution functions [*Titterington et al.*, 1985] but even this approach seems unsatisfactory for the task at hand. Moreover, the S approach lacks cohesion since the flood characteristics such as the peak and volume are studied through flood frequency analysis, while the flood event shapes are modeled separately using a probability distribution function. The new method proposed in this paper brings forward an integrated approach in which hydrographs are modeled as functions in a probabilistic framework. This ensures statistical coherence between important characteristics of hydrographs, like flood peaks and flood volumes, and the shapes of the hydrographs.

[11] In the next section, the tools of functional data analysis which we use in this study are put forward. We first describe an approach based on landmark registration to make the individual hydrographs of a given river similar on

the time domain. We then set up a general nonparametric regression framework based on regression spline functions; this framework offers the modeling power and flexibility, which are necessary to capture the different shapes of hydrographs. The Bayesian probabilistic model is exposed in section 3 and the methodology is applied to the data in section 4.

2. Functional Data Analysis Context

[12] Functional data analysis is often concerned with modeling longitudinal data, that is data formed by a collection of repeated measurements of a response variable on a certain experimental unit or individual. Longitudinal data are frequently encountered in the life sciences where it is often the case that a response variable is studied on several individuals through time. Some examples are growth curves, the effect of a treatment as a function of time on patients, etc. In analogy with longitudinal data, we consider each year as an experimental unit for which we have repeated water flow measurements.

[13] We have, for each experimental unit i, the following observations:

$$(x_{i,1}, y_{i,1}), \ldots, (x_{i,j}, y_{i,j}), \ldots, (x_{i,n_i}, y_{i,n_i}),$$

where $x_{i,j}$ can be an explanatory variable or the time at which the response variable $y_{i,j}$ has been measured. We assume that $x_{i,j}$ is a deterministic variable, while $y_{i,j}$ is the random variable to be modeled. In our modeling context, $x_{i,j}$ is the time at which the water flow $y_{i,j}$ is measured for the year i; furthermore, we have $x_{i,j} = x_j$ and $n_i = n$ for every i since the measurements in our case are always taken at the same time increments, either every day (n = 365) or every week (n = 52). Our data for year i are therefore of the following form:

$$(x_1, y_{i,1}), \ldots, (x_j, y_{i,j}), \ldots, (x_n, y_{i,n}),$$

where i = 1,..., N and N represents the number of yearly hydrographs in our sample.

[14] As will be seen in section 2.2, each observed yearly hydrograph can be modeled with a nonparametric model. This is not the course we pursue in the present paper because we want to tackle another important issue, namely to obtain a hydrograph which is representative of a sample of hydrographs originating from a given river; in other words, we seek to model the underlying average process of a sample of hydrographs, which we refer to as a representative or reference hydrograph.

2.1. Landmark Registration

[15] The average of the four yearly hydrographs shown in Figure 1 (daily flow) is given in Figure 3a, while Figure 3c gives the average of the yearly hydrographs of Figure 2 (weekly flow). It is clear that the mean hydrographs do not have flood events representative of those illustrated in Figures 1 and 2. For most rivers in northern Quebec, the average of observed hydrographs, whether for daily or weekly measurements, cannot be used as a reference hydrograph. This average can be useful for volume analyses since it is indicative of the mean water flow during a certain period of the year, but it is not indicative of peak flows or of flood events shapes.

[16] In order to model a reference hydrograph, we use landmark registration which has been studied by *Kneip and Gasser* [1988, 1992] in a statistical context. The key idea behind registration is to transform the independent variable x in the present context, in order to make the yearly hydrographs similar on the domain of the transformed variable. For our purposes, this comes down to performing a time transformation such that the yearly hydrographs have important features occurring at simultaneous times; for example, it is possible to perform time registration which makes all the flood peaks of the yearly hydrographs happen at the same time of the year. Specifically, landmark registration consists in identifying salient features of a sample of functions and using these landmarks to execute the registration. We want to go from the original time x to a registered time t, and therefore from the observations $(x_j, y_{i,j})$ to the registered observations $(t_{i,j}, y_{i,j})$, where $t_{i,j} = g_i(x_j)$ and $g_i(\cdot)$ is the registration function for year i. We note that the registration function should, at least intuitively, contain information on the climatic conditions of a given year i, a possibility which we are currently studying.

[17] For the transformations to be one-to-one, the registration functions need to be monotonically increasing. Furthermore, we constrain the functions to transform the times at which important features happen to specified times. We thus have a sequence of constraints of the following form:

$$t_{i,\nu} = g_i(x_\nu) = \tau_\nu, \tag{1}$$

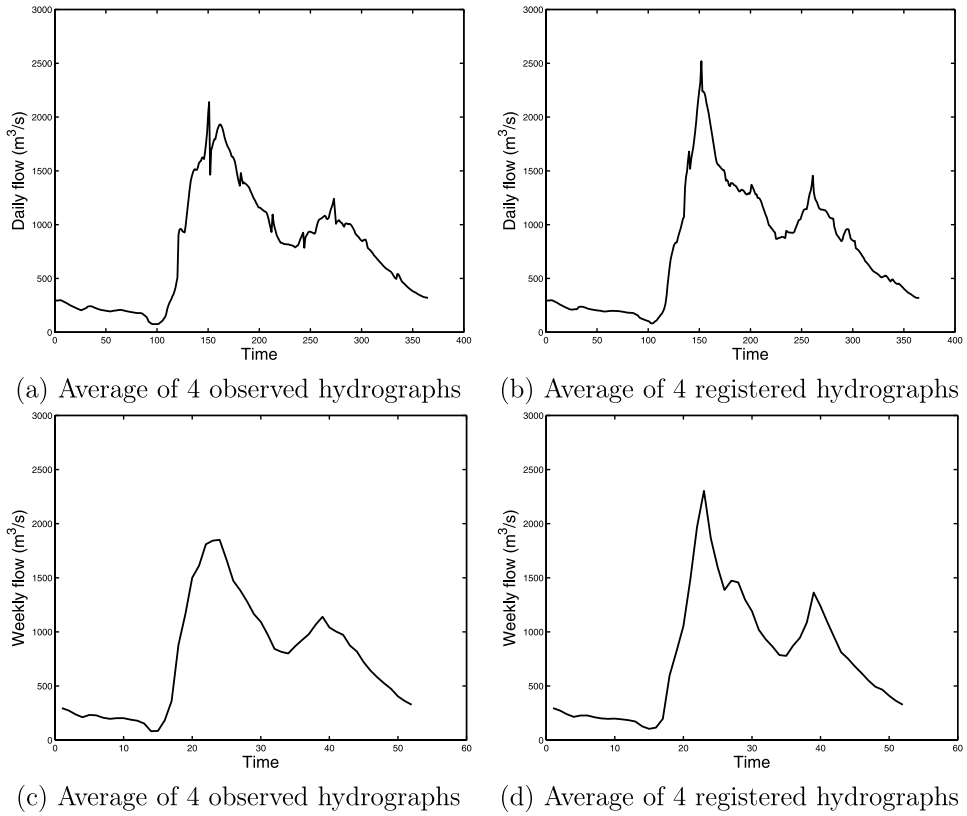
where x_ν represents the time at which the landmark ν occurs for year i and τ_ν is the specified time at which the landmark ν happens, for all years, in the transformed time domain.

[18] The registration functions can be modeled by several methods: a Taylor expansion approximation [*Angers et al.*, 2004], interpolating splines [*Kneip and Gasser*, 1992] or an approach such as the one suggested by *Ramsay and Li* [1998]. Here we consider each function to be made up of linear parts $L_p(x)$ and we then have

$$g_i(x) = \sum_{p=1}^{P} L_p(x) I_{D_p}(x) = \sum_{p=1}^{P} (c_{p,0} + c_{p,1}x) I_{D_p}(x), \tag{2}$$

where D_p is the domain for which the linear function $L_p(x)$ is nonzero, $I_{D_p}(x) = 1$ for $x \in D_p$ and 0 otherwise, and P represents the number of parts of the registration function. This simple model possesses an exact solution when the continuity of the registration function is imposed; it also satisfies the monotonicity criterion as long as the landmarks are events that occur in the same sequence every year. Furthermore, this type of registration function generally performs well for preserving flood event volumes [*Merleau et al.*, 2005].

[19] We will now illustrate the use of landmark registration to obtain a reference hydrograph for the hydrographs shown in Figures 1 and 2. We choose the four following events as landmarks: the first measurement of the year, the peak of the spring flood, the peak of the fall flood and the



(a) Average of 4 observed hydrographs (b) Average of 4 registered hydrographs

(c) Average of 4 observed hydrographs (d) Average of 4 registered hydrographs

Figure 3. Daily measurements: (a) average of four observed hydrographs, (b) average of the same four hydrographs after registration. Weekly measurements: (c) average of four observed hydrographs, (d) average of the same four hydrographs after registration.

last measurement of the year. We then have the following constraints for the registration function of year i:

$$g_i(L_x) = L_x, g_i(x_s) = \tau_s, g_i(x_f) = \tau_f, g_i(U_x) = U_x, \qquad (3)$$

where L_x and U_x are, respectively, the lower and upper bounds of the domain of x; x_s and x_f are the times, for year i, at which the peak of the spring flood and the peak of the fall flood, respectively, happened; and τ_s and τ_f are the specified times at which the spring flood peak and the fall flood peak are fixed to occur in the domain of the synchronous time τ. We fix τ_s and τ_f to be the median values of the observed x_s and x_f. Figure 4a shows the registration function for the yearly hydrograph given in Figure 2b (weekly flow). Figure 4b illustrates the effect of the registration function on the observed hydrograph. From the constraints given in equation (3), the registration function given in equation (2) is made up of three linear parts. The slope of a given part, $c_{p,1}$, determines if the corresponding section of the hydrograph is stretched ($c_{p,1} > 1$) or contracted ($c_{p,1} < 1$). In Figure 4, the middle section is stretched, while the first and last sections are contracted.

[20] Figure 3b shows the average obtained after the registration for the hydrographs of Figure 1 (daily flow), and Figure 3d shows the average of the registered hydrographs of Figure 2 (weekly flow). If we compare the average registered hydrographs with their observed counterparts, it is clear that registration makes the average hydrograph more representative of a sample of hydrographs. The spring floods in Figures 3b and 3d are much better defined and closer to the observed ones than those illustrated in Figures 3a and 3c. Furthermore, the peak value of the average spring floods, after registration, is the real average of the four observed hydrographs because of the way the registration is performed. We also notice the presence of secondary spring flood peaks in Figures 3b and 3d, which can also be seen in Figures 1 and 2.

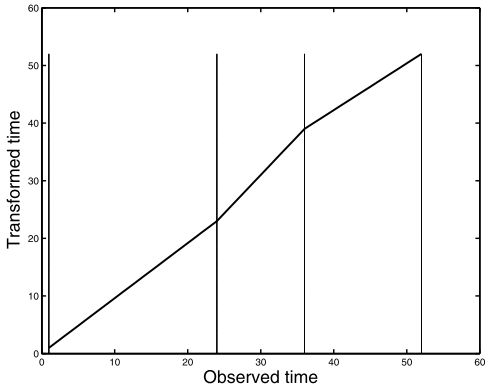
2.2. Nonparametric Regression With Spline Functions

[21] In our functional data analysis context, we assume that

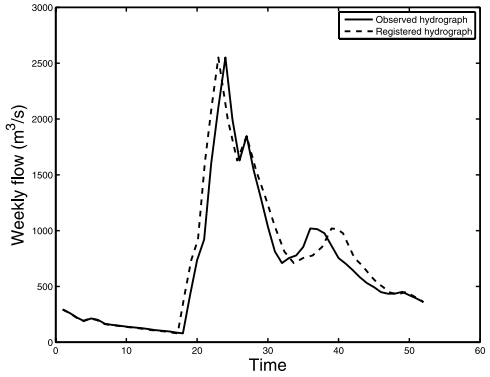
$$y_{i,j} = h_i(t_{i,j}) + \varepsilon_{i,j}, \qquad (4)$$

where $h_i(t_{i,j})$ is a continuous function evaluated at $t_{i,j}$ and $\varepsilon_{i,j}$ is an error term. We therefore go from the data points $(t_{i,j}, y_{i,j})$ $(j = 1,\ldots, n)$ to a functional representation: $(t, h_i(t))$, for $t \in D_t = [L_t, U_t]$ where L_t and U_t represent, respectively, the lower and upper bounds of the t domain. In the present paper, we seek to model the average process which underlies yearly hydrographs and we therefore assume that $h_i(\cdot) = h(\cdot)$ for all i.

[22] Several methods exist to estimate the function $h(\cdot)$: kernel methods [*Hastie and Tibshirani*, 1990], Fourier series, spline based methods [*Ramsay and Silverman*, 2005], wavelet methods [*Ogden*, 1997], etc. We choose to work with regression polynomial spline functions as a basis to evaluate the functions of interest because this type of



(a) Registration function



(b) Effect of registration on the observed hydrograph

Figure 4. (a) Registration function for 1982 hydrograph; (b) effect of the registration function on the observed 1982 hydrograph. The vertical lines in Figure 4a represent L_x, x_s, x_f and U_x (see equation (3)).

basis possesses good mathematical properties such as differentiability and integrability, the latter property being useful in the present context as will be seen shortly. It also offers good flexibility and leads to parsimonious models when free-knots are used.

[23] The function $h(t)$ is estimated by

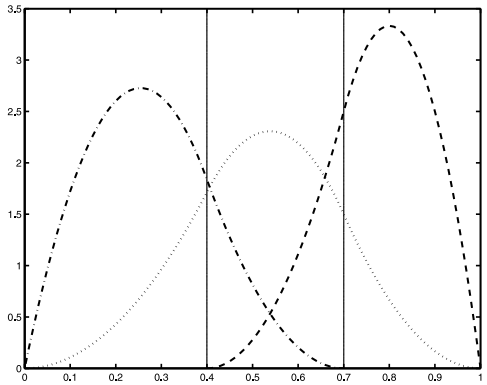
$$h_{\boldsymbol{\omega}}(t) = \sum_{k=1}^{K_\omega} \beta_{k,\boldsymbol{\omega}} b_{k,\boldsymbol{\omega}}(t), \tag{5}$$

where $\beta_{k,\boldsymbol{\omega}}$ represents the coefficient of the basis element $b_{k,\boldsymbol{\omega}}(\cdot)$. Once the order l of the spline polynomial functions is fixed, the basis elements $b_{k,\boldsymbol{\omega}}(\cdot)$ are determined by the number of interior knots, m, and the ordered location of these knots, $\boldsymbol{r}^{(m)} = (r_1, \ldots, r_m)$; the number of elements in the basis is given by $K_\omega = l + m$. We summarize this information in the model parameter $\boldsymbol{\omega} = (m, \boldsymbol{r}^{(m)})$. The determination of $\boldsymbol{\omega}$ is a model selection problem, which is discussed in section 3.2. We note that this model can be understood as a linear model such as those encountered in linear regression. The model given in equation (4) can now be written as

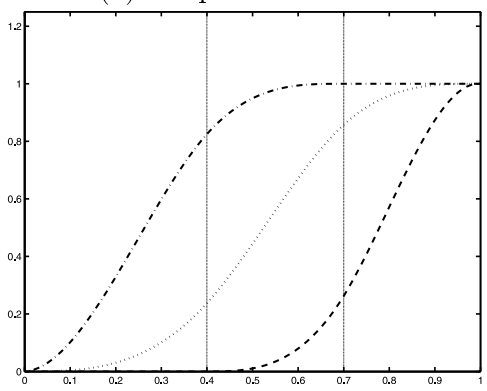
$$\boldsymbol{y}_i = \boldsymbol{B}_\omega \boldsymbol{\beta}_\omega + \boldsymbol{\varepsilon}_i, \tag{6}$$

where $\boldsymbol{y}_i = (y_{i,1}, \ldots, y_{i,n})'$ is an $n \times 1$ vector; $\boldsymbol{B}_\omega = (\boldsymbol{b}_\omega(t_{i,1}), \ldots, \boldsymbol{b}_\omega(t_{i,n}))'$ is an $n \times K_\omega$ matrix, with $\boldsymbol{b}_\omega(t_{i,j}) = (b_{1,\omega}(t_{i,j}), \ldots, b_{K_\omega,\omega}(t_{i,j}))'$ as a $K_\omega \times 1$ vector of the basis elements evaluated at $t_{i,j}$; $\boldsymbol{\beta}_\omega = (\beta_{1,\omega}, \ldots, \beta_{K_\omega,\omega})'$ is a $K_\omega \times 1$ vector of parameters; and $\boldsymbol{\varepsilon}_i = (\varepsilon_{i,1}, \ldots, \varepsilon_{i,n})$ is an $n \times 1$ vector of error terms.

[24] Figure 5 shows M-spline and I-spline functions for fixed order ($l = 3$) and fixed model parameter: $\boldsymbol{\omega} = (2, (0.4, 0.7))$, which corresponds to two interior knots positioned at 0.4 and 0.7. In this case, there are five basis elements, i.e., $K_\omega = 5$; for clarity, only three of these elements are illustrated in Figure 5. For the current knot configuration, Figure 5a shows three M-spline functions, $b^{\mathrm{M}}_{k,\omega}(t)$, and the corresponding I-spline functions, $b^{\mathrm{I}}_{k,\omega}(t)$, are given in Figure 5b. Each M-spline function is made up of polynomial parts of order l, or degree $(l - 1)$; in the current example, the polynomial parts are quadratic. Each I-spline function is an integrated M-spline function and constitutes a monotonically increasing function; therefore I-spline functions are well suited to model monotone functions, as indicated by *Ramsay* [1988]. M-spline functions are closely related to B-spline functions which are widely used in statistics [*He and Shi*, 1998]. We note that the M-spline functions integrate to 1 and are, in that respect, equivalent to probability distribution functions used in the S methods (see Introduction). To pursue the parallel further, the S methods use only one basis element to model a flood event, while our approach uses several basis elements to give a representation of a yearly hydrograph.



(a) M-spline functions



(b) I-spline functions

Figure 5. Three M-spline and three I-spline functions for $\boldsymbol{\omega} = (2, (0.4, 0.7))$. The vertical lines indicate the positions of the interior knots. (a) M-spline functions; (b) I-spline functions.

[25] As mentioned previously, I-spline functions form a good basis to model monotone functions. Since a yearly hydrograph is a positive function, the function

$$H(\tau) = \int_{L_t}^{\tau} h(t)\mathrm{d}t, \tag{7}$$

is a monotone increasing function and it represents the integrated water flow from the beginning of the year to a certain time τ. This cumulative hydrograph is of particular interest for conducting volume analyses. If a hydrograph $h(t)$ is modeled with M-spline functions, $b^{\mathrm{M}}_{k,\omega}(t)$, in equation (5), then a model for $H(t)$ is obtained simultaneously and it is given by

$$H_\omega(t) = \sum_{k=1}^{K_\omega} \beta_{k,\omega} b^{\mathrm{I}}_{k,\omega}(t), \tag{8}$$

where each coefficient $\beta_{k,\omega}$ is the same for the two functional representations $h_\omega(t)$ and $H_\omega(t)$. This result follows from the fact that each coefficient $\beta_{k,\omega}$ is independent of t and each I-spline function is an integrated M-spline.

3. Bayesian Statistical Model

[26] We adopt the Bayesian paradigm instead of the frequentist approach for several reasons, some of which we now put forward. It enables the statistician to take into account, in a coherent probabilistic framework, the uncertainty related to all the parameters of the model and thus gives a more adequate representation of the uncertainty concerning a model; frequentist approaches can often underestimate this uncertainty. As will be seen in section 3.2, it makes the model selection procedure formal in the sense that the selection follows directly from the initial assumptions about the probabilistic model; therefore it does not rely on some ad hoc procedure. If we would choose to do so, it is easy in the Bayesian framework to include constraints on the parameter space through the a priori statistical distributions, thus making the implementation of constraints straightforward. Furthermore, the inclusion of expert opinion can also be included through the prior statistical distribution; for example, it would be possible to consult hydrologists to obtain a given shape of a hydrograph and to transform this shape into the coefficient space of the spline functions. Although we have not done this in the paper, we are currently thinking of incorporating this aspect in our model.

[27] In order to treat a certain function $h(t)$ as a random functional event, we can regard the parameters ($\beta_{k,\omega}$) of equation (5) as random variables. In a Bayesian framework, the parameters, $\boldsymbol{\theta}$, of a given model are considered to be random variables which are drawn from a certain probability distribution. A Bayesian statistical model is made up of the following two elements [e.g., *Lee*, 1989; *Bernardo and Smith*, 1994]: a prior probability distribution for the model parameters, $\pi(\boldsymbol{\theta})$, and a probability distribution function, $f(\boldsymbol{y}|\boldsymbol{\theta})$, from which the observations arise. A prior distribution is a probabilistic formulation of the information available before observations are collected. From these two probability distributions, the posterior distribution associated with the model parameters can be obtained by applying Bayes' theorem:

$$\pi(\boldsymbol{\theta}|\boldsymbol{y}) = \frac{f(\boldsymbol{y}|\boldsymbol{\theta})\pi(\boldsymbol{\theta})}{\int_{\Theta} f(\boldsymbol{y}|\boldsymbol{\theta})\pi(\boldsymbol{\theta})\mathrm{d}\boldsymbol{\theta}} = \frac{f(\boldsymbol{y}|\boldsymbol{\theta})\pi(\boldsymbol{\theta})}{\mathrm{m}(\boldsymbol{y})}, \tag{9}$$

where $\mathrm{m}(\boldsymbol{y})$ is the marginal distribution of $\boldsymbol{y}$, i.e., the statistical distribution of $\boldsymbol{y}$ after the model parameters have been integrated, and $\boldsymbol{\Theta}$ is the parameter space. All statistical inferences concerning the parameters are made from $\pi(\boldsymbol{\theta}|\boldsymbol{y})$, which represents a probabilistic model for the parameters that has been updated by the empirical information.

3.1. Distributional Hypotheses

[28] We assume that the distribution of each vector $\boldsymbol{\varepsilon}_i$ is a multivariate normal distribution (see Appendix A for the definitions of the probability distributions used in this section):

$$\boldsymbol{\varepsilon}_i \sim \mathrm{N}_n\left(0, \sigma^2\boldsymbol{\Sigma}_\varepsilon\right), \tag{10}$$

where σ^2 is a variance proportionality constant and $\boldsymbol{\Sigma}_\varepsilon$ is the covariance matrix which captures the covariance, or correlation, structure between the elements of $\boldsymbol{\varepsilon}_i$. We note that σ^2 and $\boldsymbol{\Sigma}_\varepsilon$ are taken to be the same for every year i. In the Application section, we consider the elements of each $\boldsymbol{\varepsilon}_i$ to be independent and identically distributed; more formally, we have $\mathrm{Cov}(\varepsilon_{i,j}, \varepsilon_{i,j'}) = 0$ for $j \neq j'$ and $\varepsilon_{i,j} \sim \mathrm{N}_1(0, \sigma^2)$ for all j, where Cov represents the covariance operator and N_1 indicates a one-dimensional normal distribution. This leads to the following prescription, A1: $\boldsymbol{\Sigma}_\varepsilon = \mathbb{I}_n$, where $\mathbb{I}_n$ is an n-dimensional unit diagonal matrix and A1 indicates the first application assumption. It should be noted that under A1, the variance is taken to be the same throughout the domain of the hydrograph, an assumption which is discussed further in the conclusion.

[29] The probability distribution function of each vector of observations, $\boldsymbol{y}_i$, is then a multivariate normal distribution which we can write as:

$$\boldsymbol{y}_i|\boldsymbol{\beta}_\omega, \sigma^2 \sim \mathrm{N}_n\left(\boldsymbol{B}_\omega\boldsymbol{\beta}_\omega, \sigma^2\boldsymbol{\Sigma}_\varepsilon\right). \tag{11}$$

In the notation given above, we have the vector of observations $\boldsymbol{y} = (\boldsymbol{y}'_1, \ldots, \boldsymbol{y}'_N)$, and the vector of parameters $\boldsymbol{\theta} = (\boldsymbol{\beta}'_\omega, \sigma^2)'$ since we make the hypothesis that the design matrix $\mathbf{B}_\omega$ is fixed and that the covariance matrix $\boldsymbol{\Sigma}_\varepsilon$ is known. Since the yearly hydrographs are considered to be independent, the joint probability distribution of the observations is then given by: $f(\boldsymbol{y}|\boldsymbol{\theta}) = \prod_{i=1}^{N} f(\boldsymbol{y}_i|\boldsymbol{\theta})$. Our interest lies in the vector of regression parameters $\boldsymbol{\beta}_\omega$ which represents the reference hydrograph in functional space.

[30] We assume that the prior distribution can be written as

$$\pi\left(\boldsymbol{\beta}_\omega, \sigma^2\right) = \pi\left(\boldsymbol{\beta}_\omega|\sigma^2\right)\pi\left(\sigma^2\right), \tag{12}$$

and choose a conjugate model for the parameters, i.e., a probabilistic model with prior and posterior distributions from the same family of distributions. For normally

distributed observations, we have the following conjugate prior distributions:

$$\boldsymbol{\beta}_\omega|\sigma^2 \sim N_{K_\omega}\left(\boldsymbol{\beta}_\omega^0, \sigma^2\boldsymbol{\Sigma}_\omega\right), \tag{13}$$

$$\sigma^2 \sim \text{I}\Gamma\left(\frac{\alpha_\omega}{2}, \frac{\gamma_\omega}{2}\right), \tag{14}$$

where K_ω denotes the dimension of the spline basis (see equation (5)). Our a priori knowledge should be used to determine the hyperparameters: $\boldsymbol{\beta}_\omega^0$, the mean of the multivariate normal distribution, $\boldsymbol{\Sigma}_\omega$, the covariance structure between the components of $\boldsymbol{\beta}_\omega$, and the shape parameters of the inverse gamma distribution, α_ω and γ_ω.

[31] Because of our lack of prior knowledge concerning the hyperparameters and because the first N_0 yearly hydrographs are data of lesser quality, because they have been reconstructed, than the rest of the data set for the province of Quebec, we use these data to determine the prior distributions; the N remaining historical hydrographs are considered the effective sample. The hydrographs of the first N_0 years are of lesser quality but they nonetheless contain information about hydrographs for a given site. Although we do not want to treat these hydrographs as part of the effective sample, it is reasonable to use this information in our model but to weigh it properly. For the prior distribution of $\boldsymbol{\beta}_\omega$, the N_0 years are used to determine the location vector $\boldsymbol{\beta}_\omega^0$ (see section 4.2 for details). This is the second application assumption (A2). Concerning the covariance matrix, we assume that for a certain model defined by $\boldsymbol{\omega}$, we have A3: $\boldsymbol{\Sigma}_\omega = \frac{1}{n_0}(\boldsymbol{B}'_\omega\boldsymbol{\Sigma}_\varepsilon^{-1}\boldsymbol{B}_\omega)^{-1}$. This type of covariance structure was suggested by *Zellner* [1986] and can be justified in several ways [see *Robert*, 1994]. From a practical point of view, it makes the implementation of the model fairly straightforward since only one parameter, n_0, needs to be specified, instead of a covariance matrix of dimension $K_\omega \times K_\omega$; furthermore, this type of structure is well adapted to take into account multicollinearity, which can be the case when spline functions are used as a basis, since it allows for a large prior variance on the components affected by multicollinearity. As with the location vector $\boldsymbol{\beta}_\omega^0$, the determination of α_ω and γ_ω is done with the first N_0 hydrographs and also depends on n_0 (see section 4.2 for details). This is the fourth application assumption (A4). The factor n_0 can be viewed as an indicator of our confidence in the prior information, therefore we have $n_0 = zN_0$, where $0 \leq z \leq 1$. If one chooses $z = 0$, then it is assumed that the prior information contains no information and the prior distributions are improper; while if $z = 1$ is chosen, each of the N_0 yearly hydrographs contributes as much to the posterior distributions as one of the hydrographs in the effective sample. Since we know that the N_0 hydrographs are of lesser quality, z should lie somewhere between these two extreme cases. We have conducted tests, on real data, by varying the value of z and these tests have shown that its value does not have a serious impact on results. In the Application section, we use $z = 1/N_0$ which leads to $n_0 = 1$ (A5); the prior information then contributes the equivalent of one hydrograph from the effective sample.

[32] Our choice in prior distributions leads to a posterior distribution that can be written as

$$\pi\left(\boldsymbol{\beta}_\omega, \sigma^2|\boldsymbol{y}\right) = \pi\left(\boldsymbol{\beta}_\omega|\sigma^2, \boldsymbol{y}\right)\pi\left(\sigma^2|\boldsymbol{y}\right), \tag{15}$$

and by using standard Bayesian calculations for linear models [*Robert*, 1994], we have

$$\boldsymbol{\beta}_\omega|\sigma^2, \boldsymbol{y} \sim \text{N}_{K_\omega}\left(\boldsymbol{\beta}_\omega^*, \sigma^2\boldsymbol{\Sigma}_\omega^*\right), \tag{16}$$

$$\sigma^2|\boldsymbol{y} \sim \text{I}\Gamma\left(\frac{\alpha_\omega^*}{2}, \frac{\gamma_\omega^*}{2}\right). \tag{17}$$

The explicit expressions for $\boldsymbol{\beta}_\omega^*$, $\boldsymbol{\Sigma}_\omega^*$, α_ω^* and γ_ω^* are given in Appendix B1.

[33] By integrating out σ^2 in equation (16), the posterior distribution of $\boldsymbol{\beta}_\omega$, independent of σ^2, is given by a multivariate Student's t distribution:

$$\boldsymbol{\beta}_\omega|\boldsymbol{y} \sim T_{K_\omega}\left(\alpha_\omega^*, \boldsymbol{\beta}_\omega^*, \frac{\gamma_\omega^*}{\alpha_\omega^*}\boldsymbol{\Sigma}_\omega^*\right). \tag{18}$$

[34] Our main interest lies in the posterior probability distributions given in equations (17) and (18), since any statistical inference proceeds from these distributions. In the present hydrological context, in which we want to simulate hydrographs, we can generate vectors of parameters from the posterior distribution (18). A simulated hydrograph will then correspond to a simulated vector of parameters since a hydrograph is fully determined by a vector $\boldsymbol{\beta}_\omega$. Another statistical distribution that will prove useful in the following section is the marginal distribution of the observations, $\text{m}(\boldsymbol{y}|\boldsymbol{\omega})$, which has been defined in equation 9 and is given explicitly in Appendix B2.

3.2. Model Selection: Determining the Best $\boldsymbol{\omega}$

[35] In this section, a method is given to determine the spline basis model parameter $\boldsymbol{\omega} = (m, \boldsymbol{r}^{(m)})$, where m is the number of interior knots and the vector of ordered positions of these knots is $\boldsymbol{r}^{(m)}$. As pointed out in section 2.1, once the order of the spline polynomial functions is fixed, the parameter $\boldsymbol{\omega}$ fully determines the spline basis; therefore the determination of this parameter is crucial to obtain a good fit to the data. Many methods have been suggested in the literature to determine this parameter. Some simple methods position interior knots at given quantiles of the independent variable, while more sophisticated algorithms rely on forward, backward and stepwise methods to determine the best $\boldsymbol{\omega}$ [*Friedman and Silverman*, 1989; *Stone et al.*, 1997]. The different models obtained from the various parameters $\boldsymbol{\omega}$ are usually compared through a model fitting criterion such as Akaike's Information Criterion (AIC) [*Akaike*, 1973], the Schwarz criterion [*Schwarz*, 1978], cross-validation [*Hastie and Tibshirani*, 1990], etc. These criteria are all essentially structured in the same manner in that they "reward" goodness of fit and "penalize" model complexity in order to obtain a model that fits the data well but is still parsimonious.

[36] The method which we adopt to explore knot configurations, i.e., the support of $\boldsymbol{\omega}$, is based on an insight of *He and Shi* [1998] in their paper on modeling monotone

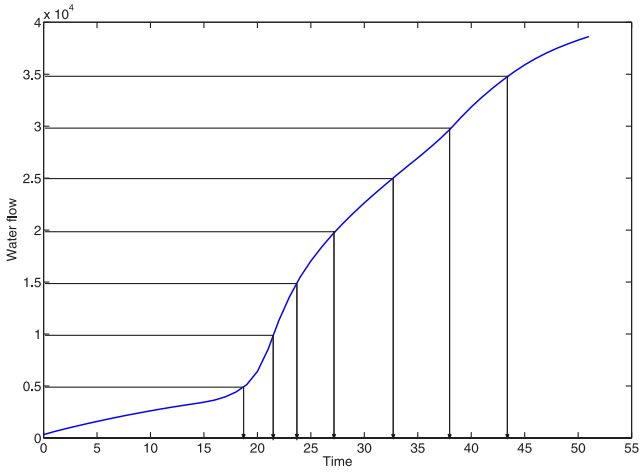


Figure 6. Illustration of He and Shi method for $m = 7$ interior knots.

functions with B-spline functions. Instead of positioning the knots at the quantiles of the independent variable, they use the quantiles of the monotone increasing function to perform a projection on the independent variable axis. The advantage of this simple method is that it positions more interior knots in the regions where the monotone function is rapidly changing; more basis elements need to be present in these regions to give more modeling flexibility where the function fluctuates the most. In the present context, the cumulative hydrograph given in equation (7), $H(\cdot)$, is used to perform this operation. An illustration of this procedure is shown in Figure 6. The water flow axis is subdivided into eight regular sections by seven markers; these are projected on the time axis using the monotone function. The interior knots are then taken to be the resulting time coordinates.

[37] A drawback with this technique is the fact that the vector $\boldsymbol{r}^{(m)}$ is solely determined by the number of interior knots m, which implies that a very limited number of knot configurations are explored. Furthermore, a fully Bayesian model would consider $\boldsymbol{\omega}$ to be a random quantity that follows a certain probability distribution. We are currently developing a fully Bayesian model similar to the approach suggested by *Denison et al.* [1998]. Nonetheless, the simple approach adopted here seems to work fairly well and it is also very effective regarding computational time.

[38] Finally, a method needs to be adopted in order to discriminate between the different knot configurations. In a frequentist modeling context, one of the approaches mentioned previously would need to be chosen arbitrarily; with the Bayesian approach, on the other hand, the model selection procedure follows directly from the hypotheses concerning the probabilistic model. The Bayesian model selection criterion, called the Bayes factor [*Kass and Raftery*, 1995], is given by the ratio of the marginal distributions of two competing models and indicates which of the models is more likely to be the best model. In the present context, let us say we are comparing model 1, $\boldsymbol{\omega}_1$, and model 2, $\boldsymbol{\omega}_2$, then the Bayes factor is given by

$$BF_{\boldsymbol{\omega}_1,\boldsymbol{\omega}_2} = \frac{\mathrm{m}(\boldsymbol{y}|\boldsymbol{\omega}_1)}{\mathrm{m}(\boldsymbol{y}|\boldsymbol{\omega}_2)}. \tag{19}$$

Model 1 is more likely to be the best model when $BF_{\boldsymbol{\omega}_1,\boldsymbol{\omega}_2} > 1$, while $BF_{\boldsymbol{\omega}_1,\boldsymbol{\omega}_2} < 1$ indicates that model 2 has a higher probability to be the best model. For the current modeling context, an explicit expression for the Bayes factor is given in Appendix B3.

3.3. Bayesian Estimator and Confidence Intervals

[39] It is a well known result of Bayesian decision theory that the decision rule concerning a parameter under a squared error loss function is given by the expected value of this parameter [see *Robert*, 1994]. Under squared error loss, the decision rule concerning $\boldsymbol{\beta}_{\boldsymbol{\omega}}$ is to choose its expected value which is given by $\boldsymbol{\beta}^*_{\boldsymbol{\omega}}$. Once the best model $\boldsymbol{\omega}$ has been determined by the method discussed in the previous section, the Bayesian estimator of a representative hydrograph will therefore be as follows:

$$h^*_{\boldsymbol{\omega}} = \mathbf{B}_{\boldsymbol{\omega}}\boldsymbol{\beta}^*_{\boldsymbol{\omega}}. \tag{20}$$

[40] Knowing that the posterior distribution of $\boldsymbol{\beta}_{\boldsymbol{\omega}}$ given in equation (18) is a multivariate Student's t distribution, it is possible to construct confidence intervals for linear combinations of the components of $\boldsymbol{\beta}_{\boldsymbol{\omega}}$ by using standard multivariate results [see *Johnson and Wichern*, 1992]. The reader can refer to Appendix B4 for the mathematical expression.

4. Application

4.1. Data

[41] The streamflow data analyzed in this section are weekly net basin supplies for two basins situated in Quebec and managed by Hydro-Québec, a public company that produces, transmits and distributes electricity throughout the province of Quebec. Hydro-Québec currently operates 54 power plants supplied by 26 large reservoirs; the major watersheds managed by Hydro-Québec are shown in Figure 7. In this paper, we focus on the statistical modeling of yearly hydrographs from two different basins: Churchill Falls, which is located in northern Quebec and has a basin area of 69,345 km^2, and Gouin, which is located in southern Quebec and has a basin area of 9376 km^2. These two watersheds serve as test basins to explore the potential use of the approach proposed in the paper. For each watershed, a sample of weekly streamflows (in m^3/s) covering the period extending from 1950 to 2002 is considered (hydrologic data post 2002 being confidential).

[42] Figure 8 shows five consecutive annual hydrographs with weekly streamflow for each watershed. Figure 8a shows the annual hydrographs observed at Churchill Falls during the 1989–1993 period. The sequence of hydrographs starts with the first week of January 1989 and ends with the last week of December 1993. Figure 8b illustrates five annual hydrographs at Gouin for the period extending from 1996 to 2000. In this case, the series starts with the first week of January 1996 and ends with the last week of December 2000. We notice that the two sequences possess some similarities and differences. They are similar in that the annual spring floods are more important than the autumn floods; but they differ in their level of "smoothness". The yearly hydrographs of Churchill Falls (Figure 8a) possess very well defined spring and autumn floods which do not show strong variations; the yearly hydrographs of Gouin

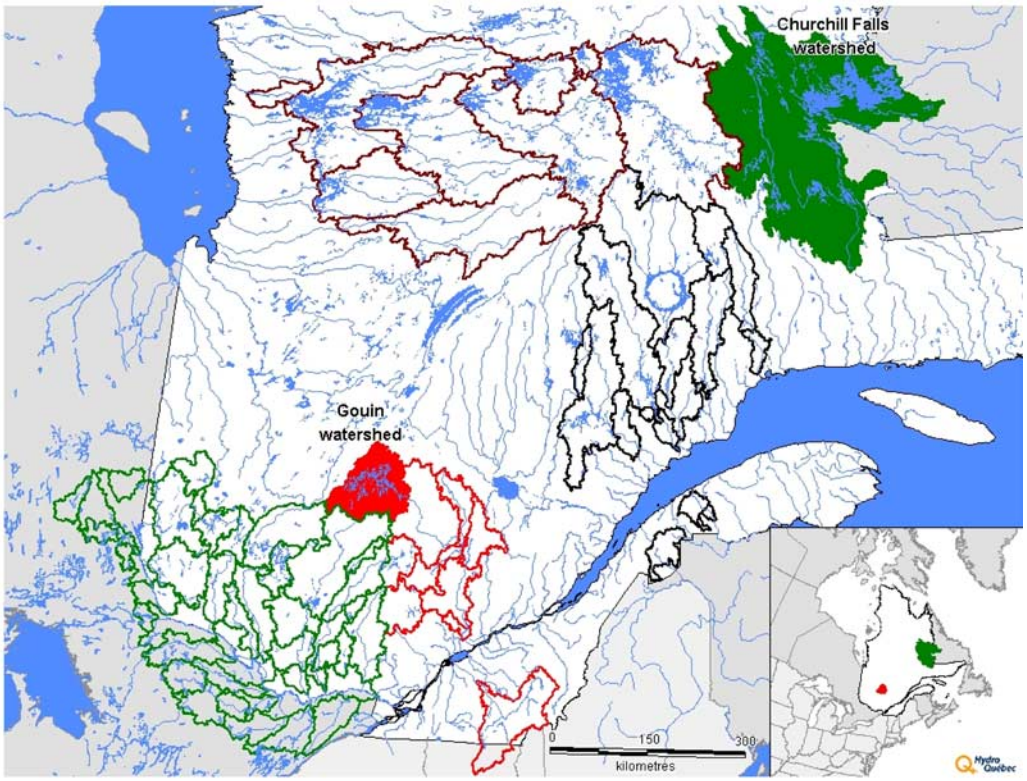


Figure 7. Location of major watersheds in the province of Quebec.

(Figure 8b) exhibit more waterflow fluctuations. On average, the spring flood accounts for about 50% of the annual total volume and is composed of melted winter snowpack and spring precipitation. Thus, with respect to dam safety, hydropower generation, operation planning and design of new power plants, a good model to simulate realistic hydrographs, particularly for this period, is certainly valuable to water resources planners and managers.

4.2. Model Specifications

[43] The order l of the spline functions which form the modeling basis needs to be specified in order to apply our method (see section 2.1). We could treat this quantity as a parameter to be estimated in the procedure, but we will consider it to be fixed as is usually done in practice. We choose to work with M-spline functions of order $l = 3$ which means that the basis elements are quadratic by parts.

[44] The registration functions are modeled by linear parts as indicated in equation (2) and the constraints on these functions are taken to be the same as those given in equation (3). The landmarks that determine these constraints could be chosen differently, but we have found that this choice gives good results.

[45] The hypotheses concerning the Bayesian probabilistic model are given in section 3.1. They are as follows:

(A1) $\boldsymbol{\Sigma}_\varepsilon = \mathbb{I}_n$ for each yearly hydrograph;

(A2) β_ω^0 determined by nonparametric least squares regression applied to the reference hydrograph for the $N_0 = 11$ historic yearly hydrographs covering the 1950 to 1960 period, leading to an effective sample formed by the $N = 42$ remaining hydrographs (1961–2002);

(A3) $\boldsymbol{\Sigma}_\omega = \frac{1}{n_0}(\mathbf{B}'_\omega \boldsymbol{\Sigma}_\varepsilon^{-1} \boldsymbol{B}_\omega)^{-1}$;

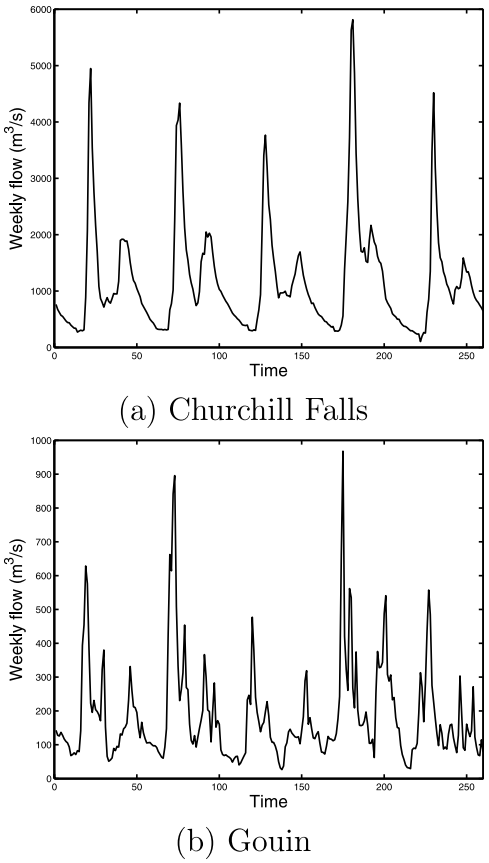


Figure 8. Five consecutive yearly hydrographs for (a) Churchill Falls (1989–1993) and (b) Gouin (1996–2000).

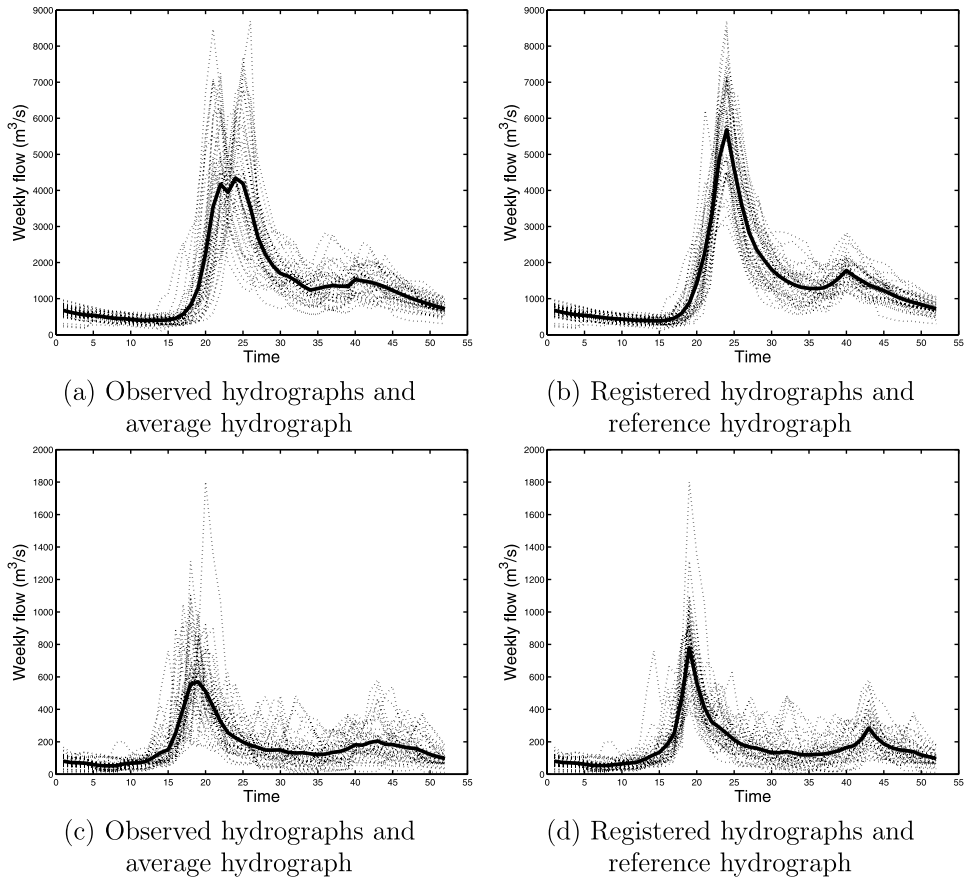


Figure 9. Churchill Falls: (a) observed hydrographs (dotted lines) and their average (full bold line), (b) registered hydrographs (dotted lines) and their average (full bold line). Gouin: (c) observed hydrographs (dotted lines) and their average (full bold line), (d) registered hydrographs (dotted lines) and their average (full bold line).

(A4) $\alpha_\omega = n_0 n$ and $\gamma_\omega = n_0 S^0_\omega$, where S^0_ω is the average of the squared residuals of the regression in A2; and

(A5) $n_0 = 1$.

4.3. Results

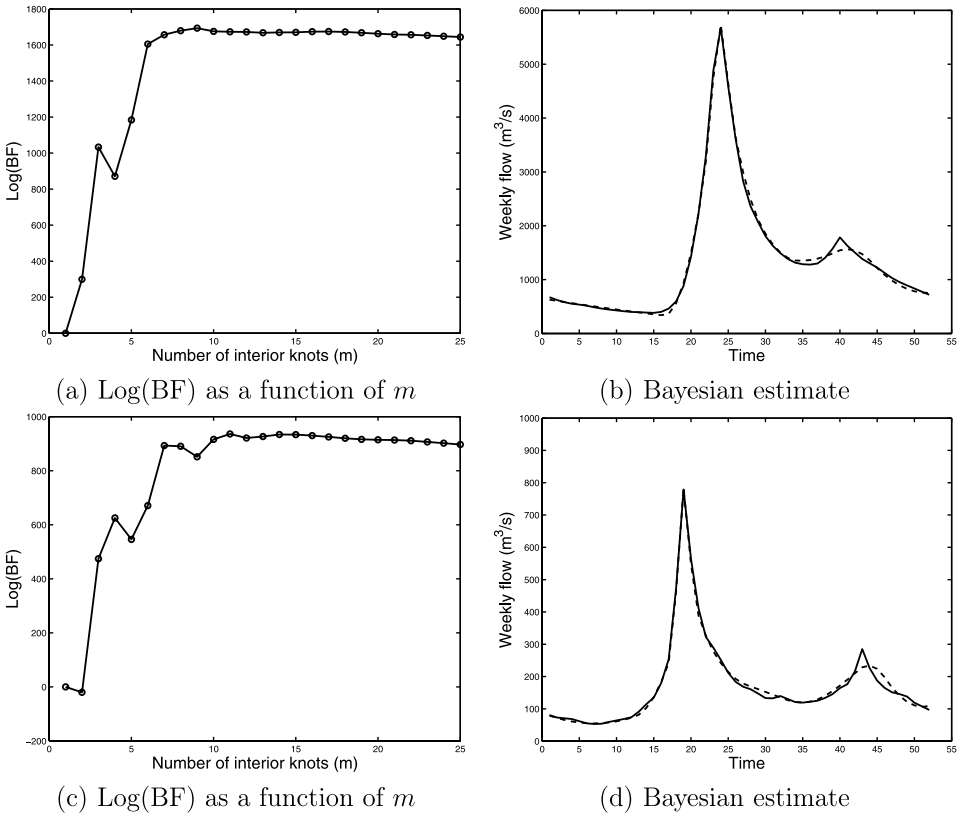
4.3.1. Registration and Reference Hydrographs

[46] We first apply landmark registration to each data set in order to be able to model the reference hydrographs of the two watersheds. Figure 9 illustrates the effect of registration on the two data sets. For Churchill Falls, Figure 9a shows the observed hydrographs as dotted lines and the solid bold line represents their weekly average, i.e., the average hydrograph, while Figure 9b shows the registered hydrographs (dotted lines) and their average, i.e., the reference hydrograph, is illustrated as the solid bold line. In Figures 9c and 9d, the same exercise is performed for Gouin. For Churchill Falls, the average hydrograph has a flood event, which is bimodal with the two modes being similar in magnitude (Figure 9a). This is very rarely observed in practice and is a result of combining yearly hydrographs, which are dissimilar regarding the moment at which the flood event occurs. For its part, the reference hydrograph for this site, shown in Figure 9b, possesses a well-defined flood event. For Gouin, we notice that the reference hydrograph (Figure 9d) has a higher peak flood than the average hydrograph (Figure 9c); furthermore, the flood event is sharper and less spread out for the reference hydrograph than for the average hydrograph. The reference hydrographs therefore better capture the characteristics of the observed flood events since they possess a steep ascent, which is usually observed in practice, and a well defined peak. Furthermore, it needs to be noted that the peak flow of the reference hydrograph is equal to the true average of the observed peak flows by construction.

4.3.2. Model for Reference Hydrographs

[47] Now that we have registered the hydrographs of each watershed, we can model the reference hydrographs with the probabilistic model exposed in section 3.1. In order to determine an adequate spline basis, the model selection approach described in section 3.2 is used. Figures 10a and 10c show the logarithm of the Bayes factors calculated with a reference model containing a single knot (the denominator of equation (19)). By writing a model which contains m interior knots as model $\boldsymbol{\omega}_m$, the Bayes factors shown in Figure 10 are given by BF_{ω_m,ω_1}, for $m = 1,\ldots, 25$. The maximum of BF_{ω_m,ω_1} for Churchill Falls is obtained for nine interior knots, i.e., a model basis containing 12 elements; for Gouin, the best model according to the Bayes factors contains 14 basis elements defined by 11 interior knots.

[48] The Bayesian estimates given by equation (20) are illustrated in Figures 10b and 10d. The reference hydrographs are shown as full lines, while dashed lines illustrate



(a) Log(BF) as a function of m (b) Bayesian estimate

(c) Log(BF) as a function of m (d) Bayesian estimate

Figure 10. Churchill Falls: (a) logarithm of the Bayes factors (see equation (B10)): nine interior knots give the best model; (b) the reference hydrograph (solid line) of Figure 9(b) and the Bayesian estimate (dashed line). Gouin: (c) logarithm of the Bayes factors: eleven interior knots give the best model; (d) the reference hydrograph (solid line) of Figure 9(d) and the Bayesian estimate (dashed line).

the estimates. We see that these estimates fit the reference hydrographs very accurately in the flood event region. Small discrepancies occur around the area of the autumn flood but globally the model performs very well. It should be noted that the reference hydrographs, which are functions of dimension 52, have had their dimensions reduced considerably by their representation in functional space. The reference hydrograph of Churchill Falls is modeled by a parameter space of dimension 12, while the reference hydrograph of Gouin is modeled by a parameter space of dimension 14. As a comparative example, a PARMA(1,1) model with a period equal to the time increment of the series would lead to a parameter space of dimension $208 = (52 \times 4)$, since for each time increment, there are four parameters to evaluate: the mean and the standard deviation of the observed hydrographs, as well as the two parameters of the ARMA process.

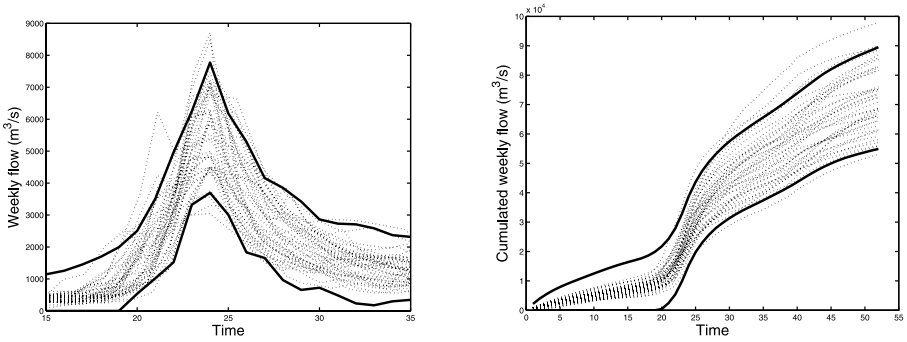
4.3.3. Confidence Intervals for Samples of Registered Yearly Hydrographs

[49] It is possible, using equation (B12) in Appendix B4, to construct simultaneous confidence intervals for the data points of a reference hydrograph and those of the cumulative reference hydrograph. Here we consider 95% confidence intervals for the flood events and for the cumulative hydrographs. For the first confidence intervals, this is done by setting $\boldsymbol{a} = \boldsymbol{b}^M_{\omega}(x_j)$ for a given data point j (see equation (5)). For the second confidence intervals, we set $\boldsymbol{a} = \mathbf{b}^I_{\omega}(x_j)$ for a given data point j (see equation (8)).

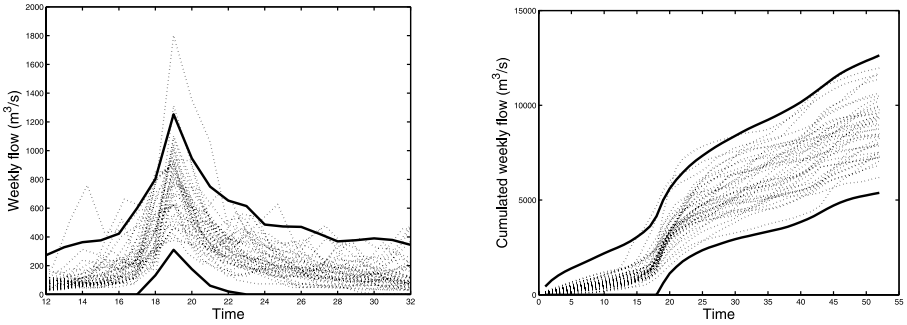
[50] The confidence intervals for the flood events of Churchill Falls are illustrated in Figure 11a, along with the registered flood events. Figure 11b shows the confidence intervals for all the data points of the cumulative hydrograph, along with the yearly registered cumulative hydrographs. The same exercise is performed for Gouin and the results are shown in Figures 11c and 11d.

[51] The effective sample has a size of $N = 42$ and there should thus be at most two or three yearly hydrographs outside a 95% confidence interval for a given data point. Figures 11a and 11c indicate that this is the case for most data points in the flood region. These confidence intervals are therefore well calibrated, or unbiased, since they seem to be able to capture the level of variability of the sample of flood events. The only data points that seem to behave differently are the ones located just before the flood peak for Churchill Falls. It should also be noted that the confidence intervals for the weeks preceding the beginning of the flood events are quite large. These two aspects are due to the constant variance assumption (A1), which causes the variance to be an average of the variances at each week. The level of variability is thus overestimated for the weeks preceding the flood event and underestimated for the weeks close to the flood peak.

[52] For Figures 11b and 11d, there are at most two or three cumulative yearly hydrographs outside the confidence intervals for the majority of data points. Although the confidence intervals seem to reproduce the level of vari-



(a) Confidence interval for flood events (b) Confidence interval for cumulative hydrographs



(c) Confidence interval for flood events (d) Confidence interval for cumulative hydrographs

Figure 11. Churchill Falls: (a) 95% confidence interval for sample of flood events; (b) 95% confidence interval for sample of cumulative hydrographs. Gouin: (c) 95% confidence interval for sample of flood events; (d) 95% confidence interval for sample of cumulative hydrographs.

ability fairly well for most of the year, they appear to be unnecessarily wide in the first few weeks of the year; this is also a result of the constant variance assumption.

5. Conclusion

[53] In this paper, we have put forward a new approach to model the average properties of a sample of yearly hydrographs. We elaborated a methodology to obtain reference hydrographs representative of given samples and it was seen that the reference hydrographs reproduce adequately the flood events encountered in the samples. Furthermore, a previous study [*Merleau et al.*, 2005] has shown that the constructed reference hydrographs also preserve the average flood event volumes of hydrograph samples. We also exposed a nonparametric regression method in a Bayesian setting to model a reference hydrograph or any particular hydrograph in the sample. The approach was applied to two samples of yearly hydrographs with weekly streamflow in order to obtain a statistical representation of two reference hydrographs of different watersheds. Using the statistical model, confidence intervals were produced for the flood events and the cumulative streamflows of each watershed hydrographs. Although we did not present an analysis of yearly hydrographs with daily measurements, we have found that the method performs just as well in this case.

[54] The methodology proposed in this paper can be related to existing methods to model and simulate hydrographs. Using our model on a single typical hydrograph, our approach would be similar to the TH methods (see section 1). One major difference though is the fact that our model is statistically based and it can therefore be used to construct confidence intervals for example. As pointed out in section 2.1, the relation between the S methods and our approach is quite clear although the latter considers the hydrographs as random functions, while the former does not. Furthermore, our model is more general since it relies on a functional representation based on spline functions, which can reproduce a wide variety of hydrograph shapes.

[55] Finally, as mentioned in section 1, the modeling techniques based on time series are mainly used to simulate hydrographs corresponding to probable scenarios. In the Bayesian statistical context, the yearly hydrographs are treated as random functional events and thus it is also possible to simulate hydrographs in this setting. Vectors of parameters can be generated from the posterior probability distribution (equation (18)) and to each vector of parameters corresponds a simulated hydrograph. If no constraints are placed on the parameters, the generated hydrographs represent random events and in this respect, our approach resembles the time series-based methods. In our statistical framework though, it is also possible to simulate hydrographs that have fixed flood characteristics, with a certain probability of occurrence, if constraints are put on the parameters.

[56] We have seen that the method proposed in this paper performs very well globally. The aspects that still require attention are the constant variance assumption and the exploration of knot configurations. It was seen in section 4.3.3 that the constant variance throughout the year leads to confidence intervals which appear to be too wide in certain periods and too narrow in other periods. We are currently working on a

method which models the variance components at each time increment; this will correct the width of the confidence intervals. Although the knot determination method used in this paper performs well, it would be preferable to use a random knot selection procedure that explores a wide variety of knot configurations; we are also working on a random knot procedure at the present time.

Appendix A: Probability Distributions

A1. Multivariate Normal Distribution

[57] If $X \sim N_q(\boldsymbol{\theta}, \boldsymbol{\Sigma})$, then the variable X is distributed according to a q dimensional multivariate normal distribution which is given by:

$$f(x|\boldsymbol{\theta}, \boldsymbol{\Sigma}) = \frac{1}{(2\pi)^{\frac{q}{2}}|\boldsymbol{\Sigma}|^{\frac{1}{2}}} \exp\left\{-\frac{1}{2}(x-\boldsymbol{\theta})'\boldsymbol{\Sigma}^{-1}(x-\boldsymbol{\theta})\right\},$$

where $\boldsymbol{\theta}$ ($q \times 1$) is the location vector of the distribution, $\boldsymbol{\Sigma}$ ($q \times q$) is the covariance matrix associated with the components of X and $|\cdot|$ represents the determinant.

A2. Inverse Gamma Distribution

[58] If $X \sim I\Gamma(\alpha, \gamma)$, then the variable X is distributed according to an inverse gamma distribution which is given by:

$$f(x \mid \alpha, \gamma) = \frac{\gamma^{\alpha}}{\Gamma(\alpha)} \frac{1}{x^{\alpha+1}} \exp\left\{-\frac{\gamma}{x}\right\}.$$

A3. Student's *t* Distribution

[59] If $X \sim T_q(\nu, \boldsymbol{\theta}, \boldsymbol{\Sigma})$, then the variable X is distributed according to a q dimensional Student's t distribution which is given by:

$$f(x|\nu, \boldsymbol{\theta}, \boldsymbol{\Sigma}) = \frac{\Gamma\left(\frac{\nu+q}{2}\right)}{|\boldsymbol{\Sigma}|^{\frac{1}{2}}(\nu\pi)^{\frac{q}{2}}\Gamma\left(\frac{\nu}{2}\right)} \left\{1 + \frac{(x-\boldsymbol{\theta})'\boldsymbol{\Sigma}^{-1}(x-\boldsymbol{\theta})}{\nu}\right\}^{-\frac{\nu+q}{2}},$$

where $\boldsymbol{\theta}$ ($q \times 1$) is the location vector of the distribution, $\boldsymbol{\Sigma}$ ($q \times q$) is the covariance matrix associated with the components of X, ν is the number of degrees of freedom and $|\cdot|$ represents the determinant.

Appendix B: Bayesian Results

B1. Parameters of the Posterior Statistical Distributions

[60] The parameters of the posterior statistical distributions are as follows:

$$\begin{aligned}\boldsymbol{\beta}^*_{\omega} &= \left(\boldsymbol{W}^G_{\omega} + \boldsymbol{W}^0_{\omega}\right)^{-1}\left(\boldsymbol{W}^G_{\omega}\boldsymbol{\beta}^G_{\omega} + \boldsymbol{W}^0_{\omega}\boldsymbol{\beta}^0_{\omega}\right) \\ &=^{A} \left(\frac{N}{N+n_0}\right)\boldsymbol{\beta}^L_{\omega} + \left(\frac{n_0}{N+n_0}\right)\boldsymbol{\beta}^0_{\omega},\end{aligned} \tag{B1}$$

$$\boldsymbol{\Sigma}^*_{\omega} = \left(\boldsymbol{W}^G_{\omega} + \boldsymbol{W}^0_{\omega}\right)^{-1} =^{A} \left(\frac{1}{N+n_0}\right)\left(\mathbf{B}'_{\omega}\mathbf{B}_{\omega}\right)^{-1}, \tag{B2}$$

$$\alpha^*_{\omega} = Nn + \alpha_{\omega} =^{A} (N+n_0)n, \tag{B3}$$

$$\gamma^*_{\omega} = NS_{\omega} + T_{\omega} + \gamma_{\omega} =^{A} NS_{\omega} + T_{\omega} + n_0 S^0_{\omega}, \tag{B4}$$

$$\begin{aligned}S_{\omega} &= \frac{\sum_i \left(\boldsymbol{y}_i - \boldsymbol{B}_{\omega}\boldsymbol{\beta}^G_{\omega}\right)'\boldsymbol{\Sigma}^{-1}_{\varepsilon}\left(\boldsymbol{y}_i - \boldsymbol{B}_{\omega}\boldsymbol{\beta}^G_{\omega}\right)}{N} \\ &=^{A} \frac{\sum_i \left(\boldsymbol{y}_i - \boldsymbol{B}_{\omega}\boldsymbol{\beta}^L_{\omega}\right)'\left(\boldsymbol{y}_i - \boldsymbol{B}_{\omega}\boldsymbol{\beta}^L_{\omega}\right)}{N},\end{aligned} \tag{B5}$$

$$T_{\omega} = \left(\boldsymbol{\beta}^{\mathrm{G}}_{\omega} - \boldsymbol{\beta}^0_{\omega}\right)'\left\{\left(N\mathbf{B}'_{\omega}\boldsymbol{\Sigma}^{-1}_{\varepsilon}\mathbf{B}_{\omega}\right)^{-1} + \boldsymbol{\Sigma}_{\omega}\right\}^{-1}\left(\boldsymbol{\beta}^{\mathrm{G}}_{\omega} - \boldsymbol{\beta}^0_{\omega}\right), \tag{B6}$$

$$=^{A} \left(\frac{n_0 N}{N+n_0}\right)\left(\boldsymbol{\beta}^L_{\omega} - \boldsymbol{\beta}^0_{\omega}\right)'\mathbf{B}'_{\omega}\mathbf{B}_{\omega}\left(\boldsymbol{\beta}^L_{\omega} - \boldsymbol{\beta}^0_{\omega}\right), \tag{B7}$$

where $=^{A}$ indicates the given quantity evaluated under assumptions A1, A3 and A4.

[61] The posterior mean of the parameters, $\boldsymbol{\beta}^*_{\omega}$, is given by a weighted average of the generalized least squares estimator $\boldsymbol{\beta}^G_{\omega} = (\boldsymbol{B}'_{\omega}\boldsymbol{\Sigma}^{-1}_{\varepsilon}\boldsymbol{B}_{\omega})^{-1}\boldsymbol{B}'_{\omega}\boldsymbol{\Sigma}^{-1}_{\varepsilon}\bar{\boldsymbol{y}}$ and of the prior location vector $\boldsymbol{\beta}^0_{\omega}$. The weights are given by $\boldsymbol{W}^G_{\omega} = N\boldsymbol{B}'_{\omega}\boldsymbol{\Sigma}^{-1}_{\varepsilon}\boldsymbol{B}_{\omega}$ and $\boldsymbol{W}^0_{\omega} = \boldsymbol{\Sigma}^{-1}_{\omega}$. Under the application assumptions, $\boldsymbol{\beta}^*_{\omega}$ simplifies to a weighted average of the ordinary least squares estimator $\boldsymbol{\beta}^L_{\omega} = (\boldsymbol{B}'_{\omega}\boldsymbol{B}_{\omega})^{-1}\boldsymbol{B}'_{\omega}\bar{\boldsymbol{y}}$ and of the prior location vector, where the weights are proportional to the number of observations used to evaluate each of these quantities.

[62] We also note that the posterior probability distribution of the variance parameter σ^2 depends on S_{ω}, which is proportional to the sum of squares of the residuals, and T_{ω} which captures the discrepancy between the two vectors of parameters $\boldsymbol{\beta}^{\mathrm{G}}_{\omega}$ and $\boldsymbol{\beta}^0_{\omega}$.

B2. Marginal Distribution

[63] For the model exposed in section 3, the marginal distribution is given by

$$\mathrm{m}(\boldsymbol{y}|\boldsymbol{\omega}) = \left(\frac{|\boldsymbol{\Sigma}^*_{\omega}|}{|\boldsymbol{\Sigma}_{\varepsilon}||\boldsymbol{\Sigma}_{\omega}|}\right)^{1/2}\left(\frac{\Gamma(\alpha^*_{\omega}/2)}{\pi^{Nn/2}\Gamma(\alpha_{\omega}/2)}\right)\left(\frac{(\gamma_{\omega})^{\alpha_{\omega}/2}}{(\gamma^*_{\omega})^{\alpha^*_{\omega}/2}}\right), \tag{B8}$$

$$=^{A} \left(\frac{n_0}{N+n_0}\right)^{K_{\omega}/2}\left(\frac{\Gamma(\alpha^*_{\omega}/2)}{\pi^{Nn/2}\Gamma(\alpha_{\omega}/2)}\right)\left(\frac{(\gamma_{\omega})^{\alpha_{\omega}/2}}{(\gamma^*_{\omega})^{\alpha^*_{\omega}/2}}\right), \tag{B9}$$

where $|\cdot|$ represents the determinant.

B3. Bayes Factor

[64] Using the marginal distribution given in equation (B8), the Bayes factor for our model is as follows:

$$\begin{aligned}BF_{\omega_1,\omega_2} = &\left(\frac{|\boldsymbol{\Sigma}^*_{\omega_1}|}{|\boldsymbol{\Sigma}^*_{\omega_2}|}\frac{|\boldsymbol{\Sigma}_{\omega_2}|}{|\boldsymbol{\Sigma}_{\omega_1}|}\right)^{1/2}\left(\frac{\Gamma(\alpha^*_{\omega_1}/2)}{\Gamma(\alpha^*_{\omega_2}/2)}\frac{\Gamma(\alpha_{\omega_2}/2)}{\Gamma(\alpha_{\omega_1}/2)}\right) \\ &\times \left(\frac{(\gamma_{\omega_1})^{\alpha_{\omega_1}/2}}{(\gamma_{\omega_2})^{\alpha_{\omega_2}/2}}\frac{(\gamma^*_{\omega_2})^{\alpha^*_{\omega_2}/2}}{(\gamma^*_{\omega_1})^{\alpha^*_{\omega_1}/2}}\right),\end{aligned} \tag{B10}$$

$$\begin{aligned}=^{A} &\left(\frac{n_0}{N+n_0}\right)^{(K_{\omega_1}-K_{\omega_2})/2}\left(\frac{\Gamma(\alpha^*_{\omega_1}/2)}{\Gamma(\alpha^*_{\omega_2}/2)}\frac{\Gamma(\alpha_{\omega_2}/2)}{\Gamma(\alpha_{\omega_1}/2)}\right) \\ &\times \left(\frac{(\gamma_{\omega_1})^{\alpha_{\omega_1}/2}}{(\gamma_{\omega_2})^{\alpha_{\omega_2}/2}}\frac{(\gamma^*_{\omega_2})^{\alpha^*_{\omega_2}/2}}{(\gamma^*_{\omega_1})^{\alpha^*_{\omega_1}/2}}\right),\end{aligned} \tag{B11}$$

where K_{ω_1} and K_{ω_2} are the number of parameters in models ω_1 and ω_2, respectively, and the other quantities are defined in Appendix B1.

B4. Bayesian Confidence Intervals (Credible Sets)

[65] Simultaneously for all vectors $\boldsymbol{a}$, a $100(1-\delta)\%$ confidence interval is given by

$$a'\beta_{\omega}^{*} \pm \left\{ K_{\omega}\left(\frac{\gamma_{\omega}^{*}}{\alpha_{\omega}^{*}}\right) a'\boldsymbol{\Sigma}_{\omega}^{*} a F_{K_{\omega},\alpha_{\omega}^{*}}(\delta) \right\}^{1/2}, \qquad \text{(B12)}$$

where $F_{K_{\omega},\alpha_{\omega}^{*}}(\delta)$ represents the $100(1-\delta)$th percentile of Fisher's F distribution with degrees of freedom K_{ω} and α_{ω}^{*}, and the other quantities are defined as before.

[66] **Acknowledgments.** The authors would like to thank Frédéric Guay for providing the map shown in Figure 7 and the referees for constructive comments, which improved the paper substantially.

References

Akaike, H. (1973), Information theory and an extension of the maximum likelihood principle, in *Second International Symposium on Information Theory (Tsahkadsor, 1971)*, pp. 267–281, Akadémiai Kiadó, Budapest.

Angers, J.-F., J. Merleau, and L. Perreault (2004), Landmark registration of hydrographs and Bayesian estimation of a mean hydrograph, in *Proceedings of the International Sri Lankan Statistical Conference: Visions of Futuristic Methodologies, Kandy, Sri Lanka*, edited by B. M. de Silva, and N. Mukhopadhyay, pp. 47–60.

Bernardo, J.-M., and A. F. M. Smith (1994), *Bayesian Theory, Wiley Series in Probability and Mathematical Statistics: Probability and Mathematical Statistics*, xiv+586 pp., John Wiley, Hoboken, N. J.

Chow, V. T. (1964), *Handbook of applied hydrology*, McGraw-Hill, New York.

Chow, V. T., D. R. Maidment, and L. W. Mays (1988), *Handbook of applied hydrology*, McGraw-Hill, New York.

Denison, D. G. T., B. K. Mallick, and A. F. M. Smith (1998), Automatic Bayesian curve fitting, *J. R. Stat. Soc., Ser. B Stat. Methodol.*, *60*(2), 333–350.

Doodge, J. (1959), A general theory of the unit hydrograph, *J. Geophys. Res.*, *64*, 241–256.

Friedman, J. H., and B. W. Silverman (1989), Flexible parsimonious smoothing and additive modeling, *Technometrics*, *31*(1), 3–39with discussions by Trevor Hastie and Douglas M. Hawkins and a reply by the authors.

Hastie, T. J., and R. J. Tibshirani (1990), *Generalized Additive Models, Monographs on Statistics and Applied Probability*, vol. 43, xvi+335 pp., Chapman and Hall Ltd., London.

He, X., and P. Shi (1998), Monotone *B*-spline smoothing, *J. Amer. Statist. Assoc.*, *93*(442), 643–650.

Javelle, P., T. B. M. J. Ouarda, M. Lang, B. Bobe, J. Gala, and J.-M. Grsillon (2002), Development of regional flow-duration-frequency curves based on the index-flood method, *J. Hydrol.*, *258*, 249–259.

Johnson, R. A., and D. W. Wichern (1992), *Applied Multivariate Statistical Analysis*, 3rd ed., xiv+642 pp., Prentice Hall, Upper Saddle River, N. J.

Kass, R. E., and A. E. Raftery (1995), Bayes factors, *J. Am. Stat. Assoc.*, *90*, 773–795.

Kneip, A., and T. Gasser (1988), Convergence and consistency results for self-modeling nonlinear regression, *Ann. Stat.*, *16*(1), 82–112.

Kneip, A., and T. Gasser (1992), Statistical tools to analyze data representing a sample of curves, *Ann. Stat.*, *20*(3), 1266–1305.

Lee, P. M. (1989), *Bayesian Statistics: An Introduction. A Charles Griffin Book*, x+294 pp., Oxford Univ. Press, New York.

Merleau, J., G. Evin, L. Perreault, A.-C. Favre, D. Tremblay, and J.-F. Angers (2005), Analyse descriptive des prvisions hydrologiques d'ensemble et modlisation par un mlange de deux lois gamma. Cration d'hydrogrammes de base et d'hydrogrammes prvisionnels pas de temps journalier, *Tech. Rep. R819*, INRS-Eau, Terre et Environnement, Que., Canada.

Nezhikhovsky, R. A. (1971), Channel network of the basin and runoff formation, *Tech. rep.*, Hydrometeorological, Leningrad, Russia.

Ogden, R. T. (1997), *Essential Wavelets for Statistical Applications and Data Analysis*, xviii+206 pp., Springer, New York.

Ouhib, L. (2005), Modlisation des apports naturels de rservoirs, Master's thesis, Université de Montréal, Montréal, Quebec.

Perreault, L., and M. Latraverse (2001), Modélisation des apports naturels pour la prise en compte de leur aléa dans la méthode SDDP de planification de la production, *Tech. rep.*, Institut de recherche d'Hydro-Qubec, Varennes, Quebec.

Pilgrim, D. H., and I. Cordery (1993), *Handbook of Hydrology*, chap. Flood runoff, McGraw-Hill, New York.

Ramsay, J. O. (1988), Monotone regression splines in action, *Stat. Sci.*, *3*, 425–461.

Ramsay, J. O., and X. Li (1998), Curve registration, *J. R. Stat. Soc. Ser., B Stat. Methodol.*, *60*(2), 351–363.

Ramsay, J. O., and B. W. Silverman (2005), *Functional Data Analysis, Springer Series in Statistics*, 2nd ed., xx+426 pp., Springer, New York.

Rasmussen, P. F., J. D. Salas, L. Fagherazzi, J.-C. Rassam, and B. Bobe (1996), Estimation and validation of contemporaneous PARMA models for streamflow simulation, *Water Resour. Res.*, *32*, 3151–3160.

Robert, C. P. (1994), *The Bayesian Choice: A Decision-Theoretic Motivation, Springer Texts in Statistics* (Translated and revised from the French original by the author), xiv+436 pp., Springer, New York.

Salas, J. D., J. Delleur, V. Yevjevich, and W. L. Lane (1980), *Applied Modelling of Hydrologic Time Series*, Water Resour. Publ., Highlands Ranch, Colo.

Salas, J. D., D. C. Boes, and R. A. Smith (1982), Estimation of ARMA models with seasonal parameters, *Water Resour. Res.*, *18*, 1006–1010.

Schwarz, G. (1978), Estimating the dimension of a model, *Ann. Stat.*, *6*(2), 461–464.

Sherman, L. K. (1932), Streamflow from rainfall by the unit-graph method, *Eng. News-Rec.*, *108*, 501–505.

Snyder, F. F. (1938), Synthetic unit-graphs, *Trans. - Am. Geophys. Union*, *19*, 447–454.

Sokolov, A. A., S. E. Rantz, and M. Roche (1976), *Flood Computation Methods Compiled from World Experience*, chap. Methods of Developing Design-Flood Hydrographs, UNESCO, Paris.

Stone, C. J., M. H. Hansen, C. Kooperberg, and Y. K. Truong (1997), Polynomial splines and their tensor products in extended linear modeling, *Ann. Stat.*, *25*(4), 1371–1470 with discussion and a rejoinder by the authors and Jianhua Z. Huang.

Titterington, D. M., A. F. M. Smith, and U. E. Makov (1985), *Statistical Analysis of Finite Mixture Distributions*, *Wiley Series in Probability and Mathematical Statistics: Applied Probability and Statistics*, x+243 pp., John Wiley, Hoboken, N. J.

U.S. Soil Conservation Service US-SCS (1985), *National Engineering Handbook*, chap. Hydrology, U.S. Dep. of Agricul., Washington, D. C.

Vecchia, A. V., J. T. B. Obeysekera, J. D. Salas, and D. C. Boes (1983), Aggregation and estimation of low-order periodic ARMA models, *Water Resour. Res.*, *19*, 1297–1306.

Yue, S., and M. Hashino (2000), Unit hydrographs to model quick and slow runoff components of streamflow, *J. Hydrol.*, *227*, 195–206.

Yue, S., T. B. M. J. Ouarda, B. Bobe, P. Legendre, and P. Bruneau (1999), The Gumbel mixed model for flood frequency analysis, *J. Hydrol.*, *226*, 88–100.

Yue, S., T. B. M. J. Ouarda, B. Bobe, P. Legendre, and P. Bruneau (2002), Approach for describing statistical properties of flood hydrograph, *J. Hydrol. Eng.*, *7*(2), 147–153.

Zellner, A. (1986), On assessing prior distributions and Bayesian regression analysis with *g*-prior distributions, in *Bayesian Inference and Decision Techniques, Stud. Bayesian Econom.*, vol. 6, pp. 233–243, Elsevier, New York.

J.-F. Angers, Département de mathématiques et de statistique, Université de Montréal, CP 6128 succ Centre-Ville, Montréal, QC, Canada, H3C 3J7. (angers@dms.umontreal.ca)

A.-C. Favre, Chaire en Hydrologie statistique, INRS, Eau, Terre et Environnement, 490, rue de la Couronne, Quebec, QC, Canada, G1K 9A9. (anne-catherine_favre@ete.inrs.ca)

J. Merleau and L. Perreault, Institut de recherche d'Hydro-Québec, 1800, boul. Lionel-Boulet, Varennes, QC, Canada, J3X 1S1. (merleau.james@ireq.ca; perreault.luc@ireq.ca)