Landmark Registration of Hydrographs and Bayesian Estimation of a Mean Hydrograph

Jean-François Angers[1], James Merleau[1] and Luc Perreault[2]

[1]Université de Montréal, Canada, [2]Institut de recherche d'Hydro-Québec, Canada

Presenter: Jean-François Angers, Département de mathématiques et de statistique, Université de Montréal, C.P. 6128 Succ. Centre-ville, Montréal, Québec, H3C 3J7, Canada. E-mail: jean-francois.angers@umontreal.ca

Key words: Statistical hydrology, seasonal variability, hydrograph, curve data, estimation of curve, B-spline.

Abstract

One problem encountered in statistical hydrology is to model and to generate hydrographs which are similar to the observed ones. In order to obtain representative hydrographs, one should be able to average observed hydrographs. Because of seasonal variability, one cannot use the average of the observed hydrographs as a good template since the major features of the curve happen at different times each year. Consequently, before averaging the observed hydrographs, one has to register them based on some fixed features such as the maximum Spring and Autumn floods. Using a Bayesian nonparametric model based on B-splines, a template hydrograph is obtained. A real example using data from a river in Northern Québec is also presented.

1. Introduction

The modeling of hydrographs (curves describing the water flow of a river as a function of time) is important in order to plan production of hydro-electricity and construction of new hydro-electricity facilities. It can also be used to evaluate the performance of hydro-electricity facilities under extreme simulated conditions.

Hydrographs are usually modeled using time series. However, due to the seasonal variability, a large number of parameters are required in order to obtain a satisfying fit. The

main objective of this paper is to find a model with fewer parameters which provides a fit as good or better than the current approach used at Hydro-Québec. In order to do so, landmark registration (*cf.* Ramsay and Li, 1998) of observed hydrographs is required. The landmark registration corresponds to a nonlinear transformation of the time axis such that the main characteristics of the observed hydrographs are comparable. Once these curves are registered and averaged, a Bayesian model is used to obtain a representative hydrograph. The Bayesian model is based on B-spline functions (*cf.* De Boor, 1978, Schumaker, 1981).

The paper is divided as follows. In Section 2, motivation of the problem and the different methods proposed to obtain a representative hydrograph are presented. In Section 3, curve registration is introduced. The Bayesian model proposed in this paper is given in Section 4. The methodology introduced in Sections 3 and 4 are illustrated by means of a real example.

2. Motivation

Observed hydrographs show a large seasonal variability (see Figure 1). As seen in Figure 1, the position of the maximum Spring flood varies from weeks 19 to 27. The same variability can be observed for the maximum Autumn flood. Because of this variability, averaging the different curves will underestimate the maximum and flatten the area around it. For an illustration of this phenomenon, see Figure 2. If we average the values in Figure 2, it is clear that the maximum will be driven down by the curves corresponding to 1972 and 1982 since their maximum Spring floods occur later than in 1986 and 1987. Hence, because of seasonal variability, using the mean of the observed hydrographs to estimate a hydrograph will not lead to a curve which is likely to be observed.

The method currently used by Hydro-Québec to estimate the representative hydrograph and to model the seasonal variability is to use a periodic autoregressive model of order p (PAR(p)) given by

$$\left(\frac{x_{\nu,t}-\mu_t}{\sigma_t}\right)=\sum_{k=1}^{p}\phi_{k,t}\left(\frac{x_{\nu,t-k}-\mu_{t-k}}{\sigma_{t-k}}\right)+\beta_t V_{\nu-1}+\varepsilon_{\nu,t},$$

where $x_{\nu,t}$ is the value of the water flow for year ν at week t, μ_t, the mean of the flow for week t, σ_t, the standard deviation of the flow for week t, $\phi_{k,t}$ represents the dependence

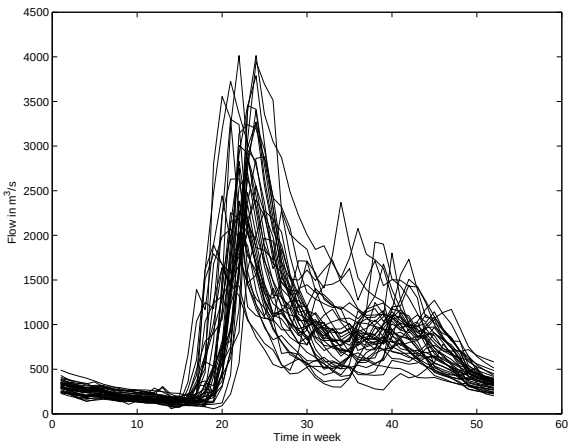


Figure 1: Observed hydrographs for a river in Northern Québec from 1961 to 1999

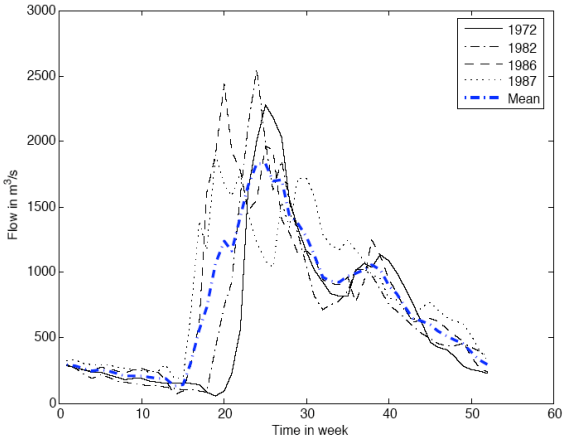


Figure 2: Seasonal variability and mean hydrograph (thick dashed-dotted line) based on the years 1972, 1982, 1986 and 1987

parameter between the periods t and $t-k$, β_t, the influence of the total cumulative flow of the previous year ($V_{\nu-1} = \sum_{t=1}^{T} x_{\nu-1,t}$) on the t^{th} week and $\varepsilon_{\nu,t}$ is the error term for year ν and week t (see Perreault and Latraverse, 2001). Hence, we have $52 \times (p+3)$ parameters to estimate. This method provides an adequate fit for the observed hydrographs but it is not parsimonious. Alternative methods to model hydrographs are presented in Yue *et al.* (1999, 2001, 2002).

The method proposed in this paper will produce a more parsimonious model. The first step is to register all the hydrographs (see Figure 1) such that the main characteristics of the registered curves, that is the maximum Spring and Autumn floods, occur at the same time for all hydrographs. This is done by performing a nonlinear transformation of the time axis. More details are given in Section 3. A representative hydrograph is then obtained using a Bayesian regression model on the average registered curves. This is presented in Section 4.

3. Landmark Registration

In order to eliminate seasonal variability, one can do a nonlinear transformation (contraction and/or expansion) of the time axis (*cf.* Ramsay and Li, 1998). This transformation is such that the main characteristics of the curve happen at predetermined times.

3.1. The registration function

Let $X_{\nu,t} = h_\nu(t)$ be the "true" value of the hydrograph for the t^{th} week ($t = 1, 2, \ldots, T = 52$) of year ν. If $g_\nu(t)$ represents the registration function for the year ν, then $X_{\nu,t}$ can be written as

$$X_{\nu,t} = h_\nu(t) = h(g_\nu(t)) = h(\tau),$$

where $h(\cdot)$ is the registered or "typical" hydrograph and τ is the transformed time for year ν, that is $\tau = g_\nu(t)$.

To allow one to go back and forth from the registered to the observed time axes, the registration function $g_\nu(t)$ should be bijective and monotonously increasing. The function $g_\nu(t)$ should also satisfy:

$$g_\nu(1) = 1,\ g_\nu(t_{\nu,S}) = \tau_S,\ g_\nu(t_{\nu,A}) = \tau_A,\ g_\nu(T) = T, \tag{3.1}$$

where $t_{\nu,S}$ and $t_{\nu,A}$ are the weeks where the maximum Spring and Autumn floods occur for year ν, τ_S and τ_A, the pre-specified standardized time for the maximum Spring and Autumn floods. Note that week 1 corresponds to the first week of January while week 52 is the last week of December. The values of $t_{\nu,S}$ and $t_{\nu,A}$ have been chosen such that $t_{\nu,S}$ is the time of the maximum observation while $t_{\nu,A}$ is the time of the last mode of the observed hydrograph of year ν. (For the present data, τ_S and τ_A were chosen to be the median time of the observed maximum Spring and Autumn floods, that is $\tau_S = 23$ (roughly the second week of June) and $\tau_A = 40$ (roughly the first week of October).) Hence, all registered hydrographs are such that the first and last values are the same as the unregistered hydrographs while the maximum Spring and Autumn floods are moved to pre-specified times. The effect of landmark registration on the time axis is illustrated in Figure 3, where the data for the year 1996 are displayed. The registration function (dashed line) illustrated in (a) is typical of the data analyzed in this paper although sometimes three regions (instead of two as illustrated here) can be tranformed. Mainly, some regions of the time axis are expanded while others are contracted in order to have the maximum Spring and Autumn floods occurring at pre-specified values.

3.2. An approximation of the registration function

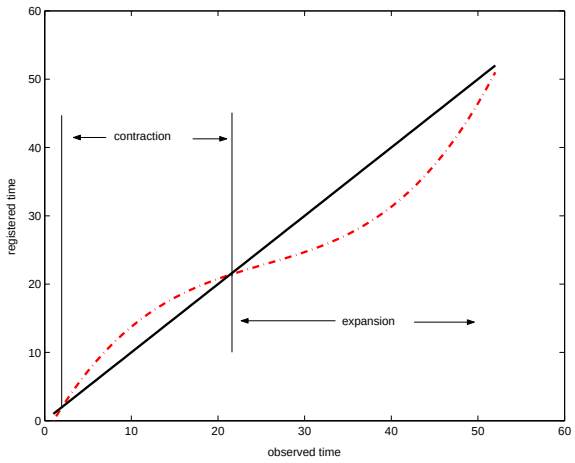
The registration function is only a tool to make the observed hydrographs comparable. Hence, the choice of the "best" possible function $g_\nu(t)$ is not the goal of this paper. We want an efficient way to compute the function which does not put a burden on the evaluation of the Bayesian estimator of the representative hydrograph. In order to obtain such a function, the registration function $g_\nu(t)$ will be replaced by a polynomial which can be seen as a Taylor expansion of the "true" registration function around $t_{\nu,S}$. Hence, let

$$g_\nu(t) = g_\nu(t_{\nu,S}) + g_\nu^{(1)}(t_{\nu,S})(t - t_{\nu,S}) + g_\nu^{(2)}(t_{\nu,S})\frac{(t - t_{\nu,S})^2}{2} + g_\nu^{(3)}(t_{\nu,S})\frac{(t - t_{\nu,S})^3}{6},$$

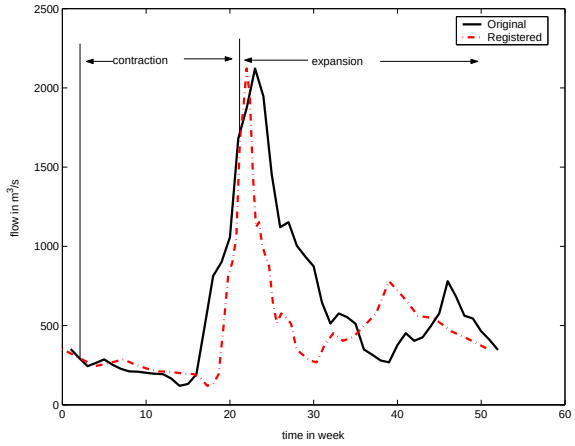
where

$$g_\nu^{(\ell)}(t_{\nu,S}) = \frac{\partial^\ell}{\partial t^\ell} g_\nu(t)\bigg|_{t=t_{\nu,S}},$$

for $\ell = 1, 2, 3$.



(a) Registration function



(b) Effect of the landmark registration on the hydrograph

Figure 3: Registration function and its effect on the observed hydrograph for 1996

The registration function can be estimated by a polynomial of degree 3, since the number of constraints $g_\nu(t)$ has to satisfy is equal to 4 (see (3.1)). The coefficients can be obtained easily by solving the linear system of equations obtained from (3.1). In order to be monotone, the first derivative of $g_\nu(t)$ has to be nonnegative for all $t \in [1, 52]$, that is

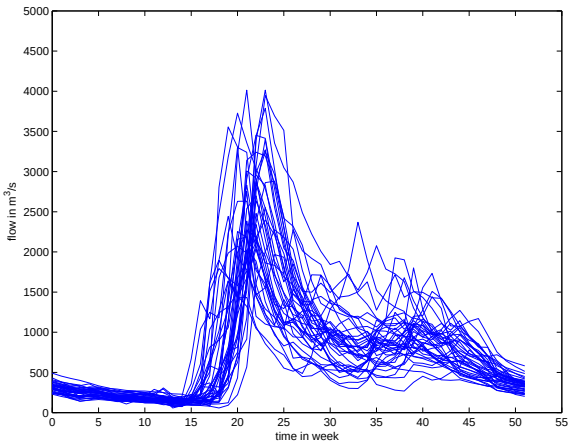
$$g_\nu^{(1)}(t_{\nu,S}) + g_\nu^{(2)}(t_{\nu,S})(t - t_{\nu,S}) + \frac{g_\nu^{(3)}(t_{\nu,S})}{2}(t - t_{\nu,S})^2 \geq 0 \ \ \forall t \in [1, 52]. \tag{3.2}$$

For the present data, (3.2) is always satisfied. In general, if the number of constraints is equal to $p + 1$, the registration function is then approximated by a polynomial of degree p.

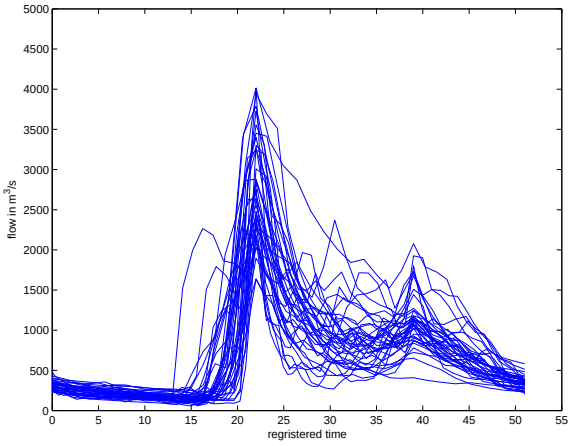
The effect of the registration functions on the observed hydrographs, given in Figure 1, is illustrated in Figure 4. Comparing both graphs in Figure 4, it can be seen that the observed maximum Spring floods are aligned together in part (b) which is not the case in part (a). The same can be said for the maximum Autumn flood although there is some variation left. This is mostly due to the fact that the maximum Autumn flood does not happen every year and is smaller in magnitude than the maximum Spring flood. From Figure 4 (b), it is clear that it is easier to obtain a representative mean hydrograph from the registered curves than from the observed ones. Figure 5, where the averages for the observed and registered hydrographs are plotted, illustrates this. One can see that the maximum Spring and Autumn floods are sharper for the mean registered hydrograph than for the mean observed one. Furthermore, the registered one is narrower than the observed one around each maximum. These two characteristics make the mean registered hydrograph a better representative hydrograph than the average based on the unregistered ones.

4. Bayesian regression model for the mean registered hydrograph

In this section, the Bayesian model used to estimate the representative registered hydrograph is presented. The model used is based on B-spline functions. The choice of knots is very important when splines are used to model a general function. However, if the function is monotonic, this choice is less crucial (*cf.* He and Shi, 1998). Consequently, since the hydrograph is a nonnegative function, we will first model the cumulative hydrograph which



(a) Observed hydrographs



(b) Registered hydrographs

Figure 4: Observed (a) and registered (b) hydrographs for a river in Northern Québec from 1961 to 1999

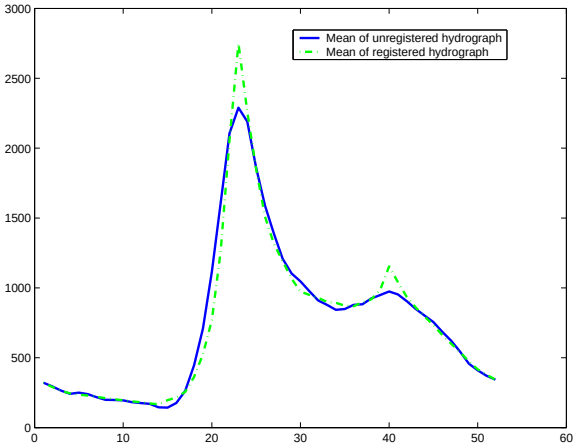


Figure 5: Mean observed (full line) and registered (dash line) hydrographs

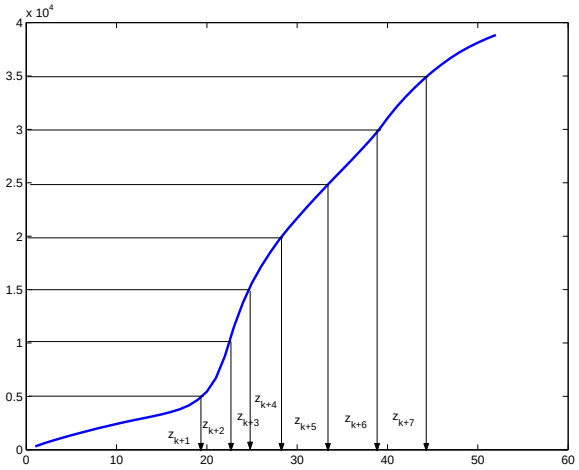


Figure 6: Choice of the interior knots for the cumulative hydrograph

is a monotone function. Hence, let

$$H_\nu(t) = \sum_{\ell=1}^{t} X_{\nu,\ell} = \sum_{\ell=1}^{t} h_\nu(\ell),$$

be the cumulative hydrograph for year ν at week t. Using a B-spline basis of order k with m interior knots (and k knots at each boundary) $z_1 = \ldots = z_k < z_{k+1} < \ldots < z_{k+m} < z_{k+m+1} = \ldots = z_{2k+m}$, the cumulative hydrograph can be written as

$$H_{\nu,m}(t) = \sum_{i=1}^{k+m} \theta_{\nu,i} B_{i,k}(t) + \epsilon_{\nu,t},$$

where $\epsilon_{\nu,t} \sim \mathrm{N}(0; \sigma_\nu^2)$ independent for $\nu = 1, 2, \ldots, N$. Taking the average over all years and using matrix notation, we obtain the following linear model

$$\underset{\sim}{H} = B_k \underset{\sim}{\theta} + \underset{\sim}{\epsilon}, \tag{4.1}$$

where $\underset{\sim}{\theta} = (\theta_1, \theta_2, \ldots, \theta_{k+m})'$, $\theta_i = N^{-1} \sum_{\nu=1}^{N} \theta_{\nu,i}$, $\underset{\sim}{\epsilon} = (\epsilon_1, \epsilon_2, \ldots, \epsilon_T)'$, $\epsilon_i = N^{-1} \sum_{\nu=1}^{N} \epsilon_{\nu,i} \sim \mathrm{N}(0; \sigma^2)$, and $\sigma^2 = N^{-2} \sum_{\nu=1}^{N} \sigma_\nu^2$.

To compute the matrix B_k, the knots $z_{k+1}, \ldots, z_{k+m}$ have to be specified. (Note that $z_1 = \ldots = z_k = 1$ and $z_{k+m+1} = \ldots = z_{2k+m} = T$.) Using He and Shi (1998), this can be done by dividing the y-axis into $m+1$ equally spaced intervals and by projecting these points on the x-axis. This process is illustrated in Figure 6. This method is easy to implement and it will put more knots where the change in the function is maximum while putting fewer

knots where the function is flatter. Only the first part of the He and Shi (1998) algorithm has been used. The step-wise deletion algorithm has not been implemented yet.

4.1. Bayesian estimator of the cumulative hydrograph

In order to simplify calculations, a conjugate prior model is used. For the regression model given in equation (4.1), the conjugate prior corresponds to a g-prior (*cf.* Zellner, 1971) which is given by

$$\underset{\sim}{\theta} \mid \sigma^2 \sim \mathrm{N}_{k+m}\left(\underset{\sim}{\theta}_0; \frac{\sigma^2}{n_0}\left(B_k' B_k\right)^{-1}\right), \tag{4.2}$$

$$\sigma^2 \sim \mathrm{I\Gamma}(\alpha/2; \gamma/2). \tag{4.3}$$

(It is assumed that all the hyperparameters are known. However, $\underset{\sim}{\theta}_0$, n_0, α and γ have to be specified. For the present data, we used a noninformative prior on σ^2 by choosing $\alpha = \gamma = 0$. The prior mean of $\underset{\sim}{\theta}$ is chosen to be equal to the least squares estimator obtained using the first three years. Since n_0 represents our confidence in the prior information, we used $n_0 = 3/36$, that is the ratio of the number of years used to estimate $\underset{\sim}{\theta}_0$ over the number of remaining years.)

Using this prior model (see equations (4.2) and (4.3)), the posterior densities are then given by (*cf.* Zellner, 1971)

$$\underset{\sim}{\theta} \mid \sigma^2, \underset{\sim}{H} \sim \mathrm{N}_{k+m}\left(\underset{\sim}{\theta}_{OLS} - \frac{n_0}{n_0+1}\left[\underset{\sim}{\theta}_{OLS} - \underset{\sim}{\theta}_0\right]; \frac{\sigma^2}{n_0+1}\left(B_k' B_k\right)^{-1}\right);$$

$$\sigma^2 \mid \underset{\sim}{H} \sim \mathrm{I\Gamma}([T+\alpha]/2; \gamma^*/2),$$

where

$$\underset{\sim}{\theta}_{OLS} = \left(B_k' B_k\right)^{-1} B_k' \underset{\sim}{H},$$

$$\gamma^* = S + \gamma + \frac{n_0}{n_0+1}(\underset{\sim}{\theta}_{OLS} - \underset{\sim}{\theta}_0)' B_k' B_k (\underset{\sim}{\theta}_{OLS} - \underset{\sim}{\theta}_0),$$

$$S = (\underset{\sim}{H} - B_k \underset{\sim}{\theta}_{OLS})'(\underset{\sim}{H} - B_k \underset{\sim}{\theta}_{OLS}).$$

It can be shown that the posterior density of $\underset{\sim}{\theta}$ is a multivariate Student-T density with $T + \alpha$ degrees of freedom, location vector and scale matrix given by

$$\underset{\sim}{\theta}_{OLS} - \frac{n_0}{n_0+1}\left[\underset{\sim}{\theta}_{OLS} - \underset{\sim}{\theta}_0\right],$$
$$\frac{\gamma^*}{(n_0+1)(T+\alpha)}\left(B_k'B_k\right)^{-1},$$

respectively. Since the cumulative hydrograph is a nonnegative function, the Bayesian estimator of $\underset{\sim}{\theta}$, under the squared error loss, is given by

$$\widehat{\underset{\sim}{\theta}} = \mathbf{E}\left[\underset{\sim}{\theta} \times \mathbf{I}_{\{\theta_i \geq 0 | i=1,2,\ldots,k+m\}}(\underset{\sim}{\theta}) \mid \underset{\sim}{H}\right], \tag{4.4}$$

where $\mathbf{I}_{\{\theta_i \geq 0 | i=1,2,\ldots,k+m\}}(\underset{\sim}{\theta}) = 1$ if all the components of $\underset{\sim}{\theta}$ are nonnegative and equal to 0 otherwise.

The Bayesian estimator of the cumulative hydrograph is then given by

$$\widehat{H}_m(t) = \sum_{i=1}^{k+m} \widehat{\theta}_i B_{i,k}(t). \tag{4.5}$$

(It should be noted that for the present data, using equation (4.4) or the posterior location vector in equation (4.5) lead to similar values for $\widehat{H}_m(t)$.)

4.2. Bayesian estimator of the mean representative hydrograph

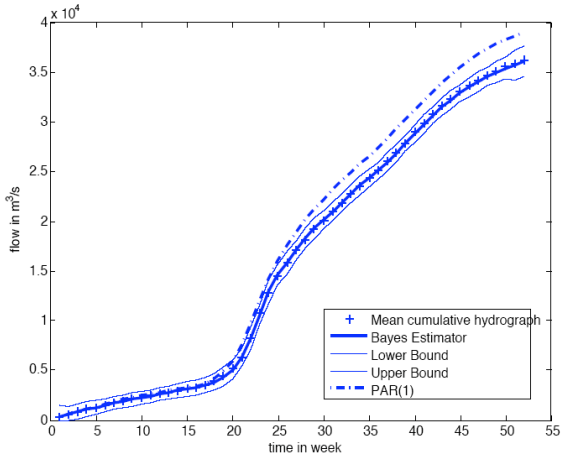
It can be shown, using the definition of B-splines, that

$$\frac{\partial}{\partial t}\left(\sum_{i=1}^{k+m} \theta_i B_{i,k}(t)\right) = \sum_{i=1}^{k+m-1} k\frac{(\theta_{i+1} - \theta_i)}{(z_{i+k+1} - z_{i+1})} B_{i+1,k-1}(t) \tag{4.6}$$

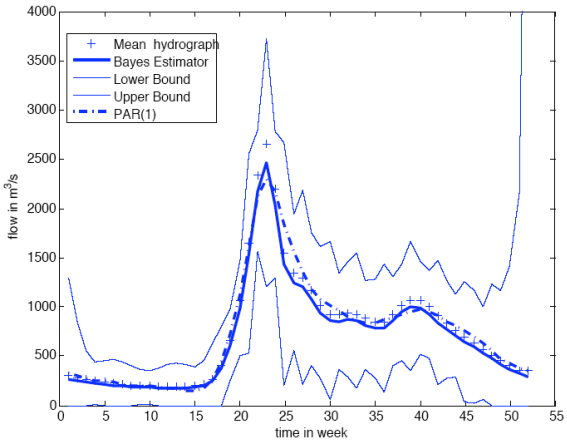
(see De Boor, 1978). Hence, the mean representative hydrograph can be estimated by taking the derivative with respect to t of equation (4.5). Consequently, it is given by

$$\begin{aligned}
\widehat{h}_m(t) &= \frac{\partial}{\partial t}\widehat{H}_m(t) \\
&= \frac{\partial}{\partial t}\sum_{i=1}^{k+m} \widehat{\theta}_i B_{i,k}(t) \\
&= k\sum_{i=1}^{k+m-1} \frac{(\widehat{\theta}_{i+1} - \widehat{\theta}_i)}{(z_{i+k+1} - z_{i+1})} B_{i+1,k-1}(t).
\end{aligned} \tag{4.7}$$

Note that, since $\widehat{h}_m(t)$ is no longer a monotone function, the lack of importance of the knot selection might not hold anymore. However, for the present data, the fit is reasonably good (see Figure 7).



(a) Cumulative hydrograph



(b) Representative hydrograph

Figure 7: Estimator of the cumulative (a) and representative (b) hydrograph for a river in Northern Québec

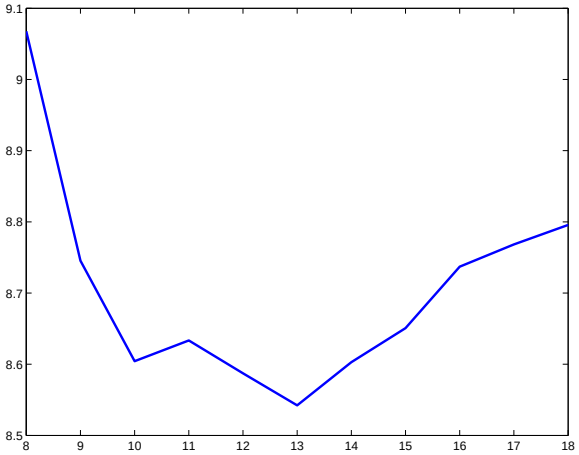


Figure 8: IC(m)

A method to choose the number of interior knots is proposed in He and Shi (1998). This method is based on the following information criterion

$$\mathrm{IC}(m) = \log\left(\sum_{t=1}^{T} |h(t) - \widehat{h}_m(t)|\right) + 2\frac{(m+2)}{T}.$$

The values of IC(m), for $m = 8, 9, \ldots, 18$ are given in Figure 8. Clearly, the best value of m corresponds to 13 interior knots. Using $m = 13$ and $k = 5$, in equation (4.5), the Bayesian estimator of the cumulative hydrograph is obtained and is displayed in part (a) of Figure 7. Its derivative, that is $\widehat{h}_m(t)$ (see equation (4.7)), is also shown in part (b) of Figure 7. From these graphs, it can be seen that the Bayesian estimators of $H(t)$ and $h(t)$ given by equations (4.5) and (4.7) respectively are very close to the registered cumulative and mean hydrographs. We also display the estimator obtained using a PAR(1) model which corresponds to the model currently used at Hydro-Québec. It can be seen that the estimator obtained from the PAR(1) model is similar to the Bayesian estimator but it does not fit the data as well. 95% confidence bands for the estimators of the cumulative and mean hydrographs are also displayed in Figure 7.

5. Conclusion

Hydrographs, which measure the quantity of water flowing through some point at a given time, show a lot of seasonal variability due to precipitations (rain and snow), current temperatures, time of year, etc. Two different approaches can be used to account for this variability.

The one currently used by Hydro-Québec is to model this variability directly. This involves a complex model with a large number of parameters to account for "random" variability. The method proposed in this paper is to register the observed hydrographs by using a deterministic function based on the location of the maximum Spring and Autumn floods. Once all the observations are registered, a representative hydrograph is computed by taking the average. Using a Bayesian regression model and a B-spline basis, a representative hydrograph is estimated. Only one practical example has been presented in this paper. However, several rivers have been studied and similar results have been obtained.

References:

1. De Boor, C. (1978). *A Practical Guide to Splines*, Springer: New York.

2. He, X. and Shi, P. (1998). Monotone B-spline smoothing. *Journal of the American Statistical Association*, **93**, 643-650.

3. Perreault, L. and Latraverse, M. (2001). Modélisation des apports naturels pour la prise en compte de leur aléa dans la méthode SDDP de planification de la production. *Technical Report*, Institut de recherche d'Hydro-Québec, Montréal

4. Ramsay, J.O. and Li, X. (1998). Curve registration. *Journal of the Royal Statistical Society, Series B*, **60**, 351-363.

5. Schumaker, L. (1981). *Spline Functions: Basic Theory*, Wiley: New York.

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

7. Yue, S., Ouarda, T.B.M.J., Beaulieu, C. and Bobée, B. (2001). Construction des hydrogrammes types annuels. *Technical Report*, INRS-Eau, Québec City.

8. Yue, S., Ouarda, T.B.M.J., Bobée, B., Legendre, P. and Bruneau, P. (2002). Approach for describing statistical properties of flood hydrograph. *Journal of Hydrologic Engineering*, **March/April**, 147-153.

9. Zellner, A. (1971). *An Introduction to Bayesian Inference in Econometrics*, Wiley: New York.