RECURSION-BASED MULTIPLE CHANGEPOINT DETECTION IN MULTIVARIATE LINEAR REGRESSION AND APPLICATION TO RIVER STREAMFLOWS

Rapport de recherche No R-843 *Mars 2006*

Recursion-based Multiple Changepoint Detection in Multivariate Linear Regression and Application to River Streamflows

O. Seidou*, T.B.M.J. Ouarda,

INRS-ETE, Hydro-Québec NSERC Chair in Statistical Hydrology/Canada Research Chair on the Estimation of Hydrological Variables, 490 rue de la Couronne, Quebec (QC), Canada G1K 9A9

Rapport de recherche R-843

Submitted to Water Resources Research
March 2006

***Corresponding author:** Tel: (418) 654-2542
Fax: (418) 654-2600
Email: ousman_seidou@ete.inrs.ca

Abstract

A large number of models in hydrology and climate sciences rely on multivariate linear regression to explain the link between key variables. The relationship in the physical world may experiment sudden changes due to climatic, environmental or anthropogenic perturbations. To deal with this issue, a Bayesian method of multiple changepoint detection in multivariate linear regression is proposed in this paper. It is an adaptation of the recursion-based multiple changepoint method of *Fearnhead* [2005] to the classical multivariate linear model. A new class of priors for the parameters of the multivariate linear model is introduced and useful formulas are derived that permit straightforward computation of the posterior distribution of the changepoints. The proposed method is numerically efficient and does not involve time consuming Monte-Carlo Markov Chain simulation as opposed to other Bayesian changepoint methods. It allows fast and straightforward simulation of the probability of each possible number of changepoints as well as the posterior probability distribution of each changepoint conditional on the number of changes. The approach is validated on simulated data sets and then compared to the methodology of *Asselin and Ouarda* [2005] on two practical problems: a) the changepoint detection in the multivariate linear relationship between mean basin scale precipitation at different periods of the year and the summer-autumn flood peaks of the Broadback River located in Northern Quebec, Canada; and b) the detection of trend variations in the streamflows of the Ogoki River located in the province of Ontario, Canada.

Keywords: Bayesian analysis, changepoint, hydrology, streamflows, multivariate linear regression.

1. Introduction

An increasing number of papers point out shifts or trends in hydrologic time series [e.g. *Burn and Elnur,* 2002; *Woo and Thorne,* 2003; *Salinger,* 2005]. A change of mentality is taking place in the whole scientific community and it is probable that hydrologic time series models which do not hold account of a possible change in the statistical distribution of the data will no longer be regarded as credible. Detection of eventual changes in collected data sets is thus obviously an important step before performing any descriptive or predictive analysis.

Changepoint analysis is addressed both in Classical and Bayesian statistics. Methods in classical statistics usually consist of performing several kinds of tests to confirm or reject the hypothesis of change. Most of them address slope or intercept change in linear regression models [*Solow*, 1987; *Easterling and Peterson*, 1995; *Vincent*, 1998; *Lund and Reeves*, 2002; *Wang*, 2003].

In Bayesian statistics, one is interested in obtaining a statistical distribution for the dates of change and eventually a distribution for the other model parameters. Bayesian changepoint analysis models are the subject of a large number of papers [e.g. *Booth and smith,* 1982; *Bruneau et Rassam*, 1983; *Gelfand et al.* 1990; *Barry and Hartigan,* 1992, 1993; *Stephens*, 1994; *Perreault et al.,* 2000a,b,c; *Rasmussen,* 2001]. More recently, *Asselin and Ouarda* [2005] developed an approach to changepoint detection in multivariate linear relationships and *Fearnhead* [2005] proposed a recursion-based inference procedure based on the theory of product-partition models [*Barry and Hartigan*, 1992,1993] for multiple changepoint problems. In the latter paper, a set of recursive relations are used to infer the posterior probabilities of different numbers of changepoints. A particularity of this approach is that it focuses only on the number and positions of changes.

The aim of this paper is to adapt the methodology of *Fearnhead* [2005] to multiple changepoint detection in multivariate linear relations. In particular, a special class of priors for the parameters of the multivariate linear model is introduced and useful formulas are derived that permit straightforward computation of the posterior distribution of the changepoints. The proposed methodology is validated on simulated data sets to prove its ability to infer the number and location of changepoints. It is then applied to two case studies. In the first case study, the summer-autumn flood peaks of the Broadback River located in the province of Quebec, Canada, are investigated for the eventual changes due to forest fires. The second case study deals with the detection of eventual trend variations in the streamflow data of the Ogoki River located in the province of Ontario, Canada.

As the first case study has already been investigated with a changepoint detection approach using Gibbs sampling [*Asselin and Ouarda,* 2005; *Seidou and Ouarda*, 2005], the results obtained with the two methodologies will be compared and discussed in this paper. The approach of *Asselin and Ouarda* [2005] will also be applied to the second case study in order to highlight the importance of having a methodology designed to handle several changepoints.

The outline of the paper is as follow: Section 2 is a quick survey of changepoint detection methodologies with an emphasis on Bayesian methodologies with application to hydrological problems. Section 3 is devoted to the methodology of *Asselin and Ouarda* [2005] which will be compared to the proposed approach. Recursion based changepoint inference models are introduced in Section 4, and the model of *Fearnhead* [2005] is adapted to multivariate linear regression. The simulation of changepoints given the conditional posterior probabilities of the dates of change is presented in Section 5. The simulation-based validation methodology is presented in section 6. Section 7 presents the results of the simulation studies and the applications

on real data are carried out in section 8. A conclusion and some recommendations are finally presented in Section 9.

2. Changepoint models

Changepoint detection has received a great deal of attention in statistical literature because modification of model structure and/or parameters is commonly encountered in applied statistics (e.g in finance, pharmacology, econometrics, hydrology, etc.). The change detection can be off-line (or retrospective) or online (or sequential) when it is important that the change be detected as soon as it occurs. Examples of online changepoint detection methods can be found in [*Lai*, 1995; *Beibel*, 1997; *Daumer and Falk*, 1998; *Gut and Steinebach*, 2002; *Daumer and Falk*, 1997; *Moreno et al*, 2005].

Most applications in hydrology are used for retrospective changepoint detection, except a few ones [e.g. *Moreno et al*, 2005]. Retrospective changepoint detection methods often use classical statistical methods to detect changes in slopes or intercepts of linear regression models [*Solow*, 1987; *Easterling and Peterson*, 1995; *Vincent*, 1998; Rasmussen, 2001; *Lund and Reeves*, 2002; *Wang*, 2003]. Other curve fitting methods are used in some rare cases [e.g. *Sagarin and Micheli*, 2001; *Bowman et al.*, 2004].

A growing number of methodologies use Bayesian statistics. *Gelfand et al* [1990] discussed Bayesian analysis of a variety of normal data models, including regression and ANOVA-type structures, where they allowed for unequal variances. *Barry and Hartigan* [1992, 1993] used product-partition models to develop a Bayesian analysis for a multiple changepoint problem that can be exactly solved using a finite number of operations. The multiple changepoint component was introduced by a normal random variable that can be added anytime to the mean of the series, but only with a certain probability. *Stephens* [1994] implemented Bayesian analysis of a multiple

changepoint problem where the number of changepoints is assumed known, but the times of occurrence of the changepoints remain unknown. Other authors emphasized on the single changepoint problem. We cite for example *Carlin et al.* [1992] who applied a three-stage hierarchical Bayesian analysis to a simple linear changepoint model for normal data: $Y_t \sim N[a_1 + b_1 x_t, \delta_1^2]$, $t = 1,...,\tau$, $Y_t \sim N[a_2 + b_2 x_t, \delta_2^2]$, $t = \tau + 1,...,n$. *Perreault et al.* [2000a; 2000b] gave Bayesian analyses of several changepoint models of univariate normal data. All of these authors implemented their analyses using Gibbs sampling. *Rasmussen* [2001] considered a single changepoint in a simple linear regression model with noninformative priors and derived the exact analytical posterior distribution of the regression parameters. His model assumes that the changepoint occurred with certainty, and does not allow a clear diagnosis of the existence of the change. *Perreault et al.* [2000c] developed an exact analytical Bayesian analysis of a changepoint in the mean of a series of multivariate normal random variables.

More recently, *Asselin and Ouarda* [2005] developed a practical and general approach to the single changepoint inference problem relying on Bayesian multivariate regression analysis. Their model can handle multivariate data and/or missing values and can be used with both informative and noninformative priors on the regression parameters. It was shown to be more performing than other approaches recently published in the hydrological literature [*Seidou and Ouarda,* 2005]. However, the approach presented in *Asselin and Ouarda* [2005] considers only one possible changepoint and involves relatively long MCMC simulations. The method presented in this paper is expected to handle theses two issues.

3. The changepoint model of Asselin and Ouarda [2005]

The model of *Asselin and Ouarda* [2005] is designed to infer the position of a change in the parameters of a multivariate regression equation. They assume that the $(r \times 1)$ response vector $\mathbf{Y}_t$

is related to the $(r \times d^*)$ matrix $\mathbf{X}_t$ by

$$\mathbf{Y}_t = \mathbf{X}_t \boldsymbol{\theta}_t^{(\tau_c)} + \boldsymbol{\upsilon}_t \qquad [1a]$$

where

$$\boldsymbol{\theta}_t^{(\tau_c)} = \begin{cases} \boldsymbol{\beta}_1^*, & 1 \le t \le \tau_c, \\ \boldsymbol{\beta}_2^*, & \tau_c < t \le n, \end{cases} \qquad [1b]$$

under the constraints

$$\boldsymbol{\beta}_1^* = (\boldsymbol{\beta}_1, \boldsymbol{\beta}_0)^T \text{ and } \boldsymbol{\beta}_2^* = (\boldsymbol{\beta}_2, \boldsymbol{\beta}_0)^T. \qquad [1c]$$

In these equations as well as in the remainder of the paper, bold letters indicate vectors and matrices while the superscript ^{T} indicates the transpose. In equation [1b], τ_c is the last point of the segment before the changepoint, and $\tau_c = n$ means that there is no change in the data series. The dimensions of the vectors $\boldsymbol{\theta}_t^{(\tau_c)}$, $\boldsymbol{\beta}_1^*$, $\boldsymbol{\beta}_,^*$ $\boldsymbol{\beta}_0$, $\boldsymbol{\beta}_1$, $\boldsymbol{\beta}_2$ are respectively $(d^* \times 1)$, $(d^* \times 1)$, $(d^* \times 1)$, $(d_0^* \times 1)$, $(d_1^* \times 1)$ and $(d_1^* \times 1)$. Equation [1c] implies that $d^* = d_0^* + d_1^*$. It is also assumed that error terms $\{\boldsymbol{\upsilon}_t\}$ are independent and identically distributed following $N[0, \Sigma_y]$.

The model assumes a changepoint in the $(d^* \times 1)$ vector $\boldsymbol{\theta}_t^{(\tau_c)}$ from the $(d_1^* \times 1)$ subvector $\boldsymbol{\beta}_1$ to the $(d_1^* \times 1)$ subvector $\boldsymbol{\beta}_2$. The $(d_0^* \times 1)$ subvector $\boldsymbol{\beta}_0$ is assumed to remain part of $\boldsymbol{\theta}_t^{(\tau_c)}$ throughout the observation series.

In *Asselin et Ouarda* [2005], some algebraic transformations allowed to apply some known results on Bayesian piecewise linear regression to Model [1] and to infer its parameters. The MCMC algorithm was also designed to account for missing data in the observations record and/or in the explanatory variables. Finally, they considered a general prior specification for regression parameters as well as for the variance structure, and used Gibbs sampling to obtain empirical

posterior distributions for each parameter. For extensive details on prior specification and MCMC inference for model [1] we refer the reader to the original paper.

4. Recursion based changepoint inference

Although recursions have been used to make inference on the number of changepoints [*Yao*, 1984; *Barry and Hartignan*, 1992, 1993], this kind of approach has been less widely used than MCMC based inference. Yao [1984] was the first to show that Bayesian inference for a single shift in a normally distributed sample can be performed in a finite number of recursive operations. As the number of operations grows quickly when the length of the data series increases, he also proposed an approximate inference for which the number of operations is reduced to the order of sample size. *Barry and Hartignan* [1992, 1993] showed that the changepoint problem can be elegantly handled using product-partition models and generalized the results of Yao [1984] to multiple changepoints and more general prior assumptions. Product partition models assume that observations in a random partition of the data are independent, and allow the data to weight the partitions that hold. The methodologies presented in these papers under this approach allow for an efficient computation of the posterior probability of different number of changepoints using recursive relations. *Fearnhead* [2005] used this kind of recursive relations to develop a general inference procedure for the number and positions of the changepoints.

4.1 General Inference procedure for the number and positions of the changepoints

Fearnhead [2005] considered a class of multiple changepoint models for which the number of changes is unknown. Let $\{y_1, y_2, ..., y_n\}$ be the sample, n the sample size, m the number of changepoints, $\tau_0 = 0, \tau_1, ..., \tau_{m+1} = n$ the changepoints and $\mathbf{Y}_{i:j}$ the observations from time i to time

j. We also denote $g(.)$ the probability distribution of the time interval between consecutive changepoints and $g_0(.)$ the probability distribution of the first changepoint. The j^{th} segment is then $y_{(\tau_{j-1}+1):\tau_j}$ with parameter Φ_j.

Assuming that the observations are independents conditional on the changepoints and parameter values, *Fearnhead* [2005] derived the posterior probability of the changepoints:

$$\begin{cases} \Pr(\tau_1 \mid \mathbf{Y}_{1:n}) = P(1,\tau_1)Q(\tau_1+1)g_0(\tau_1)/Q(1) \\ \Pr(\tau_j \mid \tau_{j-1}, \mathbf{Y}_{1:n}) = P(\tau_{j-1}+1,\tau_1)Q(\tau_j+1)g(\tau_j-\tau_{j-1})/Q(\tau_{j-1}+1) \end{cases} \quad [2]$$

where $P(t,s), \quad s \geq t$ is the probability that t and s be in the same segment:

$$\begin{aligned} P(t,s) &= \Pr(\mathbf{Y}_{t:s}; t,s \text{ in the same segment}) \\ &= \int \prod_{i=t}^{s} f(y_i \mid \Phi)\pi(\Phi)d\Phi \end{aligned} \quad [3]$$

and $Q(t)$ is the likelihood of the segment $\mathbf{Y}_{t:n}$ given a changepoint at $t-1$. $Q(t) \quad t=1,..,n$ and $P(t,s), \quad s \geq t$ are linked by these recursive equations:

$$\begin{cases} Q(1) = \sum_{s=1}^{n-1} P(1,s)Q(s+1)g_0(s) + P(1,n)(1-G_0(n-1)) \\ Q(t) = \sum_{s=1}^{n-1} P(t,s)Q(s+1)g_0(s+1-t) + P(t,n)(1-G(n-t)) \end{cases} \quad [4]$$

where $G(t) = \sum_{i=1}^{t} g(i)$ and $G_0(t) = \sum_{i=1}^{t} g_0(i)$.

4.2. Adaptation of the changepoint inference procedure to multivariate linear regression

Consider the n_p+1 series of data $y_j, j=1,\ldots n$ and $x_{ij}, i=1,\ldots,d^*; j=1,\ldots n$ where x_{ij} is the j^{th} value of the i^{th} series of explanatory variables. The multivariate linear relationship can be represented by

$$y_j = \sum_{k=1}^{d^*} \theta_k x_{ij} + \varepsilon_i \qquad i = 1,\ldots,n \qquad [5]$$

or

$$\mathbf{y} = \mathbf{X}\boldsymbol{\theta} + \boldsymbol{\varepsilon} \qquad [6]$$

The parameter vector Φ is thus given by $\Phi = [\theta_1\ \theta_2 \ldots \theta_{d^*}\ \sigma]$ and we have:

$$f(y_i \mid \Phi) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left(-0.5 \left(\frac{y_i - \sum_{j=1}^{d^*} \theta_j x_{ij}}{\sigma} \right)^2 \right) \qquad [7]$$

Following *Rasmussen* [2001], we have:

$$\Pr(\mathbf{y}_{t:s} \mid \Phi) = \prod_{i=1}^{s} f(y_i \mid \Phi) = (2\pi\sigma^2)^{-n/2} \exp\left[-\frac{(\mathbf{Y}_{t:s} - \mathbf{X}_{t:s}\boldsymbol{\theta})^T (\mathbf{Y}_{t:s} - \mathbf{X}_{t:s}\boldsymbol{\theta})}{2\sigma^2} \right] \qquad [8]$$

From [3] and [8] we have:

$$\begin{aligned} P(t,s) &= \Pr(\mathbf{Y}_{t:s}; t,s \text{ in the same segment}) = \int\prod_{i=t}^{s} f(y_i \mid \Phi)\pi(\Phi)d\Phi \\ &= \int (2\pi\sigma^2)^{-n/2} \exp\left[-\frac{(\mathbf{Y}_{t:s} - \mathbf{X}_{t:s}\boldsymbol{\theta})^T (\mathbf{Y}_{t:s} - \mathbf{X}_{t:s}\boldsymbol{\theta})}{2\sigma^2} \right] \pi(\Phi)d\Phi \end{aligned} \qquad [9]$$

Assume that the prior depends only on σ and has this particular form:

$$\pi(\Phi) = \pi(\sigma) = p(\sigma \mid a, C) = \frac{\sigma^{-a} \exp(-\frac{c}{2\sigma^2})}{2^{\frac{a-3}{2}} c^{-\frac{a-1}{2}} \Gamma(\frac{a-1}{2})}, a > 1, c > 0 \qquad [10]$$

In equation [10], the denominator $2^{\frac{a-3}{2}} c^{-\frac{a-1}{2}} \Gamma(\frac{a-1}{2})$ is only a normalizing constant that ensures that $\int_0^{+\infty} \pi(\sigma)d\sigma = 1$. Note that when σ is very large, $p(\sigma)$ tends towards a multiple of σ^{-a}.

Jeffrey's non informative prior for linear regression is $p(\boldsymbol{\theta},\sigma) \propto \sigma^{-\frac{d^*}{2}}$ [*Minka*, 2001], and it is sometimes assumed in Bayesian linear regression that $p(\sigma) \propto \sigma^{-1}$ [e.g. *Rasmussen*, 2001]. Unfortunately, these kinds of priors are improper contrarily to the one proposed in equation [10]. Basic properties of $p(\sigma \mid a,c)$ are derived in Appendix 1.

Finally, the expression of $P(s,t)$ is obtained after substituting equation [10] in equation [9] and integrating out σ and $\boldsymbol{\theta}$ in equation [9]:

$$P(t,s) = (2\pi)^{\frac{d^*}{2}} \frac{\left(\pi(\boldsymbol{\varepsilon}_{s:t}^T \boldsymbol{\varepsilon}_{s:t} + c)\right)^{-\frac{(t-s+a)}{2}}}{(c\pi)^{-\frac{a-1}{2}} \left|\mathbf{X}_{s:t}^T \mathbf{X}_{s:t}\right|^{1/2}} \frac{\Gamma\left(\frac{t-s+a}{2}\right)}{\Gamma(\frac{a-1}{2})} \qquad [11]$$

Exhaustive details on how the expression of $P(s,t)$ is obtained are given in Appendix 2.

5. Simulation of changepoints given the conditional posterior probabilities of the changepoints

The relations presented in Section 4 give only the posterior probability mass of the first changepoint, and the conditional probability mass of subsequent changepoints. To make inference on the positions of changepoints, we simulate a set $E = \{S_k, k = 1:M\}$ of M possible scatter schemes of the changepoints on the segment using the posterior probability mass of the first changepoint, and the conditional probability mass of subsequent changepoints. Indeed, M should be large enough to obtain a reliable distribution for the positions of the changepoints. The k^{th} element of E is a set of m_k changepoints $S_k = \{\tilde{t}_1^k, \tilde{t}_2^k, ..., \tilde{t}_{\tilde{m}_k}^k\}$. An efficient simulation algorithm for E is given by *Fearnhead* [2005]:

1. For a sample of size M, initiate M samples with a changepoint at $t = 0$.

2. For $t = 0,...,n-2$, repeat the following steps:

   a) Compute the number n_t of samples for which the last changepoint was at time t;

   b) If $n_t > 0$, compute $\Pr(\tau \mid \tau_{j-1} = t, \mathbf{y}_{1:n})$;

   c) Sample n_t times from $\Pr(\tau \mid \tau_{j-1} = t, \mathbf{y}_{1:n})$ and use the values to update the n_t samples of changepoints which have a changepoint at time t;

This algorithm is very efficient since $\Pr(\tau \mid \tau_{j-1} = t, \mathbf{y}_{1:n})$ has to be computed only one time regardless of the number of samples required from it. Inference on the number and positions of the changepoints is readily carried out using the M samples. For instance, the probability of having i changepoints is approximated by:

$$\Pr(m = i) \approx card(\{k \mid card(S_k) = i\}) / M \qquad [12]$$

The posterior probability of having the k^{th} changepoint at position t given m changepoints can be approximated by:

$$\Pr(\tau_i = t \mid m) \approx \frac{card\left(\left\{k \mid \left(card(S_k) = m\right) \& \left(\tilde{t}_i^k = t\right)\right\}\right)}{card(\{k \mid card(S) = m\})} \qquad [13]$$

where $card(S)$ stands for the number of elements of the set S. The estimators of the number and positions of changepoints are the modes of their posterior distributions, i.e:

$$\hat{m} = \underset{t}{Max}\{card(\{k \mid card_k(S) = t\}) / M\} \qquad [14]$$

$$\hat{\tau}_i = \underset{t}{Max}\left\{\frac{card\left(\left\{k \mid \left(card(S_k) = \hat{m}\right) \& \left(\tilde{t}_i^k = t\right)\right\}\right)}{card\left(\left\{k \mid card(S) = \hat{m}\right\}\right)}\right\} \qquad [15]$$

Other estimators can be defined using the posterior distributions but in Bayesian analysis the mode of the posterior distribution is generally the best estimator.

6. Validation methodology

The validation of the proposed method requires large data sets in which all the characteristics of the changepoints are known. These data sets were obtained by simulation using a procedure that mimics the ranges of shifts and trends that are usually observed in streamflow data. The ability of the method to correctly detect the number and position of changes was assessed using four performance measures that are described further in the text.

6.1. Simulated data sets

Artificial shifts and trends with random magnitudes and positions were inserted in three sets of simulated normal series. The first set contains series which only display shifts in the mean. The series in the second set contain abrupt changes of trend, while the changepoints in the third set can be either shifts or changes in trend.

The series in the first data set were simulated in the following manner:

1) Set the number of series to generate (N), the minimum number of points between changepoints ($l_{\min}$) and the maximum magnitude of the shift $\delta_{\max}$;
2) Set u to 1;
3) Simulate a set $\{\mathbf{Y}\}_u = \{y_i, i = 1,...,n\}$ of n random numbers from the normal distribution with mean 0 and standard deviation 1;
4) Simulate the number of changes by uniformly drawing a number m in $\{0,1,...,m_{\max}\}$;
5) For each i in $\{1,...,m\}$, if $n - l_{\min} - \tau_{i-1} > 0$, uniformly draw a changepoint position τ_i in $\{\tau_{i-1} + l_{\min},...,n\}$. Repeat this step until τ_m is sampled;
6) For each i in $\{1,...,m\}$, if $n - l_{\min} - \tau_{i-1} > 0$, uniformly draw a shift magnitude δ_i in $[-\delta_{\max}, \delta_{\max}]$;

7) For each i in $\{1,...,m\}$, set $y_k = y_k + \delta_i, k = \tau_i + 1,...,n$;

8) If $u < N$, increment u and return to step 3, otherwise end the simulation procedure.

The second data set is generated in the same manner except that trend changes rather than shifts are introduced in the series. In that case, if we denote tr_i the trend in the (i+1)th first segment, all the above listed steps hold, except the seventh step that should be replaced by this one:

7.a) For each i in $\{0,...,m\}$, set $y_v = y_v + tr_i\left(x_k - x_{\tau_i+1}\right), v = \tau_i + 1,...,n$.

In the third data set, the changes can either be a shift in the mean or a change of trend. The type of change is randomly selected using a binomial distribution with parameter 0.5.

6.2. Performance measures

Let's denote m_u the number of changepoints in the u^{th} generated sample $\{\mathbf{Y}\}_u$ and $\{t_i^k, i = 1:m_u\}$ their positions. Let $\hat{m}_u$ be the estimate of m_u, and $\{\hat{t}_i^k, i = 1:\hat{m}_u\}$ the estimates of the positions of the $\hat{m}_u$ detected changepoints. Two simple measures of the ability of the proposed approach to detect the number of changepoints are the Percentage of Correct Detections of the Number of changepoints (*PCDN*) and the Root Mean Square Error (*RMSE*) of the estimations of the number of changepoints defined as follow:

$$PCDN = \frac{1}{M}\sum_{u=1}^{M} 1_{\{\hat{m}_u = m_u\}} \qquad [16]$$

$$RMSE = \sqrt{\frac{1}{M}\sum_{u=1}^{M}(\hat{m}_u - m_u)^2} \qquad [17]$$

Another measure of the capability of the method to correctly estimate the number of changepoints is the Ranked Probability Score (*RPS*): if F_u denotes the empirical cumulative probability

distribution of m_u obtained with the application of the changepoint detection method, the *RPS* can be defined as follow:

$$RPS = \frac{1}{M}\sum_{u=1}^{M}\sum_{i=1}^{n}(F_u(i) - 1_{i \geq m_u})^2 \text{ where } 1_{i \geq m_u} = \begin{cases} 1 \; if \; i \geq m_u \\ 0 \; if \; i < m_u \end{cases} \quad [18]$$

The *RPS* is usually used to rate ensemble forecasts [e.g *Buiza and Palmer*, 1998; *Hamil*, 2001]. The *RPS* values are within $[0, n-1]$ and a value of zero is obtained for perfect forecasts.

Unfortunately, the *RPS* is designed to rate the prediction for a single variable and cannot be easily applied to the estimators of the positions of changepoints, as the number of detected changepoints may be different from the real number of changepoints. A new performance measure was thus developed as follows: let $\{\mathbf{Y}\}_u$ be a series generated as described in Section 6.1 with m_u changepoints $\{t_j^u, j = 1:m_u\}$. The application of the changepoint detection approach to $\{\mathbf{Y}\}_u$ will give a set $E = \{S_k, k = 1:M\}$ of M possible scatter schemes where $S_k = \{\tilde{t}_1^k, \tilde{t}_2^k, ..., \tilde{t}_{\tilde{m}k}^k\}$ has $\tilde{m}_k$ elements. $\tilde{m}_k$ may be different from the real number of changes m_u in $\{\mathbf{Y}\}_u$. Given k and u, consider $\{a_i, i = 1, ..., \min(\tilde{m}_k, m_u)\}$ and $\{b_i, i = 1, ..., \min(\tilde{m}_k, m_u)\}$ such that $i \neq j \Rightarrow a_i \neq a_j$, $i \neq j \Rightarrow b_i \neq b_j$ and $\sum_{i=1}^{\min(\tilde{m}_k, m_u)} \left(t_{a_i}^u - \tilde{t}_{b_i}^k\right)^2$ is minimal. The performance of the changepoint detection method when applied to the generated series $\{\mathbf{Y}\}_u$ can be measured with the Multiple Change Detection Performance Index (*MCDPI*) defined as

$$MCDPI_k = \begin{cases} \frac{1}{\tilde{m}_k}\sum_{i=1}^{\tilde{m}_k}\left(t_{a_i}^u - \tilde{t}_{b_i}^k\right)^2, & \tilde{m}_k = m_u \\ \frac{1}{\tilde{m}_k}\left(\sum_{i=1}^{m_u}\left(t_{a_i}^u - \tilde{t}_{b_i}^k\right)^2 + \sum_{j\neq b_i, i=1,\ldots,m_u} \tilde{t}_j^k\left(n - \tilde{t}_j^k\right)\right), & \tilde{m}_k > m_u \\ \frac{1}{m_u}\left(\sum_{i=1}^{\tilde{m}_k}\left(t_{a_i}^u - \tilde{t}_{b_i}^k\right)^2 + \sum_{j\neq a_i, i=1,\ldots,\tilde{m}_k} t_j^u\left(n - t_j^k\right)\right), & \tilde{m}_k < m_u \end{cases} \quad [19]$$

The introduction of a_i and b_i is motivated by the need to associate as much as possible each element of the set of real changepoints to an element of the set of detected changepoints. Note that $\{a_i, i = 1,\ldots,\min(\tilde{m}_k, m_u)\}$ and $\{b_i, i = 1,\ldots,\min(\tilde{m}_k, m_u)\}$ are different for each pair (u,k). This association is performed using a minimum square distance criterion. The penalty term for the false detection of a change $\tilde{t}_j^k$ is $\tilde{t}_j^k\left(n - \tilde{t}_j^k\right)$; the penalty for the non detection of the change t_j^u is $t_j^u\left(n - t_j^u\right)$. These penalty terms have the interesting property of not over-penalising false detections at the beginning and at the end of the series. They are consistent with the practice of discarding detected changes that are close to the end or the beginning of the series [*Beaulieu et al.*, 2005].

The overall performance is the mean of the criterion over the set of generated series

$$MCDPI = \frac{1}{N}\sum_{k=1}^{N} MCDPI_k \quad [20]$$

7. Settings and results of the simulation studies

The prior for σ and the parameters for the data generation algorithms were first chosen to have a noninformative prior. Three data sets were generated according to the procedure described in Section 6.1 and changepoints are identified with the proposed procedure. A two-column vector of

explanatory variables was considered, the first one containing only ones and the second containing the date of the observation.

7.1. Prior specification for σ

As pointed out in Section 4.2, the prior variance of σ (equation 1.5 of Appendix 1) is infinite when $a < 3$. Any value lower than 3 is thus a relatively noninformative prior. We chose $a = 2$ to be consistent with the classical $p(\sigma) \propto \sigma^{-2}$ usually used in Bayesian linear regression. As in equation [11] c has the dimension of a variance, it was set to the variance obtained by least square estimates of the linear regression equations, i.e.:

$$c = \boldsymbol{\varepsilon}_{1:n}^T \boldsymbol{\varepsilon}_{1:n} = Y_{1:n}^T Y_{1:n} - X_{1:n}(X_{1:n}^T X_{1:n})^{-1} Y_{1:n}^T Y_{1:n}) . \qquad [21]$$

7.2. Parameters of the simulations

The number of series in each of the three simulated data sets was set to 1000. The length of the series was fixed to 75. The number of changepoints varies from zero to three with at least ten epochs between changepoints, and the shifts were assumed to have a magnitude ranging between zero and five times the standard deviation of the data series. The magnitudes of the trends are assumed inferior to three standard deviations per ten epochs. These values are consistent with the authors experience with changes observed in streamflows data series.

7.3. Performance of the proposed method on simulated data sets

The changepoint detection method was applied to each simulated data set with a two-column vector of explanatory variables. The first column of this vector contains only ones while the second column contains the dates of the observations. Including the dates of observations in the vector of explanatory variables allows the detection of changes in trend in the data series. The performance of the changepoint detection method on the first two simulated data sets was

compiled as a function of the number of real changepoints and the minimum magnitude of the change in a given series. Similar results were compiled for the third simulated data set, but only using the number of changepoints since the series contained two kinds of changes with different definitions of the magnitude. These results are summarized in Tables 1 and 2 (resp. Tables 3 and 4) and plotted in Figure 1 (resp. Figure 2) for the first (resp. second) simulated data set. The same results are presented in Table 5 and Figure 3 for the third simulated data set. Analysis of these results allows drawing the following conclusions:

a) The rate of false detection is very low since the *PCDN* is close to 100% when $m_u = 0$ (Figures 1a, 1u, 2a, 2u and 3a). The *PCDN* is remains very high when there is only one real change $m_u = 1$ (Figures 1e, 2e) and, as expected, it increases when the minimum magnitude of the change increases. The same conclusions can be drawn from all the other performance measures considering that a good forecast means small *RMSE*, *RPS* and *MCDPI* values.

b) The performance indices (except the *MCDPI*) decrease with the number of changepoints (c.f. Figures 1 and 2);

c) It seems easier for the method to detect shifts than changes in trend (Figure 1 vs Figure 2), although the relative performance depends on the range of change of magnitude in each set. This conclusion holds only if we consider that the range of magnitudes that were generated is representative of the real world.

Results suggest that in this particular case (series of 75 years) the method can be trusted if the shifts in the data set have the order of magnitude of the standard deviation, and if the number of changes is known to be inferior to three. Indeed, the performance should not be the same for other data sets with different lengths and different statistical characteristics. However, since the data

sets were generated to cover the range of magnitudes generally encountered in streamflow records, the method proposed in this paper will be useful for detecting changes in river discharges. It can also be used in several other problems involving multivariate linear regression, such as data homogenization or signal processing.

8. Application to cases studies

The methodology is applied herein to two case studies to illustrate its behaviour on real data and to compare it to the approach of *Asselin and Ouarda* [2005]. The first case study deals with change detection in the linear regression describing the relationship between Summer-Autumn flood peaks and precipitations on the Broadback River basin. *Seidou and Ouarda* [2005] studied this data set using the Bayesian single changepoint detection method of *Asselin and Ouarda* [2005] and found that the relation has significantly changed after 1972 ($\tau_c = 1972$). As in their paper, the changepoint τ_c corresponds to the last point on the segment before the change and differs from the definition that was used in this paper (first point of the segment after the change), the expected value of τ with the approach proposed in this paper should be 1973.

The second case study is an example drawn from the Canadian Reference Hydrometric Basin Network (RBHN) data base [*Brimley et al.*, 1999]. The case was selected because it displayed a relatively large number of changes.

8.1. Changepoint detection in the linear regression describing the relationship between Summer-Autumn flood peaks and precipitations on the Broadback River basin

8.1.1. The data

The Broadback River has a catchment of 17100 km^2 and experiences forest fire bursts from time to time (Figure 4). According to the Canadian Large Fire Database [*Stocks et al.*, 2002; *Natural*

Resources Canada, 2005], major forest fires occurred during the summer of 1971, burning 506 km^2 in the upper parts of the catchment (1/34 of the total basin area). It can be hypothesized that the deforestation due to these fires can change the basin response function to meteorological inputs. In order to perform the analysis, the 1961-1981 daily flood discharges at station 80801 were obtained from Quebec Ministry of the Environment. The Broadback River is subject to two types of floods: spring floods, which are dominated by snowmelt, and summer-autumn floods which are caused by direct liquid precipitations. Figure 5a presents the mean daily discharge at this station for the 1961-1981 period. It appears that the summer-autumn maximum flood peak is generally observed at the end of October (Figure 5a). Daily precipitations for the July-October period from 1961 to 1981 were obtained by interpolation from the neighbouring weather stations on a regularly spaced grid of 100*100 points and averaged to have a time series representing precipitation at the catchment scale. This time series was then used to obtain the mean precipitation on the Broadback river catchment for every half month period from July to October. Exploratory analysis of the linear relationship between observed flood discharge and the obtained precipitation series led to the choice of four explanatory variables for the flood peak values: 1) the mean precipitations of the 16^{th}-31^{st} of July period, 2) the sum of precipitations of the 1^{st}-15^{th} of August period, 3) the sum of precipitations of the 16^{th}-31^{st} of August period and 4) total precipitations for the September-October period. The values of the 1961-1981 summer-autumn flood peaks are presented in Figure 5b and those of the chosen explanatory variables in Figure 5c. Figure 5d presents the burned areas on the catchment for each year of the period of study. The series of explanatory variables as well as the maximum flood peaks are summarized in Table 6.

8.1.2. Results

The application of the changepoint detection method leads to a probability of 0.2 for the absence of changepoints and 0.8 for the existence of a unique changepoint (Figure 6a). A small weight (<0.01) is attributed to the existence of two changes. The posterior probability distribution of the changepoint τ is illustrated in Figure 6b. The posterior probability distribution of τ_c obtained with the same data set by *Seidou and Ouarda* [2005] with the Bayesian method of *Asselin and Ouarda* [2005] is also presented in Figure 6c. The two methods agree that the changepoint occurred probably between 1973 and 1974, with however different weights for these two dates. The weight differences may be due to the differences in the prior specifications of the two methods, and to the uncertainty introduced by the use of limited samples when computing the posterior distribution with the two approaches.

8.2. Shifts and trend change detection in the flood peaks of the Ogoki river

8.2.1. The data

The Ogoki River is a 480 km long river located in the province of Ontario, Canada. It flows northeast from lakes west of Lake Nipigon to join the Albany River which ends into the James Bay. Station 04GB004 (Ogoki River above Whiteclay Lake) is part of the Canadian Reference Hydrometric Basin Network (RHBN) which comprises stations that have been carefully selected for climate change detection and assessment studies [*Brimley et al.*, 1999]. The RHBN network comprises stations that are pristine. Station 04GB004 was selected because it displays a relatively large number of changepoints. The location of this station is given in Figure 7.

8.2.2. Results

The results of the changepoint analysis of the Ogoki River streamflows with the method proposed in this paper are presented in Figure 8. The results obtained with the approach of *Asselin and Ouarda* [2005] are provided in Figure 9. The posterior probability distribution of the number of changepoints obtained with the proposed method is plotted in figure 8a. Up to 4 changepoints are plausible ($\Pr(m = 4) > 0$), but the most probable number of changepoints is two. Figures 8b and 8c provide the posterior probability distributions of the first and second changepoints, conditional to $m = 2$. The position of each of these changepoints is chosen to be the mode of the posterior distribution: 1961 for the first changepoint and 1971 for the second changepoint. Given these positions, the posterior means of the three segments in the data series are readily computed (Figure 8d). According to the analysis, the flows of the Ogoki River displayed a negative downward trend from 1951 to 1961, and increased regularly from 1960 to 1970. From 1960 to the present date, the streamflow record displayed a small downward trend.

Figure 9a illustrates the posterior probability distribution of the changepoint obtained with the methodology of *Asselin and Ouarda* [2005]. This method gives less than 0.01 probability of no change (with this method, the probability of no change is equal to the probability that the changepoint is at the end of the data series). The mode of the posterior distribution of the date of change corresponds to 1967. This date corresponds grossly to the mean of the two changepoints detected with the methodology presented in this paper. This indicates that the results of the two methods are consistent. Although the method of Asselin and Ouarda [2005] is designed to detect only one change, a multimodal posterior distribution is often the sign of the existence of more than one changepoint. In this example, the fact that the posterior distribution is bimodal suggests

that there may be another changepoint in 1955. However, this seems to have been caused rather by the high discharge observed in 1954 than by a real change of trend in the data series.

Since the causes of trend change in the streamflow record are not known, it is impossible to decide whether the results of one or the other of the two methods correspond to the reality. The main advantage of the proposed approach is that it has less constraints and gives a larger chance for the data to influence the posterior distributions. The proposed approach is thus preferable in cases where there is only one response variable, where no data is missing and where more than one change is plausible. The results presented in this work are also easier to interpret than those of the approach proposed by *Asselin and Ouarda* [2005]

9. Conclusions and recommendations

A Bayesian method of multiple changepoint detection in multivariate linear regression is developed and validated with both simulated data and real data sets. The paper also proposes a new class of priors for the parameters of the multivariate linear model, as well as useful formulas that permit straightforward computation of the posterior distribution of the positions of changepoints. Results suggest that, in the particular case of series with 75 observations, the proposed method can be trusted if the shifts in the data set have the order of magnitude of the standard deviation, and if the number of changes is known to be inferior to three. It is also shown that in cases where there is only one response variable, where no data is missing and where more than one change is plausible, it is better to use the proposed methodology instead of *Asselin and Ouarda* [2005].

The extension of the work presented in this paper to much more general models is straightforward since the most important equations were obtained without assumptions on model structure. An interesting direction for future work is the development of similar approaches for hidden Markov

chain Models. Much more complex changepoint problems can be handled in the framework of hidden Markov chain models, especially those which display serial dependence structure in the observations [e.g *Thyer and Kuczera*, 2003a,b].

10. Acknowledgements

The financial support provided by the Natural Sciences and Engineering Research Council of Canada (NSERC), and the Nordic Study Center (CEN) is gratefully acknowledged. The authors are also grateful to the Quebec Ministry of the Environment for having provided the data sets used in the case studies.

List of symbols

$\boldsymbol{\beta}_1^*$	Regression parameters before the changepoint in the methodology of *Asselin and Ouarda* [2005]
$\boldsymbol{\beta}_2^*$	Regression parameters after the changepoint in the methodology of *Asselin and Ouarda* [2005]
$\boldsymbol{\beta}_0$	Component of the vector of regression parameter that does not change in the methodology of *Asselin and Ouarda* [2005]
$\boldsymbol{\beta}_1$	Component of the vector of regression parameter that change to $\boldsymbol{\beta}_2$ after τ_c t in the methodology of *Asselin and Ouarda* [2005]
$\boldsymbol{\beta}_2$	Component of the vector of regression parameter that change replaces $\boldsymbol{\beta}_2$ after τ_c in the methodology of *Asselin and Ouarda* [2005]
$\boldsymbol{\varepsilon}$	Vector of random errors in the linear regression equation (one response variable)
$\boldsymbol{\varepsilon}_{s:t}$	Part of the vector of random errors between s and t
$\boldsymbol{\upsilon}_t$	Vector of random errors in the linear regression equation (several response variables)
Φ	Parameters of the linear regression equation
Σ_y	Variance-covariance matrix of the distribution of $\boldsymbol{\upsilon}_t$
τ_c	Last point of the segment before the change (methodology of *Asselin and Ouarda* [2005])
τ_k	k^{th} changepoint in the proposed methodology
$\boldsymbol{\theta}$	Vector of regression parameters
$\boldsymbol{\theta}_t^{(\tau_c)}$	Vector of regression parameters at date t given τ_c (methodology of *Asselin and Ouarda* [2005])
a	Parameter of the prior distribution of Φ

c	Parameter of the prior distribution of Φ
d^*	Number of explanatory variables
d_0^*	Number of explanatory variables for which the regression coefficients do not change
d_1^*	Number of explanatory variables for which the regression coefficients display a change (methodology of *Asselin and Ouarda* [2005])
E	Set of generated scatter schemes
$G(t)$	Cumulative probability distribution of the time interval between consecutive changepoints
$g(t)$	Probability distribution of the time interval between consecutive changepoints
$G_0(t)$	Cumulative probability distribution of the first changepoint
$g_0(t)$	Probability distribution of the first changepoint
k	Number of generated scatter schemes $S_k = \{\tilde{t}_1^k, \tilde{t}_2^k, ..., \tilde{t}_{mk}^k\}$ in the inference procedure
M	Number of scatter schemes to generate with the posterior distributions of the positions of changepoints
MCDPI	Multiple Change Detection Performance Index
m_u	number of changes in the u^{th} generated series
$\hat{m}_u$	Estimate of the number of changes in the u^{th} generated series
$\tilde{m}_k$	Number of changes in the k^{th} generated scatter scheme during the simulation of the changepoints
n	Length of the data series
N	Number of sets to generate
$P(t,s), \quad s \geq t$	Probability that t and s be in the same segment.
PCDN	Percentage of Correct Detections of the Number of changepoints
$Q(t)$	Likelihood of the segment $\mathbf{Y}_{t:n}$ given a changepoint at $t-1$
r	Number of response variables (methodology of *Asselin and Ouarda* [2005])
RMSE	Root Mean Square Error
RPS	Ranked Probability Score
$S_k = \{\tilde{t}_1^k, \tilde{t}_2^k, ..., \tilde{t}_{m_k}^k\}$	k^{th} scatter scheme generated with the posterior distributions of the positions of changepoints
t	Time
$\tilde{t}_i^k$	Estimate of the i^{th} change in the k^{th} generated scatter scheme
t_i^k	i^{th} change in the k^{th} generated Scatter scheme
u	Number of the generated series $\{\mathbf{Y}\}_u$ in the validation procedure
$\mathbf{X}$	Vector of explanatory variables
$\mathbf{X}_t$	t^{th} row of the vector of explanatory variables
$\mathbf{X}_{t:s}$	Rows t to s of the vector of explanatory variables

$\mathbf{Y}_t$	Rows t to s of the vector of response variables
$\{\mathbf{Y}\}_u$	u^{th} generated series in the validation procedure

References

Asselin, J.J, Ouarda, T.B.M.J. (2005). Bayesian Multivariate Linear Regression with Application to Changepoint Models in Hydrometeorological Variables. Part I. Model Development. Submitted to *Water Resources Research.*

Barry, D., and Hartigan, J. A. (1992). Product Partition Models for Change Point Models. *The Annals of Statistics* **20**:260-279.

Barry, D., and Hartigan, J. A. (1993). A Bayesian Analysis for Change Point Problems. *Journal of the American Statistical Association* **88:**309-319.

Beaulieu, C., Ouarda, T.B.M.J. and Seidou, O. (2005). Comparative study of homogenization techniques for precipitation data series (*in French*). Progress report No 3 (Project on the homogenization of precipitation data). Ouranos Consortium, Montreal.

Beibel, M. (1997). Sequential change-point detection in continuous time when the post-change drift is unknown. Bernoulli Journal of Mathematical Statistics and Probability 3(4): 457-478

Booth, N.B. and Smith, A.F.M. (1982). A Bayesian approach to retrospective identification of change-points. Journal of Econometrics **19** :7–22.

Brimley, B., Cantin, J.F., Harvey, D., Kowalchuk, M., Marsh, P., Ouarda, T.B.M.J., Phinney, B., Pilon, P., Renouf, M., Tassone, B., Wedel, R. and T. Yuzyk (1999). Establishment of the reference hydrometric basin network (RHBN). Research report, Environment Canada, 41p.

Bruneau, P. and Rassam, J.-C. (1983). Application d'un modèle bayésien de détection de changements de moyennes dans une se´rie. *Journal des Sciences Hydrologiques* **28**: 341–354.

Buizza, R. and Palmer, T. N. (1998). Impact of Ensemble Size on Ensemble Prediction. Monthly Weather Review **126**(9): 2503-2518.

Burn, D.H. and Hag Elnur, M.A. (2002). Detection of hydrologic trends and variability. *Journal of hydrology* **255**: 107-122.

Daumer, M. and Falk, M. (1997). On-line detection (for state space models) using multi-process Kalman filters. Linear Algebra and its applications 284: 125:135.

Easterling D.R. and Peterson T.C. (1992) Techniques for detecting and adjusting for artificial discontinuities in climatological time series: a review. *Proc. of the Fifth International Meeting on Statistical Climatology, 22-26 june 1996, Toronto, Ontario, Canada.*

Fearnhead. P. (2005). *Exact and Efficient Bayesian inference for Multiple Changepoint Problems.* Preprints, online [http://www.maths.lancs.ac.uk/~fearnhea/PScpt.ps].

Gelfand, A. E., Hills, S. E., Racine-Poon, A., and Smith, A. F. M. (1990). Illustration of Bayesian Inference in Normal Data Models Using Gibbs Sampling. *Journal of the American Statistical Association* ,**85**: 972-985.

Gut, A. and Steinebach, J. (2002). Truncated Sequential Change–point Detection Based on Renewal Counting Processes.Scandinavian Journal of Statistics 29(4):693-719.

Hamil, T.M. (2001). Interpretation of Rank Histograms for Verifying Ensemble Forecasts. Monthly Weather Review **129**(3): 550-560.

Lai, T.L. (1995) Sequential change point detection in quality control and dynamical systems. *Journal of the Royal Statistical Society,* (Serie B) **57**:613–658

Lund, R. and Reeves, J. (2002) Detection of undocumented changepoints: A revision of the two-phase regression model. *Journal of Climate* **15**, 2547-2554.

Minka, P. (2001). Bayesian linear regression. Unpublished paper. Online [https:// research.microsoft.com/~minka/papers/minka-linear.ps.gz]

Moreno, E., Casella, G., and Garcia-Ferrer, A. (2005). An objective Bayesian analysis of the change point problem. Stochastic Environmental Research and Risk Assessment (SERRA) 19(3):191 - 204

Natural Resources Canada. (2005). *Canadian Large Fires Database.* Online document [http://fire.cfs.nrcan.gc.ca/Downloads/LFDB/LFD_5999_e.ZIP]. downloaded on August 2005.

Perreault, L., Bernier, J., Bobée, B., and Parent, E. (2000a). Bayesian change-point analysis in hydrometeorological time series 1. Part 1. The normal model revisited. *J. of Hydrology* **235**: 221-241.

Perreault, L., Bernier, J., Bobée, B., and Parent, E. (2000b). Bayesian change-point analysis in hydrometeorological time series 2. Part 2. Comparison of change-point models and forecasting. *Journal of Hydrology* **235**: 242-263.

Perreault, L., Haché, M., Slivitzky, M., and Bobée, B. (1999). Detection of changes in precipitation and runoff over eastern Canada and U.S. using a Bayesian approach. *Stochastic Environmental Research and Risk Assessment* **13**:201-216.

Perreault, L., Parent, É., Bernier, J., and Bobée, B. (2000c). Retrospective multivariate Bayesian change-point analysis: A simultaneous single change in the mean of several hydrological sequences. *Stochastic Environmental Research and Risk Assessment*,**14**: 243-261.

Rasmussen, P. (2001). Bayesian estimation of change points using the general linear model. *Water Resources Research* **37**:2723-2731.

Salinger, M. (2005). Climate Variability and Change: Past, Present and Future – An Overview. *Climatic Change* **70**: 9-29

Seidou, O., Ouarda, T.B.M.J. (2005). Bayesian Multivariate Linear Regression with Application to Changepoint Models in Hydrometeorological Variables. Part II. Cases studies. Submitted to *Water Resources Research.*

Stephens, D. A. (1994). Bayesian Retrospective Multiple-changepoint Identification. *Applied Statistics* **43**: 159-178.

Solow, A.R. (1987). Testing for climate change: an application of the two-phase regression model. *Journal of Applied Meteorology* **26**, 1401-1405.

Stocks, B.J.; Mason, J.A.; Todd, J.B.; Bosch, E.M.; Wotton, B.M.; Amiro, B.D.; Flannigan, M.D.;Hirsch, K.G.; Logan, K.A.; Martell, D.L.and Skinner, W.R. (2002). Large forest fires in Canada, 1959–1997. *Journal of Geophysical Research* (107,8149,doi:10.1029/2001 JD000484).

Thyer, M and Kuczera, G. (2003a). A hidden Markov model for modelling long-term persistence in multi-site rainfall time series 1. Model calibration using a Bayesian approach. *Journal of Hydrology* 275:12–26

Thyer, M and Kuczera, G. (2003b). A hidden Markov model for modelling long-term persistence in multi-site rainfall time series. 2. Real data analysis. *Journal of Hydrology* 275:27–48.

Vincent, L.A. (1998). A technique for the identification of inhomogeneities in Canadian temperature series. *Journal of climate Climate* **11**, 1094-1105.

Wang, X.L. (2003). Comments on 'Detection of Undocumented Changepoints: A revision of the Two-Phase regression model'. *Journal of Climate* **16**, 3383-3385.

West, M. (1984). Outlier models and prior distributions in Bayesian linear regression. Journal of the Royal Statistical Society (Ser. B), **46:** 431-439

Woo, M. and Thorne, R. (2003). Comment on 'Detection of hydrologic trends and variability' by Burn, D.H. and Hag Elnur, M.A., 2002. Journal of Hydrology 255, 107-122. *Journal of hydrology* **277**:150-160.

Yao, Y. (1984). Estimation of a noisy discrete-time step function: Bayes and empirical Bayes approaches. *The Annals of Statistics* 12: 1434:1447

Table 1 : Performance of the changepoint detection procedure as function of the number of real changepoints and the minimum magnitude of the shift for the first simulated set.

Number of changes	Minimum magnitude of the shift	*PCDN* (%)	*RMSE*	*RPS*	*MCDPI*
0	-	96.62	0.28	0.08	28.54
1	0.5 σ	82.50	0.57	0.23	27.31
1	1 σ	85.94	0.51	0.18	23.58
1	1.5 σ	92.35	0.38	0.10	16.28
1	2 σ	95.24	0.30	0.07	12.49
1	2.5 σ	97.20	0.23	0.05	10.35
1	3 σ	97.39	0.23	0.05	8.62
1	3.5 σ	97.78	0.21	0.04	7.47
1	4 σ	99.42	0.11	0.03	6.93
1	4.5 σ	99.07	0.14	0.03	7.21
1	5 σ	100.00	0.00	0.01	6.37
2	0.5 σ	61.63	0.96	0.55	22.97
2	1 σ	68.43	0.81	0.43	18.87
2	1.5 σ	74.90	0.70	0.34	16.02
2	2 σ	81.81	0.58	0.25	13.65
2	2.5 σ	88.41	0.47	0.17	11.04
2	3 σ	91.77	0.40	0.13	9.63
2	3.5 σ	91.50	0.40	0.14	8.99
2	4 σ	95.12	0.31	0.09	6.79
2	4.5 σ	100.00	0.00	0.01	4.55
2	5 σ	100.00	0.00	0.00	3.63
3	0.5 σ	37.28	1.53	1.06	22.62
3	1 σ	43.08	1.40	0.92	20.45
3	1.5 σ	48.30	1.23	0.79	18.67
3	2 σ	60.82	1.00	0.57	15.35
3	2.5 σ	67.29	0.91	0.48	14.22
3	3 σ	72.76	0.75	0.38	12.08
3	3.5 σ	55.47	0.83	0.45	13.20
3	4 σ	70.71	0.71	0.37	10.04
3	4.5 σ	100.00	0.00	0.00	3.37

Table 2 : Performance of the changepoint detection procedure as function of the number of real changepoints for the first simulated set.

Number of changes	*PCDN* (%)	*RMSE*	*RPS*	*MCDPI*
0	96.62	0.28	0.08	28.54
1	82.50	0.57	0.23	27.31
2	61.63	0.96	0.55	22.97
3	37.28	1.53	1.06	22.62

Table 3: Performance of the changepoint detection procedure as function of the number of real changepoints and the minimum magnitude of the shift for the second simulated set.

Number of changes	Minimum magnitude of the trend (per ten epochs)	*PCDN* (%)	*RMSE*	*RPS*	*MCDPI*
0	-	97.33	0.23	0.06	27.03
1	0.5 σ	80.71	0.59	0.26	33.56
1	1 σ	88.95	0.46	0.15	25.86
1	1.5 σ	93.42	0.36	0.09	20.38
1	2 σ	96.80	0.25	0.06	16.97
1	2.5 σ	98.13	0.19	0.05	13.42
1	3 σ	97.65	0.22	0.04	12.31
2	0.5 σ	36.59	1.10	0.74	28.03
2	1 σ	42.64	1.04	0.65	25.22
2	1.5 σ	54.29	0.93	0.54	21.80
2	2 σ	56.11	0.89	0.49	19.78
2	2.5 σ	63.96	0.77	0.40	17.56
2	3 σ	91.29	0.41	0.19	17.27
3	0.5 σ	12.57	1.82	1.47	27.27
3	1 σ	16.55	1.71	1.35	25.80
3	1.5 σ	18.79	1.61	1.26	23.64
3	2 σ	17.96	1.56	1.21	21.60
3	2.5 σ	44.72	1.61	1.04	20.78
3	3 σ	70.71	1.41	0.88	20.41

Table 4 : Performance of the changepoint detection procedure as function of the number of real changepoints for the second simulated set.

Number of changes	*PCDN* (%)	*RMSE*	*RPS*	*MCDPI*
0	97.33	0.23	0.06	27.03
1	80.71	0.59	0.26	33.56
2	36.59	1.10	0.74	28.03
3	12.57	1.82	1.47	27.27

Table 5 : Performance of the changepoint detection procedure as function of the number of real changepoints for the third simulated set.

Number of changes	*PCDN* (%)	*RMSE*	*RPS*	*MCDPI*
0	97,19	0,26	0,05	26,08
1	80,85	0,59	0,27	31,46
2	53,61	1,06	0,66	26,85
3	21,74	1,74	1,30	26,20

Table 6: Basin scale precipitation and summer-autumn flood peak time series for the Broadback river basin.

Year	Total precipitation for the July 16th-31st period (mm)	Total precipitation for the August 1st-15th period (mm)	Total precipitation for the August 16th-31st period (mm)	Total precipitation for the September-October period (mm)	Summer-Autumn maximum flood peak (m^3/s)
1961	47.60	24.99	29.85	110.71	535
1962	79.61	45.34	70.96	90.98	714
1963	46.52	55.41	55.76	101.69	433
1964	69.96	30.52	36.23	132.00	762
1965	56.37	49.07	53.60	146.21	572
1966	44.56	59.93	33.27	213.33	796
1967	37.91	34.25	13.84	216.20	847
1968	49.04	52.02	54.45	152.14	745
1969	102.94	88.15	57.50	157.51	702
1970	53.04	55.06	68.32	102.24	586
1971	38.67	38.29	76.19	157.44	399
1972	29.98	61.48	50.26	137.10	552
1973	75.31	39.16	71.57	135.31	612
1974	33.14	59.81	48.58	168.72	1140
1975	66.11	43.33	59.15	104.56	493
1976	42.46	41.89	60.29	69.45	603
1977	57.16	61.02	41.64	126.90	759
1978	56.95	57.92	37.51	97.12	632
1979	59.22	49.73	73.62	143.59	1060
1980	66.02	20.74	61.98	124.47	478
1981	70.38	27.73	88.40	123.76	705

FIGURES CAPTION

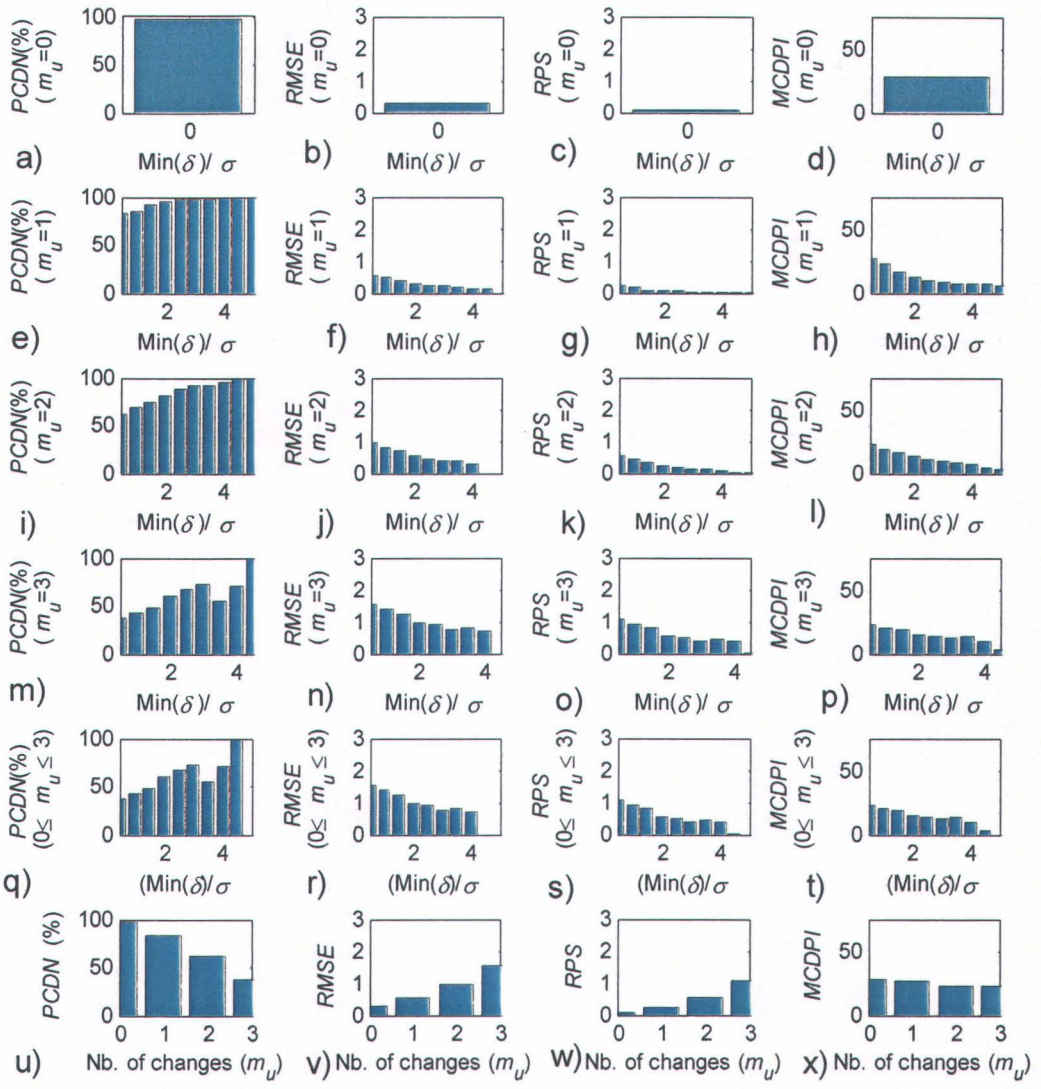


Figure 1: Performance of the changepoint detection procedure as function of the number of real changepoints and the minimum magnitude of the shift for the first simulated set.

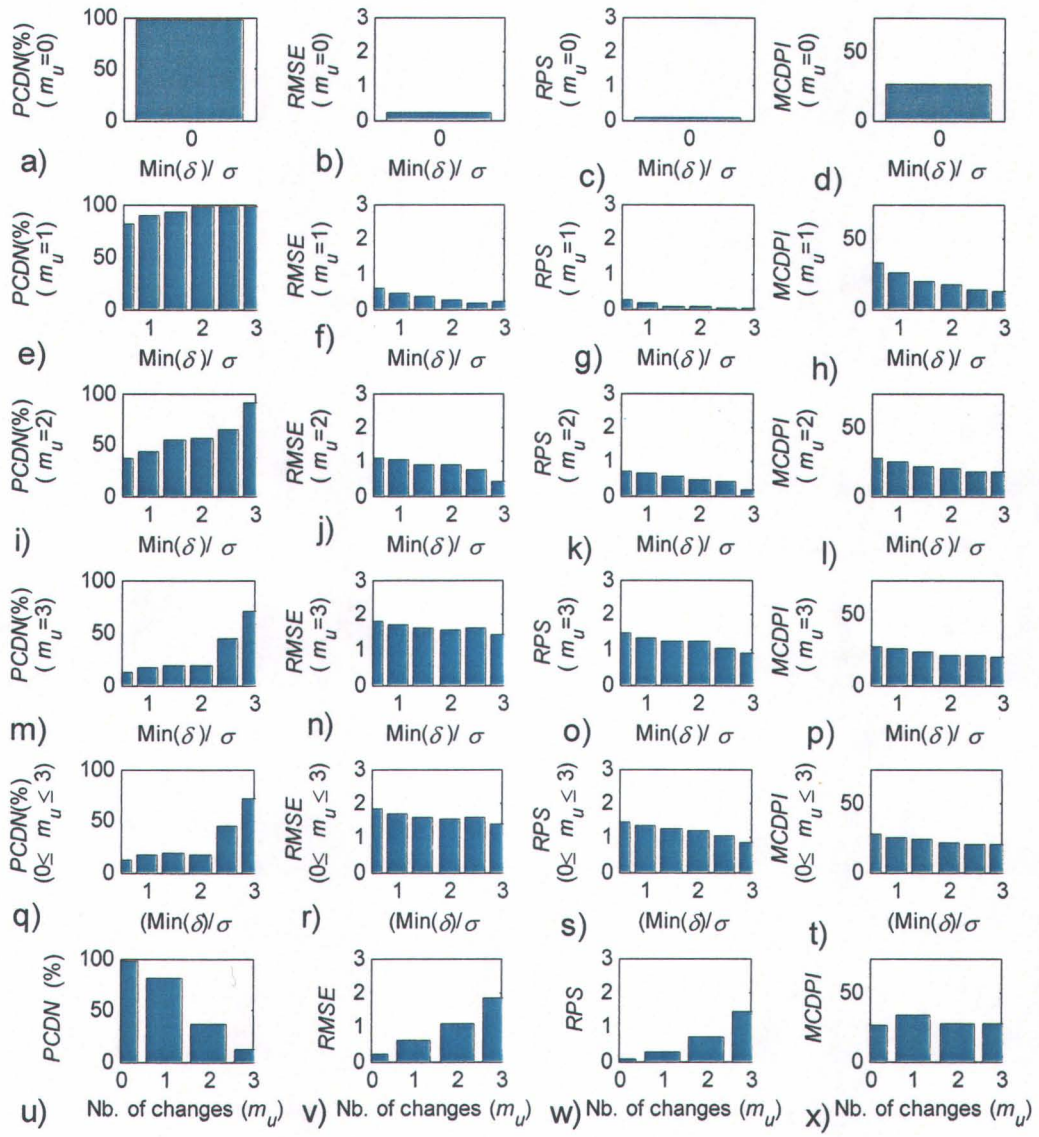


Figure 2: Performance of the changepoint detection procedure as function of number of real changepoints and the minimum magnitude of the shift for the second simulated set.

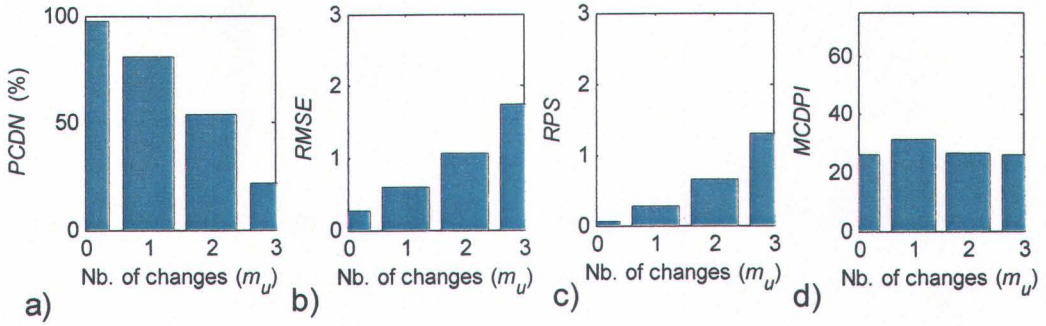


Figure 3: Performance of the changepoint detection procedure as function of the number of real changepoints for the third simulated set.

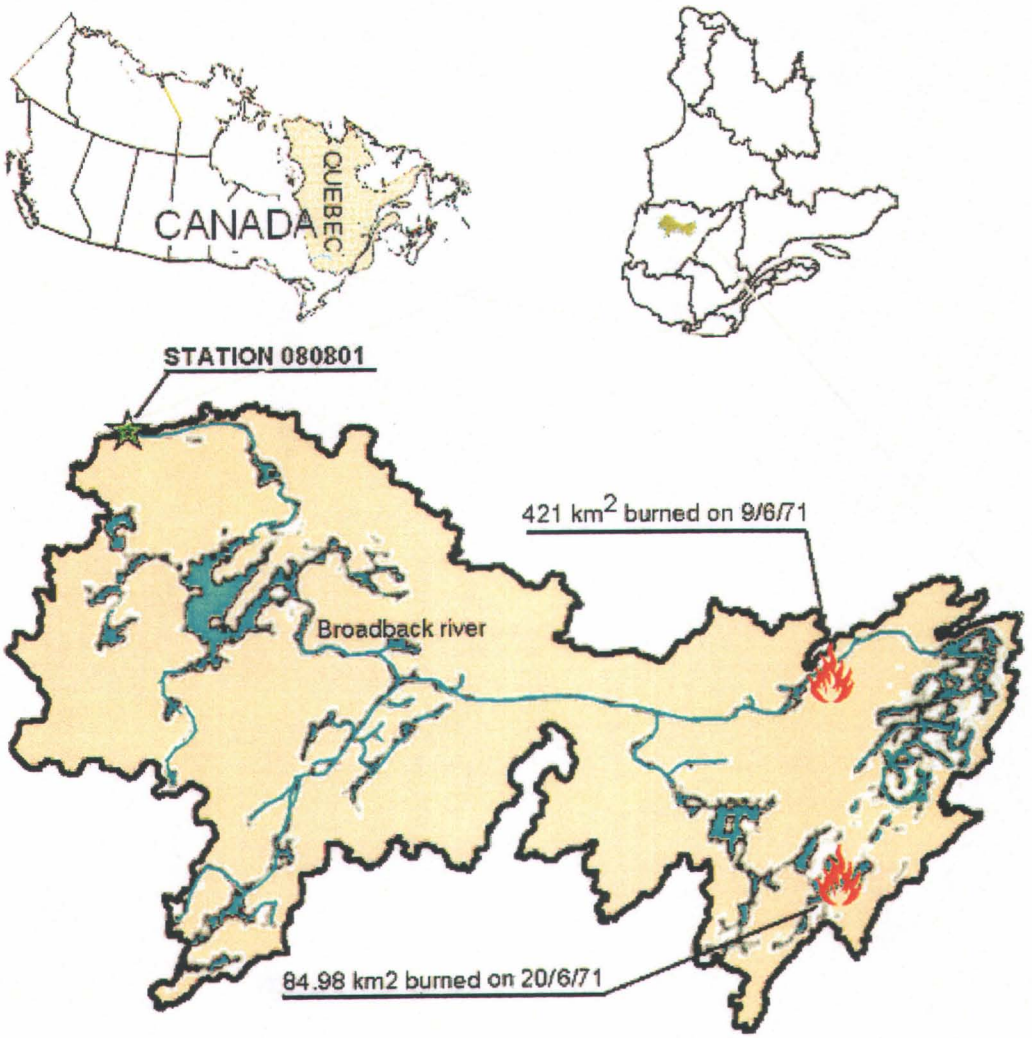
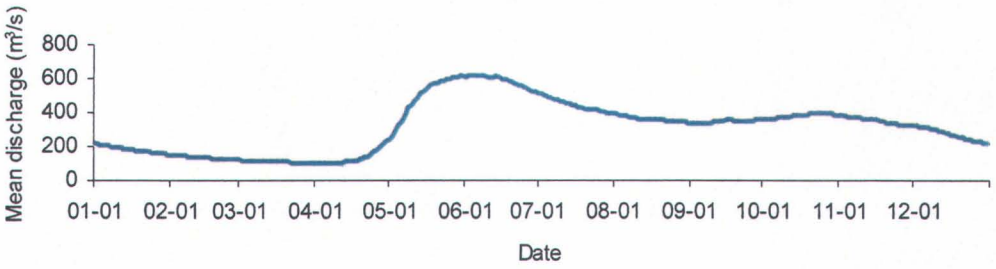
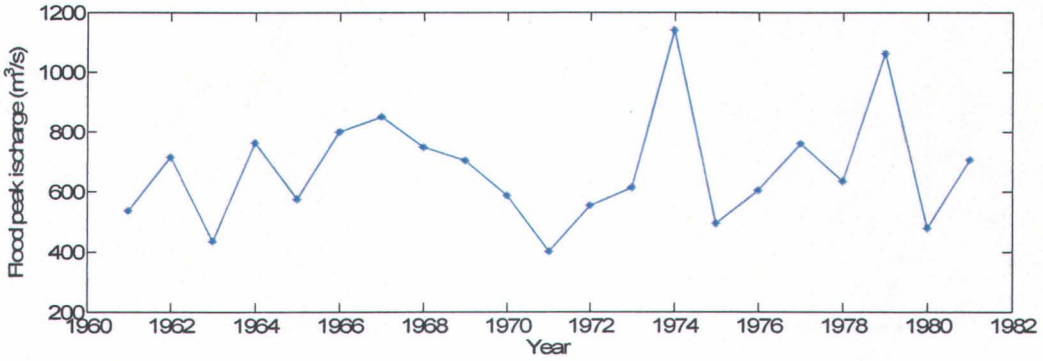


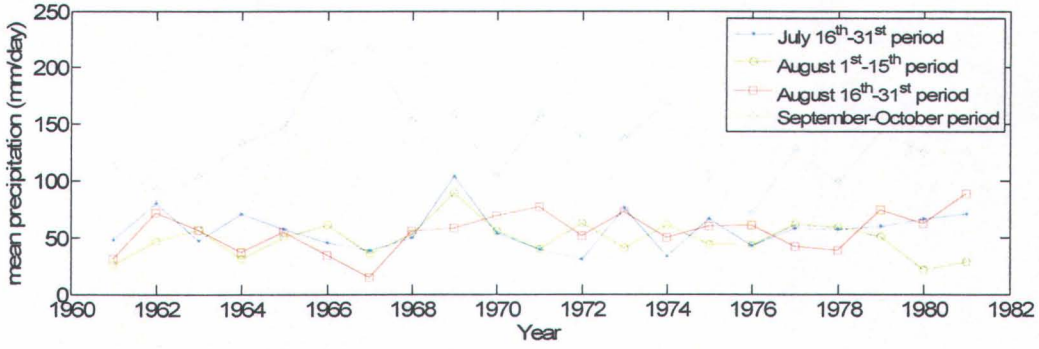
Figure 4: Location map of station 080801 (broadback River).



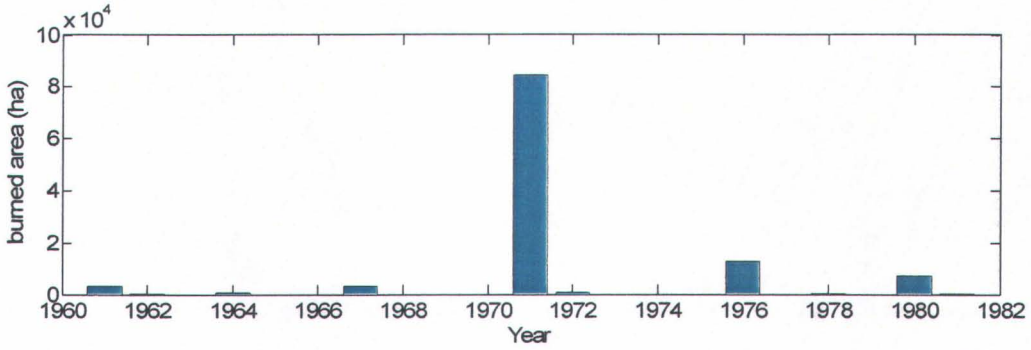
a)



b)

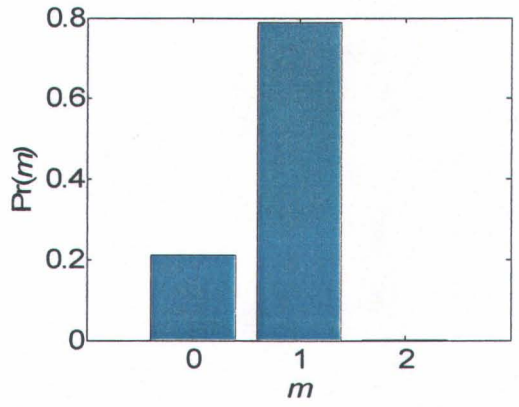


c)

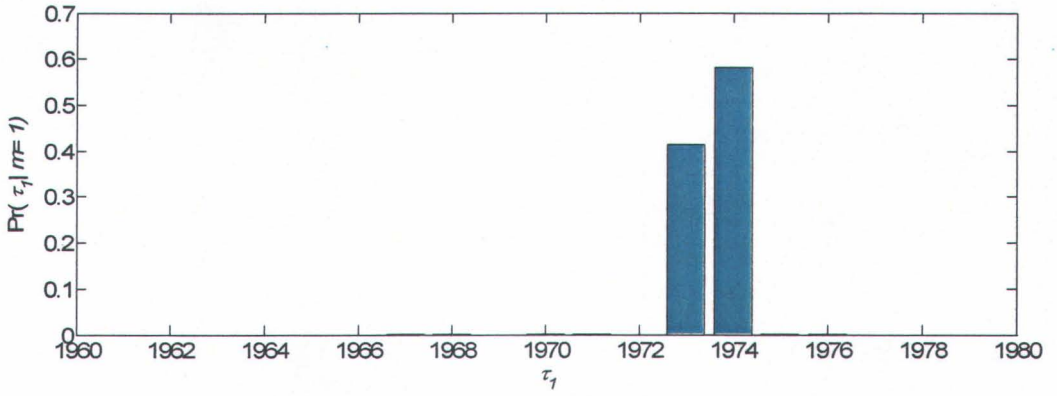


d)

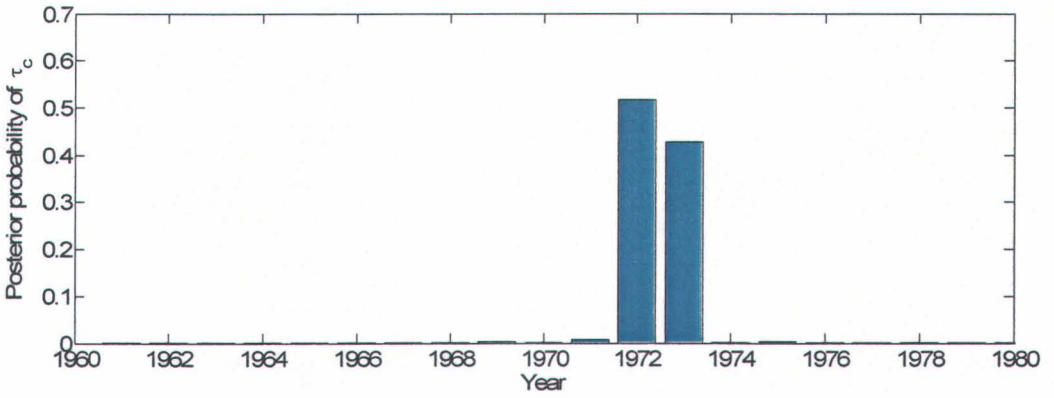
Figure 5: Data for changepoint detection in summer-autumn flood peaks for the Broadback river: a) mean hydrograph; b) summer-autumn flood peak time series; c) precipitation time series; d) burned catchment area time series.



a)



b)



c)

Figure 6 : Changepoint detection in summer-autumn flood peaks of the Broadback river: a) posterior probability of the number of changepoints; b) posterior probability of the first point of the segment after the changepoint obtained with the proposed methodology; c) posterior probability of the last point of the segment before the change obtained with the methodology of *Asselin and Ouarda* [2005].

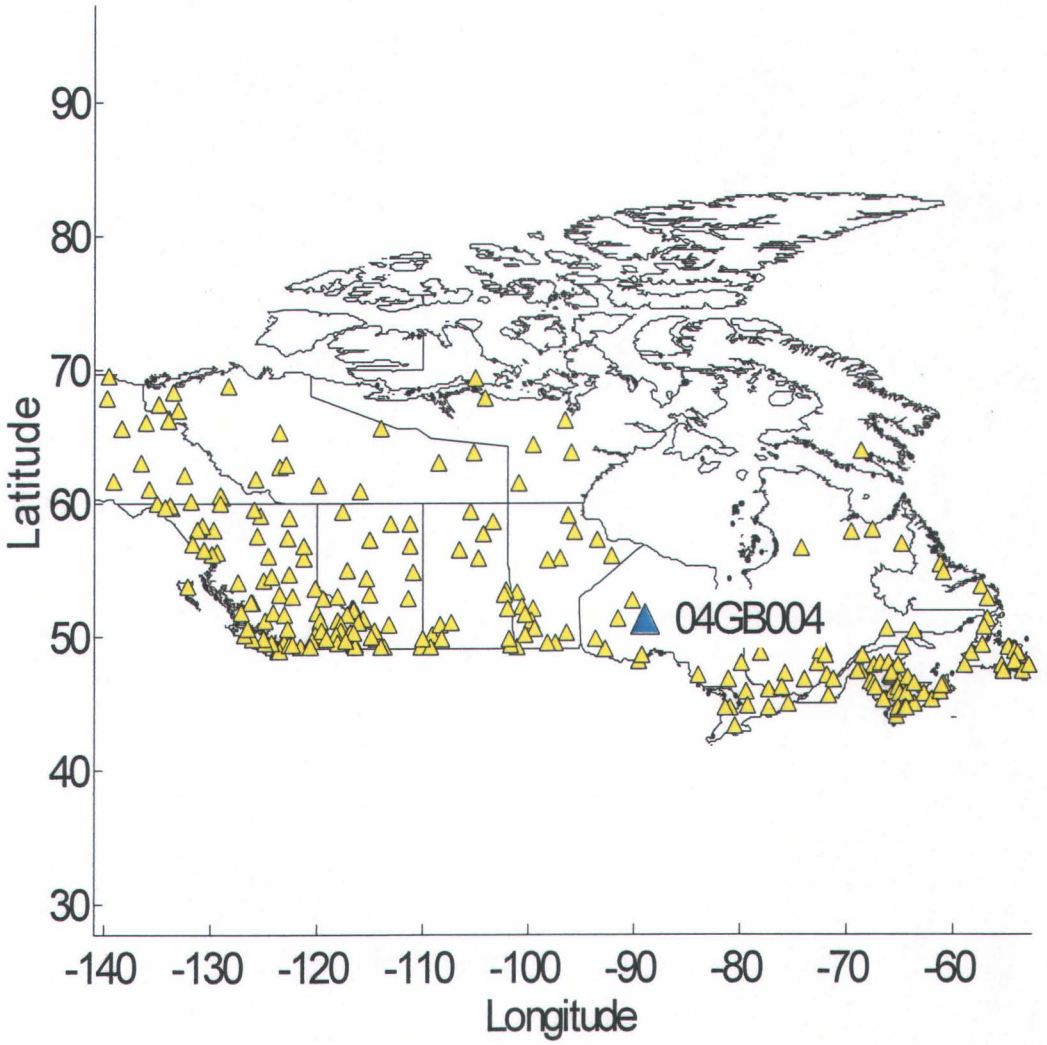


Figure 7: Location of station 04GB004 (Ogoki River above Whiteclay Lake).

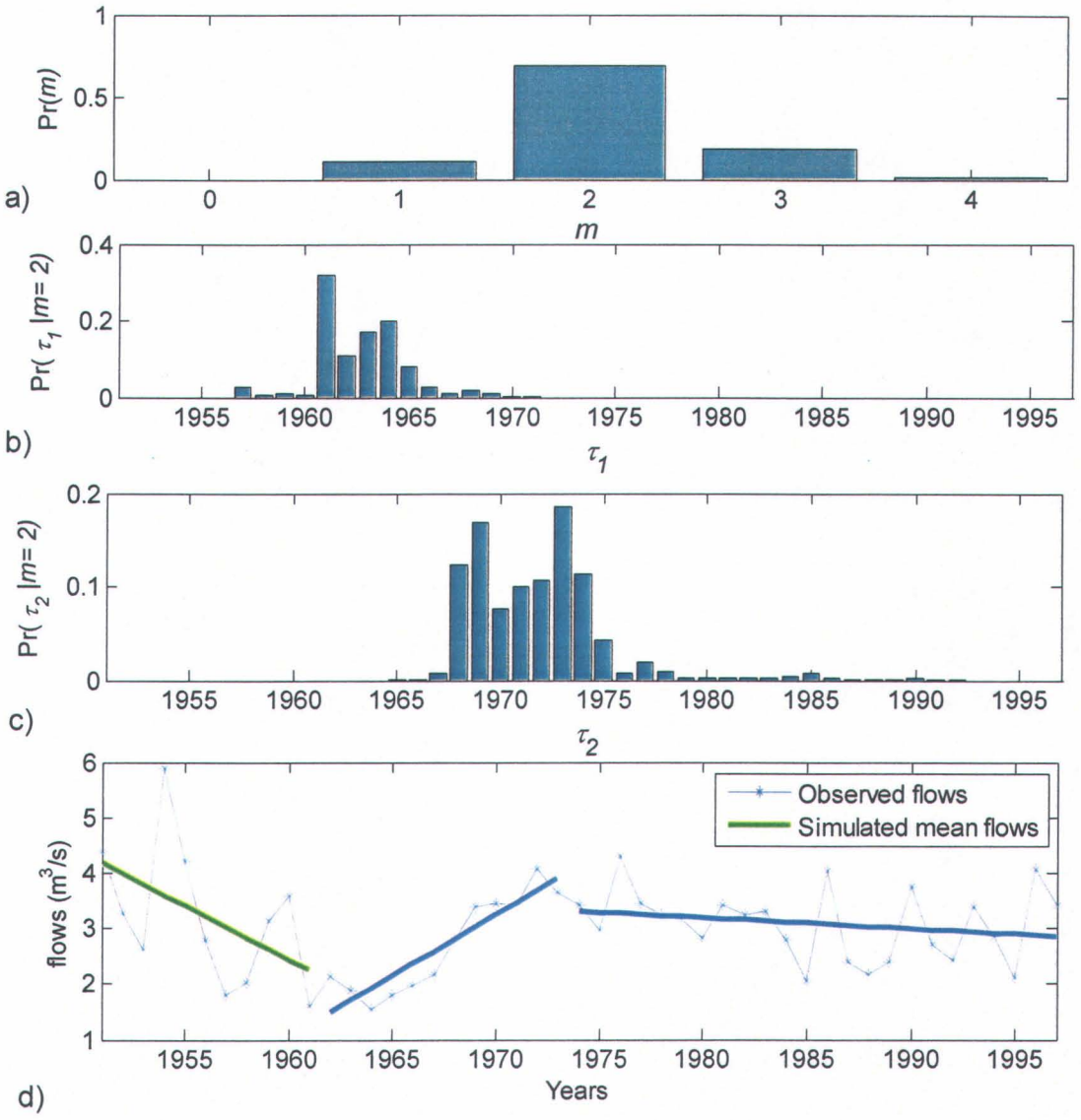


Figure 8: detection of trend changes at station 04GB004 (Ogoki River above Whiteclay Lake) with the proposed methodology.

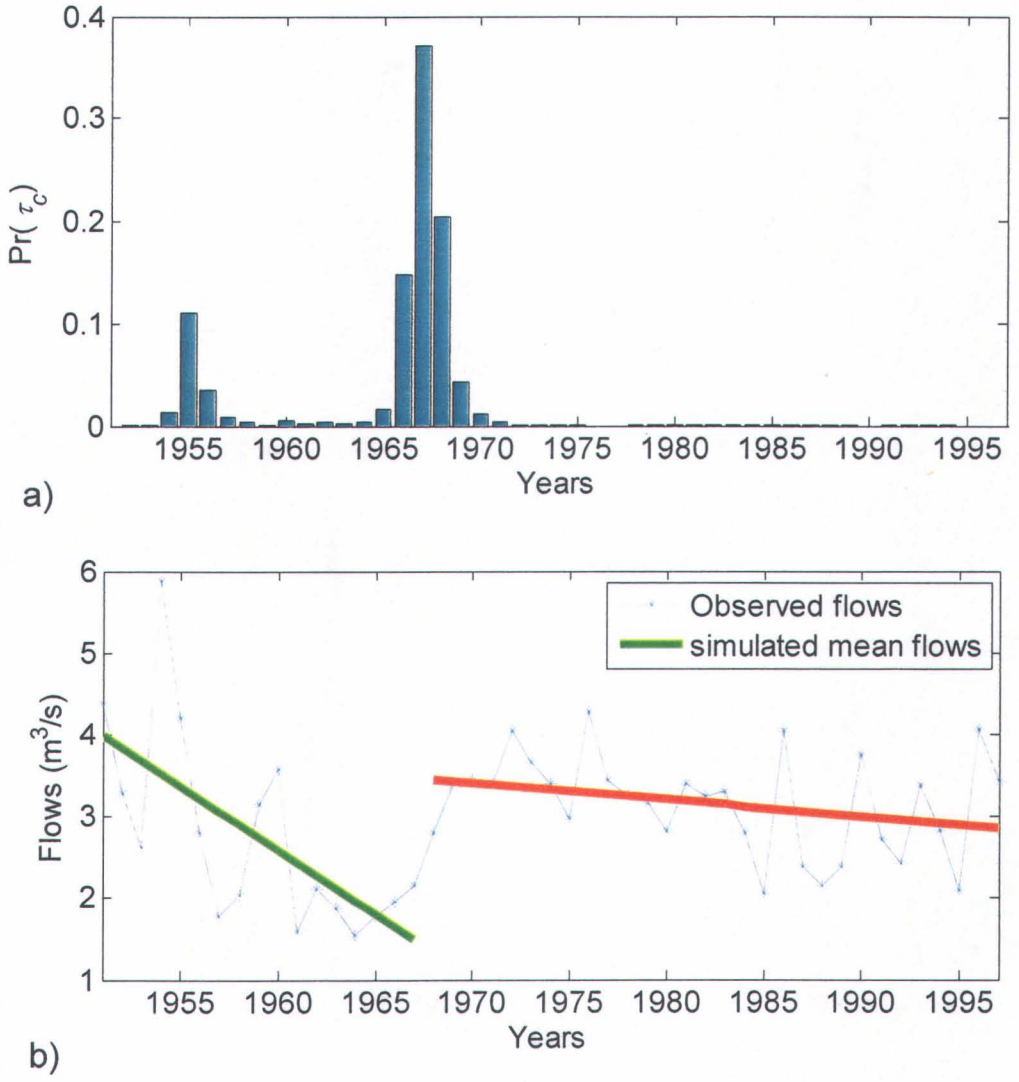


Figure 9: detection of trend changes at station 04GB004 (Ogoki River above Whiteclay Lake) with the methodology of *Asselin and Ouarda* [2005].

Appendix 1: properties of $p(\sigma \mid a,c) \propto \sigma^{-a} \exp(-\frac{c}{2\sigma^2}),\quad a>1, c>0$

$$\text{Let } I(a) = \int_{\sigma=0}^{\infty} \sigma^{-a} \exp(-\frac{c}{2\sigma^2}) d\sigma \qquad [1.1]$$

$$\text{Let } t = \frac{c}{2\sigma^2} \Rightarrow \sigma = \left(\frac{c}{2t}\right)^{1/2} \Rightarrow d\sigma = -\frac{1}{2} t^{-\frac{3}{2}} \left(\frac{c}{2}\right)^{1/2} dt = -2^{-\frac{3}{2}} c^{\frac{1}{2}} t^{-\frac{3}{2}} dt$$

$$I(a) = \int_{\sigma=0}^{\infty} \left(\frac{c}{2t}\right)^{-a/2} \exp(-t) \left(-2^{-\frac{3}{2}} c^{\frac{1}{2}} t^{-\frac{3}{2}}\right) dt \qquad [1.2]$$

$$I(a) = 2^{\frac{a-3}{2}} c^{\frac{1-a}{2}} \int_{\sigma=0}^{\infty} t^{(a-3)/2} \exp(-t) dt = 2^{-\frac{a-3}{2}} c^{-\frac{a-1}{2}} \Gamma(\frac{a-1}{2}) \qquad [1.3]$$

If $a>2$, the expectation of σ is finite:

$$E(\sigma) = \frac{I(a-1)}{I(a)} = \frac{2^{-\frac{a-4}{2}} c^{-\frac{a-2}{2}} \Gamma(\frac{a-2}{2})}{2^{-\frac{a-3}{2}} c^{-\frac{a-1}{2}} \Gamma(\frac{a-1}{2})} = \frac{(2c)^{-\frac{1}{2}} \Gamma(\frac{a-2}{2})}{\Gamma(\frac{a-1}{2})} \qquad [1.4]$$

if $a>3$, the variance of σ is finite:

$$E(\sigma^2) = \frac{I(a-2)}{I(a)} = \frac{2^{-\frac{a-5}{2}} c^{-\frac{a-3}{2}} \Gamma(\frac{a-3}{2})}{2^{-\frac{a-3}{2}} c^{-\frac{a-1}{2}} \Gamma(\frac{a-1}{2})} = \frac{\Gamma(\frac{a-2}{2})}{(2c)\Gamma(\frac{a-1}{2})} \qquad [1.5]$$

$$Var(\sigma) = E(\sigma^2) - (E(\sigma))^2 = \frac{\Gamma(\frac{a-3}{2})\Gamma(\frac{a-1}{2}) - \left(\Gamma(\frac{a-2}{2})\right)^2}{\left(2c\Gamma\left(\frac{a-1}{2}\right)^2\right)} \qquad [1.6]$$

The case $a<3$ leads to an infinite variance for σ, i.e. $\lim_{x\to+\infty} \int_0^x p(\sigma) d\sigma = +\infty$. Any value of a less than 3 can thus be used as a non informative prior. Note that when σ is very large, $p(\sigma) \propto \sigma^{-a}$.

Appendix 2: Derivation of $P(t,s)$

In this section we will derive the expression for $P(t,s)$. Let $\boldsymbol{\theta}_0$ be the ordinary least square solution of the equation $\mathbf{Y}_{t:s} = \mathbf{X}_{t:s}\boldsymbol{\theta}$ and $\boldsymbol{\varepsilon}_{t:s} = \mathbf{Y}_{t:s} - \mathbf{X}_{t:s}\boldsymbol{\theta}_0$. Note that $\boldsymbol{\varepsilon}_{s:t}$ does not depend on $\boldsymbol{\theta}$ or σ. It's well known from linear algebra that $\boldsymbol{\theta}_0 = (\mathbf{X}_{t:s}^T \mathbf{X}_{t:s})^{-1} \mathbf{X}_{t:s}^T \mathbf{Y}_{t:s}$. We also suppose that

$p(\sigma \mid a,c) = \dfrac{\sigma^{-a}\exp(-\frac{c}{2\sigma^2})}{2^{\frac{a-3}{2}} c^{-\frac{a-1}{2}} \Gamma(\frac{a-1}{2})}, a>1, c>0$. We have:

$$P(t,s) = \int_{\sigma} (2\pi\sigma^2)^{-(t-s+1)/2} \pi(\sigma) \int_{\mathbf{b}} \exp\left[-\frac{(\mathbf{Y}_{t:s} - \mathbf{X}_{t:s}\boldsymbol{\theta})^T (\mathbf{Y}_{t:s} - \mathbf{X}_{t:s}\boldsymbol{\theta})}{2\sigma^2} \right] d\sigma d\boldsymbol{\theta} \qquad [2.1]$$

Equation [2.1] can be simplified since

$$\frac{(\mathbf{Y}_{t:s} - \mathbf{X}_{t:s}\boldsymbol{\theta})^T (\mathbf{Y}_{t:s} - \mathbf{X}_{s:t}\boldsymbol{\theta})}{2\sigma^2} = \frac{(\boldsymbol{\varepsilon}_{t:s} - \mathbf{X}_{t:s}(\boldsymbol{\theta}-\boldsymbol{\theta}_0))^T (\boldsymbol{\varepsilon}_{t:s} - \mathbf{X}_{t:s}(\boldsymbol{\theta}-\boldsymbol{\theta}_0))}{2\sigma^2}$$
$$= \frac{1}{2\sigma^2}\left(\boldsymbol{\varepsilon}_{t:s}^T\boldsymbol{\varepsilon}_{t:s} - \boldsymbol{\varepsilon}_{t:s}^T \mathbf{X}_{t:s}(\boldsymbol{\theta}-\boldsymbol{\theta}_0)(\boldsymbol{\theta}-\boldsymbol{\theta}_0)^T \mathbf{X}_{t:s}^T \boldsymbol{\varepsilon}_{t:s} + (\boldsymbol{\theta}-\boldsymbol{\theta}_0)^T (\mathbf{X}_{t:s}^T \mathbf{X}_{t:s})(\boldsymbol{\theta}-\boldsymbol{\theta}_0) \right) \qquad [2.2]$$

and

$$\mathbf{X}_{t:s}^T \boldsymbol{\varepsilon}_{t:s} = \left(\boldsymbol{\varepsilon}_{t:s}^T \mathbf{X}_{t:s} \right)^T = \mathbf{X}_{t:s}^T (\mathbf{Y}_{t:s} - \mathbf{X}_{t:s}(\mathbf{X}_{t:s}^T \mathbf{X}_{t:s})^{-1} \mathbf{X}_{t:s}^T \mathbf{Y}_{t:s}) = 0, \qquad [2.3]$$

thus:

$$P(t,s) = \int (2\pi\sigma^2)^{-(t-s+1)/2} p(\sigma) \exp\left[-\frac{\boldsymbol{\varepsilon}_{t:s}^T \boldsymbol{\varepsilon}_{t:s} + (\boldsymbol{\theta}-\boldsymbol{\theta}_0)^T (\mathbf{X}_{t:s}^T \mathbf{X}_{t:s})(\boldsymbol{\theta}-\boldsymbol{\theta}_0)}{2\sigma^2} \right] d\sigma d\boldsymbol{\theta} \qquad [2.4]$$

$$P(t,s) = \int (2\pi\sigma^2)^{-(t-s+1)/2} \pi(\sigma) \exp\left(-\frac{\boldsymbol{\varepsilon}_{t:s}^T \boldsymbol{\varepsilon}_{t:s}}{2\sigma} \right) \int_{\boldsymbol{\theta}} -\frac{(\boldsymbol{\theta}-\boldsymbol{\theta}_0)^T \left(\mathbf{X}_{t:s}^T \mathbf{X}_{t:s} \right) (\boldsymbol{\theta}-\boldsymbol{\theta}_0)}{2\sigma} d\sigma d\boldsymbol{\theta} \qquad [2.5]$$

let Σ be $\sigma^2 \left(\mathbf{X}_{t:s}^T \mathbf{X}_{t:s} \right)^{-1} \Rightarrow |\Sigma| = \dfrac{\sigma^2}{\left| \mathbf{X}_{t:s}^T \mathbf{X}_{t:s} \right|}$;

$$P(t,s) = \int_{\sigma} (2\pi\sigma^2)^{-(t-s+1)/2} p(\sigma) \exp(-\frac{\boldsymbol{\varepsilon}_{t:s}^T \boldsymbol{\varepsilon}_{t:s}}{2\sigma^2}) (2\pi)^{m^*/2} |\Sigma|^{-1/2}$$
$$\times \int_{b} \underbrace{\frac{(2\pi)^{-d^*/2}}{|\Sigma|^{1/2}} \exp\left(-\frac{(\boldsymbol{\theta}-\boldsymbol{\theta}_0)^T \Sigma^{-1} (\boldsymbol{\theta}-\boldsymbol{\theta}_0)}{2} \right)}_{\int_{\boldsymbol{\theta}} N_{s-t+1}(\boldsymbol{\theta}\mid\boldsymbol{\theta}_0,\Sigma)=1} d\sigma d\boldsymbol{\theta} \qquad [2.6]$$

$$P(t,s) = (2\pi)^{-(t-s+1-d^*)/2} \left| \mathbf{X}_{t:s}^T \mathbf{X}_{t:s} \right|^{-1/2} \int_{\sigma} \sigma^{-(t-s+1)} \exp(-\frac{\boldsymbol{\varepsilon}_{t:s}^T \boldsymbol{\varepsilon}_{t:s}}{2\sigma^2}) p(\sigma) d\sigma \qquad [2.7]$$

$$P(t,s)=\frac{(2\pi)^{-(t-s+1-d^*)/2}\left|\mathbf{X}_{t:s}^T\mathbf{X}_{t:s}\right|^{-1/2}}{2^{\frac{a-3}{2}}c^{-\frac{a-1}{2}}\Gamma(\frac{a-1}{2})}\int_\sigma \sigma^{-(t-s+1+a)}\exp(-\frac{\boldsymbol{\varepsilon}_{t:s}^T\boldsymbol{\varepsilon}_{t:s}+c}{2\sigma^2})p(\sigma)d\sigma \qquad [2.8]$$

$$P(t,s)=\frac{(2\pi)^{-(t-s+1-d^*)/2}\left|\mathbf{X}_{t:s}^T\mathbf{X}_{t:s}\right|^{-1/2}}{2^{\frac{a-3}{2}}c^{-\frac{a-1}{2}}\Gamma(\frac{a-1}{2})}2^{\frac{t-s+a-2}{2}}\left(\boldsymbol{\varepsilon}_{t:s}^T\boldsymbol{\varepsilon}_{t:s}+c\right)^{-\frac{(t-s+a)}{2}}\Gamma\left(\frac{t-s+a}{2}\right) \qquad [2.9]$$

$$P(t,s)=(2\pi)^{\frac{d^*}{2}}\frac{\left(\pi(\boldsymbol{\varepsilon}_{t:s}^T\boldsymbol{\varepsilon}_{t:s}+c)\right)^{-\frac{(t-s+a)}{2}}}{(c\pi)^{-\frac{a-1}{2}}\left|\mathbf{X}_{t:s}^T\mathbf{X}_{t:s}\right|^{1/2}}\frac{\Gamma\left(\frac{t-s+a}{2}\right)}{\Gamma(\frac{a-1}{2})} \qquad [2.10]$$

as $\boldsymbol{\varepsilon}_{t:s}^T\boldsymbol{\varepsilon}_{t:s}=Y_{t:s}^TY_{t:s}-X_{t:s}(X_{t:s}^TX_{t:s})^{-1}Y_{t:s}^TY_{t:s})$ we finally obtain the expression for $P(s,t)$:

$$P(t,s)=(2\pi)^{\frac{d^*}{2}}\frac{\left(\pi(\boldsymbol{\varepsilon}_{t:s}^T\boldsymbol{\varepsilon}_{t:s}+c)\right)^{-\frac{(t-s+a)}{2}}}{(c\pi)^{-\frac{a-1}{2}}\left|\mathbf{X}_{t:s}^T\mathbf{X}_{t:s}\right|^{1/2}}\frac{\Gamma\left(\frac{t-s+a}{2}\right)}{\Gamma(\frac{a-1}{2})} \qquad [2.11]$$