

Bayesian estimation of change points using the general linear model

Peter Rasmussen

Department of Civil Engineering, University of Manitoba, Winnipeg, Manitoba, Canada

Abstract. Bayesian analysis is applied to the general linear model to develop a framework for studying different types of change in the mean value of time series and linear regressions. The output of the Bayesian analysis is the posterior distribution of change point location and amplitude. This information provides a rational and relatively objective basis for making decisions as to where to locate a change point. Several examples of hydrological applications are presented to demonstrate the utility of the methodology.

1. Introduction

In most statistical analyses of hydrological time series an assumption is made that the phenomenon under consideration is stationary over time. For example, it is not uncommon to assume that the population mean and variance of a random variable are constant. While in most situations this assumption is reasonable, cases arise where data suggest that at some point, there has been a change in some of the basic statistical characteristics of the process. If there is reason to believe that a change has taken place, a statistical analysis should be undertaken to examine the time and nature of the change.

A closely related problem may arise in regression analyses where data sometimes suggest that it would be preferable to divide the predictor space into two or more regions and fit different models to observations in each region. In both cases it would be preferable to employ an objective data-based method to identify the most likely point of change. Certain nonparametric regression methods such as recursive partitioning regression and multivariate adaptive regression splines have been developed in recent years to address this issue [Friedman, 1991].

We employ here a Bayesian approach to investigate the location and nature of changepoints in time series and linear regressions. Change point determination has been studied extensively in the literature, using both classical statistical approaches and Bayesian approaches. Some recent references to Bayesian work on the subject include *Carlin et al.* [1992], *Bernier* [1994], *Stephens* [1994], *Ó Ruanaidh and Fitzgerald* [1996], and *Perreault et al.* [1999, 2000a, 2000b]. The specific purpose of this paper is to show that the generalized linear model used in conjunction with Bayesian analysis provides a convenient framework for describing changepoints associated with a variety of change types. The output of the Bayesian analysis is the posterior distribution of the time of change and the marginal distribution of the change amplitude.

The paper is organized as follows: In section 2 we review properties of the general linear model and the Bayesian frame-

work used to make inference about linear models. In sections 3–5 the general procedure for studying change points in time series and regression relationships is outlined. Section 6 presents three hydrological applications of the methodology.

2. General Linear Model

Our point of departure is the general linear model, which is frequently employed in statistical analyses and which also has seen many applications in the field of water resources. Any data set that can be described by a linear combination of basis functions and an additive Gaussian noise satisfies the general linear model:

$$y_i = \sum_{k=1}^M b_k g_k(i) + \varepsilon_i \quad i = 1, \dots, N, \quad (1)$$

where y_i is the dependent variable and $g_k(i)$ is the k th basis function which is a function of the explanatory variables associated with the i th observation. In the case where y_i is a time series, $g_k(\cdot)$ will typically be a function of the observation time, that is, $g_k(t_i)$. Coefficients b_k are associated with the basis functions, and ε_i is an independent random noise assumed here to have a Gaussian distribution with zero mean and variance σ^2 . The assumptions regarding the noise term may at first seem restrictive. However, approximate normality can often be achieved by an appropriate transformation of the data. Transformation may also serve to render the residuals homoscedastic. A small to moderate serial correlation of observations is not expected to have any major impact on the approach to be described in this paper.

In matrix form, (1) becomes

$$\mathbf{y} = \mathbf{G}\mathbf{b} + \boldsymbol{\varepsilon}, \quad (2)$$

where \mathbf{y} is a N -dimensional vector containing the observed y data, \mathbf{G} is a $N \times M$ matrix with columns representing basis functions and rows representing observations, and $\boldsymbol{\varepsilon}$ is an N -dimensional vector of Gaussian noise with zero mean and covariance matrix $\sigma^2\mathbf{I}$.

The model parameters are the noise variance σ^2 , the coefficients \mathbf{b} associated with the basis functions, and a set of parameters $\{\omega\}$ describing the basis functions. For the above model the likelihood function L of $\Phi = [\sigma, \mathbf{b}, \{\omega\}]$ is given by

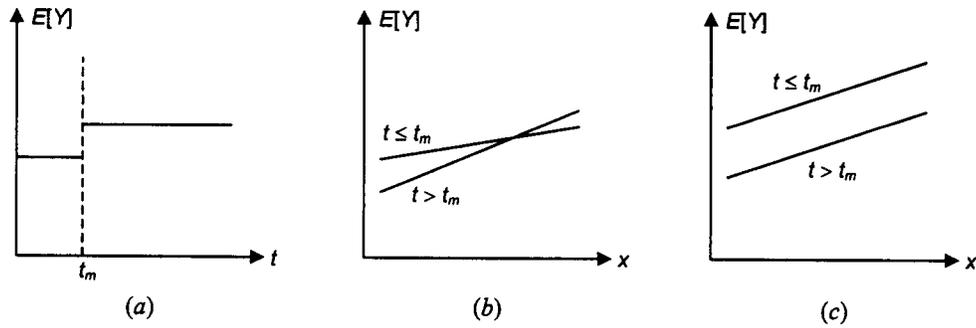


Figure 1. Models for change point in time series. (a) Change in mean. (b) Change in linear relationship between two variables. (c) Change in intercept of linear relationship.

$$L(\Phi|y) = p(y|\Phi) = p(\epsilon|\Phi) = (2\pi\sigma^2)^{-N/2} \exp\left[-\frac{\epsilon^T \epsilon}{2\sigma^2}\right]$$

$$= (2\pi\sigma^2)^{-N/2} \exp\left[-\frac{(y - Gb)^T (y - Gb)}{2\sigma^2}\right]. \quad (3)$$

To perform a Bayesian analysis of model parameters, a prior distribution of the parameters must be specified. The prior distribution should reflect any knowledge of Φ that is not related to the data. In practice, such information is often not available or is so vague compared with the information conveyed by the data that it can be neglected. Therefore as prior distribution of Φ we consider Jeffrey's noninformative prior

$$p(\Phi) \propto \sigma^{-1}, \quad (4)$$

which corresponds to assuming a uniform distribution of b , $\{\omega\}$, and $\log \sigma$. Bayes' theorem can then be used to obtain the posterior distribution of Φ :

$$p(\Phi|y) \propto L(\Phi|y)p(\Phi). \quad (5)$$

As will be shown in sections 3 and 4, the basis functions can be employed to represent various forms of change in time series and regressions. Therefore we will be concerned with the marginal distribution of the parameters $\{\omega\}$. The marginal posterior distribution of $\{\omega\}$ is obtained by integrating out the parameters b and σ from the joint posterior distribution. With the assumptions made above the joint posterior distribution may be written as

$$p(\{\omega\}, \sigma, b|y) \propto (2\pi\sigma^2)^{-N/2} \cdot \exp\left[-\frac{(y - Gb)^T (y - Gb)}{2\sigma^2}\right] \frac{1}{\sigma}. \quad (6)$$

After integrating out the parameters b and σ , the following expression for the marginal distribution of $\{\omega\}$ is obtained:

$$p(\{\omega\}|y) \propto \frac{[y^T y - y^T G(G^T G)^{-1} G^T y]^{-(N-M)/2}}{|G^T G|^{1/2}}, \quad (7)$$

where $|G^T G|$ is the determinant of $G^T G$ and M is the dimension of b . In change point analyses we will generally assume that there are a limited number of possible changepoint locations and each of the candidates will be described by a particular G matrix. Therefore $p(\{\omega\}|y)$ will be a discrete distribu-

tion, and the normalizing constant required in (7) can be easily determined by summation. Note that the determination of the marginal posterior distribution of $\{\omega\}$ does not require explicit estimation of the regression coefficients b and the noise variance σ^2 .

3. Determination of Change Points in Time Series

3.1. Change in the Mean Value of a Series of Independent Normal Variables

The case of a change in the mean value of a series of independent normal random variables is the simplest case considered here. This problem has been studied extensively in the literature, both from a classical statistical perspective and from a Bayesian viewpoint. Let y_1, y_2, \dots, y_N be a series of observations of a normal distributed random variable Y , observed at times t_1, t_2, \dots, t_N . It is assumed that the observations are independent; that is, there is no serial correlation. We hypothesize that at some point in the sequence, there has been a shift in the population mean, and we want to determine the posterior distribution of the change point, t_m . It is assumed that the noise variance is constant over time. This particular problem is illustrated in Figure 1a. The posterior distribution of t_m can be readily obtained using the results from section 2. In particular, we postulate the following model,

$$y_i = \begin{cases} \mu_1 + \epsilon_i & i \leq m \\ \mu_2 + \epsilon_i & i > m \end{cases} \quad i = 1, \dots, N \quad (8)$$

corresponding to observation times $t_i, i = 1, \dots, m, m + 1, \dots, N$. Our interest is to make inference about t_m or, equivalently, about m . The above model can be easily formulated as a general linear model, i.e., in the form $y = Gb + \epsilon$. We take $b^T = (\mu_1, \mu_2)$ and define the G matrix as

$$G_m^T = \left[\begin{array}{ccc|ccc} 1 & 1 & \dots & 1 & 0 & 0 & \dots & 0 \\ 0 & 0 & \dots & 0 & 1 & 1 & \dots & 1 \end{array} \right]. \quad (9)$$

In other words, G_m has two columns, the first containing m rows of ones followed by $N - m$ rows of zeros and the second containing m rows of zeros and $N - m$ rows of ones. The only parameter of the basis functions is m . Insertion of G_m in (7) yields

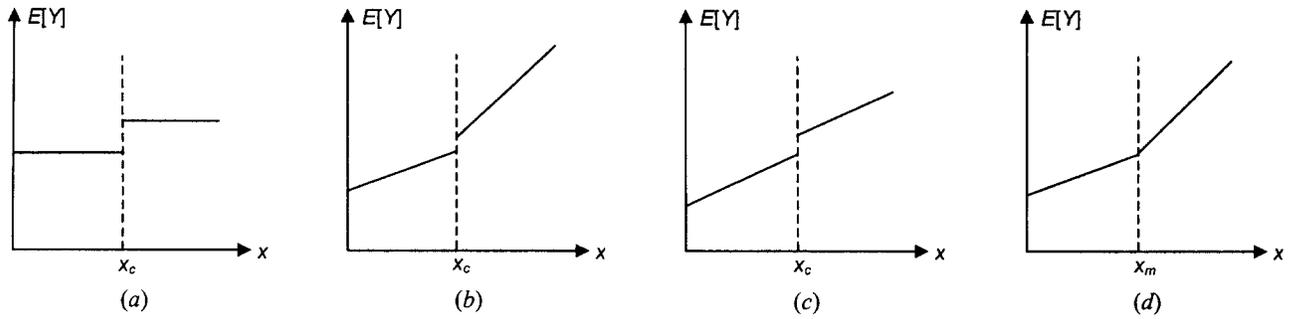


Figure 2. Models for change point in linear regressions. (a) Change in constant mean. (b) Change in both intercept and slope. (c) Change in intercept. (d) Change in linear regression with continuity at change point.

$$p(t_m|y) = \{ |G_m^T G_m|^{-1/2} [y^T y - y^T G_m (G_m^T G_m)^{-1} G_m^T y]^{-(N-M)/2} \} \cdot \left\{ \sum_{i=1}^{N-1} |G_i^T G_i|^{-1/2} [y^T y - y^T G_i (G_i^T G_i)^{-1} G_i^T y]^{-(N-M)/2} \right\}^{-1} \quad m = 1, \dots, N-1 \quad (10)$$

with $M = 2$. The denominator serves to standardize the discrete probability function so that $\sum_{m=1}^{N-1} p(t_m|y) = 1$. This expression corresponds to the probability function derived by Lee and Heghinian [1977]. Note that m is restricted to $[1; N-1]$ because there must be at least one observation before and after the change point. Candidates for a change point estimator are the mode and the mean. The mode may be preferred if $p(t_m|y)$ has a distinct peak. The mean value may be preferred if the distribution is more dispersed.

The generalization to the case of several change points is straightforward. For example, in the case of two change points (m_1, m_2) we would define $b^T = (\mu_1, \mu_2, \mu_2)$, and G_m would take the form

$$G_{m_1, m_2}^T = \left[\begin{array}{ccc|ccc} 1 & \dots & 1 & 0 & \dots & 0 \\ 0 & \dots & 0 & 1 & \dots & 1 \\ 0 & \dots & 0 & 0 & \dots & 0 \\ \hline & & & & & 1 \end{array} \right] \quad (11)$$

The change point distribution would be bivariate and its support would be all possible combinations of m_1 and m_2 .

3.2. Change in Linear Relationship Between Two Variables

Suppose now that we have N joint observations of two variables x and y , i.e., (x_i, y_i) observed at equidistant times $t_i, i = 1, \dots, N$. We assume that a linear relationship exists between x and y but that at time t_m the parameters of the linear relationship have changed. This situation is illustrated in Figure 1b. We want to make inference about t_m or, equivalently, about m . This problem is also easily formulated and solved using the general linear model. The postulated model is

$$y_i = \begin{cases} \alpha_1 + \beta_1 x_i + \varepsilon_i & i \leq m \\ \alpha_2 + \beta_2 x_i + \varepsilon_i & i > m \end{cases} \quad i = 1, \dots, N \quad (12)$$

In matrix form this corresponds to $b^T = (\alpha_1, \beta_1, \alpha_2, \beta_2)$ and

$$G_m^T = \left[\begin{array}{cccc|cccc} 1 & \dots & 1 & 0 & \dots & 0 & & \\ x_1 & \dots & x_m & 0 & \dots & 0 & & \\ 0 & \dots & 0 & 1 & \dots & 1 & & \\ \hline & & & & & & x_{m+1} & \dots & x_N \end{array} \right] \quad (13)$$

This expression for G_m , along with $M = 4$, may be plugged into (10) to give the posterior distribution of t_m , with the only modification being that the posterior distribution must be restricted to $m = 2, 3, \dots, N-2$ because at least two points are needed to establish a linear relationship. Note again that for a given set of observations the basis functions are determined entirely by m .

3.3. Change of Intercept in Linear Regression

It is straightforward to modify the model 12 to the case where there is a change in the intercept but not in the slope of the linear relationship (Figure 1c). We then have the model

$$y_i = \begin{cases} \alpha_1 + \beta x_i + \varepsilon_i & i \leq m \\ \alpha_2 + \beta x_i + \varepsilon_i & i > m \end{cases} \quad i = 1, \dots, N \quad (14)$$

with $b^T = (\alpha_1, \beta, \alpha_2)$ and

$$G_m^T = \left[\begin{array}{cccc|cccc} 1 & \dots & 1 & 0 & \dots & 0 & & \\ x_1 & \dots & x_m & x_{m+1} & \dots & x_N & & \\ 0 & \dots & 0 & 1 & \dots & 1 & & \end{array} \right] \quad (15)$$

and $M = 3$. Again, (10) yields the desired posterior density of t_m .

4. Determination of Change Points in Linear Regression

Section 3 dealt with change points in time series. A closely related problem arises when two variables x and y are linearly related within subintervals of x , but the exact extent of the subintervals is unknown. For example, we may hypothesize the existence of two regimes corresponding to x values less than and greater than some critical value x_c and desire to obtain the posterior distribution of x_c . Figure 2 illustrates different cases of change in linear regressions.

If the observations (x_i, y_i) are ordered so that $x_1 \leq x_2 \leq \dots \leq x_N$, the formulas given in sections 2 and 3 can be used without modifications. Note that the condition $i \leq m$ is equivalent to the condition $x_i \leq x_m \leq x_c$. Since in the general case the x observations are not equidistant, some precaution should be taken when interpreting the posterior distribution of m . If the mode is used to determine the most likely point of change, then one should consider the discrete probabilities associated with observed x values and, for example, select the midpoint of the interval between the mode and the next observed x value as the appropriate change point. This seems reasonable since $p(x_m|y)$ represents the probability that the change occurs in

the interval $[x_m; x_{m+1}]$. On the other hand, if the mean value is used as an estimate of the change point, then it may be preferable to consider x a continuous variable. The posterior density is constant between adjacent x observations x_m and x_{m+1} and is given by

$$f(x) = \frac{p(x_m|\mathbf{y})}{c(x_{m+1} - x_m)} \quad (16)$$

for $x_m \leq x < x_{m+1}$ and $m = 1, 2, \dots, N - 1$, where

$$c = \sum_{m=1}^{N-1} \frac{p(x_m|\mathbf{y})}{x_{m+1} - x_m}$$

As defined above, the density function is restricted to $[x_1, x_N]$. In some cases (such as the case in Figure 2b), the posterior density of x_m must be further restricted to allow for several observations before and after the change point.

A variant of the case in Figure 2b arises if one requires that the relationship between y and x be continuous over the change point. This situation is illustrated in Figure 2d.

Assuming that the change point coincides with the observation x_m , it is readily seen that (12) becomes

$$y_i = \alpha + \beta_1 x_i + \varepsilon_i \quad i \leq m \quad i = 1, \dots, N \quad (17)$$

$$y_i = \alpha + \beta_1 x_m + \beta_2 (x_m - x_i) + \varepsilon_i \quad i > m \quad i = 1, \dots, N.$$

With $\mathbf{b}^T = (\alpha, \beta_1, \beta_2)$, we obtain the following \mathbf{G}_m matrix:

$$\mathbf{G}_m^T = \begin{bmatrix} 1 & \cdots & 1 & 1 & \cdots & 1 \\ x_1 & \cdots & x_m & x_m & \cdots & x_m \\ 0 & \cdots & 0 & (x_m - x_{m+1}) & \cdots & (x_m - x_N) \end{bmatrix} \quad (18)$$

and $M = 3$.

5. Posterior Distribution of Basis Function Coefficients

In many situations the basis function coefficients provide valuable information about the amplitude and significance of the change. Therefore it is of interest to consider the posterior distribution of \mathbf{b} and subsets of \mathbf{b} . For a fixed changepoint m and after having integrated out the residual variance, the marginal posterior distribution of \mathbf{b} is a multivariate t distribution:

$$p(\mathbf{b}|\mathbf{y}, m) = \frac{\Gamma[\frac{1}{2}(\nu + M)]|\mathbf{G}_m^T \mathbf{G}_m|^{1/2}}{[\Gamma(\frac{1}{2})]^M \Gamma(\frac{1}{2})[s\sqrt{\nu}]^M} \cdot \left[1 + \frac{(\mathbf{b} - \hat{\mathbf{b}})^T \mathbf{G}_m^T \mathbf{G}_m (\mathbf{b} - \hat{\mathbf{b}})}{\nu s^2} \right]^{-(\nu+M)/2}, \quad (19)$$

where Γ is the gamma function. In the above expression, $\hat{\mathbf{b}} = (\mathbf{G}_m^T \mathbf{G}_m)^{-1} \mathbf{G}_m^T \mathbf{y}$ is the least squares estimate of \mathbf{b} , $s^2 = \sum (y_i - \hat{y}_i)^2 / (N - M)$ is the unbiased estimate of the noise variance, and $\nu = N - M$ is the number of degrees of freedom.

Typically, a change is modeled by a subset of the elements of \mathbf{b} . In that case it is of interest to consider the marginal distribution of the particular subset of \mathbf{b} that reflects the change in model parameters. Consider the following partitioning:

$$\mathbf{b} = \begin{bmatrix} \mathbf{b}_1 \\ \mathbf{b}_2 \end{bmatrix} \quad \hat{\mathbf{b}} = \begin{bmatrix} \hat{\mathbf{b}}_1 \\ \hat{\mathbf{b}}_2 \end{bmatrix} \quad (\mathbf{G}_m^T \mathbf{G}_m)^{-1} = \begin{bmatrix} \mathbf{C}_{11} & \mathbf{C}_{12} \\ \mathbf{C}_{21} & \mathbf{C}_{22} \end{bmatrix}, \quad (20)$$

where it is assumed that \mathbf{b}_1 has r elements and \mathbf{b}_2 has $M - r$ elements and the \mathbf{C} matrices have corresponding dimensions. Then the marginal distribution of \mathbf{b}_1 is

$$p(\mathbf{b}_1|\mathbf{y}, m) = \frac{\Gamma[\frac{1}{2}(\nu + r)]|\mathbf{C}_{11}^{-1}|^{1/2}}{[\Gamma(\frac{1}{2})]^r \Gamma(\frac{1}{2})[s\sqrt{\nu}]^r} \cdot \left[1 + \frac{(\mathbf{b}_1 - \hat{\mathbf{b}}_1)^T \mathbf{C}_{11}^{-1} (\mathbf{b}_1 - \hat{\mathbf{b}}_1)}{\nu s^2} \right]^{-(\nu+r)/2}. \quad (21)$$

This result, along with (19), is useful for quantifying the amount of change given a specific change point. The distribution can be made unconditional upon m by the following summation:

$$p(\mathbf{b}_1|\mathbf{y}) = \sum_{i=1}^N p(\mathbf{b}_1|\mathbf{y}, m) p(m|\mathbf{y}), \quad (22)$$

where the general form of $p(m|\mathbf{y})$ is given in (7).

Although (19) and (21) provide the general formulas for computing the posterior distribution of \mathbf{b} and \mathbf{b}_1 , we shall illustrate their use by determining the posterior distribution of the amplitude of change in an otherwise constant mean value of a series of random normal variables (equation (8)). Since, in this case, interest focuses on the amount of change $\delta = \mu_2 - \mu_1$ rather than the mean values themselves, we recast (8) as

$$y_i = \begin{cases} \mu + \varepsilon_i & i \leq m \\ \mu + \delta + \varepsilon_i & i > m \end{cases} \quad i = 1, \dots, N \quad (23)$$

which corresponds to $\mathbf{b} = (\delta, \mu)^T$ and

$$\mathbf{G}_m^T = \left[\underbrace{0 \ 0 \ \cdots \ 0}_m \mid \underbrace{1 \ 1 \ \cdots \ 1}_{N-m} \right] \quad (24)$$

It is readily seen that

$$\mathbf{G}_m^T \mathbf{G}_m = \begin{bmatrix} N - m & N - m \\ N - m & N \end{bmatrix}, \quad (25)$$

and therefore \mathbf{C}_{11} as defined in (20) takes a particularly simple form, namely $\mathbf{C}_{11} = N/[m(N - m)]$. As before, \mathbf{b} is estimated by $\hat{\mathbf{b}} = (\hat{\delta}, \hat{\mu})^T = (\mathbf{G}_m^T \mathbf{G}_m)^{-1} \mathbf{G}_m^T \mathbf{y}$. When inserted into (21), we obtain

$$p(\delta|\mathbf{y}, m) = \frac{\Gamma[\frac{1}{2}(N - 1)] \sqrt{(N - m)m/N}}{[\Gamma(\frac{1}{2})]^2 \Gamma(\frac{1}{2})[N - 2] \sqrt{N - 2}s}} \cdot \left[1 + \frac{(\delta - \hat{\delta})^2 N}{(N - m)m(N - 2)s^2} \right]^{-(N-1)/2}, \quad (26)$$

which is the desired posterior density of δ , conditional upon a change at time t_m .

6. Hydrological Applications

In this section, we give some examples of how Bayesian change point analysis can be used for estimating statistical characteristics of hydrological data. The examples illustrate the ease by which the framework can be adapted to particular situations.

6.1. Estimating Trends in Hydrologic Time Series

The traditional approach to hydrologic time series analysis requires a time series to be decomposed into a trend component, a periodic component, and a random component. The deterministic trend component usually refers to the evolution of the mean value of the deseasonalized time series. Detrending therefore requires estimation of the mean value as a function of time in the presence of random noise. A common procedure is to assume a polynomial form of the mean value as a function of time and to estimate the coefficients of the polynomial by the method of least squares. It is, of course, possible to divide the timescale into two or more segments with different functional forms for the trend. In that context the Bayesian change point analysis described in sections 2-5 can provide valuable input as to how to divide the timescale in a sensible way.

For the purpose of illustration, consider the time series of annual streamflows of the St. Lawrence River at Ogdensbourg, New York, shown in Figure 3. Inspection of Figure 3 reveals that the mean value in the first years of the record is significantly higher than in later years. This could be due to low-frequency climate variability or perhaps to a change in the gauging method. We do not attempt to provide a physical explanation of the trend but will focus on how to detrend the series in a rational and objective way.

Most methods for trend removal require prior specification of the functional form of the trend. A simple model assumes that the mean is constant except for an abrupt change occurring at some point in time. Of course, if the shift in the mean is thought to be due to a change in the gauging method and one has knowledge of the time when the recording method was modified, then this information is sufficient to fix the change point. In the opposite case, where information about the change point is vague or nonexistent, a choice must be made. While a change in the mean value can be observed visually from Figure 3, it is not obvious exactly where to locate the change point. Assuming that the population variance before and after the change is the same, we can use model (8) for the time series and perform a Bayesian change point analysis. The resulting posterior distribution of t_m is shown in Figure 4a. The

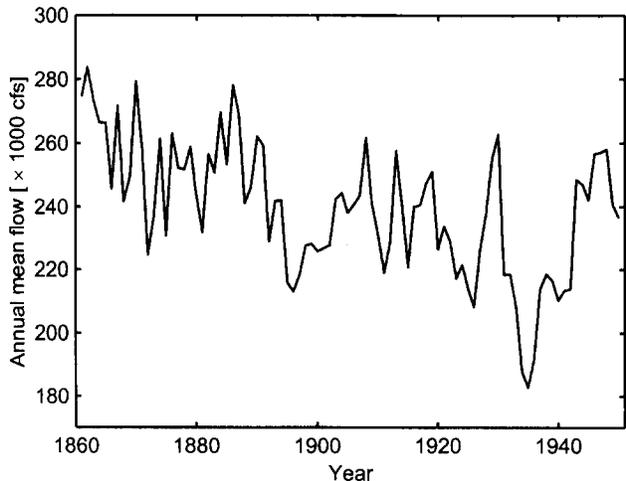


Figure 3. Annual mean flow of the St. Lawrence River at Ogdensbourg, New York, for the period 1861-1950.

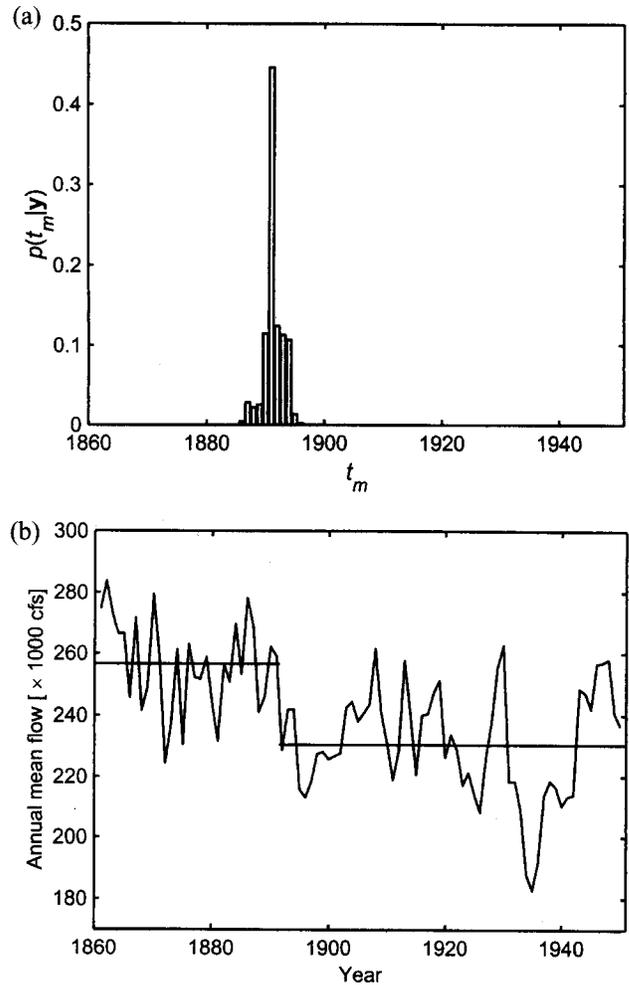


Figure 4. Abrupt change in constant mean level. (a) Posterior distribution of change point, t_m . (b) Mean value before and after 1891.

mode of the posterior distribution is equal to 1891 and the mean value is equal to 1892. The difference between the two values is negligible, and either may be used as the appropriate year of change. The mean value of annual flows before (and including) 1891 is 256,500 cubic feet per second (cfs) ($1 \text{ cfs} = 2.8317 \times 10^{-2} \text{ m}^3/\text{s}$), and after 1891 the mean value of flows is equal to 230,400 cfs. The mean values are indicated in Figure 4b. The posterior distribution of the difference in mean values conditional upon a change in 1891 can be obtained from (26). The posterior distribution of $\delta = \mu_2 - \mu_1$ is shown in Figure 5. Clearly, there is ample evidence to support the theory of a change. The density can be made unconditional upon the year of change. However, because the posterior distribution of the year of change is relatively concentrated, the conditional and unconditional densities are almost identical.

As an alternative to an abrupt change in an otherwise constant mean, one could hypothesize a linear trend before the change point, followed by a constant mean. This model was not considered in sections 2-5 but can be easily formulated. It corresponds essentially to model (17) with β_2 set equal to zero. The G_m matrix then becomes

$$G_m^T = \begin{bmatrix} 1 & \cdots & 1 & 1 & \cdots & 1 \\ x_1 & \cdots & x_m & x_m & \cdots & x_m \end{bmatrix}$$

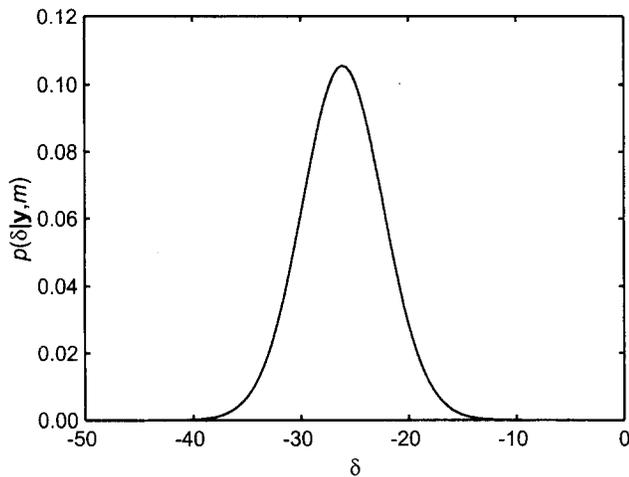


Figure 5. Posterior distribution of change amplitude, conditional upon a change in 1891.

with $\mathbf{b} = [\alpha, \beta]^T$ and $M = 2$. The Bayesian analysis results in the posterior change point distribution shown in Figure 6a. This distribution is much more spread out than the previous distributions of t_m , and the mode is clearly not the appropriate estimate of the change point. Since the distribution is reasonably symmetric, the mean value would be a reasonable choice. The mean value is equal to 1916, and the corresponding parameter estimates are $[\alpha, \beta] = [1486.5, -0.66]$. The mean value function is shown in Figure 6b and appears to provide a good fit to the data.

6.2. Determination of Two Regimes of an Intensity-Duration-Frequency Curve

A common task in hydrologic design is the estimation of intensity-duration-frequency (IDF) curves from observed rainfall data. Although maps are available from which IDF relationships can be interpolated [e.g., Hershfield, 1962; Hogg and Carr, 1985], hydrologists often prefer to determine IDF curves directly from a reliable rainfall station in the vicinity of the site of interest. The determination of a set of IDF curves involves the following steps:

1. Slide a window of a given duration (D) over the observed data and extract the series of annual maximum rainfall depths corresponding to that duration. Repeat this for different rainfall durations which should be multiples of the time resolution of the data. Convert depths to intensities (I).
2. For each duration, fit a probability density function (e.g., the Gumbel distribution) to the series of annual maximums.
3. For selected return periods, use the fitted distributions to obtain the intensities associated with the durations considered in step 1.
4. Finally, for each return period, fit an appropriate function to the set of displaying intensities and durations. This will result in a set of curves displaying intensity as a function of duration for selected return periods.

The following example deals with the last point, specifically, the fitting of a function to the (I, D) points. IDF curves are often special cases of the generalized form $I = a(D^b + c)^{-d}$, where a , b , c , and d are parameters [Koutsoyiannis et al.,

1998]. Here we will assume that $c = 0$ and $d = 1$, in which case the I - D relationship for fixed frequency becomes

$$I = a/D^b. \quad (27)$$

To estimate the value of a and b , the above expression may be linearized by a logarithmic transformation:

$$\log(I) = \alpha + \beta \log(D), \quad (28)$$

where $\alpha = \log a$ and $\beta = -b$. The coefficients can be easily determined by regressing $\log(I)$ on $\log(D)$.

For the purpose of illustration, Table 1 gives estimated values of the I - D relationship for a return period of 2 years for Baltimore, Maryland [McCuen, 1998]. Figure 7 shows the plot of $\log(I)$ versus $\log(D)$, along with the least squares fit based on all I - D data. Clearly, the fit is not particularly good, especially for short-duration rainstorms. A better fit could possibly be obtained if the points were divided into two groups, one with durations less than some value D_m and the other with durations greater than D_m , and (27) then fitted to each group of I - D values. In addition, it would be reasonable to require continuity at D_m .

The question is how to determine D_m in a rational way. Of course, in this relatively simple example, it would be straight-

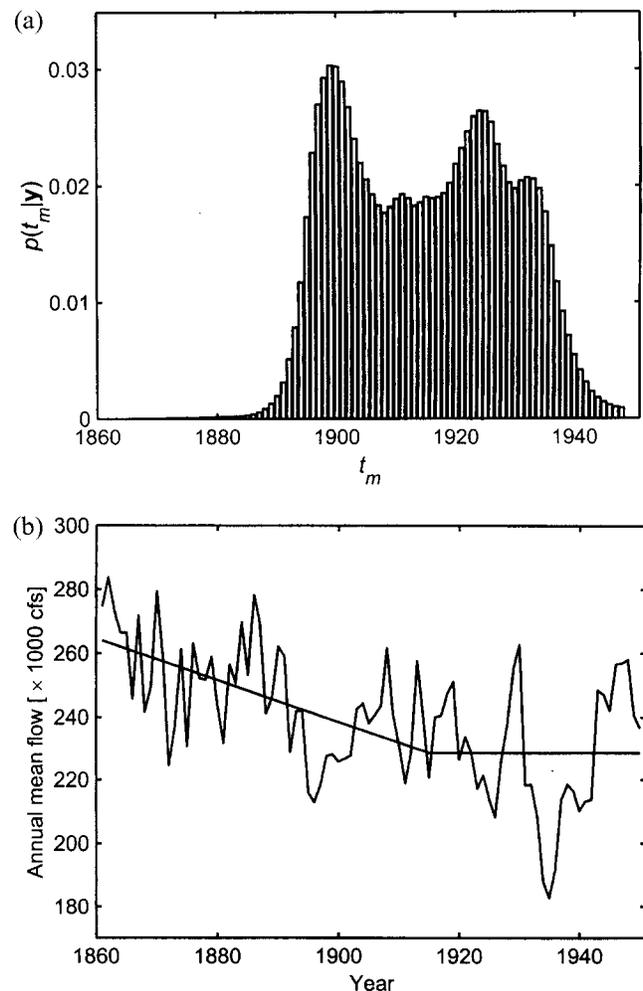


Figure 6. Linear trend followed by constant mean. (a) Posterior distribution of change point, t_m . (b) Mean value function before and after 1916.

Table 1. Estimated 2-Year Intensities for Baltimore, Maryland

| Duration, hour | Estimated 2-Year Intensity, ^a inch hr ⁻¹ |
|----------------|--|
| 0.083 | 5.20 |
| 0.10 | 5.00 |
| 0.167 | 4.11 |
| 0.25 | 3.40 |
| 0.333 | 3.00 |
| 0.50 | 2.30 |
| 0.75 | 1.65 |
| 1 | 1.35 |
| 1.5 | 1.00 |
| 2 | 0.81 |
| 4 | 0.50 |
| 6 | 0.36 |
| 8 | 0.30 |
| 10 | 0.25 |
| 12 | 0.22 |
| 18 | 0.16 |
| 24 | 0.13 |

^a1 inch = 2.54 cm.

forward to try different values of D_m and select the one that appears to provide the best overall fit of the two models. Alternatively, one could conduct a Bayesian change point analysis. Equation (17) is the appropriate model for the case considered here. With the appropriate changes in notation, we obtain the following model:

$$\begin{aligned} \log I_i &= \alpha + \beta_1 \log D_i + \varepsilon_i & D_i \leq D_m \\ \log I_i &= \alpha + \beta_1 \log D_m + \beta_2(\log D_m - \log D_i) & \\ &+ \varepsilon_i & D_i > D_m. \end{aligned} \tag{29}$$

The posterior distribution of D_m resulting from the Bayesian analysis is shown in Figure 8a. The analysis reveals that the posterior distribution has a distinct peak at $\log D_m = -1.0986$ or $D_m = 1/3$ hour = 20 min, where virtually all probability mass is concentrated. Hence we choose to divide the I - D points in two groups, one corresponding to durations

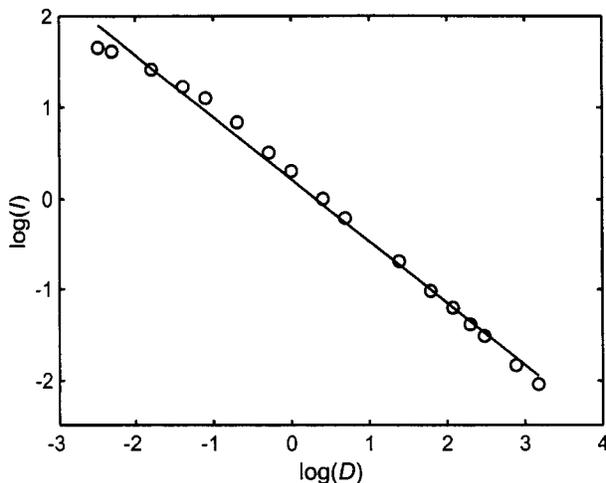


Figure 7. I - D relationship for 2-year precipitation events in Baltimore, Maryland. The straight line is the least squares fit to the points.

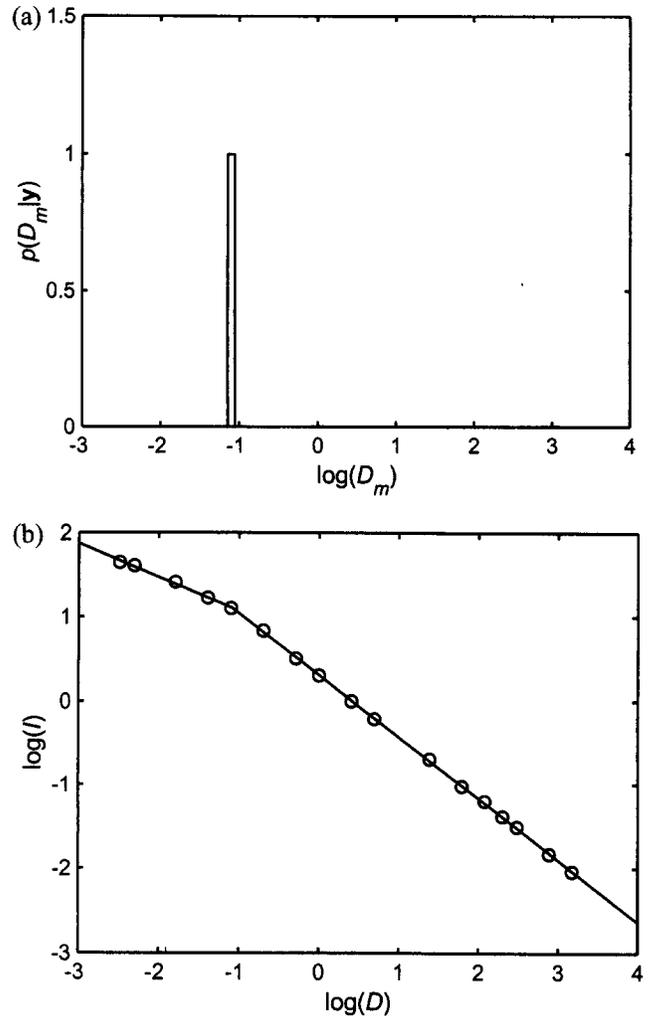


Figure 8. Change in linear relationship between $\log(I)$ and $\log(D)$. (a) Posterior distribution of change point, D_m . (b) Fitted model corresponding to change point $\log D_m = -1.0986$.

≤ 20 min and the other to durations > 20 min. The parameters of the global model corresponding to $D_m = 20$ min are $[\alpha, \beta_1, \beta_2] = [0.671, -0.402, 0.735]$. The fit is shown in Figure 8b. Figure 8b explains why the posterior probability of D_m is concentrated at 20 min; the fit is almost perfect with very little scatter around the regression lines.

6.3. Determination of Error in Rain Gauge Data

Data from precipitation networks are the main input to most hydrological models. Before using data from a particular station, the consistency of the gauge should be verified. An apparent decline in average precipitation may be due to a factual decline in precipitation but could also be the result of a change in the exposure of the gauge or a mechanical problem with the gauging device. A standard procedure for checking the consistency of a rainfall gauge is the so-called double-mass curve technique [McCuen, 1998]. To check the consistency of a gauge, the cumulative catch for the station is plotted against the cumulative catch for a neighboring station or against the sum of a set of regional stations that are known to be consistent. In the case of a consistent station, the double-mass curve should appear as an almost straight line. A sudden change in

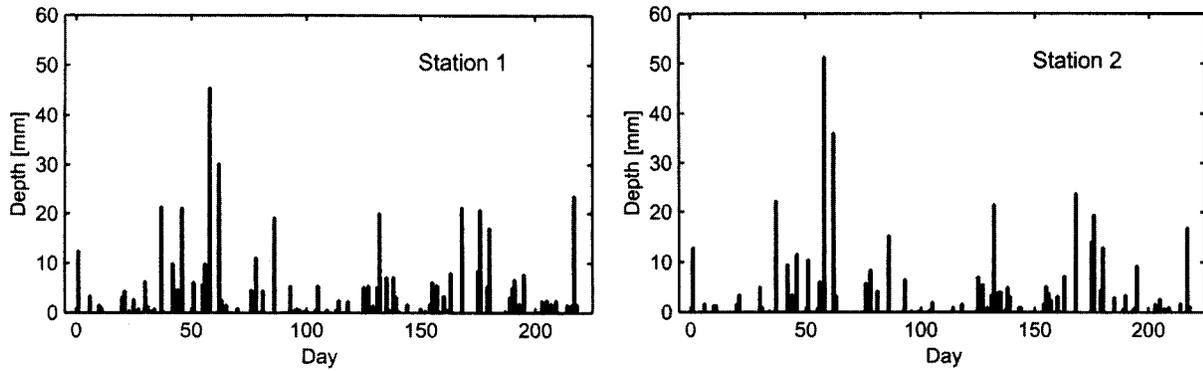


Figure 9. Hyetographs of daily rainfall for two rainfall gauges in southern Ontario (April 25 to November 30, 1993).

the slope of a double-mass curve indicates a drift in the gauge and calls for a correction. In cases where the change in slope can be related to well-documented interventions, such as a change of location or a change of gauging device, the time when the change occurred will usually be known. However, in some cases, the exact time of change may not be known, and, in addition, changes may be so small that the point in time before or after which the correction should be made is not evident from the double-mass curve. Bayesian change point analysis may be helpful to guide the choice of appropriate change point.

For the purpose of illustration we consider two rainfall gauges in southern Ontario. The distance between the two stations is ~2.5 km. The data consists of daily rainfall depths from April 25 to November 30, 1993, for a total of 220 days. The two hyetographs are shown in Figure 9. To construct the double-mass curve, we first eliminated all days for which both stations recorded no rain. This was done in order to reduce the number of data. From this reduced set of data, cumulative depths were obtained. Figure 10 shows the double-mass curve constructed by plotting the cumulative catch of station 2 versus the cumulate catch of station 1 (lower curve). To illustrate how the Bayesian analysis can provide information about the time of change, we inflated all rainfall events at station 2 after August 15 is 313 mm. The drifting series is shown as the upper curve

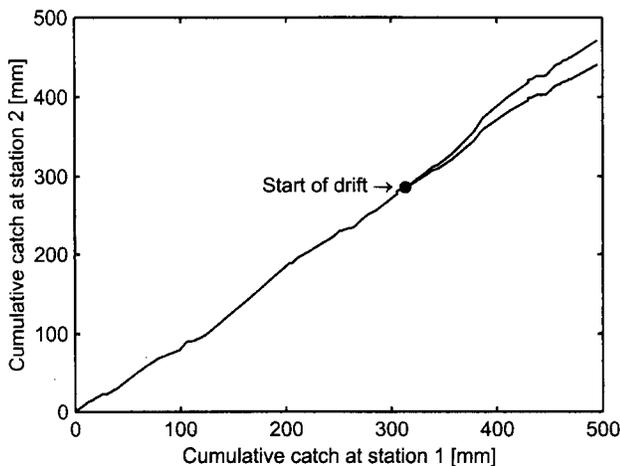


Figure 10. Double-mass curve with (upper curve) and without drift at station 2.

in Figure 10. The change in slope is barely visible and, in practice, it would not be evident where to locate the change point. Note that in virtue of the relationship between cumulated catch and time at station 1, by identifying the change point on the double-mass curve, one implicitly determines the point in time when the change took place.

The double-mass curve with a sudden change in slope can be described by a slightly modified version of (17). The double-

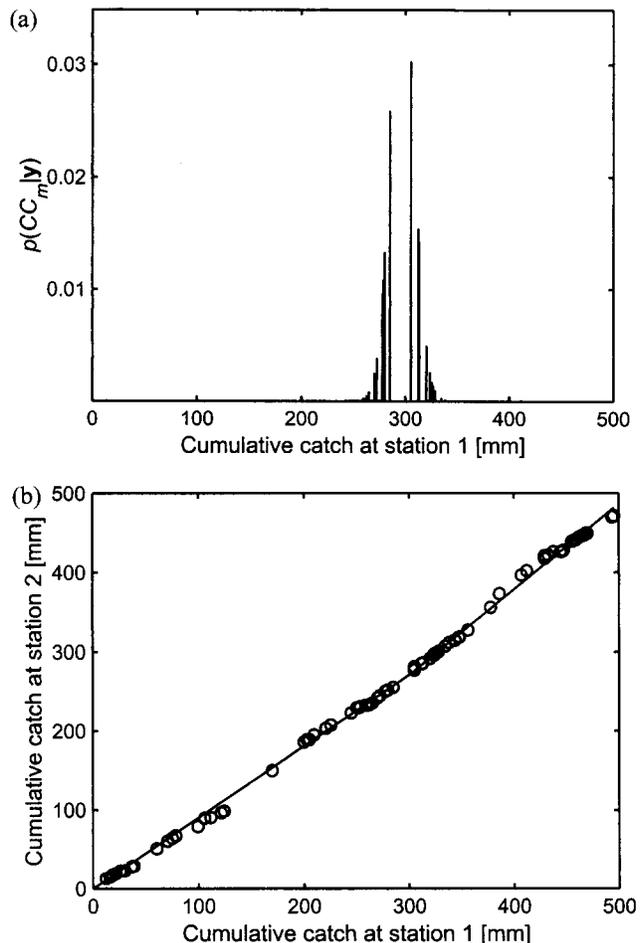


Figure 11. Change in slope of double-mass curve. (a) Posterior distribution of change point, cumulative catch after m observations (CC_m). (b) Fitted model corresponding to change point $CC_m = 303$ mm.

mass curve is known to pass through the origin, so the intercept is zero, and therefore we let $\alpha = 0$ and consider

$$\mathbf{G}_m^T = \begin{bmatrix} x_1 & \cdots & x_m & x_m & \cdots & x_m \\ 0 & \cdots & 0 & (x_m - x_{m+1}) & \cdots & (x_m - x_N) \end{bmatrix}$$

with $M = 2$. As in the previous example, since we are dealing with the problem of curve fitting, the hypothesis that the errors are uncorrelated with each other and with the explanatory variable may not be entirely verified. Nonetheless, the linear model is appropriate for describing the double-mass curve, and we can invoke the Bayesian change point analysis developed previously.

The posterior distribution of the change point resulting from the Bayesian analysis is shown in Figure 11a. Note that the distribution is discrete and irregularly spaced because of the jumps in the series of cumulative catch. The mode seems to be a reasonable choice and is equal to 303 mm at station 1. This corresponds to the change occurring on August 10, an estimate very close to the true value. Figure 11b shows the general linear model fitted to the points on the double-mass curve.

7. Conclusions

The joint use of the generalized linear model and Bayesian analysis has been found to provide a convenient framework for analyzing changes in statistical parameters. The strength of this framework is that it can be easily adapted to a variety of situations, in particular, to different hypotheses about the functional forms before and after a change point and to an arbitrary number of change points. This was demonstrated in this paper by several examples. The key formula is (7), which can be easily implemented using a matrix-based software such as Matlab. For any change point model the basis function matrix \mathbf{G} must be specified, which is usually relatively straightforward. The characteristics of the posterior change point distribution will provide helpful information for selection of the most appropriate change point.

Bayesian change point analysis requires specification of a model before and after the changepoint. This is, in fact, common to most change point analysis procedures. While this component of the analysis necessarily involves a certain amount of subjectivity, the Bayesian analysis eliminates much of the subjectivity involved in choosing the change point from a visual inspection of the data. As illustrated in this paper, the posterior change point distribution will, in many cases, lead to a clear indication of where to locate a change point.

Finally, it should be stressed that what has been presented here is not a statistical test of change versus no change. The a priori assumption is made that there is a change, and one even

prescribes the form of change. The question is when or where the change took place and how much it is. The concentration of the posterior distribution of the change point may be taken as an indicator of the likelihood of a change. A highly concentrated posterior distribution would support the hypothesis of change.

Acknowledgments. The thoughtful comments of Upmanu Lall, Caterina Valeo, and Fahim Ashkar are gratefully acknowledged. Caterina Valeo also helped develop some of the hydrological applications. Financial support for this study was provided by NSERC and by the University of Manitoba.

References

- Bernier, J., Statistical detection of changes in geophysical series, in *Engineering Risk in Natural Resources Management, NATO ASI Ser. E*, vol. 275, edited by L. Duckstein and E. Parent, pp. 159–176, Kluwer Acad., Norwell, Mass., 1994.
- Carlin, B. P., A. E. Gelfand, and A. F. M. Smith, Hierarchical Bayesian analysis of changepoint problems, *Appl. Stat.*, 41, 389–405, 1992.
- Friedman, J. H., Multivariate adaptive regression splines, *Ann. Stat.*, 19, 1–144, 1991.
- Hershfield, D. M., Rainfall frequency atlas of the United States for durations from 30 minutes to 24 hours and return periods from 1 to 100 years, *Tech. Pap. 40*, U.S. Weather Bur., Washington, D. C., 1962.
- Hogg, W. D., and D. A. Carr, *Rainfall Frequency Atlas for Canada*, Environment Canada, Ottawa, Ontario, 1985.
- Koutsoyiannis, D., D. Kozonis, and A. Manetas, A mathematical framework for studying rainfall intensity-duration-frequency relationships, *J. Hydrol.*, 206, 118–135, 1998.
- Lee, A. S. F., and S. M. Heghinian, A shift of the mean level in a sequence of independent normal random variables—A Bayesian approach, *Technometrics*, 19, 503–506, 1977.
- McCuen, R. H., *Hydrologic Analysis and Design*, 2nd ed., Prentice-Hall, Englewood Cliffs, N. J., 1998.
- Ó Ruanaidh, J. J. K., and W. J. Fitzgerald, *Numerical Bayesian Methods Applied to Signal Processing*, Springer-Verlag, New York, 1996.
- Perreault, L., M. Haché, M. Slivitsky, and B. Bobée, Detection of changes in precipitation and runoff over eastern Canada and US using a Bayesian approach, *Stochastic Environ. Res. Risk Assess.*, 12, 201–216, 1999.
- Perreault, L., J. Bernier, B. Bobée, and E. Parent, Bayesian change-point analysis in hydrometeorological time series, part 1, The normal model revisited, *J. Hydrol.*, 235, 221–241, 2000a.
- Perreault, L., J. Bernier, B. Bobée, and E. Parent, Bayesian change-point analysis in hydrometeorological time series, part 2, Comparison of changepoint models and forecasting, *J. Hydrol.*, 235, 242–263, 2000b.
- Stephens, D. A., Bayesian retrospective multiple-changepoint identification, *Appl. Stat.*, 43, 159–178, 1994.

P. Rasmussen, Department of Civil Engineering, University of Manitoba, 342 Engineering Building, 15 Gillson Street, Winnipeg, Manitoba, Canada R3T 5V6. (rasmusse@cc.umanitoba.ca)

(Received January 24, 2000; revised November 28, 2000; accepted May 31, 2001.)