WATER RESOURCES RESEARCH, VOL. 44, W02425, doi:10.1029/2006WR005615, 2008

Intercomparison of homogenization techniques for precipitation data

Claudie Beaulieu,[1] Ousmane Seidou,[1] Taha B. M. J. Ouarda,[1] Xuebin Zhang,[2] Gilles Boulet,[3] and Abderrahmane Yagouti[3]

Received 11 October 2006; revised 30 August 2007; accepted 16 October 2007; published 19 February 2008.

[1] This paper presents an intercomparison of eight statistical tests to detect inhomogeneities in climatic data. The objective was to select those that are more suitable for precipitation data in the southern and central regions of the province of Quebec, Canada. The performances of these methods were evaluated by simulation on several thousands of homogeneous and inhomogeneous synthetic series. These series were generated to reproduce the statistical characteristics of typical precipitations observed in the southern and central parts of the province of Quebec and nearby areas, Canada. It was found that none of these methods was efficient for all types of inhomogeneities, but some of them performed substantially better than others: the bivariate test, the Jaruskova's method, and the standard normal homogeneity test. Techniques such as the Student sequential test and the two-phase regression led to the worst performances. The analysis of the performances of each method in several situations allowed the design of an optimal procedure that takes advantage of the strengths of the best performing techniques.

Citation: Beaulieu, C., O. Seidou, T. B. M. J. Ouarda, X. Zhang, G. Boulet, and A. Yagouti (2008), Intercomparison of homogenization techniques for precipitation data, *Water Resour. Res.*, *44*, W02425, doi:10.1029/2006WR005615.

1. Introduction

[2] Hydroclimatic data records often undergo artificial disturbances that do not reflect the real climate variations. These disturbances can be related for instance to station relocation, instrument replacement, change in observation procedures or modification in the immediate environment of the site. Homogenization is the technique of detecting and correcting these artificial disturbances. A climate series is considered homogeneous when the measurement conditions of the station had not varied with time. Different types of homogenization techniques are presented in the literature: statistical techniques or techniques based on expert judgment, techniques based principally on metadata (archive of a station), or techniques based on the concept of relative homogeneity (it consists of a comparison of the series to homogenize with a reference series to separate the artificial change from the regional climate signal). The different issues related to the homogenization process are discussed by *Peterson et al.* [1998] and by *Aguilar et al.* [2003]. In this paper, we focus on statistical techniques which use a reference series.

[3] Different types of changes will introduce different types of inhomogeneities in the series. For a variable such as precipitation, a change of exposition or a relocation of the gauge are the type of changes that are the most likely to introduce an inhomogeneity in the series. For an exhaustive review of the degree of influence of each type of change on different variables, the reader is referred to *Heino* [1997].

[4] In various fields, the need for long and reliable climatic data series is high. During the last decades, several efforts were made to develop techniques to correct anthropogenic changes in climate series. For example, climate change studies require the creation of complete databases in order to adequately analyze the climatic signal, and estimate the future change with a minimal uncertainty. The need to develop robust homogenization techniques and to identify the most suitable method for each type of variable (e.g., temperature or precipitation) is thus obvious. A study of appropriate techniques for temperature series was conducted by *Ducré-Robitaille et al.* [2003].

[5] A comparative study of several homogenization methods for precipitation is carried out in this paper. The techniques are systematically applied to several thousands of synthetic data sets having the same statistical characteristics as the recorded series of total annual precipitation. Stations located in the southern and central regions of the province of Quebec (Canada) and nearby areas were selected to estimate these statistical characteristics. First, the homogenization techniques were selected using some practical considerations such as speed and technical characteristics. Then, several sets of homogeneous and inhomogeneous synthetic series were generated using a technique derived by *Easterling and Peterson* [1992]. Once the series were generated, each selected homogenization method was applied to each synthetic series, and the resulting performance was used as the basis of the intercomparison. Finally, recommendations are formulated

[1] Eau, Terre et Environnement, Institut National de la Recherche Scientifique, Université du Québec, Quebec, Quebec, Canada.
[2] Science and Technology Branch, Climate Research Division, Environment Canada, Toronto, Ontario, Canada.
[3] Direction du suivi de l'état de l'environnement, Ministère de l'Environnement du Québec, Quebec, Quebec, Canada.

0043-1397/08/2006WR005615$09.00

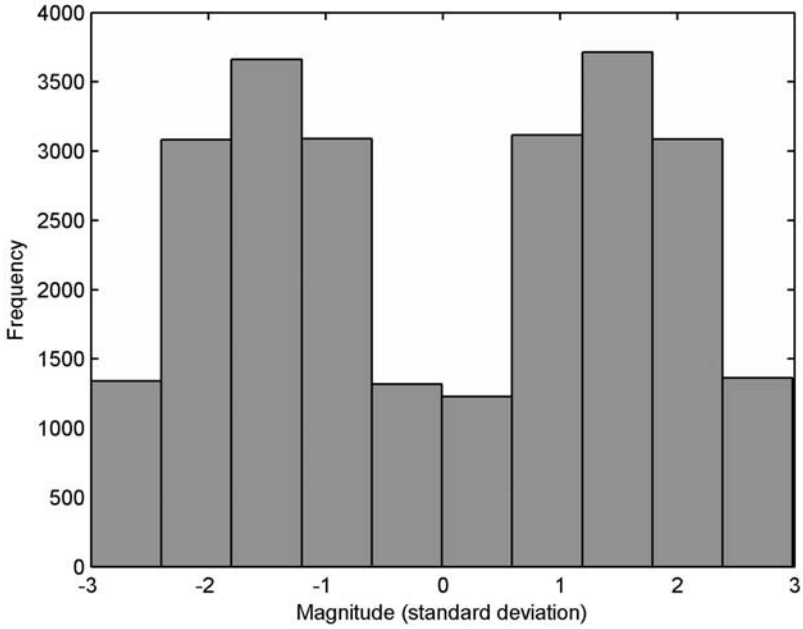

Figure 1. Histogram of the magnitudes of the shifts introduced in 100-year-long series with a single shift.

concerning the best combination of homogenization methods for the application to total annual precipitation.

2. Data

[6] The data sets are synthetic series of precipitation with the same statistical characteristics (average, standard deviation, order 1 autocorrelation and spatial cross-correlation) as typical series of annual total precipitation observed in southern and central Quebec and nearby areas (Canada). The parameters of the generation scheme were chosen to fit the statistical characteristics of precipitation data series recorded at some selected stations in the province or immediate surroundings. These stations have high-quality records (long observation series with little missing data). All the observation series of the selected stations passed the Shapiro-Wilk normality test, supporting our choice of a normal distribution in the generation scheme.

[7] Homogeneous synthetic series (mean and variance are constant) and inhomogeneous (one or multiple shifts, trend, shift in standard deviation) were first generated. Correlated neighbor series were also generated using a technique that will be described further in the text [*Easterling and Peterson*, 1992; *Vincent*, 1998; *Ducré-Robitaille et al.*, 2003].

2.1. Base Series

2.1.1. Homogeneous

[8] Homogeneous base series were generated to study the sensitivity of the techniques on homogeneous series. Lag one autoregressive variables, z_i, were first generated using the following model:

$$z_i = \phi_1 z_{i-1} + e_i \tag{1}$$

where ϕ_1 is the autocorrelation coefficient and e_i is a normally distributed residual with zero mean and variance $1 - \phi_1^2$. The mean and variance of the real total annual precipitation were introduced in the z_i series.

[9] The statistical characteristics to be reproduced are (1) a mean total annual precipitation of 1089 mm, (2) a standard deviation of 142 mm, and (3) a lag one autocorrelation of 0.02. These values are the average characteristics of the selected stations. Even though the autocorrelation was not significant, the series were generated using an autoregressive model instead of a normal independent model to represent the real data series as much as possible. A total number of 10 000 homogeneous series (5000 60-year-long and 5000 100-year-long) were generated this way.

2.1.2. Series With a Single Shift

[10] Series with a single shift in the mean were generated to study the ability of the methods to detect the position and to estimate the magnitude of a single shift. The procedure for the selection of the magnitude and position of a shift is described in the next paragraph. A series with a single shift can be represented by:

$$y_i = \begin{cases} y_i^* - \delta_{p_1}\sigma, & i = 1, \ldots, p_1 - 1 \\ y_i^*, & i = p_1, \ldots, n \end{cases} \tag{2}$$

δ_{p_1} and p_1 were randomly generated using the following distributions:

$$\delta_{p_1} = sign(u - 1/2) \cdot 3 \cdot b, \quad u \sim U(0,1), \quad b \sim BETA(2,2) \tag{3}$$

$$p_1 = 10 + ud, \quad ud \sim DUNIF(n - 20) \tag{4}$$

where y_i represents the ith observation of an inhomogeneous series of size n, y_i^* the ith observation of a homogeneous series, σ is the standard deviation of the last segment of the series, δ_{p_1} the magnitude of the shift, u is a uniform variable, b is a beta variable, p_1 is the position of the shift and ud is a discrete uniform variable. The magnitude varies randomly between -3 and 3 standard deviations (equation (3)). The

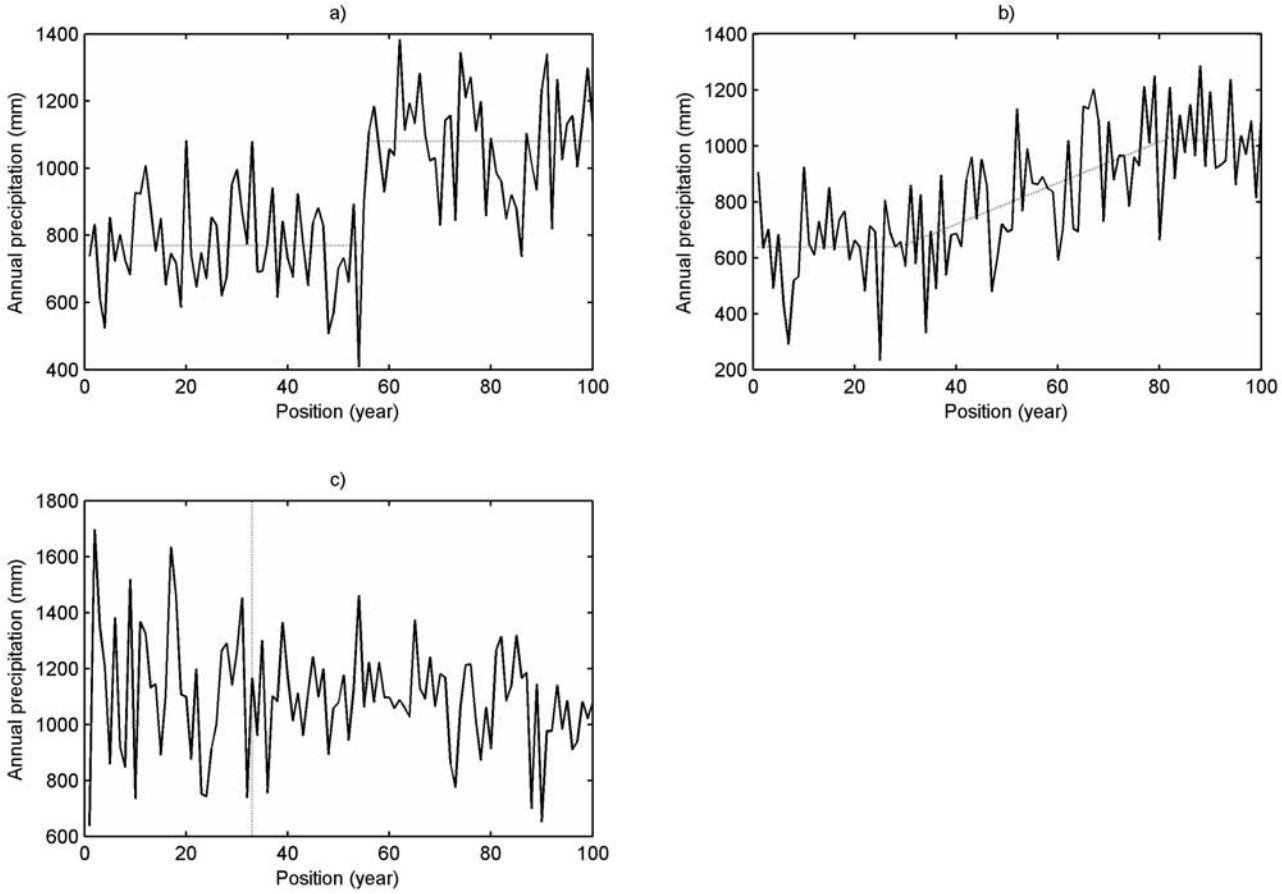

Figure 2. Example of generated synthetic candidate series (a) with a shift at position 56, (b) with a trend starting at position 30 and ending at position 83, and (c) with a change in standard deviation at position 33.

distribution of the magnitudes was chosen to have few cases for which the magnitude is close to zero, to avoid a repetition of assessing the methods when the series was close to being homogeneous, and to save computing time. Furthermore, since the distribution of the magnitudes of real inhomogeneities is never known, the distribution was chosen to represent all potential magnitudes. Figure 1 presents the distribution of the magnitudes of the shifts for the 100-year-long series with a single shift. The position was generated from a discrete uniform distribution that was truncated to avoid the presence of shifts during the first 10 or the last 10 years of the series (equation (4)). A total number of 50 000 series (25 000 60-year-long and 25 000 100-year-long series) were generated this way. Since the techniques were developed to detect a single shift, a higher number of synthetic series were generated for this case. The number of series for each case was chosen to have enough repetitions of the different scenarios and according to the interest of each type of series. Figure 2a presents an example of a series with a single shift.

2.1.3. Series With Multiple Shifts

[11] Series with multiple shifts (2 and 3) were also generated. A series with two shifts can be represented by:

$$y_i = \begin{cases} y_i^* - \delta_{p_1}\sigma, & i = 1, \ldots, p_1 - 1 \\ y_i^* - \delta_{p_2}\sigma, & i = p_1, \ldots, p_2 - 1 \\ y_i^*, & i = p_2, \ldots, n \end{cases} \quad (5)$$

The positions and magnitudes of the shifts were generated using the following distributions:

$$\delta_{p_1}, \delta_{p_2} = sign(u - 1/2) \cdot 3 \cdot b, \quad u \sim U(0,1), \quad b \sim BETA(2,2) \quad (6)$$

$$p_1 = 10 + ud, \quad ud \sim DUNIF(n - 31) \quad (7)$$

$$p_2 = 10 + p_1 + ud, \quad ud \sim DUNIF(n - 20 - p_1) \quad (8)$$

where y_i represents the ith observation of an inhomogeneous series of size n, y_i^* the ith observation of a homogeneous series, and σ the standard deviation of the last segment of the series. δ_{p_1}, δ_{p_2}, p_1 and p_2 respectively denote the magnitudes and positions of the shifts. u, b and ud respectively denote a uniform variable, a beta variable and a discrete uniform variable. The magnitudes were generated in the same way as for a single shift. Discrete uniform distributions were used to generate the positions, but their parameters were adapted to the number of shifts to generate. The minimum interval between two consecutive shifts was set to 10 years. Series with three shifts were generated using a similar procedure. A total number of 15 000 series with two shifts were generated and 15 000 series with three shifts in the mean.

2.1.4. Series With a Trend

[12] When there is a trend in the data, the homogenization methods may interpret it as one or several consecutive shifts. Series with trends were generated to study the behavior of the homogenization methods on such series. In spite of the fact that most of the homogenization methods are not developed to detect trends, we were interested in their performance to identify a change inside the trend. A series with a trend can be represented by:

$$y_i = \begin{cases} y_i^* - \delta_{p_1:p_2}\sigma, & i = 1, \ldots, p_1 - 1 \\ y_i^* - \delta_{p_1:p_2}\sigma - mi, & i = p_1, \ldots, p_2 - 1 \\ y_i^*, & i = p_2, \ldots, n. \end{cases} \quad (9)$$

The following distributions were used to generate the trended series:

$$\delta_{p_1:p_2} = sign(u - 1/2) \cdot 3 \cdot b, \quad u \sim U(0,1), \quad b \sim BETA(4,2) \quad (10)$$

$$p_1 = 10 + ud, \quad ud \sim DUNIF(n - 31) \quad (11)$$

$$p_2 = 10 + p_1 + ud, \quad ud \sim DUNIF(n - 20 - p_1) \quad (12)$$

where y_i represents the ith observation of an inhomogeneous series of length n, $y_i^*$the ith observation of a homogeneous series, σ the standard deviation of the last part of the series, $\delta_{p_1:p_2}$ the magnitude of the trend, m and the slope. u, b and ud respectively denote a uniform variable, a beta variable and a discrete uniform variable. p_1 and p_2 represent the beginning and the end of the trend. The magnitude was chosen to lie between -3 and 3 standard deviations. The positions of the beginning and the end of the trend were generated with the same technique as that for series with two shifts. A total number of 10 000 synthetic series with a random trend were generated. Figure 2b presents an example of series with this type of discontinuity.

2.1.5. Series With a Shift of Variance

[13] Series with a shift of variance were generated to determine which methods are sensitive to this type of discontinuity. The studied methods were not initially designed to detect changes in the variance and most of them are based on the hypothesis that the variance is constant. It is thus interesting to check their robustness to violations of the latter postulate. The position of the variance shift was randomly selected from a discrete uniform distribution (equation (4)). The magnitudes are generated from a BETA(8, 2) distribution multiplied by a random sign and divided by 2. The magnitudes of the generated variance shifts lie between 0 and 50% of the standard deviation. A total number of 10 000 series with a shift of variance were generated. Figure 2c represents an example of a synthetic series with a shift of variance.

2.2. Neighbor Series

[14] For every base series three correlated neighbor series were generated in two steps. First, three homogeneous series (independent of the base series) were generated:

$$w_i = \phi_1 w_{i-1} + e_i \quad (13)$$

where ϕ_1 is the lag one autocorrelation coefficient and e_i is a normally distributed residual with zero mean and variance $1 - \phi_1^2$. Then, a correlation structure was introduced between the base series and the neighbor series:

$$w_i = \psi z_i + w_i \quad (14)$$

where z_i represents the standardized total precipitation at the base station for year i, w_i the standardized total precipitation at a neighbor station for year i, and ψ is a correlation coefficient between the neighbor series and the base series. w_i was then standardized to ensure it has zero mean and standard deviation 1. The neighbor series possess the same statistical characteristics as the base series. The correlation coefficient (0.7) was determined by simulation to reproduce a spatial cross-correlation of 0.55. This value was chosen because the mean spatial cross-correlation in the set of selected stations that are located at a distance less than 300 km is 0.55. This distance was chosen to test the techniques in the worst conditions. Correlation between stations varies enormously and is likely to affect the performance of the homogenization methods, which are expected to perform better when the base and neighbor series are highly correlated. However, the station network density in Quebec is relatively low given the large size of the province (1 542 056 km^2). Consequently, it can be difficult in some regions to find several neighbors for the same base series. To represent this reality as closely as possible, the number of neighbor stations was set to 3.

3. Methods

[15] Homogenization techniques compared in this work were selected according to the following set of criteria. First, the methods should be objective. Since the techniques were applied to thousands of series, subjective methods could not be used. Second, they should be able to detect multiple shifts (as these can be observed in practice) and estimate them. Some methods developed for one shift were nevertheless selected, but they were adapted for multiple shifts using a segmentation approach. Third, the techniques should allow the use of one or several neighbor series. Finally, the algorithms must be available in the literature and have a reasonable running time.

[16] An extensive literature review was performed [*Beaulieu et al.*, 2007] and eight methods were selected: (1) standard normal homogeneity test (SNHT) [*Alexandersson*, 1986; *Khaliq and Ouarda*, 2007], (2) multiple regression (MREG) [*Vincent*, 1998], (3) two-phase regression (REG2) [*Easterling and Peterson*, 1995; *Lund and Reeves*, 2002], (4) bivariate test (BIVT) [*Maronna and Yohai*, 1978; *Potter*, 1981], (5) sequential Wilcoxon test (WILS) [*Karl and Williams*, 1987; *Lanzante*, 1996; *Ducré-Robitaille et al.*, 2003], (6) sequential t test (STUS) [*Gullett et al.*, 1990], (7) Jaruskova's method (JARU) [*Jaruskova*, 1996], and (8) Bayesian approach (BAYE) [*Rasmussen*, 2001]. The hypothesis of normality has to be respected to apply these tests except for the sequential Wilcoxon test which is nonparametric. All methods were coded to ignore the shifts detected among the first ten or last ten observations. The authors consider that it is not reasonable to estimate the

magnitude of an inhomogeneity with less than ten observations on each side of the inhomogeneity.

3.1. Standard Normal Homogeneity Test

[17] A series of ratios between the base and the neighbor series is created:

$$q_i = y_i \Big/ \left[\left(\sum_{j=1}^{k} \rho_j^2 x_{ij} \overline{y}_{1:n} / \overline{x}_{1:n,j} \right) \Big/ \sum_{j=1}^{k} \rho_j^2 \right] \quad \begin{matrix} i = 1, \ldots, n \\ j = 1, \ldots, k \end{matrix} \tag{15}$$

where the value of the year i of the base series is represented by y_i, x_{ij}, is the ith observation of the neighbor series j. There are k nearby sites with n observations each. The correlation coefficient between the base series and the neighbor series j is noted ρ_j. The hypothesis that the standardized ratios follow a normal distribution with zero mean and a variance of 1 is tested against the hypothesis that there is a shift in the mean of the series. To find the position of the change, a weighted average series is created:

$$Q_i = i\overline{v}_{1:i}^2 + (n-i)\overline{v}_{i+1:n}^2, \quad i = 1, \ldots, n-1 \tag{16}$$

where $\overline{v}_{1:i}$ and $\overline{v}_{i+1:n}$ are the average of the standardized ratios for segments 1:i and i+1: n. The test statistic, $Q_{p_1} = \max_{i=1,\ldots,n-1}\{|Q_i|\}$, is significant if it exceeds the associated critical value [*Khaliq and Ouarda*, 2007] and the shift is located at position p_1. For this study, this test was successively applied to all synthetic series with a critical level of 5%.

3.2. Multiple Linear Regression

[18] This approach is based on the application of four regression models representing different types of inhomogeneities [*Vincent*, 1998]. In this study only two models are used. The first one represents a homogeneous base series. When residuals are independent, the model provides a good fit to the data and the series is considered homogeneous. The homogeneity of the series is verified with an independence test on the residuals (e.g., a confidence interval on the lag one autocorrelation of the residuals). On the other hand, if residuals are autocorrelated, it indicates that the model does not fit the data well and that the base series could be inhomogeneous. In this case, the model describing a shift in the base series is applied:

$$y_i = \begin{cases} \tau + \beta_1 x_{i1} + \ldots + \beta_k x_{ik} + e_i & i = 1, \ldots, p_1 - 1 \\ \tau + \delta_{p_1} + \beta_1 x_{i1} + \ldots + \beta_k x_{ik} + e_i & i = p_1, \ldots, n \end{cases} \tag{17}$$

where y_i represents the base series, x_{ij} represents the neighbor series j, and τ and β_j (j = 1, …, k) are the least squares parameters. There are k neighbor series with n observations each. The residuals, e_i, follow a normal distribution with zero mean and constant variance. The position of the shift, p_1, is determined by fitting the model for all possible positions and by selecting the results with the smallest residual sum of squares. Again, if residuals are independent, then there is probably a shift at position p_1. A Fisher test comparing the fit of the homogeneous model with the model representing a shift is applied. If the model with a shift is better, the significance level of the shift is evaluated with a Student statistic. The same process is continued on each side of the shift until all segments are considered homogeneous. This method was applied to all synthetic series with a 5% critical level.

3.3. Two-Phase Regression

[19] Several versions of the REG2 can be found in the literature [*Solow*, 1987; *Easterling and Peterson*, 1995; *Lund and Reeves*, 2002; *Wang*, 2003]. For the purpose of this work, the model proposed by *Lund and Reeves* [2002] was used. Two regression models are fitted with time as the explanatory variable. The first model represents a homogeneous series. The second model represents a discontinuous series displaying a shift at time p_1:

$$y_i = \begin{cases} \tau_1 + \lambda_1 i + e_i, & i = 1, \ldots, p_1 - 1 \\ \tau_2 + \lambda_2 i + e_i, & i = p_1, \ldots, n \end{cases} \tag{18}$$

where y_i represents the ith observation of the base series, τ_1, τ_2, λ_1, and λ_2 are respectively the intercepts and slopes before and after the change. The position of the shift is chosen by fitting the model for all possible values of p_1, by computing the Fisher statistics, comparing all fitted models to that of the homogeneous one, and then by choosing the one which gives the maximum Fisher statistic (*Fmax*). The Fisher test gives information regarding the contribution of a new variable in a regression model. In this case, it means that the significance of the introduction of a step at position p_1 is verified. The critical values of the Fmax statistic were obtained by simulation and provided by *Lund and Reeves* [2002]. The same process is repeated until all segments are found homogeneous or have less than ten observations. The method described in the above section was applied to the difference series between the base series and the neighbor series, with a critical level of 5%.

3.4. Bivariate Test

[20] This method was developed by *Maronna and Yohai* [1978] and applied to homogenization problems by *Potter* [1981]. The technique is based on the postulate that the base series (y_i, i = 1, …, n) and a reference series (x_{i1}, i = 1, …, n) belong to the same bivariate normal distribution. It is hypothesized that there is a shift in the base series that does not occur in the reference series. The method is based on the following series of ratios:

$$q_i = \frac{i(n-i)\delta_i^2 F_i}{S_x S_y - S_{xy}^2}, \quad i = 1, \ldots, n-1 \tag{19}$$

where

$$\delta_i = \frac{\left[S_x \left(\overline{y}_{1:n} - \sum_{j=1}^{i} y_j / i \right) - S_{xy} \left(\overline{x}_{1:n,1} - \sum_{j=1}^{i} x_{j1} / i \right) \right] n}{(n-i)F_i}, \quad i = 1, \ldots, n-1 \tag{20}$$

$$F_i = S_x - (x_i - \overline{x})^2 ni/(n-i) \tag{21}$$

$$S_{xy} = \sum_{i=1}^{n} \left(x_{i1} - \overline{x}_{1:n,1} \right) \left(y_i - \overline{y}_{1:n} \right) \tag{22}$$

$$S_x = \sum_{i=1}^{n} \left(x_{i1} - \overline{x}_{1:n,1}\right)^2 \quad (23)$$

$$S_y = \sum_{i=1}^{n} \left(y_i - \overline{y}_{1:n}\right)^2 \quad (24)$$

The statistic of the test is given by $Q_{p_1} = \max_{i=1,\ldots,n-1}\{|q_i|\}$. The critical values of Q_{p_1} are obtained by simulation [*Maronna and Yohai*, 1978]. When the test is positive, it is assumed that a shift occurred at year p_1. The approach was iteratively applied on all sets of synthetic series, assuming a critical level of 5%. Since the technique allows the use of a single reference series, the average of the three synthetic neighbor series was used as reference series.

3.5. Sequential Student Test

[21] The sequential Student test consists of using a moving window and testing successively the equality of the means of the first half and the second half of the observations in the window [*Gullett et al.*, 1990]. Following the recommendations given by *Ducré-Robitaille et al.* [2003], the size of the moving window was increased to 20 years to obtain a better performance. Then, we test the equality of the means by using 10 years before and after every potential position:

$$Q_i = \frac{\overline{q}_{i-10:i-1} - \overline{q}_{i:i+9}}{\sqrt{s^2_{i-10:i-1}/10 + s^2_{i:i+9}/10}}, \quad \mathrm{i} = 11, .., \mathrm{n} - 9 \quad (25)$$

where $\overline{q}_{i-10:i-1}$ $\overline{q}_{i:i+9}$, $s^2_{i-10:i-1}$ and $s^2_{i:i+9}$ are the means and variances of the ten observations located before and after the position i in the ratios series. The maximum value of the Student statistic corresponds to the position of the shift. The Student statistic is significant if it exceeds the critical value of the Student distribution with 18 degrees of freedom. When the statistic is significant, the series is split into two segments and the process is repeated until all shifts are detected. This procedure was substituted for the original one because, by extracting simultaneously all the significant statistics, the same shift is identified several times since successive Student statistics are highly correlated. Furthermore, since the test is applied to the same series several times, the probability to meet a type 1 error (to reject the null hypothesis while it is true) is increased. That is why the critical levels used to get a global critical level of 5% were computed by simulation. The critical levels used were 0.225% and 0.0875% for series of length 60 and 100 years, respectively. Finally, the method was applied to the series of ratios between the base series and the neighbor series, since the ratios are usually used with precipitation.

3.6. Sequential Wilcoxon Test

[22] The Wilcoxon test has been extensively used for the homogenization of climate data [*Karl and Williams*, 1987; *Lanzante*, 1996; *Ducré-Robitaille et al.*, 2003]. The most recent version of the method was used in this work. It consists of computing successively the Wilcoxon statistic and estimating its significance level using a normal approximation:

$$Q_i = \frac{R_i - i(n+1)/2}{\sqrt{i(n-i)(n+1)/12}}, \quad i = 11, \ldots, n-9 \quad (26)$$

where n represents the length of the tested series, $R_i = \sum_{j=1}^{i} r_i$ and r_i are the ranks of the first part of the series. The maximum of the series and its position are then extracted ($Q_{p_1} = \max_{i=11,..,n-9}\{|Q_i|\}$). If the statistic is significant, then there is a change of mean at this position. The series is then split into two segments, and the same procedure is applied on each of the new series. The same operations are performed on each obtained segment until all segments are found homogeneous or have a length smaller than 10 observations. As in the case of the sequential Student test, the critical levels were modified to have a critical level of 5%. The critical levels used were 0.44% and 0.289% for series of length 60 and 100 years, respectively. In the work of *Ducré-Robitaille et al.* [2003], this method was applied to a series of differences between the base series and a reference series because the variable of interest was the temperature. For the purpose of this work, we used a series of ratios between the base series and the neighbor series.

3.7. Jaruskova's Method

[23] This method was proposed by *Jaruskova* [1996] to detect a shift in a meteorological series. Several alternative approaches were presented in the work of *Jaruskova* [1996], but the model for which the date of change is unknown was selected in this work. The method consists of building a difference series between the base series and a reference series, and then testing the hypothesis that there is a change in the mean of the difference series. The following statistic is computed for all possible positions for a shift in the series.

$$Q_i = \sqrt{\frac{(n-i)i}{n}} \frac{\left(\overline{q}_{1:i} - \overline{q}_{i+1:n}\right)}{s_i} \qquad i = 1, \ldots, n-1 \quad (27)$$

where

$$s_i^2 = \frac{1}{n-2}\left[\sum_{j=1}^{i}\left(q_j - \overline{q}_{1:i}\right)^2 + \sum_{j=i+1}^{n}\left(q_j - \overline{q}_{i+1:n}\right)^2\right] \quad (28)$$

and q_i represents the difference between the base station and a reference station for the year i and n is the length of the series. We assume that the difference series is normally distributed. The maximum of the series, $Q_{p_1} = \max_{i=1,\ldots,n-1}\{|Q_i|\}$, is extracted. The shift is significant if the statistic exceeds the critical value of the distribution [*Jaruskova*, 1996]. This method was applied to all sets of synthetic series at the critical level of 5%, ignoring the inhomogeneities in the first ten or last 10 years. The reference series is the mean of the three synthetic neighbor series.

3.8. Bayesian Method

[24] Many Bayesian models have been proposed in the literature [*Asselin et al.*, 1999; *Perreault et al.*, 1999; *Rasmussen*, 2001; *Ouarda et al.*, 2005]. In this study we used the technique presented by *Rasmussen* [2001]. It consists of inferring the parameters of a linear regression model using an analytical Bayesian approach. Several alternative models were considered by *Rasmussen* [2001], but the one that was used in this study is:

$$y_i = \begin{cases} \tau_1 + \beta_1 x_{i1} + \beta_2 x_{i2} + \beta_3 x_{i3} + e_i & i = 1, \ldots, p_1 - 1 \\ \tau_2 + \beta_1 x_{i1} + \beta_2 x_{i2} + \beta_3 x_{i3} + e_i & i = p_1, \ldots, n \end{cases} \quad (29)$$

where y_i represents the base series, x_{ij} represents the ith observation of the jth neighbor series, τ_1, τ_2, and β_j (j = 1, 3) respectively represent the intercept point before the change, the intercept point after the change and the neighbor series coefficients. The model assumes that the data are independent and normally distributed. The prior probability densities on the regression parameters and on the position of change are noninformative (uniformly distributed with bounds: $(-\infty, \infty)$). The position of the change is chosen to be the mode of the posterior distribution. For the purpose of this work, the Bayesian inference is performed on the position of the change (p_1) as well as on the parameter vector $\boldsymbol{\theta} = [\tau_1, \tau_2, \beta_1, \beta_2, \beta_3]^T$. Let us denote the base series by the vector Y in this description. The posterior probability density of the position of the change is given by:

$$pr(p_1|Y) = \left\{ \left| G_{p_1-1}^T G_{p_1-1} \right|^{-1/2} \left[Y^T Y - Y^T G_{p_1-1} \left(G_{p_1-1}^T G_{p_1-1} \right)^{-1} G_{p_1-1}^T Y \right] \right\} * \left\{ \sum_{i=1}^{n-1} \left| G_i^T G_i \right|^{-1/2} \left[Y^T Y - Y^T G_i \left(G_i^T G_i \right)^{-1} G_i^T Y \right]^{-(n-5)/2} \right\}^{-1} \quad (30)$$

where

$$G_{p_1-1}^T = \begin{pmatrix} 1 & 0 & x_{1,1} & x_{1,2} & x_{1,3} \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ 1 & 0 & x_{p_1-1,1} & x_{p_1-1,2} & x_{p_1-1,3} \\ 0 & 1 & x_{p_1,1} & x_{p_1,2} & x_{p_1,3} \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ 0 & 1 & n_{n,1} & x_{n,2} & x_{n,3} \end{pmatrix} \quad (31)$$

The first two columns of the G_i matrix contain the indicative variables. The other columns contain the observations of the neighbor series. n is the length of the vector and x_{ij} the observations of the neighbor series j. The posterior density of the parameters of $\boldsymbol{\theta}$ is given by:

$$pr(\boldsymbol{\theta}|Y, p_1) = \frac{\left\{ \left[\Gamma(\upsilon + 5/2) \left| G_{p_1-1}^T G_{p_1-1} \right|^{1/2} \right] / \left[(\Gamma(1/2))^5 \Gamma(\upsilon/2) (c\sqrt{\upsilon})^5 \right] \right\}}{\left\{ 1 + \frac{\left(\boldsymbol{\theta} - \hat{\boldsymbol{\theta}} \right)^T G_{p_1-1}^T G_{p_1-1} \left(\boldsymbol{\theta} - \hat{\boldsymbol{\theta}} \right)}{\upsilon c^2} \right\}^{(\upsilon+5)/2}} \quad (32)$$

with

$$\hat{\boldsymbol{\theta}} = \left(G^T G \right)^{-1} G^T Y \quad (33)$$

$$c^2 = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 / (n-2) \quad (34)$$

and υ is the number of degrees of freedom (n-5), Γ represents the Gamma function and c^2 is the unbiased estimate of the noise variance. As there is no analytical expression for the posterior distribution of the magnitude of change, it was computed with Monte-Carlo simulations. A Bayesian credibility interval was used to verify the significance of the shift. This method was applied to all sets of synthetic series.

4. Performance Evaluation

[25] The selected homogenization methods were developed to detect a single shift in a series. In the presence of multiple shifts, the estimation of the magnitude may be biased. To avoid this, the magnitudes can be obtained by computing the difference of the means of the segments before and after the shift. For all sets of synthetic series, the homogenization methods were applied to find the position of the shifts. The magnitudes were then estimated by difference of means. Furthermore, the performance of each technique was evaluated differently according to the type of synthetic series (homogeneous, with a single step, with multiple steps, with a change of standard deviation and with a trend). In this section, the criteria used to evaluate the performance of the homogenization techniques for each set of synthetic series are presented. The reader must note that since the performance of the different techniques is compared inside each set of series, with the same number of repetitions, there is no bias in the statistics that are presented in the result section.

4.1. Homogeneous Series

[26] The homogenization methods were applied to two sets of homogeneous series (60-year-long and 100-year-long) and their performance was evaluated by the percentage of type 1 error (the number of cases for which the homogeneity hypothesis is rejected while it is true).

4.2. Series With a Single Shift

[27] The selected methods were applied to two sets of synthetic series containing a shift with random position and magnitude. The numbers of correctly identified, well-identified, and well-positioned shifts were computed. We consider that a shift is correctly identified when its position is exact and the relative difference between the estimated magnitude and the real magnitude is less than 20% of the real magnitude. A shift is well identified when the estimated position is less than 2 years from the real position and the absolute error on the estimation of the magnitude is lower or equal to 50% of the real magnitude. A well-positioned shift is located between 0 and 2 years from the exact location of the shift and there is no measure of the accuracy of the magnitude of the shift. Furthermore, the differences between the real position and magnitude and the estimated position and magnitude were computed independently. Four types of error series were produced: the position error, the magnitude error, the absolute position error and the absolute magnitude error. When a technique did not detect a shift, the position error and absolute position error are fixed to the length of the series (either 60 or 100) while the magnitude error and absolute magnitude error are fixed to 3 (the highest possible magnitude).

4.3. Series With Multiple Shifts

[28] Most of the methods compared in this work are not designed to detect multiple shifts. This issue is addressed in practice using a segmentation approach. The performance of the methods to identify multiple shifts on synthetic series with two or three shifts was thus tested. A performance criterion was designed to estimate the capacity of the

Table 1. Falsely Detected Shifts by Each Technique When Applied to 60-Year-Long Homogeneous Series[a]

Magnitude (Standard Deviation)	SNHT	MREG	REG2	BIVT	STUS	WILS	JARU	BAYE
0–0.25	0.1	0.1	1.5	0.0	1.7	0.3	0.0	0.7
0.25–0.5	0.4	0.4	1.2	0.4	1.9	1.1	0.1	11.8
0.5–1	2.0	0.4	1.2	1.7	1.5	2.1	0.9	8.5
1–2	0.0	0.0	0.1	0.1	0.1	0.3	0.0	0.2
>2	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
Total	2.5[b]	0.9[b]	4.0[b]	2.2[b]	5.2	3.8[b]	1.0[b]	21.2[b]

[a]Shifts are in percent.
[b]Significantly different from the expected percentage of type I error (5% critical level).

techniques to correctly position all the shifts, without omission or false detection. This criterion measures a distance between the positions of the real shifts and the detected shifts. It can be expressed as follows:

$$C = \begin{cases} \frac{1}{nd}\sum_{i=1}^{nd}\left(p_i^d - p_i\right)^2, & nr = nd \\ \frac{1}{nr}\left[\sum_{i=1}^{nd}\left(p_i^d - p_i\right)^2 + |nr - nd|(n-1)^2\right], & nr > nd \\ \frac{1}{nd}\left[\sum_{j=1}^{nr}\left(p_j^d - p_j\right)^2 + |nr - nd|(n-1)^2\right], & nr < nd \end{cases} \quad (35)$$

where p_i^d, $i = 1, \ldots, nd$ and p_j, $j = 1, \ldots, nr$ represent respectively the positions of the detected and real shifts and n is the length of the series. The pairs (p_i^d, p_j) are chosen to minimize the criterion. When the number of detected and real shifts is the same, the criterion is the sum of squares of the differences between the pairs which minimize the criterion. When the number of detected shifts is different from the number of real shifts, the value $(n-1)^2$ is added for every wrongly detected shift. This value corresponds to the square of the maximum possible distance between two shifts. C is equal to zero when all shifts are correctly positioned. When C is close to zero, the detected shifts are located near the real positions. A high value of C indicates that some shifts in the series are not detected or are wrongly detected. The performance criterion C was computed for the sets of synthetic series with two and three shifts.

4.4. Series With a Shift of Variance

[29] The techniques presented in this paper make the assumption that the variance is constant throughout the series. Therefore a change of variance could affect the results of a homogenization procedure. To investigate the robustness of the methods regarding this postulate, synthetic series (10 000) with a shift in standard deviation were generated. As in the case of homogeneous series, the percentage of falsely detected shifts was computed.

4.5. Series With a Trend

[30] An inhomogeneity can also take the form of a trend. However, it is impossible to compare the methods selected for this work to identify trends because only MREG and REG2 are developed to detect this type of inhomogeneity. The number of shifts that are positioned inside the trend (two positions before the beginning and two after the end of the trend) were rather computed. This aims to show that gradual inhomogeneities can be interpreted as one or several consecutive shifts by most techniques.

5. Results

5.1. Homogeneous Series

[31] The percentages of falsely detected shifts were approximately between 1% (MREG, JARU) and 5% (STUS) for the classical techniques while BAYE gave a higher percentage (more than 20%) of false detections (Tables 1 and 2). The high number of wrongly detected shifts can be explained by the fact that the Bayesian model of *Rasmussen* [2001] makes the implicit hypothesis that there is necessarily a shift in the series. Indeed, the prior probability of no change is automatically fixed to $1/n$, which is very negligible. Therefore this method practically assumes that there is always a change in the series, and forces to position it somewhere since the sum of probabilities of all possible positions is constrained to be 1. In the case of homogeneous series, the probabilities for a change are concentrated toward the extremities of the series. These probabilities would be near 100% if the extremities were not ignored. However, in spite of the decision to ignore shifts close to the two extremities, the percentage of false detections remained very high. Most of the percentages of falsely detected shifts presented in Tables 1 and 2 are significantly different from the type I error of 5% that was used to apply the tests. If the false detection at the extremities were not removed, then the percentage of falsely detected shifts would be around 5%. For STUS, the extremities have less impact on the detection rate since a moving window is used.

[32] Figure 3 presents the magnitudes and positions of wrongly detected shifts on 100-year-long homogeneous series. These magnitudes are presented according to the position of the real shift. Figures corresponding to the 60-year-long series are not presented as the results were very similar to those corresponding to the 100-year series. It can be noticed that for all methods, the magnitudes of falsely detected shifts rarely exceeded one standard deviation.

[33] In a similar study dealing with the homogenization of temperature series [*Ducré-Robitaille et al.*, 2003], REG2 and WILS displayed high false detection rates in opposition to the results presented in this paper. For REG2, this can be explained by the fact that the Fisher revised statistic [*Lund and Reeves*, 2002] was used in the present work. This

Table 2. Falsely Detected Shifts (%) by Each Technique When Applied to 100-Year-Long Homogeneous Series[a]

Magnitude (Standard Deviation)	SNHT	MREG	REG2	BIVT	STUS	WILS	JARU	BAYE
0–0.25	0.3	0.2	2.1	0.2	2.3	0.8	0.1	1.7
0.25–0.5	2.0	0.7	1.5	2.0	1.8	1.7	0.5	18.3
0.5–1	1.7	0.3	0.8	1.7	0.8	2.2	0.6	7.4
1–2	0.0	0.0	0.0	0.0	0.1	0.1	0.0	0.1
>2	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
Total	4.0[b]	1.2[b]	4.4	3.9[b]	5.0	4.8	1.2[b]	27.5[b]

[a]Shifts are in percent.
[b]Significantly different from the expected percentage of type I error (5% critical level).

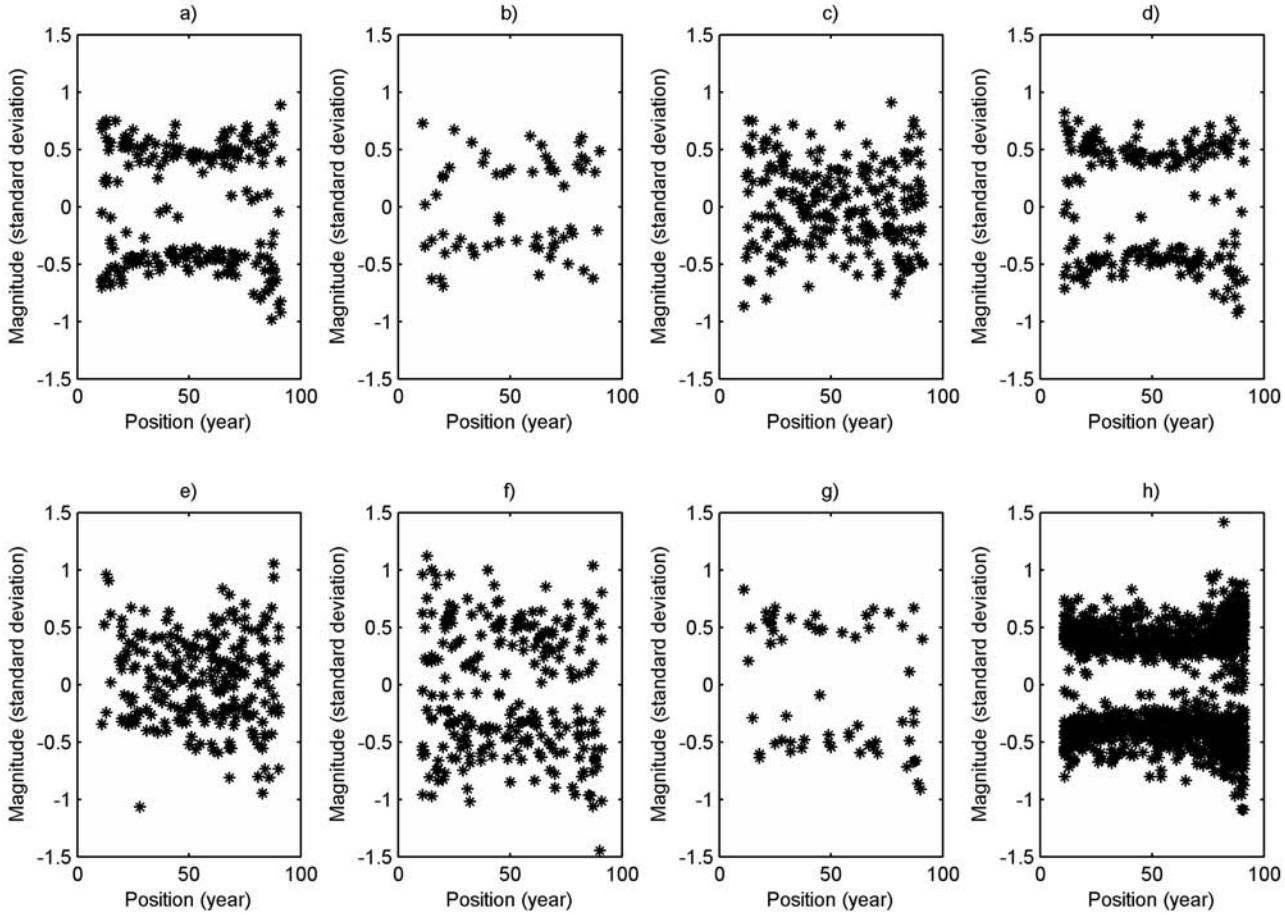

Figure 3. Position and magnitude of falsely detected shifts by each technique when applied to 100-year-long homogeneous series: (a) SNHT, (b) MREG, (c) REG2, (d) BIVT, (e) STUS, (f) WILS, (g) JARU, and (h) BAYE.

revised statistic has higher critical values than the original one [*Solow*, 1987; *Easterling and Peterson*, 1995], which was used in the study of *Ducré-Robitaille et al.* [2003], and hence gives more conservative results and decreases the percentage of false detection. For WILS, this result can be explained by the fact that we reduced the critical level of the Wilcoxon test to have a global critical level around 5% in this study. The critical levels used for WILS and STUS were obtained by simulation and depend on the length of the series.

5.2. Series With a Single Shift

[34] Tables 3 and 4 present the total percentages of correctly identified, well-identified, and well-positioned shifts for all methods. It can be seen in Tables 3 and 4 that most techniques were able to position the shift. Indeed, the percentage of well-positioned shifts was of the order of 75% for most methods except for REG2 and STUS. Thus the homogenization methods were efficient to approximately identify the position of a shift. However, the correct estimation of magnitudes appeared to be more problematic. Indeed, the difference of percentage between the correctly identified and well-identified shifts varies between 13% and 39%. To investigate the magnitudes and\or problematic positions, the percentages of correctly and well-identified shifts were analyzed for various classes of magnitude and position. Figures 4 and 5 present the results of this analysis for the 100-year-long series. Results for the 60-year-long series are very similar to those of the 100-year-long series and are consequently not presented.

[35] All these techniques identified well the shifts with magnitudes greater than two standard deviations. For shifts with magnitudes less than a standard deviation, the percentage of well-identified shifts decreased very quickly. Furthermore, the performance of STUS and REG2 dropped down for shifts with a magnitude of 1.5 standard deviations and less. This was probably due to the narrow moving window, which degraded the performance of STUS. Similar results were reported by *Ducré-Robitaille et al.* [2003] on temperature series. On the other hand, there was no position that seemed to affect the performance of the techniques. In the work of *Ducré-Robitaille et al.* [2003], shifts located at position 5 were less easily identified than shifts located in the middle of the series. By introducing shifts starting at position 10 instead of 5, the effect of the position on the percentage of identified shifts was attenuated.

[36] The errors and absolute errors in the position and magnitude were also analyzed separately. Tables 5 and 6 present the descriptive statistics of the absolute errors in position and magnitude for each technique. The Kruskal-

Table 3. Correctly Identified, Well-Identified, and Well-Positioned Shifts by Each Technique When Applied to 60-Year-Long Series With a Single Shift[a]

	SNHT	MREG	REG2	BIVT	STUS	WILS	JARU	BAYE
Correctly identified	60.0	56.0	51.9	61.6	32.1	57.3	61.7	58.6
Well identified	81.1	71.8	64.7	81.7	70.8	80.5	81.0	81.9
Well positioned	81.6	72.0	64.8	82.2	71.2	80.9	81.4	83.3

[a]Defined in section 4.2. Shifts are in percent.

Wallis test [*Lehman and D'Abrera*, 1998] with a 5% critical level was used to verify the significance of the differences in absolute errors for the various techniques. Since the differences were significant, the Conover-Inman [*Conover*, 1999] procedure was used to make multiple pairwise comparisons between the absolute errors obtained from the various techniques (Tables 7 and 8).

[37] The absolute errors in position and magnitude are significantly different from one technique to another. Furthermore, for 60-year-long series, the absolute errors in the position and magnitude obtained with BAYE are significantly the smallest (Tables 5 and 7). BAYE is followed by the methods BIVT, JARU and SNHT which are not significantly different. The techniques WILS, REGM, REG2 and STUS follow and lead to absolute errors in the position and magnitude that are significantly different from all other techniques.

[38] For 100-year-long series, the absolute errors in the position and magnitude obtained with BAYE and BIVT are the smallest and are not significantly different from each other (Tables 6 and 8). They are followed by JARU, SNHT, WILS, REGM, REG2 and STUS. Figure 6 presents the magnitude and position errors on 100-year-long series with one shift obtained from each technique. The errors rather than absolute errors are presented as they are easier to visualize. The shifts that were not detected are not presented in Figure 6. For all techniques, the errors are concentrated around the origin. For most techniques, the magnitude errors lie between -1 and 1 standard deviation with few cases outside this interval. The position errors interval ranges as far as -80 and 80 for most techniques. The technique REG2 has the errors that are the most scattered.

5.3. Series With Multiple Shifts

[39] Figures 7a and 7b present a histogram of the performance criterion (equation (35)) obtained on both sets of synthetic series. For every class of the criterion, eight bands representing the various techniques are presented. The criterion has some preferential values since the classes of the criterion represent different cases and the peaks mean that some cases occur more often. The most successful methods are those with the lowest values of the performance criterion. Tables 9 and 10 present the descriptive statistics of C (equation (35)). Table 9 indicates that the criterion median for five methods is very low. This means that in half of the synthetic series with two shifts, these methods well positioned all the shifts without detecting nonexistent shifts or omitting real ones. For the series with three shifts, the median criterion was higher. Indeed, when the number of real shifts increases, it becomes more difficult to identify all of them. The maximum criterion was 9801 and corresponds to the case that all the real shifts were not detected. An analysis of variance of Kruskal-Wallis was realized with a critical level of 5% to compare the criteria obtained with the various methods. Since the criteria obtained with the different techniques are significantly different, the Conover-Inman procedure was applied to make multiple pairwise comparisons. Tables 11 and 12 present the results of the multiple pairwise comparisons. For the series with two shifts, the smaller criteria were obtained with BIVT (Table 9) and for the series with three shifts, BAYE gave the smaller criteria (Table 10).

[40] The performance of the various methods according to the distance separating both shifts was then investigated. For every method, the mean criterion was computed according to various classes of distance between two shifts (Figure 8). It was noticed that the distance between the shifts seems to have an influence on the performance criterion. It seems that close or far shifts were less easily identified than two shifts of a mean distance. When the two shifts are too close, the segment between the two shifts is short. Then, it can be more difficult to detect a change with few observations. When the two shifts are far away, it means that they are located in the extremities of the series and it is more problematic to detect a step in this case.

[41] Notably, for STUS, the distance did not have any effect because of the moving window. The technique REG2 seemed to perform better when positioning two far shifts than two close ones. We were also interested to verify if the signs of the shifts have an effect on the performance of the techniques: Is it easier to identify shifts of opposite signs (a positive shift followed by a negative one) or shifts with the same sign? The performance of the methods regarding the signs of the shifts was analyzed. For every method, the mean criterion was calculated according to all possible combinations of signs (Figure 9). It was effectively noticed that the combination of the signs seems to influence the performance criterion. Indeed, in both cases of three shifts of the same sign, the average criterion was higher than in the other cases. When a shift was followed by another one of an opposite sign, the performance seems better except for BAYE. Finally, the cases where two shifts of the same sign were followed or preceded by a shift of an opposite sign led to a better performance than three shifts of the same sign.

Table 4. Correctly Identified, Well-Identified, and Well-Positioned Shifts by Each Technique When Applied to 100-Year-Long Series With a Single Shift[a]

	SNHT	MREG	REG2	BIVT	STUS	WILS	JARU	BAYE
Correctly identified	61.5	59.3	57.4	62.7	30.3	57.9	63.0	57.8
Well identified	83.5	76.8	73.5	83.9	67.4	82.4	83.4	83.2
Well positioned	83.9	76.9	73.6	84.3	67.5	82.7	83.6	85.2

[a]Defined in section 4.2. Shifts are in percent.

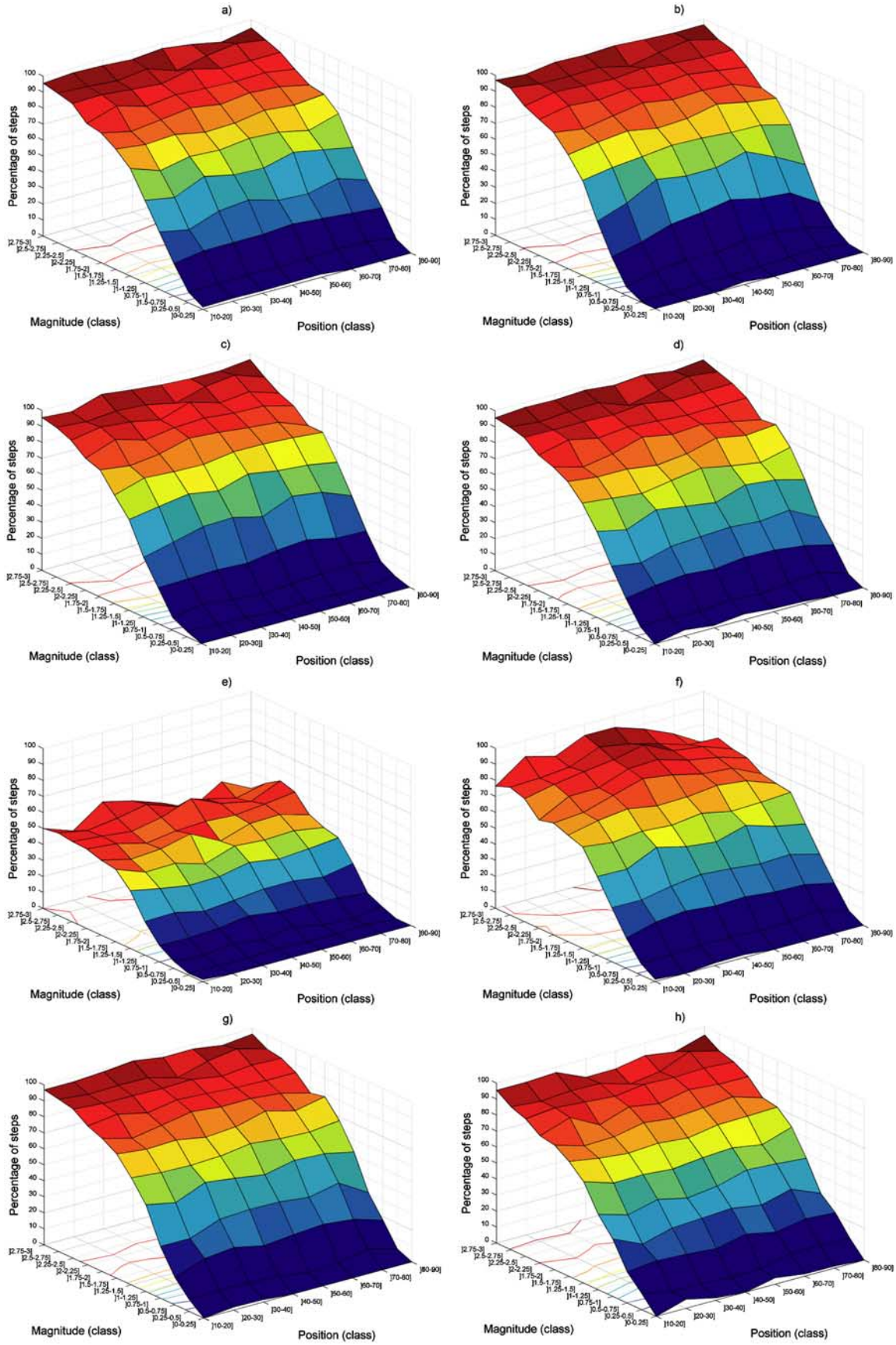

Figure 4. Percentage of correctly identified shifts, as defined in section 4.2, according to their position and magnitude obtained by each technique when applied to 100-year-long series with a single shift: (a) SNHT, (b) MREG, (c) REG2, (d) BIVT, (e) STUS, (f) WILS, (g) JARU, and (h) BAYE.

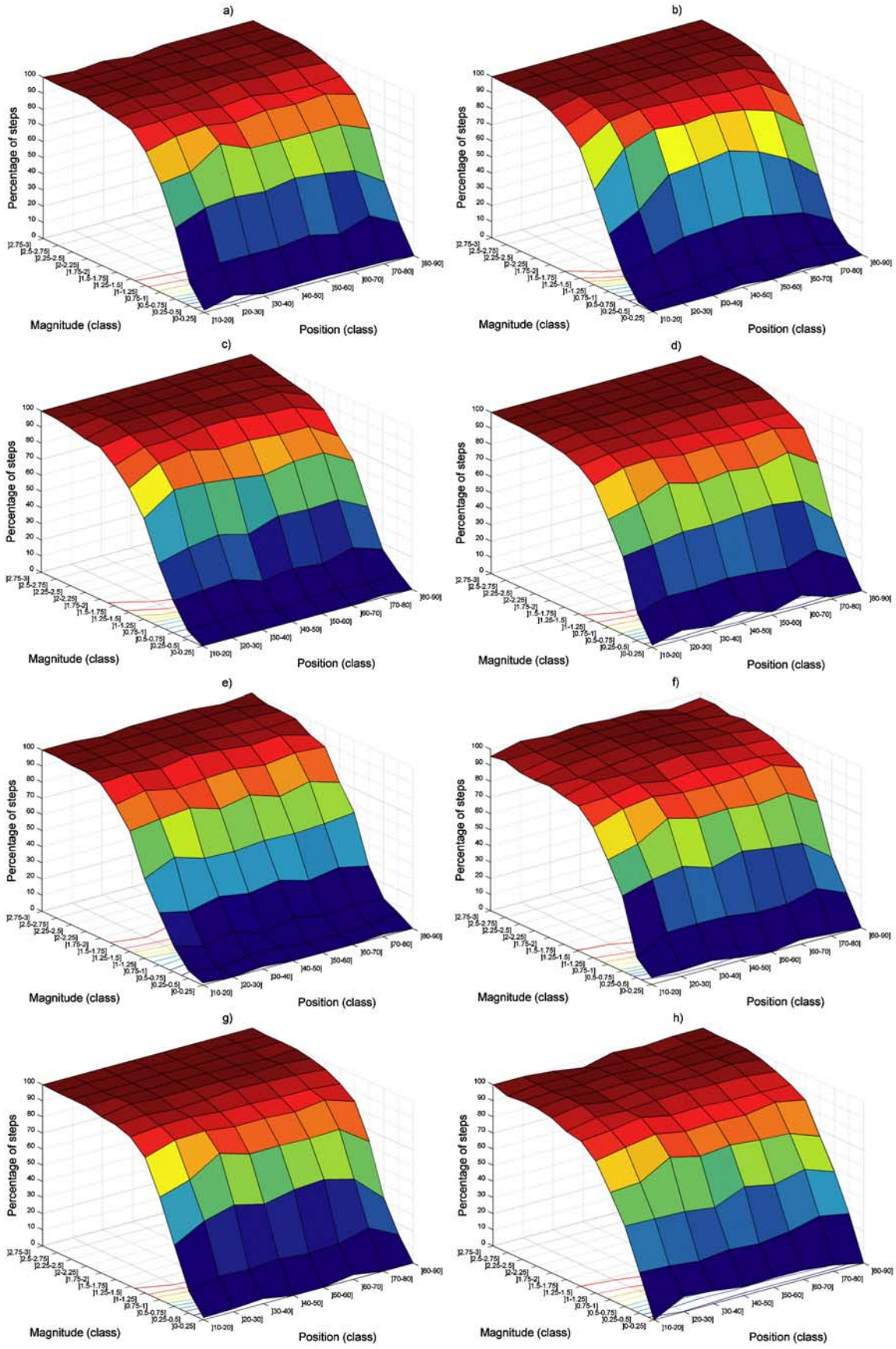

Figure 5. Percentage of well-identified shifts, as defined in section 4.2, according to their position and magnitude obtained by each technique when applied to 100-year-long series with a single shift: (a) SNHT, (b) MREG, (c) REG2, (d) BIVT, (e) STUS, (f) WILS, (g) JARU, and (h) BAYE.

Table 5. Descriptive Statistics of the Absolute Errors in Position and Magnitude for Each Technique When Applied to 60-Year-Long Series With a Single Shift[a]

Statistic	SNHT	REGM	REG2	BIVT	STUS	WILS	JARU	BAYE
Absolute Error in Position								
Mean	6.9	14.8	17.7	6.8	14.8	6.4	7.7	5.0
Median	0	0	0	0	1	0	0	0
Standard Deviation	18.0	25.4	26.5	18.0	25.0	17.2	19.2	15.0
Mean Rank[b]	93840	103110	109340	92380	120780	95960	93180	91140
Absolute Error in Magnitude								
Mean	0.3	0.7	0.9	0.3	0.8	0.3	0.4	0.2
Median	0	0	0	0	0.1	0	0	0
Standard Deviation	0.9	1.3	1.3	0.9	1.2	0.9	1.0	0.7
Mean Rank[b]	93120	102970	109280	91710	124890	95190	92600	90250

[a]Defined as in section 4.2.
[b]The absolute errors in position and magnitude differ significantly according to the techniques used (Kruskal-Wallis test, 5% critical level).

Table 6. Descriptive Statistics of the Absolute Errors in Position and Magnitude for Each Technique When Applied to 100-Year-Long Series With a Single Shift[a]

Statistic	SNHT	REGM	REG2	BIVT	STUS	WILS	JARU	BAYE
Absolute Error in Position								
Mean	7.7	18.3	18.7	7.5	28.9	7.5	9.2	5.2
Median	0	0	0	0	1	0	0	0
Standard Deviation	25.0	38.0	37.4	24.7	44.5	24.4	27.6	19.6
Mean Rank[b]	93750	100370	102900	92720	127630	97110	93340	92180
Absolute Error in Magnitude								
Mean	0.2	0.6	0.6	0.2	0.9	0.2	0.3	0.1
Median	0	0	0	0	0.1	0	0	0
Standard Deviation	0.7	1.1	1.1	0.7	1.3	0.7	0.8	0.6
Mean Rank[b]	92890	100170	102860	91840	132130	96430	92720	90980

[a]Defined as in section 4.2.
[b]The absolute errors in position and magnitude differ significantly according to the techniques used (Kruskal-Wallis test, 5% critical level).

Table 7. Pairwise Comparison of the Absolute Errors in Position and Magnitude for Each Technique When Applied to 60-Year-Long Series With a Single Shift[a]

Technique	SNHT	REGM	REG2	BIVT	STUS	WILS	JARU	BAYE
Absolute Error in Position								
SNHT	0	1	1	1	1	1	0	1
REGM		0	1	1	1	1	1	1
REG2			0	1	1	1	1	1
BIVT				0	1	1	0	1
STUS					0	1	1	1
WILS						0	1	1
JARU							0	1
BAYE								0
Absolute Error in Magnitude								
SNHT	0	1	1	1	1	1	0	1
REGM		0	1	1	1	1	1	1
REG2			0	1	1	1	1	1
BIVT				0	1	1	0	1
STUS					0	1	1	1
WILS						0	1	1
JARU							0	1
BAYE								0

[a]Defined as in section 4.2. Here 1, significantly different; 0, not significantly different (Conover-Inman test, 5% critical level).

Table 8. Pairwise Comparison of the Absolute Errors in Position and Magnitude for Each Technique When Applied to 100-Year-Long Series With a Single Shift[a]

Technique	SNHT	REGM	REG2	BIVT	STUS	WILS	JARU	BAYE
				Absolute Error in Position				
SNHT	0	1	1	1	1	1	0	1
REGM		0	1	1	1	1	1	1
REG2			0	1	1	1	1	1
BIVT				0	1	1	0	0
STUS					0	1	1	1
WILS						0	1	1
JARU							0	1
BAYE								0
				Absolute Error in Magnitude				
SNHT	0	1	1	1	1	1	0	1
REGM		0	1	1	1	1	1	1
REG2			0	1	1	1	1	1
BIVT				0	1	1	0	0
STUS					0	1	1	1
WILS						0	1	1
JARU							0	1
BAYE								0

[a]Defined as in section 4.2. Here 1, significantly different; 0, not significantly different (Conover-Inman test, 5% critical level).

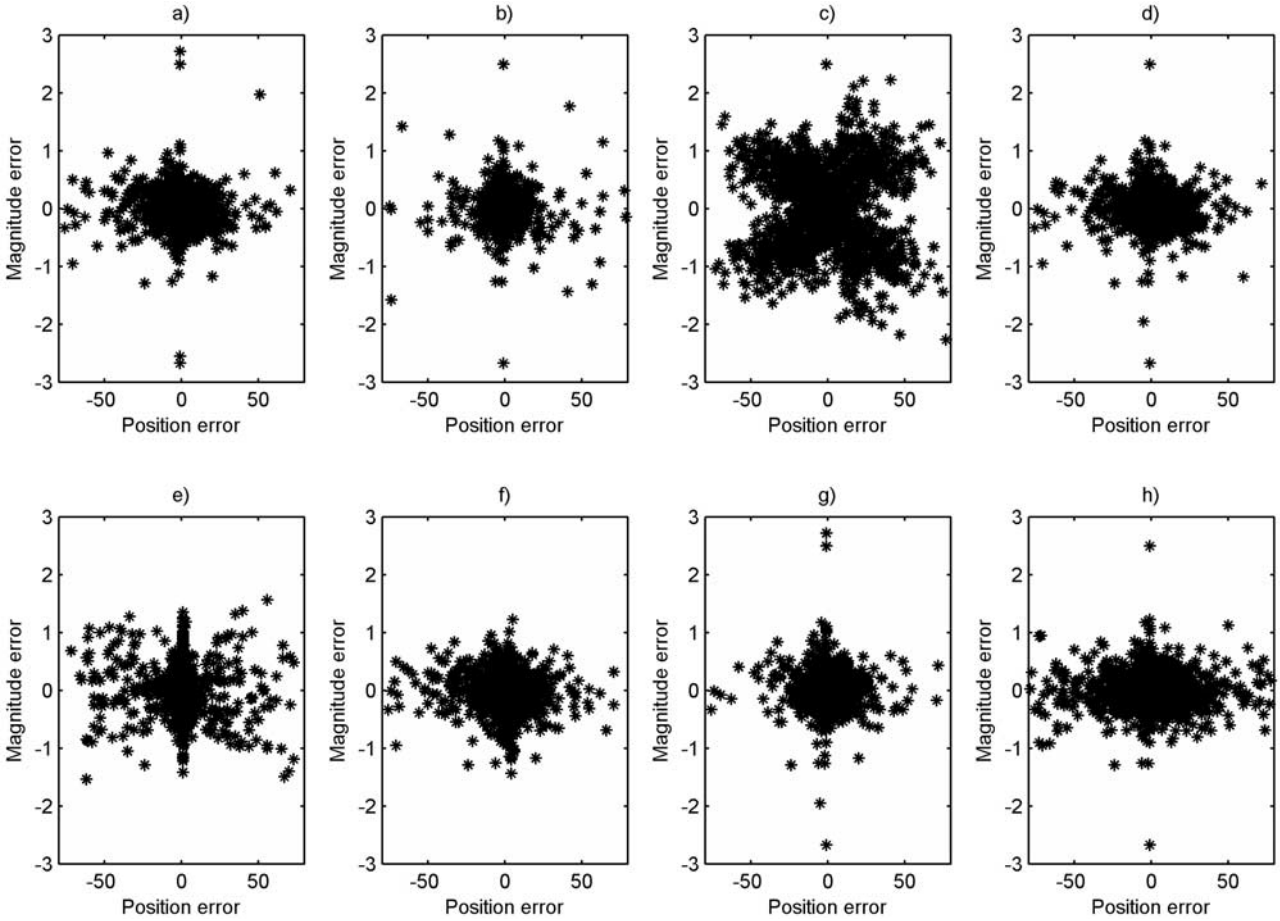

Figure 6. Magnitude and position errors, calculated as described in section 4.2, obtained by each technique when applied to 100-year-long series with a single shift: (a) SNHT, (b) MREG, (c) REG2, (d) BIVT, (e) STUS, (f) WILS, (g) JARU, and (h) BAYE. Shifts that were not detected by the techniques are not represented in this figure.

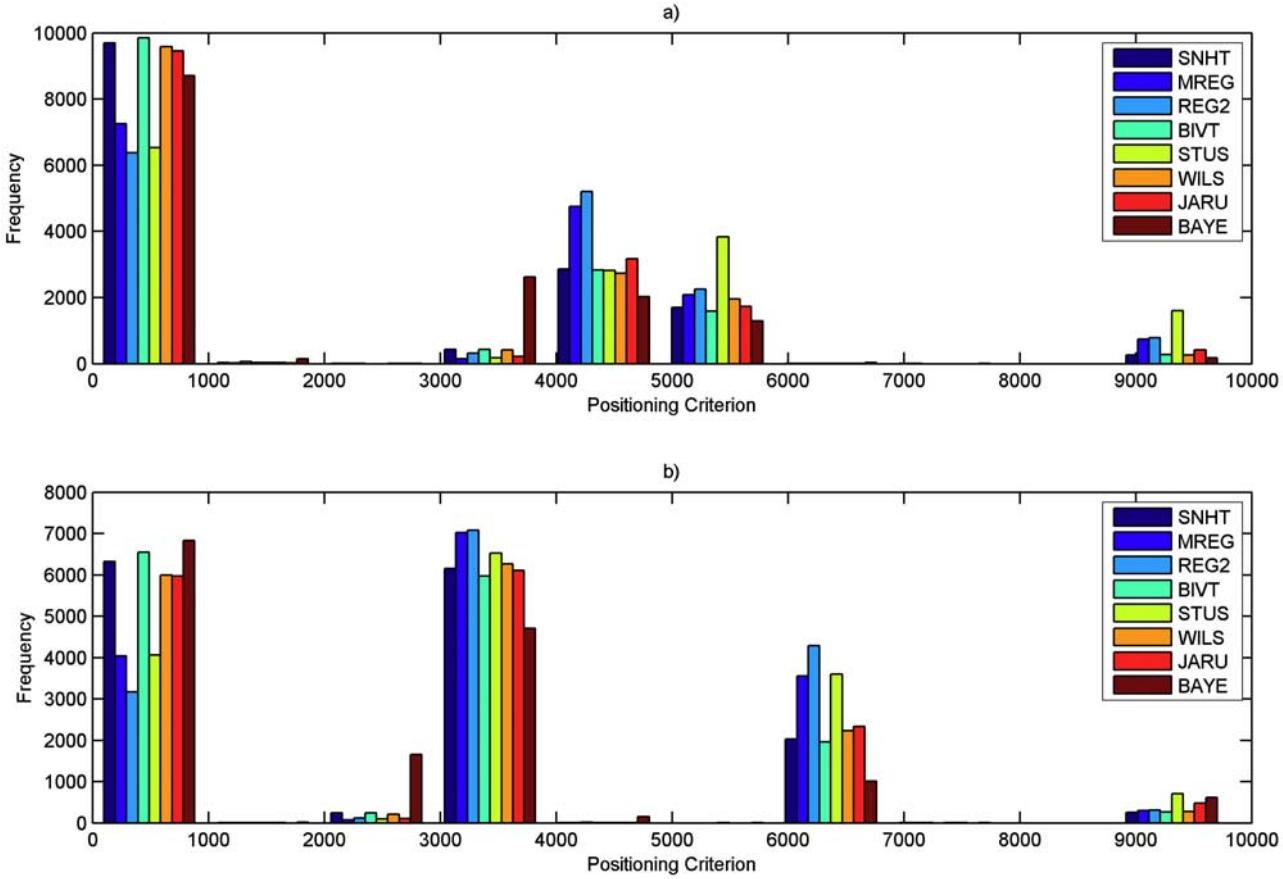

Figure 7. Histogram of the positioning criterion C (equation (35), section 4.3) obtained by all techniques when applied to series with multiple shifts: (a) two shifts and (b) three shifts.

5.4. Series With a Shift of Variance

[42] Table 13 presents the wrongly detected shifts on the synthetic series with a change in variance. It is considered that a method that is robust to a change of variance should give approximately the same rate of false detection in the presence of a shift of standard deviation. The rates of false detection were compared with those obtained from the homogeneous series by a confidence interval on the difference between the two proportions:

$$\hat{p}_H - \hat{p}_I - Z_{\alpha/2}\sqrt{\frac{\hat{p}_H(1-\hat{p}_H)}{n_H} + \frac{\hat{p}_I(1-\hat{p}_I)}{n_I}} \le p_H - p_I \le \hat{p}_H - \hat{p}_I + Z_{\alpha/2}\sqrt{\frac{\hat{p}_H(1-\hat{p}_H)}{n_H} + \frac{\hat{p}_I(1-\hat{p}_I)}{n_I}} \tag{36}$$

where $\hat{p}_H$ and $\hat{p}_I$ are the rates of false detection in the homogeneous series and in the series with a change in variance, n_H and n_I are the number of homogeneous series and inhomogeneous series respectively and $Z_{\alpha/2}$ is the standard normal value corresponding to a 5% critical level. When the interval contains the value zero, there is no significant difference between the two proportions. The results are presented in Table 14. Most of the differences of proportions were significant at a 5% critical level except for STUS, WILS and BAYE. This was expected for WILS because it is a nonparametric technique. On the other hand, STUS is based on the assumption of the equality of variances, and hence the results for this method were unexpected. Once again, this can be a result of the moving window given that the change of standard deviation has less impact on a small part of the series. Finally, it seems that a change in variance in the base series increases the probability of the type 1 error for most of the methods, but this increase is relatively minor. In summary, the shifts of variance seem to raise slightly the percentage of false detection.

5.5. Series With a Trend

[43] Tables 15 and 16 present respectively the number of cases for which one shift and two shifts or more were detected inside the trend. It was noticed that in most cases, the trend was interpreted as an abrupt shift of mean (Table 15). This occurs less often when the trend has a weak magnitude (0–0.5 standard deviation). It also happened that the trend was interpreted as several consecutive shifts when the magnitude is high (Table 16). To avoid this kind of mistake, a graphical method combined with an objective method could be used. Graphically, it is easier to identify the type of change (abrupt or gradual).

6. Discussion

[44] When the metadata is incomplete, there is little information about the presence, number, type, position and magnitude of potential inhomogeneities in the base series.

Table 9. Descriptive Statistics of the Positioning Criterion C for Each Technique When Applied to Series With Two Shifts[a]

Statistic	SNHT	MREG	REG2	BIVT	STUS	WILS	JARU	BAYE
Mean	1768	2757	3044	1730	3268	1812	1928	1813
Median	2	4901	4901	2	4901	3	2	18
Standard Deviation	2483	2864	2831	2489	3219	2503	2643	2264
Minimum	0	0	0	0	0	0	0	0
Maximum	9801	9801	9801	9801	9801	9801	9801	9801
Mean rank	54607	62960	66917	53537	74179	56446	54808	56550

[a]The mean positioning criterion C differs significantly according to the techniques used (Kruskal-Wallis test, 5% critical level). See also equation (35), section 4.3.

Table 10. Descriptive Statistics of the Positioning Criterion C for Each Technique When Applied to Series With Three Shifts[a]

Statistic	SNHT	MREG	REG2	BIVT	STUS	WILS	JARU	BAYE
Mean	2439	3291	3654	2378	3476	2565	2683	2216
Median	3267	3267	3267	3267	3267	3267	3267	2451
Standard Deviation	2443	2500	2451	2451	2723	2479	2633	2488
Minimum	0	0	0	0	0	0	0	0
Maximum	9801	9801	9801	9801	9801	9801	9801	9801
Mean rank	54941	64350	69104	53625	71305	57850	56548	52280

[a]The mean positioning criterion C differs significantly according to the techniques used (Kruskal-Wallis test, 5% critical level). See also equation (35), section 4.3.

Table 11. Pairwise Comparison of the Positioning Criterion C for Each Technique When Applied to Series With Two Shifts[a]

Technique	SNHT	REGM	REG2	BIVT	STUS	WILS	JARU	BAYE
SNHT	0	1	1	1	1	1	0	1
REGM		0	1	1	1	1	1	1
REG2			0	1	1	1	1	1
BIVT				0	1	1	1	1
STUS					0	1	1	1
WILS						0	1	0
JARU							0	1
BAYE								0

[a]Here 1, significantly different; 0, not significantly different (Conover-Inman test, 5% critical level). See also equation (35), section 4.3.

Table 12. Pairwise Comparison of the Positioning Criterion C for Each Technique When Applied to Series With Three Shifts[a]

Technique	SNHT	REGM	REG2	BIVT	STUS	WILS	JARU	BAYE
SNHT	0	1	1	1	1	1	1	1
REGM		0	1	1	1	1	1	1
REG2			0	1	1	1	1	1
BIVT				0	1	1	1	1
STUS					0	1	1	1
WILS						0	1	1
JARU							0	1
BAYE								0

[a]Here 1, significantly different; 0, not significantly different (Conover-Inman test, 5% critical level). See also equation (35), section 4.3.

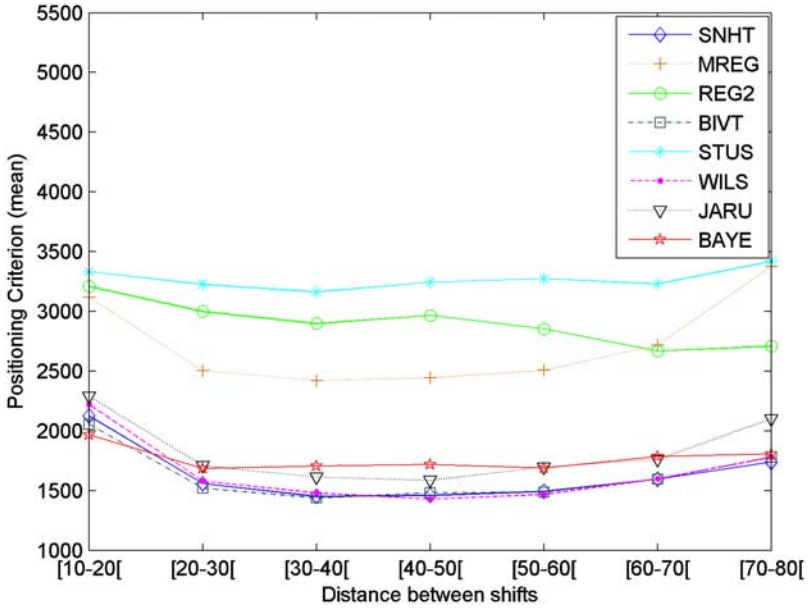

Figure 8. Mean positioning criterion C (equation (35), section 4.3) according to the distance between two shifts obtained by all techniques.

Therefore this work aimed to identify methods able to detect homogeneous and inhomogeneous series. Several types of synthetic series (homogeneous, one shift, multiple shifts, a trend, a shift of variance) representing the typical annual total precipitation of the southern and central regions of the province of Quebec (Canada) and nearby areas were generated.

[45] The performance of the studied methods can be summarized according to some criteria. First, it is very important to have a small percentage of false detection (5% and less), to avoid introducing inhomogeneities in a series that is homogeneous in reality. Also, a homogenization method should be able to identify a shift in a series. The technique should also be able to position a shift in at least 75% of the cases. Shifts of high magnitudes (2 standard deviations and more), are expected to be identified in nearly 100% of the cases. It is also expected to observe a reasonable percentage of well-identified shifts of 1 standard deviation and more. For multiple shifts, the method should also be able to position several shifts without omission or false detection. Finally, the methods that can be applied in real conditions without additional modifications are advantageous because they allow a gain of time.

[46] The application on series with a shift of variance showed that most of the methods being compared in this

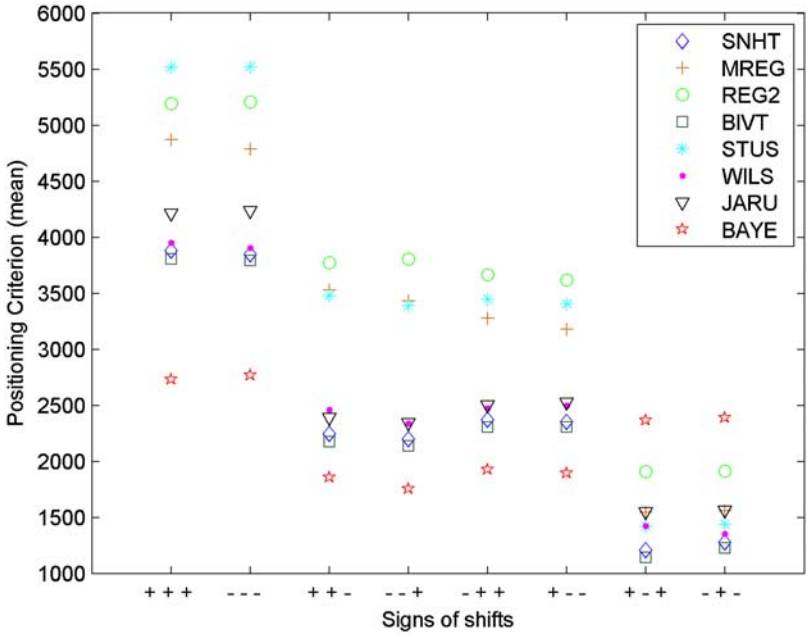

Figure 9. Mean positioning criterion C (equation (35), section 4.3) according to the order of signs of the three shifts obtained by all techniques.

Table 13. Falsely Detected Shifts by Each Technique When Applied to Series With a Change in Variance

Technique	Rejected, %
SNHT	5.4
MREG	2.9[a]
REG2	7.7[a]
BIVT	5.4
STUS	4.7
WILS	5.1
JARU	2.4[a]
BAYE	27.7[a]

[a]Significantly different from the expected percentage of type I error (5% critical level).

work are not very sensitive to this type of inhomogeneity. The results of the application on synthetic series containing a trend showed that a gradual change is often interpreted as one or several consecutive shifts inside the trend. This is not a weakness because the compared methods were not designed for this purpose.

[47] Since our results indicate that none of the methods performed well in all cases, the design of an optimal procedure using the strengths of some of the techniques was considered. This procedure consists of the sequential application of some selected techniques. If we judge that it is better to omit inhomogeneities in a series than to introduce new artificial ones, JARU should be applied first because of its good capacity to identify homogeneous series (Tables 1 and 2). Although MREG gave an equivalent performance, JARU is preferred because on series with a single shift, JARU was less conservative than MREG (the absolute errors in position and magnitude obtained with JARU are significantly smaller than those obtained with MREG, section 5.2). If the series is found inhomogeneous with JARU, BIVT can be applied. In the case of series with a single shift, BIVT had the second best performance for the absolute errors in position and magnitude (Tables 5 and 6). On series with two shifts, BIVT had the best performance (Table 9) and for series with three shifts, BAYE was better (Table 10). Given that BAYE also detected a high number of nonexistent shifts in the homogeneous series, BIVT was preferred. Indeed, it has a weak percentage of false detection (Tables 1 and 2). With such a procedure, the probability of identifying all existing shifts is increased, and the risk of false detection is reduced. It is important to mention that large shifts (greater than one standard deviation), have a very low false detection rate with all methods. It is nevertheless recommended to confirm the results with the metadata. It should be stressed that this proposed procedure is rather conservative and is based on the idea that it is better to omit existing shifts than to correct a series for shifts that are not real.

Table 14. Difference of Proportions Between Falsely Detected Shifts in 100-Year-Long Homogeneous Series and Series With a Change in Variance With the Associated 95% Confidence Interval[a]

Technique	Lower Bound	Difference	Upper Bound
SNHT[b]	0.7	1.4	2.1
MREG[b]	1.2	1.7	2.1
REG2[b]	2.5	3.3	4.1
BIVT[b]	0.8	1.5	2.2
STUS	−1	−0.3	0.4
WILS	−0.5	0.3	0.9
JARU[b]	0.8	1.2	1.6
BAYE	−1.3	0.2	1.7

[a]Differences are in percent. See also equation (36), section 5.4.
[b]Significant (5% critical level).

Table 15. Number of Cases for Which One Shift Is Detected Inside the Trend[a]

Magnitude (Standard Deviation)	SNHT	MREG	REG2	BIVT	STUS	WILS	JARU	BAYE
0–0.5	39.1	4.3	13.0	30.4	8.7	30.4	30.4	52.2
0.5–1	84.2	54.1	28.5	84.9	13.2	83.5	83.0	79.4
1–2	77.3	89.8	65.7	76.0	47.0	76.8	81.5	63.2
2–3	56.1	80.9	77.0	55.3	64.5	52.6	58.2	52.2

[a]Cases are in percent.

[48] This comparison included only objective techniques because the application of subjective methods (involving the judgment of an expert) on thousands of synthetic series is an impracticable task. Nevertheless, the use of subjective approaches should not be automatically rejected as these approaches may sometimes be appropriate to analyze the data and interpret the results. We also insist on the use of metadata, when available, to validate and identify the cause of the detected inhomogeneities.

[49] A limitation of the presented techniques is that they require the presence of homogeneous neighboring stations, while it may not be always the case in practice. The authors believe that it is important to use neighbor series to avoid a misinterpretation of a regional climate change. Nevertheless, in the cases where neighbor series are not available or inhomogeneous, techniques developed to homogenize isolated stations and to create homogeneous reference series from neighbor series can be used. For a review of these techniques, see *Peterson et al.* [1998].

[50] The compared techniques require data to be normally distributed (except WILS). The synthetic series were generated from a normal distribution. Nevertheless, the introduction of inhomogeneities in the series could have affected the distribution and hence the performance of the techniques.

[51] The techniques selected for this work are based on the analysis of either differences or ratios between the base series and neighbor series. Since some tests use ratios and others use differences, the performance could be affected because of the choice of variable. The objective of this work was to compare the techniques as they are presented in the literature. The sensitivity of the techniques to the use of ratios or differences was not studied in this work. Future work can focus on the study of the sensitivity of the various techniques to the choice of variables.

[52] Finally, the various homogenization methods were applied under the specific conditions of the province of Quebec, Canada, and the results of this study may only be valid under these conditions. The same techniques could lead to different performances on series with a different distribution, a different autocorrelation structure and/or different correlation with neighbor series.

7. Conclusions

[53] Homogeneous precipitation series are essential, particularly when data are used in climate models or to assess

Table 16. Number of Cases for Which at Least Two Shifts Are Detected Inside the Trend[a]

Magnitude (Standard Deviation)	SNHT	MREG	REG2	BIVT	STUS	WILS	JARU	BAYE
0–0.5	0.0	0.0	0.0	4.3	0.0	0.0	0.0	4.3
0.5–1	2.6	1.2	0.0	2.6	0.0	4.3	1.4	13.6
1–2	21.9	6.1	1.8	23.4	2.7	22.3	17.7	36.5
2–3	43.9	19.0	6.4	44.7	12.3	47.3	41.7	47.8

[a]Cases are in percent.

climate change and associated environmental and socioeconomics impacts. The performance of eight homogenization techniques on synthetic precipitation series with similar characteristics to typical series observed in southern and central Quebec and surrounding areas in Canada were compared in this work. The results of this study will be of use for future activities dealing with the homogenization of precipitation series in Canada. It was found that techniques which gave a good performance on temperature series like the multiple regression [*Ducré-Robitaille et al.*, 2003], were not necessarily appropriate for precipitation data.

[54] Three methods had similar performances with all sets of synthetic series (BIVT, JARU and SNHT). Some techniques cannot be applied efficiently to all types of series. For instance, MREG performed well for the identification of a homogeneous series and was good to identify a single shift. However, in the presence of multiple shifts, the performance of this method was poor. The technique BAYE performed well for the identification of one or multiple shifts, but detected too many nonexistent shifts. Finally, STUS and REG2 were able to detect homogeneous series, but did not perform well on series with single and multiple shifts. An optimal procedure using the strengths of the various methods was proposed.

[55] The mathematics of the studied methods were first developed to detect a single shift. A sequential application of these techniques can disadvantage them because they are not designed to detect multiple shifts. However, it is common to have several inhomogeneities in hydroclimatic series. The development of methods that are able to identify one or multiple changes of several types (shift or trend) is desirable.

[56] In this work, it was shown that the selected homogenization techniques are robust in presence of a change of variance. Nevertheless, the homogenization methods are based on other assumptions such as the normality and the homogeneity of neighbor series. The violation of these assumptions may also alter the performance of the techniques. There is no existing work which addresses these aspects. Future work in these directions is desirable.

Notation

β_j coefficient of the jth neighbor series in the regression model l.
Γ Gamma function.
δ_{pi} magnitude of the shift at position p_i.
θ vector of parameters.
$\hat{\theta}$ vector of estimated parameters.
λ_l time coefficient in the regression model l.
ρ_j correlation coefficient between the base series and neighbor series j.
σ standard deviation of the last segment of the base series.
τ_l intercept of the regression model l.
υ number of degrees of freedom.
ϕ_1 lag one autocorrelation coefficient.
ψ correlation constant between the base series and the neighbor series.
b Beta variable.
$BETA$ Beta distribution.
C position criterion.
$DUNIF$ Discrete Uniform distribution.
e_i residual at time i.
k number of neighbor series.
m slope of the trend.
n length of the base series.
nd number of shifts detected.
nr number of real shifts.
p_i^d position of the ith detected shift.
p_i position of the ith real shift.
q_i difference/ratios series between the base series and the neighbors at time i.
Q_{pi} test statistic associated with the ith shift.
r_i rank of the ith observation.
R_i sum of the ranks of observations 1 to i.
s^2 variance in a series.
S sum of squares of the differences.
$sign$ sign (positive or negative).
u Uniform variable.
U Uniform distribution.
ud Discrete Uniform variable.
w_i standardized neighbor series at time i.
$x_{1:i,j}$ observations from 1 to i of the neighbor series j.
$y_{1:i}$ observations from 1 to i of the base series.
$\bar{y}_{1:i}$ mean of the segment from 1 to i of variable y.
y_i^* homogeneous base series at time i.
z_i standardized base series at time i.

[57] **Acknowledgments.** The authors wish to thank the National Sciences and Engineering Research Council of Canada (NSERC), the OURANOS Consortium, and the Canada Research Chair Program for funding this research. The authors would also like to thank Lucie Vincent of the Meteorological Service of Canada and Paul Whitfield of Environment Canada for their useful comments. The authors would also like to thank the Associate Editor and four anonymous reviewers for their valuable comments and suggestions.

References

Aguilar, E., I. Auer, M. Brunet, T. C. Peterson, and J. Wieringa (2003), Guidelines on climate metadata and homogenization, *Rep. WMO-TD 1186*, 50 pp., World Meteorol. Organ., Geneva, Switzerland.

Alexandersson, H. (1986), A homogeneity test applied to precipitation data, *J. Climatol.*, *6*, 661–675.

Asselin, J., T. B. M. J. Ouarda, V. Fortin, and B. Bobée (1999), Une procédure Bayésienne bivariée pour détecter un décalage de la moyenne, *Res. Rep. R-528*, 33 pp., Eau, Terre et Environ., Inst. Natl. de la Rech. Sci., Quebec, Que., Canada.

Beaulieu, C., T. B. M. J. Ouarda, and O. Seidou (2007), Synthèse des techniques d'homogénéisation des séries climatiques et analyse d'applicabilité aux séries de précipitations, *Hydrol. Sci. J.*, *52*(1), 18–37.

Conover, W. J. (1999), *Practical Nonparametric Statistics*, 3rd ed., 584 pp., John Wiley, New York.

Ducré-Robitaille, J. F., G. Boulet, and L. A. Vincent (2003), Comparison of techniques for detection of discontinuities in temperature series, *Int. J. Climatol.*, *23*, 1087–2003.

Easterling, D. R., and T. C. Peterson (1992), Techniques for detecting and adjusting for artificial discontinuities in climatological time series: A review, paper presented at the Fifth International Meeting on Statistical Climatology, Am. Meteorol. Soc., Toronto, Ont., Canada, 22–26 June.

Easterling, D. R., and T. C. Peterson (1995), A new method for detecting undocumented discontinuities in climatological time series, *Int. J. Climatol.*, *15*, 369–377.

Gullett, D. W., L. A. Vincent, and P. J. F. Sajecki (1990), Testing for homogeneity in temperature time series at Canadian climate stations, *Can. Clim. Cent. Rep. 90–4*, 43 pp., Atmos. Environ. Serv., Downsview, Ont., Canada.

Heino, R. (1997), Metadata and their role in homogenization, paper presented at First Seminar for Homogenization of Surface Climate Data, Hung. Meteorol. Serv., Budapest.

Jaruskova, D. (1996), Change-point detection in meteorological measurement, *Mon. Weather Rev.*, *124*, 1535–1543.

Karl, T. R., and C. N. Williams Jr. (1987), An approach to adjusting climatological time series for discontinuous inhomogeneities, *J. Clim. Appl. Meteorol.*, *26*, 1744–1763.

Khaliq, M. N., and T. B. M. J. Ouarda (2007), A note on the critical values of the standard normal homogeneity test (SNHT), *Int. J. Climatol.*, *27*, 681–687.

Lanzante, J. R. (1996), Resistant, robust and non-parametric techniques for the analysis of climate data: Theory and examples, including applications to historical radiosonde station data, *Int. J. Climatol.*, *16*, 1197–1226.

Lehman, E. L., and H. J. M. D'Abrera (1998), *Nonparametrics: Statistical Methods Based on Ranks*, revis. 1st ed., 463 pp., Prentice-Hall, Upper Saddle River, N. J.

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

Maronna, R., and V. J. Yohai (1978), A bivariate test for the detection of a systematic change in mean, *J. Am. Stat. Assoc.*, *73*, 640–645.

Ouarda, T. B. M. J., J. Asselin, and O. Seidou (2005), Bayesian multivariate linear regression with application to changepoint models in hydrometeorological variables. Model development, *Res. Rep. R-838*, 39 pp., Eau, Terre et Environ., Inst. Natl. de la Rech. Sci., Quebec, Que., Canada.

Perreault, L., M. Haché, M. Slivitzsky, and B. Bobée (1999), Detection of changes in precipitation and runoff over eastern Canada and U.S. using a Bayesian approach, *Stochastic Environ. Res. Risk Assess.*, *13*, 201–216.

Peterson, T. C., et al. (1998), Homogeneity adjustments of in situ atmospheric climate data: A review, *Int. J. Climatol.*, *18*, 1493–1517.

Potter, K. W. (1981), Illustration of a new test for detecting a shift in mean in precipitation series, *Mon. Weather Rev.*, *109*, 2040–2045.

Rasmussen, P. (2001), Bayesian estimation of change points using the general linear model, *Water Resour. Res.*, *37*, 2723–2731.

Solow, A. R. (1987), Testing for climate change: An application of the two-phase regression model, *J. Clim. Appl. Meteorol.*, *26*, 1401–1405.

Vincent, L. A. (1998), A technique for the identification of inhomogeneities in Canadian temperature series, *J. Clim.*, *11*, 1094–1105.

Wang, X. L. (2003), Comments on "Detection of undocumented changepoints: A revision of the two-phase regression model," *J. Clim.*, *16*, 3383–3385.

C. Beaulieu, T. B. M. J. Ouarda, and O. Seidou, Eau, Terre et Environnement, Institut National de la Recherche Scientifique, Université du Québec, 490, de la Couronne, Québec, QC, Canada G1K 9A9. (claudie_beaulieu@ete.inrs.ca)

G. Boulet and A. Yagouti, Direction du suivi de l'état de l'environnement, Ministère de l'Environnement du Québec, Québec, QC, Canada G1R 5V7.

X. Zhang, Climate Research Division, Science and Technology Branch, Environment Canada, Toronto, ON, Canada M3H 5T4.