New techniques for detection and adjustment of shifts in daily precipitation data series

Xiaolan L. Wang[1,2], Hanfeng Chen[3], Yuehua Wu[2], Yang Feng[1] and Qiang Pu[2]

1. Climate Research Division, ASTD, STB, Environment Canada, Toronto, Ontario, Canada

2. Department of Mathematics and Statistics, York University, Toronto, Ontario, Canada

3. Department of Mathematics and Statistics, Bowling Green State University, Bowling Green, Ohio, USA

Received ______________________; accepted ______________________

Short title:

J. APPL. METEOR. CLIMATOL.

(ACCEPTED SUBJECT TO REVISIONS ON 10 JANUARY 2010)

(REVISED 10 MARCH 2010)

Abstract.

This study integrates a Box-Cox power transformation procedure into a common trend two-phase regression model based test (the PMFred algorithm) for detecting changepoints, to make the test applicable to non-Gaussian data series, such as non-zero daily precipitation amounts or wind speeds. The detection power aspects of the transformed method (transPMFred) are assessed by a simulation study, which shows that this new algorithm is much better than the corresponding untransformed method for non-Gaussian data; the transformation procedure can increase the hit rate by up to about 70%. Examples of application of this new transPMFred algorithm to detect shifts in real daily precipitation series are provided using non-zero daily precipitation series recorded at a few stations across Canada that represent very different precipitation regimes. The detected changepoints are in good agreement with documented times of changes in for all the example series.

This study clarifies that it is essential for homogenization of daily precipitation data series to test the non-zero precipitation amount series and the frequency series of precipitation occurrence (or non-occurrence), separately. The new transPMFred can be used to test the series of non-zero daily precipitation (which are non-Gaussian and positive), while the existing PMFred algorithm can be used to test the frequency series. A software package for using the transPMFred algorithm to detect shifts in non-zero daily precipitation amounts has been developed and made available online free of charge, along with a quantile matching (QM) algorithm for adjusting shifts in non-zero daily precipitation series, which is applicable to all positive data. In addition, a similar QM algorithm has also been developed for adjusting Gaussian data such as temperatures.

It is noticed that frequency discontinuities are often inevitable due to changes in the measuring precision of precipitation, and that they could complicate the detection of shifts in non-zero daily precipitation data series and void any attempt to homogenize the

series. In this case, one must account for all frequency discontinuities before attempting to adjust the measured amounts. This study also proposes approaches to account for detected frequency discontinuities, for example, to fill in the missed measurements of small precipitation or the missed reports of trace precipitation. It stresses the importance of testing the homogeneity of the frequency series of reported zero precipitation and of various small precipitation events, along with testing the series of daily precipitation amounts that are larger than a small threshold value, varying the threshold over a set of small values that reflect changes in measuring precision over time.

1. Introduction

Climate data series usually contain artificial shifts due to inevitable changes in observing instrument (or observer), location, environment, and observing practises/procedures taking place in the period of data collection. Data discontinuities also arise from the continuously evolving technology of climate monitoring. It is important to detect artificial changepoints in climate data series, because these artificial changes could considerably bias the results of climate trends, variability and extremes analysis.

Many changepoint detection methods have been developed (Wang et al. 2007, Wang 2008a,b and 2003, Vincent 1998, Alexandersson 1986, among many others). However, most of the commonly used methods (including those listed above) assume that the data are normally distributed; and those that do not need the normality assumption, such as nonparametric methods (see Reeves et al. 2007) or empirical likelihood based methods, are at most comparable with a method based on normality assumption if the data can be transformed to well approximate a normal distribution. Unfortunately, the normality assumption is often invalid for daily precipitation data, which is one of the most important climate variables. Apart from its departure from normality, daily precipitation is not a continuous variable. Therefore, the most common approach to modelling daily precipitation has been to use models describing the occurrence (non-occurrence) process and to describe the distribution of the non-zero amounts independently (Woolhiser 1992; Katz and Melange 1996 and 1993; Wang and Cho 1997; among others).

For homogenization of daily precipitation data series, it is essential to model the occurrence process and the non-zero amounts separately. Otherwise, estimates of adjustments needed for non-zero amounts would be biased; particularly, adding an adjustment value to all days, including days of zero precipitation, will make all days of reported zero precipitation disappear from the series or result in negative daily precipitation amounts. Changes in measuring device usually would not introduce any sudden change in the reported zero precipitation, unless they are accompanied with a change in the measuring precision (which could affect the frequency of small amounts measured and hence the frequency of reported precipitation occurrence or non-occurrence). One should not change the zeros in a daily precipitation series, unless there are corresponding reports of trace occurrence or there is a change in the measuring precision. In the latter cases the amount to be added to some of the days of reported zero precipitation should be very small, not exceeding the smallest measured amount (i.e., the highest measuring precision during the period of data record).

In data homogenization context, precipitation occurrence process can be represented by frequency of precipitation occurrence (or non-occurrence). For series of monthly or annual relative frequency (i.e., count divided by the total number of observations in the month or year) or a logistic transformation of the count series (Wang 2006), normality is not a big concern. Some existing methods, such as the penalized maximal F (PMF) test and its extended version PMFred (which accounts for the first order autocorrelation; see Wang 2008a,b), can be used to test homogeneity of frequency series. For the non-zero daily amounts series, however, a more data-adaptive transformation is necessary (more details in section 3 below).

The main objective of this study is to develop a method for detecting changepoints in series of non-zero daily precipitation amounts, which are positive and typically non-Gaussian. The focus here is on integrating a Box-Cox power transformation procedure into the PMFred algorithm, a common trend two-phase regression model based test for detecting changepoints (Wang 2008a, Wang 2008b, Wang 2003). However, with minor modification, the transformation approach is also applicable to other two-phase regression model settings, as detailed in section 2 below, and to other positive data (such as non-zero wind speeds). This study also proposes methods for adjustment of shifts in non-zero daily precipitation and other series of positive values; one of which can also be used for Gaussian data with slight modification (see section 5 below).

The paper is arranged as follows: Section 2 briefly reviews the maximal F tests for changepoint detection. Section 3 describes the new changepoint detection procedure with details. Section 4 reports the results of detection power assessment. Section 5 describes the newly proposed methods for adjustment of shifts. Section 6 gives examples of application of the new algorithm to real daily precipitation data series from different precipitation regimes across Canada. We complete this article with some concluding remarks in Section 7.

2. The maximal F tests for changepoint detection

Two maximal F tests for changepoint detection have been developed: the test of Lund and Reeves (2002) for a mean shift that may be accompanied with a trend-change (TPR4 test), and the test of Wang (2003) for a mean shift that is not accompanied with a trend-change (TPR3 test). We briefly review these tests in this section.

Let $\{X_i,\ i = 1, \ldots, N\}$ denote a data series observed at times $t_1 < \cdots < t_i < \cdots < t_N$. Assuming that there is at most one changepoint in this time series, the following two-phase regression model has been developed to test if there is a sudden change at time t_c:

$$X_i = \begin{cases} \mu_1 + \beta_1 t_i + \epsilon_i, & 1 \leq i \leq c \\ \mu_2 + \beta_2 t_i + \epsilon_i, & c+1 \leq i \leq N \end{cases} \tag{1}$$

where the errors $\{\epsilon_i\}$ are usually assumed to be identically and independently distributed (IID) normal random variables with zero mean and common variance σ^2 (Lund and Reeves 2002; Wang 2003). Denote $\boldsymbol{\theta} = (\mu_1, \mu_2, \beta_1, \beta_2)$ and the expectation of X_i by $E(X_i) = \mu(t_i, c, \boldsymbol{\theta})$, i.e.,

$$\mu(t_i, c, \boldsymbol{\theta}) = \begin{cases} \mu_1 + \beta_1 t_i, & 1 \leq i \leq c, \\ \mu_2 + \beta_2 t_i, & c+1 \leq i \leq N. \end{cases} \tag{2}$$

If $(\mu_1, \beta_1) \neq (\mu_2, \beta_2)$ for any $c \in \{N_{min}, N_{min}+1, ..., N - N_{min}\}$ (N_{min} is the pre-set minimum segment length; the series can be split further only if its length $N \geq 2N_{min}$), there exists a sudden change at time t_c (between t_c and t_{c+1}), in which case t_c is called a changepoint (and c, the changepoint parameter) and

$$\begin{cases} X_i \sim IID\ \mathcal{N}(\mu_1 + \beta_1 t_i, \sigma^2), & 1 \leq i \leq c \\ X_i \sim IID\ \mathcal{N}(\mu_2 + \beta_2 t_i, \sigma^2), & c+1 \leq i \leq N \end{cases}$$

where $\mathcal{N}(\mu, \sigma^2)$ denotes a normal distribution with mean μ and variance σ^2.

Due to incomplete or inaccurate metadata, the changepoint time t_c is often unknown. Two maximal F tests have been developed to estimate the changepoint time t_c (or the parameter c) and to test for its statistical significance. These are to test the full model (1) against the null model:

$$X_i = \mu + \beta t_i + \epsilon_i \text{ for } 1 \leq i \leq N, \tag{3}$$

using the following F_{max} statistic:

$$F_{\max} = \max_{N_{min} \leq c \leq N - N_{min}} F_c, \quad \text{with } F_c = \frac{[S_0^2 - S^2(c)]/\kappa_d}{S^2(c)/(N - \kappa)},$$

where S_0^2 denotes the sum of residual squares of the null model (3) fit, and $S^2(c)$, that of the full model (1) fit with a changepoint at time t_c; and κ is the number of free regression parameters under the full model (1) and κ_d is equal to the difference in the number of free regression parameters between the full and null model (Seber 1977; Lund and Reeves 2002). Note that the distribution function of $F_{\max}$ is complicated and unknown; and hence the $F_{\max}$ threshold values are estimated via Monte Carlo simulations (Lund and Reeves 2002; Wang 2003). As shown in Wang (2008b), these empirical threshold values not only depend on N, κ_d, and κ, but also depend on the minimum segment length N_{min}. Here $N_{min} \geq 2$ because there are two regression parameters (μ, β) in the null model (3). We set $N_{min} = 10$ in this study, so that we can use the $F_{\max}$ critical values given in Wang (2008b) and in the RHtestsV3 software package (Wang and Feng 2010) and its previous version, which took several months of CPU time to obtain. Thus, there are at least 10 data available for estimating the two regression parameters (μ, β).

The integer values κ_d and κ depend on the model settings. For the setting in which $\mu_1 \neq \mu_2$ and/or $\beta_1 \neq \beta_2$ (i.e., a mean-shift that may be accompanied by a trend change), $\kappa_d = 2$ and $\kappa = 4$, and the related maximal F test (Lund and Reeves 2002) is referred to as the TPR4 test. Different constrains on the regression parameters may be imposed to meet various environmental and application conditions. For the cases in which $\mu_1 \neq \mu_2$ but $\beta_1 = \beta_2 = \beta$ (a mean-shift without an accompanying trend change), $\kappa_d = 1$ and $\kappa = 3$, and the corresponding maximal F test (Wang 2003) is referred as the TPR3 test. The latter tests the null model (3) against the following full model:

$$X_i = \begin{cases} \mu_1 + \beta t_i + \epsilon_i, & 1 \leq i \leq c, \\ \mu_2 + \beta t_i + \epsilon_i, & c+1 \leq i \leq N. \end{cases} \tag{4}$$

For the case in which $\beta_1 \neq \beta_2$ and $\mu_2 = \mu_1 + (\beta_1 - \beta_2)t_c$ (i.e., the regression function is continuous at the changepoint c; Solow 1987), $\kappa_d = 1$ and $\kappa = 3$; and the corresponding maximal F test is to test for a sudden trend change without an accompanying mean shift, i.e., to test the following full model:

$$X_i = \begin{cases} \mu + \beta_1 t_i + \epsilon_i, & 1 \leq i \leq c \\ \mu + (\beta_1 - \beta_2)t_c + \beta_2 t_i + \epsilon_i, & c+1 \leq i \leq N \end{cases} \tag{5}$$

against the null model (3). The empirical F_{max} percentiles of Wang (2003) are applicable in this case, because the number of free regression parameters in the full model is also 3 in this setting of Solow (1987). Readers are referred to Reeves et al (2007) for a comprehensive review and comparison of these maximal F tests.

Note that if c is fixed and known (no need to estimate c statistically), the standard F test should be used to test for the statistical significance of the known/documented changepoint at time t_c, that is, to compare the F_c statistic with its critical value from $F_{\kappa_d, N-\kappa}$, a central F-distribution with κ_d and $(N - \kappa)$ degrees of freedom (e.g., Wang 2008b). This is a much easier case, which is not the focus of this study, though a similar search for the best data transformation (see Section 3) can be performed while accounting for the known changepoint c.

The above changepoint detection methods require the following core assumptions:

- *Independent normality.* The errors $\{\epsilon_i\}$ are IID normally distributed.
- *Constant variance.* $\{X_i\}$ have the common variance σ^2 across the time period examined.
- *Single changepoint.* $\{X_i\}$ experience at most one changepoint over the time period examined.
- *Piecewise linearity.* The expectation of X_i is linear in time t_i, respectively, before and after the putative changepoint c.

Violation of these model assumptions can result in severe consequences and impacts on validation and efficiency of the detection procedures and even break down the procedures. As many researchers and practitioners would have recognized, these assumptions are often not met in climate applications.

It has been noticed that, as a result of the effect of unequal sample sizes, the distribution of false alarm rate (FAR) is W-shaped for the TPR3 test, and U-shaped for the TPR4 test (Wang 2008a). In order to even out the uneven FAR distribution of the TPR3 test, Wang (2008a) has proposed the PMF test, a penalized version of the TPR3 test. [Similarly, a penalized version of the TPR4 test is being developed (manuscript in preparation by Xiaolan L. Wang).] Wang (2008b) further extended this PMF test to account for the first order autocorrelation in the series being tested, proposing the PMFred algorithm, which also includes a stepwise testing algorithm for detecting multiple changepoints in a series (note that the stepwise algorithm does not test the null hypothesis of no changepoint against the alternative of one or more changepoints; rather, it repeats the test of the no changepoint null hypothesis against the alternative of at most one changepoint in a segment of the series, with the segments being determined by the results of the preceding tests; see Wang 2008b).

In the next section, we propose to integrate a data transformation procedure in the PMFred algorithm (Wang 2008b), to diminish the effects of departure from normality on the test results. Note that this data transformation procedure is also applicable to other positive data such as non-zero wind speeds, and to other two-phase regression model based tests, such as the TPR4 test or the maximal F test for data of structure as described in Solow (1987). The transformation procedure is also applicable to monthly total precipitation data series, but zero amounts need not be treated separately in this case. However, a simple log-transformation is often sufficient for monthly total precipitation series, as is recommended in the RHtestsV3 User Manual (Wang and Feng 2010).

3. The transPMFred algorithm

Let $\{Y_i,\ i = 1, \ldots, N\}$ be a climate data series observed at times t_i $(t_1 < \cdots < t_N)$. The main idea of the proposed changepoint detection procedure is to first seek an appropriate transformation, say $h(\cdot)$, such that the transformed data series $\{h(Y_i)\}$ comes near to meet the core model assumptions, namely, the assumptions of normality, constant variance, single changepoint, and piecewise linearity. The changepoint detection is then performed on the transformed data.

Selection of a transformation obviously needs to be data-based, since different climate data series examined may have different distributions, especially distribution shapes, and hence require different transformations so that the transformed data nearly satisfy the normality assumption. Following common modelling practise and techniques, a transformation is to be selected from a family that is indexed by a scalar parameter, say $h(y; \lambda)$, where h is a specific function and λ is called transformation parameter. The transformation family is expected to possess the following properties:

(a) As far as changepoint detection is concerned, the transformation must guarantee that no any information on changepoints contained in the data is altered or modified by the transformation. For each λ fixed, $h(y; \lambda)$ must be a one-to-one function in y.

(b) The transformation family is easy to handle mathematically. Specifically, the transformation does not bring in intractable computational difficulties. Common requirements include differentiability of h with respect to y for each fixed λ.

(c) The transformation family is rich and flexible enough to allow a selection to attain good approximation to normality.

Since non-zero daily precipitation data are positive and typically highly skewed, we recommend to use the Box-Cox power transformation (Box and Cox, 1964), which is defined as follows: For a response $Y_i > 0$,

$$h(Y_i;\lambda) = \begin{cases} (Y_i^\lambda - 1)/\lambda, & \lambda \neq 0 \\ \log Y_i, & \lambda = 0. \end{cases} \tag{6}$$

It is seen that the power transformation family is of the properties (a), (b) and (c). Given positive data series $\{Y_i\}$, the transformed distribution of $X_i = h(Y_i;\lambda)$ in the common trend two-phase regression model (4) is expected to have the probability density function

$$f_i(x|c,\boldsymbol{\theta}) = \begin{cases} (2\pi\sigma^2)^{-1/2}\exp\{-(x-\mu_1-\beta t_i)^2/(2\sigma^2)\}, & 1 \leq i \leq c, \\ (2\pi\sigma^2)^{-1/2}\exp\{-(x-\mu_2-\beta t_i)^2/(2\sigma^2)\}, & c+1 \leq i \leq N, \end{cases}$$

for $i = 1, 2, \cdots, N$, where $\boldsymbol{\theta} = (\mu_1, \mu_2, \beta, \sigma^2)$ as defined before. Noting that the transformation $x = h(y;\lambda)$ has Jacobian $dx/dy = y^{\lambda-1}$, in terms of the original scale, Y_i is expected to have the probability density function

$$g_i(y|\lambda, c, \boldsymbol{\theta}) = f_i(h(y;\lambda)|c,\boldsymbol{\theta})y^{\lambda-1}. \tag{7}$$

For a given transformation parameter λ, let $S^2(c|\lambda)$ denote the sum of residual squares of the least squares fit to the transformed responses $X_i = h(Y_i;\lambda)$ with a putative changepoint at time $t = t_c$. Let $P(c)$ denote the penalty function proposed in Wang (2008a) for the TPR3 setting; and let '$\hat{c}|\lambda$' denote the most probable changepoint in the transformed data series $\{X_i\} = \{h(Y_i;\lambda)\}$. Then, the PMF test statistic can be expressed as

$$PF_{max} = \max_{N_{min} \leq c \leq N - N_{min}} [P(c)F_c(\lambda)] = P(\hat{c})F_{\hat{c}}(\lambda) = P(\hat{c})\frac{[S_0^2(\lambda) - S^2(\hat{c}|\lambda)]/\kappa_d}{S^2(\hat{c}|\lambda)/(N-\kappa)}.$$

The profile log-likelihood function for λ up to a constant can be defined as

$$l(\hat{c}|\lambda) = -\frac{N}{2}\log[\frac{S^2(\hat{c}|\lambda)}{N-\kappa}] + (\lambda - 1)\sum_{i=1}^{N}\log Y_i.$$

The number of free parameters in a multi-phase common trend regression model with M_c mean shifts is $\kappa = (2 + M_c)$; thus, $\kappa = 3$ for the single mean shift full model (4). This log-likelihood function has been maximized over all admissible c's ($c = N_{min}, N_{min}+1, ..., N - N_{min}$), given the transformation parameter λ.

For any admissible changepoint parameter c, the pseudo maximum likelihood estimate $\hat{\lambda}$ for the transformation parameter λ is such that

$$l(c|\hat{\lambda}) = \max_{a \leq \lambda \leq b} l(c|\lambda),$$

where a and b are two specific numbers. In applications, $a = -1$ and $b = 1$ are commonly recommended and found to be sufficient for the maximizer $\hat{\lambda}$ to occur within the interval $[a, b]$.

In view of modelling, the proposed changepoint detection procedure can be considered as a changepoint detection procedure in the model (7) with λ being treated as a model parameter. In other words, for the observed responses, there is a λ such that the transformed data $X_i = \{h(Y_i; \lambda)\}$ approximately meet the assumptions of normality, constant variance, single changepoint, and piecewise linearity, so that the changepoint detection can be performed on the transformed data series $\{X_i\}$.

It has been a common practise in applications to re-express the observed data and use a new scale for the measurements in order to employ a standard statistical inference procedure, which dated back at least to 1964 when the paper Box and Cox (1964) was published (Atkinson, 1985). Chen and Loh (1992) and Chen (1995) show that the transformed tests are usually more efficient and have greater testing powers in context of large samples (such as daily data). A drawback of the transformed analysis is that it causes difficulties to interpret the conclusions in the new scale with $\hat{\lambda}$ [perhaps we may feel comfortable only with $\hat{\lambda} = -1, -1/2, 0,$ or $1/2$ to interpret the results in terms of $X_i = h(Y; \hat{\lambda})$]. Fortunately, this drawback does not show up in the present problem of changepoint detection since we are concerned with the question whether a changepoint occurs in the climate data during the time period examined.

In applications, the estimate $\hat{\lambda}$ is usually approximated by an exhaustive search algorithm. For convenience, a practical algorithm to search for the best λ value is outlined below. Let $a = \lambda_1 < \lambda_2 < \cdots < \lambda_J = b$ be a grid of λ over the interval $[a, b]$, with the increment δ, i.e., $\lambda_j = a + (j-1)\delta$ for $j = 1, 2, ..., J$, where $J = ((b-a)/\delta + 1)$. To save computation time, we carry out four rounds of search over the interval [-1.0, 1.0] in this study. In the first round, we search over three trial values of λ, $\{-1.0, 0.0, 1.0\}$ (i.e., set $a = -1.0$, $b = 1.0$, and $\delta = 1.0$), obtaining the first estimate of λ, say $\hat{\lambda}^o$. We refine this estimate by setting $a^o = \hat{\lambda}^o - \delta$ and $b^o = \hat{\lambda}^o + \delta$ (and if $a^o < a$, reset $a^o = a$; if $b^o > b$, reset $b^o = b$), and $\delta^o = 0.5$, to search over the $J = ((b^o - a^o)/\delta^o + 1)$ trial values of λ: $\lambda_j = a^o + (j-1)\delta^o$ for $j = 1, 2, ..., J$. For example, if $\hat{\lambda}^o = 0.0$, the second set of $J = 5$ trial values of λ are $\{-1.0, -0.5, 0.0, 0.5, 1.0\}$ (here only -0.5 and 0.5 are new trial values, for which the calculations need to be done). Similarly, we carry out two more rounds of search, with $\delta^o = 0.2$ and $\delta^o = 0.1$, respectively. The resulting final estimate of λ is of one decimal point precision; the calculations are repeated up to 11 times (instead of 21 times required by the direct search over the 21 trial values: -1.0, -0.9, ..., 0.9, 1.0). If desired, further rounds of search can be performed to refine the estimate further (e.g., from 0.1 to 0.01 precision, which would double the calculation time). Our experiments indicate that only the second decimal point of the test statistic could be affected by the λ precision change from 0.1 to 0.01, which would only affect detection of marginally significant changepoints. A λ value of one decimal point precision is often of sufficient precision. Thus we decided to perform only four rounds of search here and in the related software package RHtests_dlyPrcp, providing λ estimates of one decimal point precision.

The smaller the λ value, the shorter the upper tail of the distribution; data of a negative λ have a shorter upper tail than those of a positive λ. We noticed that for Canadian non-zero (i.e. > 0) daily precipitation data the best λ values range from -0.2 to 0.2, with -0.2 for the high Arctic stations (Alert, Resolute), -0.1 for central Canada, 0.0 for the region from the Great Lakes area eastward to the east coast, and 0.2 for the southwest-coastal stations (e.g. Port Hardy). For 11 series of non-zero daily precipitation data from stations in different regions of China, the best λ values range from -0.1 to 0.1, with $\lambda = 0.0$ for 8 out of the 11 series. We also noticed that the best λ value is 0.1 for two series from Turkey, 0.0 for one series from Jordan, and 0.1 and 0.2 for two series from Congo (a very wet area), respectively. [We are not able to estimate the λ value range for other areas because we do not have access to the daily precipitation data series.] From these results we speculate that negative λ values are typical for regions where precipitation is dominantly light, such as the Arctic and sub-Arctic regions where more than one third of the daily amounts are less than 1 mm and the daily amount rarely exceeds 25 mm (for the high Arctic station Alert, one half of the daily amounts are less than 1 mm and the daily amount rarely exceeds 20 mm). However, when the series of daily precipitation greater than 3 mm (at The Pas) was tested, the best λ value is -0.6 (see section 6 for the reason for testing such a series). These indicate that it is necessary to search for the best λ value over the interval $[-1.0, 1.0]$, allowing the comparison with the case of no transformation (i.e. $\lambda = 1.0$, in which case the results of testing the transformed series $\{X_i\} = \{h(Y_i; 1)\} = \{Y_i - 1\}$ will be the same as testing the original series $\{Y_i\}$, because $\{X_i\}$ and $\{Y_i\}$ share exactly the same shape of distribution). Note that we also include the most common case, $\lambda = 0.0$, up front in our first round of search for the best λ.

Next, we outline a practical algorithm to perform the transformed changepoint detection procedure. Here we focus on data of the TPR3 setting (i.e., common trend series: $\beta_1 = \beta_2 = \beta$). We describe the procedure for series of at most one changepoint and of multiple change changepoints, subsequently.

For common trend series of at most one changepoint (mean shift), the detection procedure is as follows:

Step A. For each trial transformation parameter λ_j $(1 \leq j \leq J)$, obtain the transformed series $\{X_i^j\} = \{h(Y_i; \lambda_j)\}$:

$$X_i^j = \begin{cases} (Y_i^{\lambda_j} - 1)/\lambda_j, & \lambda_j \neq 0 \\ \log Y_i, & \lambda_j = 0 \end{cases} \quad \text{for } i = 1, 2, .., N;$$

Step B. Use the PMFred algorithm (Wang 2008b) to find the most probable changepoint in the transformed series $\{X_i^j\}$ for each λ_j. This step involves computing $S_0^2(\lambda_j)$ and $S^2(c|\lambda_j)$ for each admissible $c \in \{N_{min}, N_{min} + 1, ..., N - N_{min}\}$ and finding $\hat{c}_j$ such that

$$PF_{max} = \max_{N_{min} \leq c \leq N - N_{min}} [P(c) F_c(\lambda_j)] = P(\hat{c}_j) F_{\hat{c}_j}(\lambda_j) = P(\hat{c}_j) \frac{[S_0^2(\lambda_j) - S^2(\hat{c}_j|\lambda_j)]/\kappa_d}{S^2(\hat{c}_j|\lambda_j)/(N - \kappa)}.$$

It also involves computing the log-likelihood function

$$l(\hat{c}_j|\lambda_j) = -\frac{N}{2} \log[\frac{S^2(\hat{c}_j|\lambda_j)}{N - 3}] + (\lambda_j - 1) \sum_{i=1}^{N} \log Y_i.$$

Step C. Maximize the profile log-likelihood function $l(\hat{c}_j|\lambda_j)$ over all $j \in \{1, 2, ..., J\}$, to find $\hat{\lambda}$ (and $\hat{c}$) such that

$$l(\hat{c}|\hat{\lambda}) = \max_{1 \leq j \leq J} l(\hat{c}_j|\lambda_j) = l(\hat{c}_m|\lambda_m),$$

where m is a fixed integer between 1 and J (inclusive). Now $\hat{\lambda} = \lambda_m$ represents the best transformation among all the trial transformations $(\lambda_1, \lambda_2, ..., \lambda_J)$, and $\hat{c} = \hat{c}_m$ is the most probable changepoint in both the transformed series $\{X_i^m\} = \{h(Y_i; \lambda_m)\}$ and the original series $\{Y_i\}$.

Step D. Compare the $PF_{\max}$ value

$$PF_{max} = P(\hat{c})F_{\hat{c}} = P(\hat{c})\frac{[S_0^2(\lambda_m) - S^2(\hat{c}|\lambda_m)]/\kappa_d}{S^2(\hat{c}|\lambda_m)/(N - \kappa)}.$$

with its critical value corresponding to the nominal significance, to determine whether or not $\hat{c}$ is a statistically significant changepoint. As noted before, some critical values of the $PF_{\max}$ for the TPR3 setting are given in Wang (2008b), with the lag-1 autocorrelation of the series being accounted for.

For common trend series that may contain more than one mean shift (multiple changepoints), the detection procedure is as follows:

Step 1. For each trial transformation parameter λ_j $(1 \leq j \leq J)$, obtain the transformed series $\{X_i^j\} = \{h(Y_i; \lambda_j)\}$.

Step 2. Use the PMFred algorithm (Wang 2008b) to identify all significant changepoints in the transformed series $\{X_i^j\}$ for each λ_j (note that the PMFred algorithm includes a step-wise algorithm for detecting multiple changepoints in a single series). For the transformed series $\{X_i^j\}$, say $M_c^j \geq 0$ changepoints, $\hat{\boldsymbol{C}}_j = (\hat{c}_1^j, \hat{c}_2^j, ..., \hat{c}_{M_c^j}^j)$, were identified to be statistically significant at the nominal level. Then $\kappa_j = N - (2 + M_c^j)$. Compute $S^2(\hat{\boldsymbol{C}}_j|\lambda_j)$, the sum of residual squares of the least squares fit to the transformed series $\{X_i^j\} = \{h(Y_i; \lambda_j)\}$ with the M_c^j changepoints, and the log-likelihood function

$$l(\hat{\boldsymbol{C}}_j|\lambda_j) = -\frac{N}{2}\log[\frac{S^2(\hat{\boldsymbol{C}}_j|\lambda_j)}{N - \kappa_j}] + (\lambda_j - 1)\sum_{i=1}^{N}\log Y_i.$$

Step 3. Maximize the profile log-likelihood function $l(\hat{\boldsymbol{C}}_j|\lambda_j)$ over all $j \in \{1, 2, ..., J\}$, to find $\hat{\lambda}$ (and $\hat{\boldsymbol{C}}$) such that

$$l(\hat{\boldsymbol{C}}|\hat{\lambda}) = \max_{1 \le j \le J} l(\hat{\boldsymbol{C}}_j|\lambda_j) = l(\hat{\boldsymbol{C}}_m|\lambda_m),$$

where m is a fixed integer between 1 and J (inclusive). Now $\hat{\lambda} = \lambda_m$ represents the best transformation among all the trial transformations $(\lambda_1, \lambda_2, ..., \lambda_J)$, and $\hat{\boldsymbol{C}} = \hat{\boldsymbol{C}}_m = (\hat{c}_1^m, \hat{c}_2^m, ..., \hat{c}_{M_c^m}^m)$ are statistically significant changepoints in both the transformed series $\{X_i^m\} = \{h(Y_i; \lambda_m)\}$ and the original series $\{Y_i\}$.

Step 4. Save the identified changepoints and their significance corresponding to the best transformation λ_m, and the best transformed series $\{X_i^m\} = \{h(Y_i; \lambda_m)\}$. Analyze the identified changepoints along with plots of the original and transformed series and all available metadata to determine whether or not they are artificial changepoints of physical significance and whether or not they should be adjusted. This procedure is the same with or without the above transformation procedure; it has been detailed in Wang and Feng (2010) and Wang (2008b). However, when one or more of the identified changepoints are determined to be insignificant and hence deleted from the list of changepoints, the best λ value for the series with the new list of changepoints being accounted for is sought; then the statistical significance of the retained changepoints is re-estimated using the new transformed series (with the new best λ value). The search for the new best λ value is conducted the same way as before. This procedure is repeated until all the retained changepoints are determined to be significant. We noticed that even adding a few changepoints to the list of changepoints usually does not change the best λ value. This is because all major changepoints that would notably bias the best λ estimate were accounted for in the earlier stage of search for the best value.

With a slight modification, the above procedure is applicable to data series of the TPR4 setting (i.e., $\mu_1 \neq \mu_2$ and/or $\beta_1 \neq \beta_2$). One only needs to replace the PMF test in the PMFred algorithm with the corresponding test statistic for the TPR4 based test in the above procedure when dealing with data of the TPR4 structure.

Note that, in the procedure described above, the λ value that corresponds to the maximal profile log-likelihood is simply chosen and is not compared to any threshold value; the position and significance of changepoints are determined by applying the PMFred test to one single transformed series, which does not involve any comparison of the test statistic from different λ values. Therefore, the critical values of the test statistic remain the same as those for the PMFred test. However, if we were to search for a most probable changepoint among all the trial values of λ and to repeat such search for the next most probable changepoint until no more significant changepoint can be found, the critical values most likely would be different from those of the PMFred algorithm, because the maximization of the test statistic for determining the most probable changepoint position involves different trial values of lambda. In this case, one must re-estimate the critical values. It is our plan to try the latter approach in a separate study.

4. Assessment of detection power aspects

In this section, we describe simulation studies to compare the detection power aspects of the new transPMFred algorithm with the corresponding untransformed version, the PMFred algorithm. Readers are referred to Reeves et al. (2007) for a detail review and comparison of eight data homogeneity tests, including the TPR3, the TPR4, and the Standard Normal Homogeneity (SNH) test (Alexandersson 1986), and to Wang (2008a) for a comparison of the PMF test with the TPR3 test. An extension of the PMF to the PMFred, along with a comparison of the PMFred with the PMF and TPR3 tests can also be found in Wang (2008b). The same study also presents an extension of the penalized maximal t (PMT) test to the PMTred, along with a comparison of the PMTred with the PMT and SNH tests; while a comparison of the PMT test with the SNH test can also be found in Wang et al. (2007). However, the SNH, PMT, and PMTred tests, which assume zero trend in the series being tested and thus are for use with a reference series, are not suitable for daily precipitation series, because the extremely high spatial variability and non-continuity of daily precipitation make it unrealistic to find a suitable reference series.

Note that an algorithm should be applied to truly homogeneous series in order to estimate its false alarm rate, and to series of known shifts (i.e., both the exact location and size of shift are known) to estimate its hit rate (both false alarm and hit rates are defined later in this section). A popular approach for conducting such a detection power assessment is to generate homogeneous surrogate series, and to apply the algorithm to be assessed/compared to these series before and after inserting one or more changepoints in these series. This is called a simulation study and is what we do in this section.

In order to generate surrogate daily precipitation series that are homogeneous and yet mimic real daily precipitation data series, we use the so-called block-bootstrapping technique in this study. Considering the length of typical synoptic scale, we choose the block size of 5 days, which would help preserve the first order autocorrelation of daily precipitation process in the homogeneous surrogate series. Specifically, we randomly take blocks of 5-day data record, including zero values, from a homogeneous real daily precipitation series to generate $M = 1000$ surrogate daily precipitation series, each of which contains 900 non-zero values [i.e. we repeat the block-bootstrapping procedure until the surrogate series contains 900 non-zero values each, so that the sample size is the same for all the surrogate series, which is important for the intended comparison of detection power over different λ values, because detection power increases with an increase in sample size (see Wang 2008b)]. Further, we add the linear trend estimated from the homogeneous real daily precipitation series to each of these surrogate series. We apply the transPMFred algorithm to each of these homogeneous surrogate series to estimate the false alarm rate. Here, a false alarm is counted when the test detects a significant changepoint in an actually homogeneous surrogate series; while a hit is counted when the test identifies a significant changepoint $\hat{c} \in [K - 10, K + 10]$ for a time series that has a true changepoint at time K (thus the hit rate is a strict measure of accurate detection power).

Also, we insert, at randomly selected positions, one or two mean shifts of different magnitudes and signs to each of the surrogate series, and then apply the transPMFred algorithm to each of these M series to estimate the hit rate. The size of the inserted shift, expressed in unit of the standard deviation (σ) of the homogeneous surrogate series, varies from 0.5σ to 2.5σ (see Table 1 and Fig. 1). Considering that the hit rate for the one shift cases would be more comparable with that for the two shifts cases (more details below) if they share the same sample size on average, or equivalently the same "mean" segment length $\overline{N_s} = N/(M_c + 1)$, we use only the first 600 non-zero values in each of these M series to assess the hit rate for one shift ($M_c = 1$) cases and the false alarm rate. That is, $N = 600$ for the one shift cases, and $N = 900$ for the two shifts cases, so that $\overline{N_s} = N/(M_c + 1) = 300$ for both the one and two shifts cases presented in Table 1. The lower bound (with subscript $_L$) of the rates in Table 1 is estimated using the upper bound of the first order autocorrelation estimate for the series being tested, and the upper bound (with subscript $_U$), using the lower bound of the autocorrelation estimate (these together represent the 95% uncertainty range). The latter is used in the RHtests_dlyPrcp package, to output also changepoints that are identified to be within the 95% uncertainty range for further analysis subjectively or along with metadata (see Wang and Feng 2010 and Wang 2008b for more detail).

We repeat the above simulation study using five homogeneous real daily precipitation data series of different λ values (-0.2, -0.1, 0.0, 0.1, and 0.2; see Table 1; they are from Alert, Coral Harbour, and Dawson in Canada, and two stations in Congo). For comparison, the PMFred algorithm is also used in place of using the transPMFred algorithm. The results are summarized in Table 1 (and some shown in Figs. 1a-b). Note that the values in Fig. 1 and Table 1 are only rough estimates, because each of them was estimated from only $M = 1000$ simulations (even so, we need to carry out a total of 180,000 simulations for the 180 values in Table 1, which is very CPU-consuming when $N = 600$ or 900).

In general, the false alarm rates of the transPMFred algorithm are around the nominal level 5% (namely, this nominal level is within the 95% uncertainty range; see Table 1); the hit rates are satisfactory for shifts of moderate to large size and are more uncertain to estimate for small shifts (e.g, of size 0.5σ; see Table 1). In terms of hit rate, as shown in Fig. 1a (see also Table 1), the transPMFred algorithm clearly outperforms the PMFred algorithm. In other words, the transformation procedure significantly improves the detection power for various non-normal data (of λ values ranging from -0.2 to 0.2); when $N = 600$ the increases in hit rates are up to 71%, 54%, 40%, and 25% for a shift of size 0.5σ, 1.0σ, 1.5σ, and 2.5σ, respectively (Table 1a, 1 shift cases). The improvement is also greater for data of negative λ values (see Fig. 1a; also compare Table 1a with Table 1b). Importantly, as shown in Fig. 1b, the hit rates of the transPMFred algorithm are roughly the same (up to estimation error) across different λ values, showing more variations only when the shifts are small (e.g., 0.5σ; see also Table 1a).

As shown in Table 1, the hit rates are slightly lower in the two shifts cases than in the corresponding one shift cases, which is due to the possibility of smaller sample size in the two shifts cases. Note that the length of the segment that contains one shift could be shorter than 600 in the two shifts cases (e.g., if the two shifts occur at the 250th and 400th point, respectively, the sample size for identifying the first shift is only 400), but it is always 600 in the one shift cases (although the mean segment length is kept the same by using $N = 600$ and $N = 900$ for the one and two shifts cases, respectively, as described above). Figure 1b also shows that the larger the shift, the higher the hit rate, as would be expected.

In order to check whether the detection power of the transPMFred is evenly distributed over changepoint position K, we also carry out the following simulations. Here, we use only the M surrogate series of $\lambda = -0.1$ and reduce the sample size from 600 to 500 (i.e. $N = 500$), just to save some CPU time. For each of the nine selected $K \in \{50, 100, 150, 200, 250, 300, 350, 400, 450\}$, we inserted a mean shift in each of the homogeneous surrogate series, setting the shift size to 0.5 and 1.0 standard deviation of the series, respectively. Then, we apply the transPMFred algorithm to each of theses surrogate series (each contains one shift), separately, to estimate the hit rates as a function of K. The results, shown in Fig. 1c, indicate that the detection power of the transPMFred algorithm is roughly the same (up to estimation error) across different K values, showing again more estimation uncertainty for small shifts (e.g., of size 0.5σ). Again, these hit rates are only rough estimates because each of them was obtained from only 1000 simulations.

5. Methods for adjustment of shifts

After identifying all significant changepoints, one wish to adjust the daily precipitation data series to diminish all significant artificial shifts, i.e., to homogenize the daily precipitation series. In this section, we propose two methods for estimating adjustments of artificial shifts identified in the non-zero daily precipitation series. The quantile matching algorithm below can be used for other non-negative data such as wind speeds and, with slight modification (see the different lower boundary conditions below), for Gaussian data as well.

First of all, one can derive the shift sizes from the "fitted mean response" of the non-zero daily precipitation series defined as

$$\hat{Y}_i^m = h^{-1}(\hat{X}_i^m; \hat{\lambda}) = \begin{cases} (\hat{\lambda}\hat{X}_i^m + 1)^{1/\hat{\lambda}}, & \hat{\lambda} \neq 0 \\ \exp(\hat{X}_i^m), & \hat{\lambda} = 0 \end{cases} \quad \text{for } i = 1, 2, .., N; \tag{8}$$

where $\hat{X}_i^m$ denotes the multi-phase regression fit to the transformed series $\{X_i = h(Y_i; \hat{\lambda})\}$ with the estimated best transformation parameter $\hat{\lambda}$, and $h^{-1}(\hat{X}_i^m; \hat{\lambda})$ denotes the inverse Box-Cox (IBC) transformation of $\hat{X}_i^m$ with $\hat{\lambda}$. The adjustments derived this way (i.e. from $\hat{Y}_i^m$) are referred to as IBC adjustments. Note that applying such IBC adjustments could result in non-positive daily precipitation amounts, unless the series is adjusted to the highest segment (i.e., the segment of highest mean value). In order to avoid non-positive daily precipitation amounts and to keep the number of "rainy" days unchanged, one can replace each non-positive value in the adjusted precipitation series with the smallest measured amount of daily precipitation. Such replacement is implemented in our RHtests_dlyPrcp software package. [Note that in this study the term "rainy days" is used to refer to precipitating days, which include days of any form of precipitation, liquid or solid such as snowfall, as is popularly used in the literature.] An example series of IBC adjustments is shown in Fig. 2a, and the resulting IBC-adjusted precipitation series shown in Fig. 3e (see section 6 below for more details). Note that the same single IBC adjustment value is applied to all data in the same segment, except for the cases where non-positive values are replaced with the smallest measured amount. In other words, an IBC adjustment is basically a mean adjustment, making little change to the shape of distribution of the data.

Alternatively, we propose the following quantile matching (QM) algorithm, in which the estimated precipitation "trend" component is preserved. Let $\{X_i^{tr} = \hat{\beta} t_i'\}$ $(i = 1, 2, ..., N)$ denote the estimated linear trend component of the transformed series $\{X_i = h(Y_i; \hat{\lambda})\}$, and $\{Y_i^{tr}\}$ is the inverse Box-Cox transformed version of $\{X_i^{tr}\}$, i.e., $Y_i^{tr} = h^{-1}(X_i^{tr}; \hat{\lambda})$. The series $\{Y_i^{tr}\}$ represents the monotone trend component in the fitted mean response of the non-zero daily precipitation series [i.e., $\hat{Y}_i^m$ as defined in (8); see the magenta trend lines in Figs. 3a-b], which we will use as an approximation to the trend component of the non-zero daily precipitation series. Then, the non-zero daily precipitation series is "detrended" as follows:

$$Y_i^{dtr} = Y_i + Y_{max}^{tr} - Y_i^{tr}$$

where $Y_{max}^{tr} = \max_{i=1}^{N} Y_i^{tr}$ (inclusion of this constant is just to prevent non-positive values to occur in the "detrended" series). The use of quotation marks here is to emphasize that this detrending procedure removes only the trend component in $\hat{Y}_i^m$, which is what we want to preserve here; they will be omitted hereafter. The detrended series $\{Y_i^{dtr}\}$ is then used to estimate the cumulative distribution function (CDF) for each segment of the data series, which is then used to estimate the QM adjustments needed to make the data series homogeneous. The procedure is detailed next.

In consideration of possible climatic change in the distribution, we allow choices of using either all data in a segment or up to a chosen number (≥ 5) of years of data before (or after) the changepoint to estimate the CDF for the segment in question. However, we do not recommend the use of less than 30 years of daily precipitation data (or 15 years of continuous data such daily temperatures) to estimate a CDF when there are more data available. This is because we noticed that an estimate of CDF (and hence the QM adjustments) using less than 30 years of daily precipitation data often contains large sampling uncertainty; the error due to sampling uncertainty is often much larger than that due to the assumption of the same distribution over the period of data record after a monotone trend has been preserved (as is done in this study). As a result of the large sampling uncertainty, the shape of the distribution could be over-adjusted (for example, the large variations in the QM adjustments for different quantiles derived with up to 10 years of data, shown in the lower-right panel of Fig. 4, could lead to an over-adjustment in the distribution shape; see section 6 for more details about the AMOS example series). We recommend the use of all available data in a segment to estimate the CDF if the assumption of the same distribution throughout the period of data record is not obviously violated (viewing the data time series plot would help determine this subjectively).

For each changepoint, there are two chosen parts of the series for use to estimate the two CDFs, respectively, before and after the changepoint, as schematically shown in Fig. 5. The data in each of these two chosen parts of the detrended series $\{Y_i^{dtr}\}$ are sorted in ascending order and then divided into M_q ascending categories "equally" (to the extent possible), in order to estimate the CDFs at M_q cumulative frequencies that are separated by the increment of $1/M_q$. For each $l \in \{1, 2, ..., M_q\}$, let $F(l) = (l - 0.5)/M_q$ denote the median empirical cumulative frequency (ECF) of the l-th category data, whose ECF falls in the interval $((l-1)/M_q, l/M_q]$. Also, let $F(0) = F(1) - 1/M_q$ and $F(M_q + 1) = F(M_q) + 1/M_q$, which are to be used for imposing the lower and upper boundary conditions for use in the spline interpolation below. Let $P_b(k,l)$ and $P_a(k,l)$ denote the mean of the l-th category data in the chosen part of the series immediately before and after the k-th changepoint, respectively (see Fig. 5). The difference in the l-th category mean between the two segments separated by the k-th changepoint are derived as

$$D(k,l) = P_a(k,l) - P_b(k,l) \quad (l = 1, 2, ..., M_q). \tag{9}$$

Note that this is the category mean of $(k+1)$-th segment minus that of the k-th segment $(k = 1, 2, ..., M_c)$, and that $P_a(k,l) = P_b(k+1,l)$ if the whole segment of data is chosen for use to estimate the CDF.

Let s_o denote the segment to which the other segments are to be adjusted (also referred to as the base segment). For each segment $s \in \{1, 2, ..., (M_c + 1)\}$, the difference in the l-th category mean between the s-th and s_o-th segments can be derived as:

$$A(s,l) = \begin{cases} \sum_{k=s}^{s_o-1} D(k,l) = \sum_{k=s}^{s_o-1}[P_a(k,l) - P_b(k,l)] & \text{if } s < s_o \\ 0 & \text{if } s = s_o \\ \sum_{k=s_o}^{s-1}[-D(k,l)] = \sum_{k=s_o}^{s-1}[-P_a(k,l) + P_b(k,l)] & \text{if } s > s_o \end{cases} \qquad (10)$$

Also, let the lower boundary value $A(s,0) = 0$ for positive data, but $A(s,0) = A(s,1)$ for Gaussian data; and let the upper boundary value $A(s, M_q+1) = A(s, M_q)$ (these boundary conditions keep the lowest and highest $50/M_q$ percent of data bounded, not to let them depart too much from the mean of the QM adjustments for the respective category). Thus, for each segment s, there are (M_q+2) data points, $(F(l), A(s,l))$ for $l = 0, 1, ..., M_q+1$ As shown in Fig. 4, a natural cubic spline is then fitted to these (M_q+2) data points for each segment s [except the base segment s_o, for which $A(s_o, l) \equiv 0$]. The QM adjustments needed to homogenize the series $\{Y_i\}$ will be derived from these fitted splines, as described next.

Let $\mathcal{F}_s(i)$ denote the ECF of the ith datum in segment s of the series $\{Y_i^{dtr}\}$. From the fitted spline for segment s, we can look up the inter-segment difference (i.e. the y-axis value, which is the difference between segments s and s_o) that corresponds to the cumulative frequency $\mathcal{F}_s(i)$. This difference, say $\mathcal{A}(s,i)$, is the amount that will be added to the ith datum in segment s of the series $\{Y_i\}$, to adjust it to segment s_o. This spline interpolation is carried out for each value in each segment that needs to be adjusted (i.e. all except the base segment). The resulting differences $\mathcal{A}(s,i)$ for $i = 1, 2, ..., N$, are referred to as the QM adjustments for segment s; the corresponding fitted spline (Fig. 4) represents these QM adjustments as a function of cumulative frequency. Similarly, non-positive values could appear in the adjusted precipitation amounts using these QM adjustments. Again, these non-positive values are replaced with the smallest measured precipitation amount (or the smallest value in the original positive data, in general). An example series of QM adjustments is shown in Fig. 2b, and the resulting QM-adjusted precipitation series shown in Fig. 3f (see section 6 below for more details).

Note that, because of the spline fit and interpolation described above, for any $M_q \geq 2$ the resulting adjustment may vary from one datum to another; even when $M_q = 2$, not only two different adjustment values are applied to the values in a segment, but each and every datum in a segment has its own adjustment that corresponds to its ECF.

The following aspects should be considered when choosing the number M_q for use to derive the QM adjustments. First, M_q shall be determined so that the shortest segment in the series has enough data in each of the M_q categories to estimate the category means. Second, the larger the M_q value, the larger the resulting adjustment to the shape of the distribution, and vice versa; the shape of the distribution is not adjusted at all when $M_q = 1$ (one single adjustment value is applied to all data in the same segment in this case, which is the usual mean adjustment). Also, the larger the M_q value, the fewer data are available for estimating the CDF and the mean inter-segment difference for each category, and thus the larger the sampling uncertainty in the estimate of the CDFs and thus the QM adjustments (see Fig. 4). Considering these, in our software package we allow users to either set the M_q value to any integer between 1 and 20 inclusive, or set $M_q = 0$ to let the codes choose an M_q value (a chosen M_q value results from either way). Further, in order to ensure that even the shortest segment (say, of length N_{srt}) has enough data to estimate the mean of each category, our codes will automatically replace a chosen M_q value with the integer $\max(N_{srt}/20, 1)$ if the chosen M_q value is larger than integer $N_{srt}/20$ or smaller than 1. This ensures that there are at least 20 daily data in each category for estimating the categorical mean, and that the minimum value for M_q is 1 (namely, setting $M_q = 1$ when there is not enough data in the shortest segment for estimating QM adjustments, and thus applying one single adjustment value to all data in the same segment). [For monthly or annual data series, the above $N_{srt}/20$ is replaced by $N_{srt}/5$, so that there are at least 5 monthly or annual data in each category for estimating the category mean.] In our software packages (RHtestsV3 and RHtests_dlyPrcp), M_q is also automatically re-set to 20 if it is larger than 20, to avoid large sampling uncertainty in the estimate of QM adjustments (and thus over-adjustments to the shape of the distribution).

Note that all quantile matching (QM) algorithms (e.g., Della-Marta and Wanner 2006, Trewin and Trevitt 1996, and the one described above) try to line up the adjustments by empirical frequencies, implicitly assuming that the frequency series are homogeneous. Thus, QM algorithms would work well for continuous variables such as air temperature or pressure; they should also work well for non-continuous variables (such as daily precipitation) in the absence of frequency discontinuity. However, when there exists a discontinuity in the frequency of measurements of a non-continuous variable (such as daily precipitation or wind speed), extra caution has to be exercised when using a QM algorithm. One must address all frequency discontinuities before estimating and applying any QM adjustments; otherwise, the QM adjustments could be largely biased by the frequency discontinuities and thus could be problematic, making the data depart more from the truth (see section 6 below for examples of, and possible solutions to deal with, frequency discontinuity).

Although the above description of the QM adjustment method is specific to non-zero daily precipitation series, this adjustment method can be used to homogenize any other climate variables. For data series that do not need the Box-Cox transformation, the detrending procedure above is also more straightforward. For example, the linear trend component as estimated by the multi-phase regression model fit to the deseasonalized base series in the PMFred (or PMTred) algorithm (Wang 2008b) can be removed before the data is used to estimate the CDF and the QM adjustments. This QM adjustment algorithm has also been added to the RHtestsV3 software package (Wang and Feng 2010) as an alternative to the mean adjustments. This is particularly important for adjusting daily climate data series. Note that, as a result of applying the QM adjustments with $M_q \geq 2$, the whole distribution of the data, not only the mean, could be adjusted. When the PMFred or PMTred algorithm is used without data transformation, the detection of shifts is done on the mean only, which indicates that any shift that occurs in the variance or higher order statistic without an accompanying significant shift in the mean may go undetected in this case. However, detection of shifts in the variance or higher order statistic without an accompanying significant shift in the mean is out of the scope of this study, although it can deal with variance seasonality (see the last paragraph of section 6).

6. Application to real daily precipitation data series

In this section, we present an application of the new transPMFred algorithm to detect changepoints in non-zero daily precipitation amounts (unit: mm) recorded at six Canadian stations in different precipitation regimes. The series for London (Ontario; joining of three stations whose climate IDs are 6144505, 6144481, and 6144475) for the period from 1 March 1883 to 31 December 2006, and the series for St. John's A (Newfoundland; climate ID: 8403506) for the period from 1 January 1942 to 31 October 2008 were found to be homogeneous at the 95% confidence level ($\hat{\lambda} = 0.0$ for both series) and hence will not be discussed further. For the other four example series, the names, climate IDs, period of data tested, and their best $\hat{\lambda}$ values are listed in Tables 2 and 3 (column 1). Except the AMOS series, these long series were formed by joining of two nearby stations' data (the two IDs listed in Tables 2 and 3).

Note that these data do not include adjustments for trace precipitation (more about trace at the end of this section), nor adjustments for the joining of stations. However, observations of snowfall were converted into their water equivalents using the adjustment factor map of Mekis and Brown (2010), and rainfall data had been adjusted for wetting loses and gauge undercatch using the methods of Mekis and Hogg (1999) but with adjustments being updated to account for new data/information (personal communication with Eva Mekis). After these adjustments, the smallest non-zero daily precipitation amount in these series is 0.20 mm.

As mentioned earlier, we propose to test the series of precipitation occurrence (or non-occurrence, i.e. zero precipitation) frequency and the series of non-zero daily precipitation amounts, separately, using different tests (namely the PMFred for the former, and the transPMFred for the latter). Thus, we collect all > 0 daily precipitation amounts to form a series of non-zero daily precipitation (a pooled precipitation series), to which we apply the transPMFred algorithm, taking into account the effects of "lag-1" autocorrelation in the series being tested (which is the pooled precipitation series, in which two consecutive data points do not necessarily correspond to two consecutive calendar days; thus the lag-1 autocorrelation is likely smaller than the autocorrelation calculated from consecutive days). We record the time index t_i at which the i-th non-zero daily precipitation in the series occurs. In other words, the i-th non-zero daily precipitation in the series occurred in the t_i-th day in the period of record. The i here is also referred to as the index number of non-zero daily precipitation day in the series. In our computing codes, the time index $t'_i = t_i/365.25$ is used (rather than the index number i or t_i) in the regression analysis to estimate the monotonic trend component in the precipitation series. Thus, the linear trend ($\hat{\beta}$) of the transformed precipitation series is expressed in "unit per year".

Wang (2008b) defines two types of changepoints: Type-1 changepoints are those identified by applying a maximal F type test (or maximal t type tests when a reference series is used) and are statistically significant at the nominal level even without metadata support. Type-0 changepoints are those whose exact times of change are documented/known (thus no need to detect them using a statistical test), for which one only needs to determine their statistical significance using the regular F test (or the t test when a reference series is used). Although focusing on Type-1 changepoints would address major discontinuities, one should look at both types of changepoints for purpose of data homogenization when metadata is available. The FindUD.dlyPrcp function in the RHtests_dlyPrcp software package can be used to narrow down the necessary metadata investigation if it has not been done (if you already know the times of changes that might cause an artificial shift in the data series, you can simply add them in the list of changepoints to test their statistical significance; see Wang and Feng 2010 for more details). Here, we focus on the detection of Type-1 changepoints, to see whether the new transPMFred algorithm is able to detect real changepoints if there were no metadata available.

The results of applying the transPMFred algorithm to the four example series of > 0 daily precipitation amounts are summarized in Tables 2-3. Two significant Type-1 changepoints were identified for each of the three series listed in Table 2, which have very different distributions ($\hat{\lambda} = 0.2$, 0.0, and -0.2 for AMOS, Dawson, and Alert, respectively); while three significant Type-1 changepoints were identified for the The Pas series of > 0 daily precipitation amounts (Table 3), whose $\hat{\lambda} = -0.1$. All these statistically identified times of change are in good agreement with the times of change indicated in the related metadata, implying a success of the transPMFred algorithm in detecting shifts in non-zero daily precipitation data of different climate regimes (reflected by different λ values, which represent different distributions). The original and transformed series of > 0 daily precipitation amounts recorded at The Pas are shown in Figs. 3a and 3c, respectively, along with the estimated trend component and the identified changepoints.

After identifying all significant changepoints, one wish to adjust the precipitation data series to diminish all significant artificial shifts, i.e., to homogenize the precipitation data series. However, homogenization of daily precipitation series is a very challenging task. Next, we use the The Pas series to showcase the challenge, proposing some practical solutions.

Since daily precipitation is not a continuous variable (it is a mixed discrete/continuous variable), discontinuities could exist in both the frequency and the amount of precipitation measured. Due to changes in the unit (e.g., from imperial to metric), in measuring precision (gauge gradation scale), and in observing practises, etc., discontinuities often exist in the frequency of measured small precipitation and hence in the frequency of reported precipitation occurrence (or non-occurrence). For example, our metadata indicates that there were changes in the measuring precision of both rainfall and snowfall, which introduced a discontinuity in the frequency of small daily precipitation amount measured at The Pas, as shown in the upper panel of Fig. 6. That is, there was not any value ≤ 0.3 (dashed line) or in the interval $(0.4, 0.5]$ (dotted line) in this daily precipitation series before 1977; there were much fewer values in the interval $(0.3, 0.4]$ before 1946 (solid line). In other words, there are two obvious discontinuities in the frequency of measured small amounts, in 1945/1946 and 1976/1977, respectively. As noted in Table 3, the 1945/1946 discontinuity is related to the joining of two stations that used different gauge types, rim heights, and observation frequencies; and the 1976/1977 discontinuity is related to changes in gauge type and rim height, and also in the snowfall measuring precision, which changed from 0.1 inch (about 2.5 mm) to 2 mm (the corresponding change in the precision of snow water equivalent would be from 0.25 mm to 0.20 mm if the snowfall to water equivalent ratio is 10:1). There are also fewer values in the interval $(0.5, 1.0]$ before 1939, which is however not necessarily a frequency discontinuity; it could be a manifestation of a shift in the measured amounts.

The results of applying the transPMFred to the series of > 0.4 mm daily precipitation (Table 3, lower part) also suggest that the 1945/1946 and 1976/1977 discontinuities are mainly discontinuities in the frequency of small amounts (≤ 0.4 mm) measured, because they are not found in the series in which all ≤ 0.4 mm amounts were excluded (i.e. the series of > 0.4 mm daily precipitation), and that the 1938/1939 changepoint is not necessarily/entirely a frequency discontinuity because it is also found in this series of > 0.4 mm daily precipitation. In general, one can, and we recommend, test the series of $> P_{thr}$ daily precipitation amounts, varying P_{thr} over a set of small values that reflect changes in the measuring precision (such as 0.3, 0.4, 0.5, ...), to find out whether or not the detected discontinuity is a frequency discontinuity. This is because a frequency discontinuity due to an increase in measuring precision, say from δ_1 to δ_2 mm ($\delta_2 < \delta_1$), shall not exist in the series of $\geq \delta_1$ daily precipitation amounts, namely, the series in which the measured amounts that are below δ_1 are excluded.

The presence of frequency discontinuity could complicate the detection of other shifts. For example, the changepoint in June 1933, which is associated with changes in gauge type/rim-height and in observing frequency, was only identified in the series of > 0.4 mm daily precipitation, not in the series of > 0 daily precipitation (Table 3).

In particular, in the presence of frequency discontinuity the QM adjustments (and even the IBC adjustments) will not work properly; they will not be good physically even if they can make the series homogeneous. For example, for the The Pas series of > 0 daily precipitation, which is affected by the frequency discontinuities in 1945/1946 and 1976/1977, the QM adjustments (with $M_q = 4$, 8, and 16) fail to homogenize it: the 1945/46 and 1976/77 changepoints always remain in the QM adjusted series. The IBC adjustments also fail to homogenize this series of > 0 daily precipitation. In this case, even if the QM or IBC adjustments can make the series homogeneous, they are not good at all physically. For example, due to the increase in measuring precision in 1976/1977, 6.12245% of the > 0 daily precipitation amounts in the period of 1977-2005 are ≤ 0.3 mm, but the lowest 6.12245% of the > 0 daily precipitation in the early period, 1910-1976, are in the range of 0.31-0.51 mm (none of them are $\leq$ 0.3 mm), which would not belong to the lowest 6.12245% percentiles if there were no missed measurements of $\leq$ 0.3 mm daily precipitation in the early period. In this case, adjusting the 1910-1976 segment of data using a QM algorithm to the 1977-2008 segment will alter the lowest 6.12245% of the data in the period of 1910-1976, while these data of 1910-1976, ranging from 0.31 to 0.51 mm, were most likely correct measurements and should not be altered in any way. The error arises from the wrong estimate of the cumulative frequency (CF) for the data of 0.31-0.51 mm, mistaking them as a lower CF category data because the true lower CF category data are missing. Such a frequency mismatch will affect the adjustments for all frequency categories.

Actually, in the presence of frequency discontinuity, any adjustments that are derived from the measured precipitation amounts are not good, regardless of how they are estimated (by QM, IBC, or the ratio based method of Alexandersson 1986), because the problem is the failure to report the small amounts, not in the amounts measured. One must account for the frequency discontinuities before estimating and applying any adjustments to the measured amounts; otherwise, correct measurements could be mistakenly adjusted to make up the missed measurements of small amounts, namely a frequency mismatch will occur and void the adjustments.

It would be the best if one can fill in the missed measurements of small daily precipitation, for example, by assigning the smallest measured amount (i.e., the smallest amount that was ever measured during the period of the data record) to the days for which precipitation occurrence was reported with zero amount (i.e., the amount is too small to be measured back then and hence reported as zero). Of course, metadata (including flags in the data record, such as the flag used to indicate trace precipitation occurrence) and/or observations of other climate variables (such as current weather reports, cloud observations, ...) are needed to find out the days of reported precipitation occurrence with zero amount recorded.

If the above best approach is not possible, one can also do the following to fill in the missed measurements of small amounts: (a) use the PMFred algorithm to identify shifts in the monthly relative frequency series of the non-zero but small (≤ 0.3 mm) precipitation events and estimate the shift size Δ_f (e.g., $\Delta_f = 6\%$ for a jump from 0.5% to 6.5%, on average, in the relative frequency); (b) estimate the number of missed measurements of small amounts for each month in the early period as $M_{msg} = \Delta_f \times N_{obs}$, where N_{obs} is the total number of daily precipitation observations in the month; (c) randomly pick M_{msg} days out of the reported zero precipitation days in the month, then assign the smallest measured amount to each of these M_{msg} days and indicate with a flag that these are estimated values. Obviously, there is no guarantee that small precipitation did occur on all or some of these M_{msg} days. Thus, warning of this uncertainty should be given to data users. With this uncertainty, the discontinuity in the frequency of measured small precipitation is diminished (adjusted). Therefore, the resulting daily precipitation series is more suitable for analyzing trends in the frequency of rainy days that are defined as days of > 0 precipitation (a popular definition of rainy day). Also, without the complication by frequency discontinuity, other shifts are more likely to be detected by the transPMFred algorithm; and more proper adjustments for other shifts detected in this series can be estimated using a quantile matching algorithm. Correctly measured values will no longer be mistakenly adjusted to make up the missed measurements of small amounts; for example, the correctly measured values of 0.31-0.51 mm in the period of 1910-1976 at The Pas will not be mistakenly adjusted to make up the missed measurements of ≤ 0.3 mm amounts in this period.

For all climate variables in general, one can also use simultaneous observations of other climate variables (predictors) that are closely related to the variable to be homogenized (predictand) to fill in the missed measurements. A statistical relationship between the predictand and the predictor(s) can be established using data from a period of more reliable observations for both the predictand and predictor(s). This relationship is then used, along with the reliable observations of the predictor(s) in the period in which observations of the predictand are missed, to estimate the missed observations of the predictand. Along this line, Dai et al. (2010) present an example of estimating missed measurements of dew point depression (DPD), a humidity variable derived from relative humidity measurements by hygrometers. Noting that the early hygrometers were considered unreliable or not able to report on humidity under cold or dry conditions, they use a statistical relationship between DPD and vapour pressure and air temperature under the cold (or dry) conditions to estimate the corresponding missed measurements of humidity (and hence DPD) under the cold (or dry) conditions, before their attempt to homogenize the DPD data series using the QM adjustment algorithm for non-negative data proposed in this study.

One can also go around frequency discontinuities in daily precipitation data by testing the series of $> P_{thr}$ daily precipitation, varying P_{thr} over a set of small values that reflect changes in the measuring precision (e.g., 0.3, 0.4, 0.5, ...), and then homogenizing the series of $> P_{thr}$ daily precipitation data that was found to be free of frequency discontinuity. For example, the series of > 0.4 mm daily precipitation at The Pas, which contains two shifts (see Table 3, lower part), can be homogenized by the QM adjustments (with $M_q = 4$ or 8) or by the IBC adjustments (Figs. 3e,f). However, all the $\leq P_{thr}$ values in the series are left unadjusted, and thus the frequency discontinuities remain in the series of > 0 daily precipitation, affecting the values between 0 and P_{thr} inclusive. In this case, warning of such frequency discontinuity effect and advice to use only the homogenized series of $> P_{thr}$ daily precipitation should be given to data users. This is the drawback of this "go-around" approach. Fortunately, some applications do not need to include small daily precipitation amounts. For example, one may define rainy days as those of > 0.5 mm daily precipitation for analyzing trends in the frequency of rainy days, which is a commonly used climate index.

Another issue relevant to precipitation data homogeneity (and accuracy) is trace precipitation, which is particularly important for regions of high northern latitudes (such as Canada, Alaska, Siberia, and northern Europe), where trace and light precipitation prevail. In Canada, a zero daily precipitation is recorded with a flag T when precipitation occurred in that day but the amount is too small to be measured and hence reported as zero. Obviously, an increase in measuring precision could make an unmeasurable amount become measurable and thus decrease the frequency of reporting trace precipitation. Changes in the practise of trace reporting that took place in the period of data record could also cause discontinuities in the frequency series of reported trace occurrence. In other words, discontinuity is inevitable in the frequency series of reported trace occurrence; adding a trace amount to days of reported trace occurrence could introduce a discontinuity to the series of daily precipitation.

As mentioned before, the PMFred algorithm can be used to test the homogeneity of relative frequency series. For example, as shown in the lower panel of Fig. 6, two Type-1 changepoints (1945/1946 and 1955/1956) were identified by the PMFred algorithm from the series of annual relative frequency of reported trace precipitation occurrence at The Pas. In this case, we can do the following to obtain the adjusted annual totals of trace precipitation: (a) correct the daily precipitation data for trace amounts over the most recent period (after 1955 in the example), using the method of Mekis and Hogg (1999); (b) calculate the mean trace rate by dividing the total trace amount with the total count of T-flags over the most recent period, so that this mean trace rate is in unit (mm) per T-flag, and it is an average over the most recent period; (c) adjust the annual relative frequencies of T flag for the early periods (1946-1955 and 1910-1945) to the most recent period using the multi-phase regression model fit of the PMFred algorithm to this series; (d) calculate the adjusted annual count of T flags for each year in the early periods by multiplying the adjusted annual relative frequency with the corresponding count of precipitation observations; (e) calculate the adjusted annual trace amount for each year in the early periods by multiplying the adjusted annual count of T flags with the mean trace rate derived in (b), and add the annual trace amount to the annual total precipitation to adjust it for the discontinuities in the frequency of reported trace occurrence. Similarly, the PMFred algorithm can be used to check the homogeneity of the series of monthly relative frequency of trace precipitation occurrence; then, adjustments to monthly total trace precipitation amounts can be derived. A different mean trace rate can be derived for each calendar month, separately, to account for any annual cycle in this quantity (or for trace rainfall and trace snowfall separately, if desired).

We would also like to point out that, for the example series from Canada, seasonal cycle is mainly in the frequency of precipitation occurrence and is negligible in the daily precipitation amounts, so that we can pool all non-zero daily precipitation amounts together for the test.

When seasonal cycle is not negligible in the daily precipitation amounts, one can pool all non-zero daily precipitation amounts in a season together, and test the pooled precipitation series for each season, separately (note that season is a general term here; it can be defined as a calendar month). This way, the seasonal cycle is accounted for by allowing a different transformation for different seasons, namely, different distributions of precipitation amounts and hence different variances in different seasons. The trend is also allowed to be different in different seasons. A disadvantage of this is that the same shift could be identified at slightly different times in the series for different seasons due to sampling variability and estimation error. Further analysis is needed to summarize the results of the tests, just like the case of testing monthly mean temperature series for each season (or each of the 12 calendar months) separately.

7. Concluding remarks

We have integrated a Box-Cox power transformation procedure into the PMFred algorithm (which is based on the common trend two-phase regression model) to make it applicable for detecting changepoints in non-Gaussian data series, such as non-zero daily precipitation amounts or wind speeds or dew point depression data (which are all non-negative and non-Gaussian). The transformation procedure is also applicable to the two-phase regression model that allows the trend to change at the time of mean shift (Lund and Reeves 2002). We have also assessed the detection power aspects of the new transPMFred algorithm by conducting a simulation study, the results of which show that the transPMFred algorithm is much better than the corresponding untransformed method, PMFred, for data of different non-Gaussian distributions (reflected by the different λ values ranging from -0.2 to 0.2). The transformation procedure can increase the hit rate by as much as 70% (see section 4).

A software package called RHtests_dlyPrcp, which contains a set of functions for use in detection and adjustments of shifts in non-zero daily precipitation amounts have also been developed and made available online at http://cccma.seos.uvic.ca/ETCCDMI/software.shtml.

This includes a quantile matching (QM) algorithm for adjusting shifts in non-zero daily precipitation series, which is applicable to other non-negative data such as wind speeds and dew point depression. Actually, we have developed a function, QMadj.nonNegativeDLY, for conducting QM adjustments to non-negative daily data, and another function, QMadj.GaussianDLY, for conducting QM adjustments to Gaussian daily data such as daily temperatures. (The major difference between these two functions is in the lower boundary condition, as described earlier in section 5.)

Noting that frequency discontinuities are often inevitable due to changes in the measuring precision, and that they could complicate the detection of shifts in non-zero daily precipitation data series and void any attempt to homogenize the series, we recommend use the transPMFred to test the series of daily precipitation amounts that are larger than a threshold value, varying the threshold over a set of small values that reflect changes in the measuring precision over time, to gain some insight into the characteristics of the discontinuities detected statistically. In the mean time, one shall also test the homogeneity of the frequency series of reported zero and various small precipitation events; and one can use the PMFred algorithm (the untransformed version) to do so. When a frequency discontinuity is present and left unaccounted for, adjustments derived from the measured daily amounts could make the data deviate more from the truth, regardless of the method used to derive the adjustments (e.g., QM or IBC or the ratio based method of Alexandersson 1986). In this case, one must account for all frequency discontinuities before attempting to adjust the measured amounts. We have also proposed approaches to account for detected frequency discontinuities, to fill in the missed measurements of small precipitation or the missed report of trace precipitation (section 6). It must be stressed that caution has to be exercised when homogenizing daily precipitation data series, not to forget or ignore the possible complication by frequency discontinuity.

The common trend assumption in the transPMFred (and PMFred) algorithm is not a problem for Canadian precipitation data series, which are mostly less than 100 years in length and mostly have no significant low-frequency variations. However, it may not be valid for other precipitation data series. The common approach to account for low-frequency variations has been to use a reference series that has the same linear trend and periodic components (including annual cycle and low-frequency variations). However, the extremely high spatial variability and non-continuity of daily precipitation make it unrealistic to find a suitable reference series for use in homogenization of daily precipitation series. A new method that allows non-linear trends (consisting of periodic components and a linear trend) in the time series being tested and thus does not need the common trend assumption is being developed and will soon be reported in a separate study. Soon we will be able to drop the common trend assumption from the PMFred and transPMFred algorithms. The related software package will be updated accordingly.

A common variance throughout the period of data record is also assumed in the SNH, PMT, and PMF tests, and thus in the PMFred and transPMFred algorithms. However, the common variance assumption may be not valid for climate variables in the changing climate; variance seasonality often exists in climate variables. As briefly discussed in the end of section 6, one can apply the test to daily precipitation series in each season, separately, thus allowing the trend, the distribution, and hence the variance to be different in different seasons of year. Alternatively, the QM adjustment algorithms can easily be modified to allow different trends in the series of data of different frequency categories, as was done in Dai et al. (2010), so that the distribution is allowed to change gradually over time. However, splicing one data series into several (say $M > 1$) series significantly decreases the sample size (from N to N/M for estimating the linear trend in each season or each of the M_q frequency categories), increasing the sampling uncertainty and hence uncertainty in the estimates of trends and the adjustments. One should seek to balance between sampling uncertainty (or estimation uncertainty) and the limitation of the assumptions. Note that the common variance assumption is also relaxed by allowing the use of a chosen part of the segment (instead of the whole segment) to estimate the CDF and hence the QM adjustments for use to homogenize the series, as detailed earlier in section 5. This could also increase the estimation uncertainty.

Acknowledgements

The authors wish to thank Lucie Vincent for her helpful internal review comments on an earlier version of this manuscript and Hui Wan for her help in preparing the 11 Chinese data series. The Climate Research Division of the Atmospheric Science and Technology Directorate of Environment Canada is acknowledged for supporting this research through Grant and Contributions ($\#KM040 - 05 - 0016 - IP$). The three anonymous reviewers are also acknowledged for their helpful comments on an earlier version of this paper.

References

Alexandersson, H., 1986: A homogeneity test applied to precipitation data. *J. Climatol.*, **6**, 661-675.

Atkinson, A.C. (1985). *Plots, Transformations and Regression.* Oxford University Press.

Box, G.E.P. and Cox, D.R. (1964). An analysis of transformation (with discussion). *Journal of Royal Statistical Society Ser. B*, **26**, 211-246.

Chen, H. (1995). Tests following transformations. *Annals of Statistics*, **23**, 1587-1593.

Chen, H. and Loh, W.Y. (1992). Bounds on AREs of tests following Box-Cox transformations. *Annals of Statistics*, **20**, 1485-1500.

Dai, A., J. Wang, P. W. Thorne, D. E. Parker, L. Haimberger, and X. L. Wang, 2010: Can we homogenize radiosonde humidity data? (in preparation for submission to *J. Clim.*).

Della-Marta, P. M. and H. Wanner, 2006: A Method of Homogenizing the Extremes and Mean of Daily Temperature Measurements. *J. Climate*, **19**, 4179-4197.

Environment Canada, 1977: Manual of surface weather observations. 7th ed. 419pp.

Llanso, P. (2003). Guidelines on climate metadata and homogenization. WMO Tech. Doc. 1186, 51 pp.

Mekis, E. and R. Brown, 2010: Derivation of an adjustment factor map for the estimation of the water equivalent of snowfall from ruler measurements in Canada. *Atmosphere-Ocean* (submitted).

Katz, R. W. and M. B. Parlange, 1996: Mixtures of stochastic processes: application to statistical downscaling. *Climate Research*, **7**, 185-193.

Katz, R. W. and M. B. Parlange, 1993: Effects of an Index of Atmospheric Circulation on Stochastic Properties of Precipitation. *Water Resources Res.*, **29**, i2335-2344.

Lund, R. and Reeves, J. (2002). Detection of Undocumented Changepoints: A Revision of the Two-Phase Regression Model. *J. Climate*, **15**, 2547-2554.

Mekis, E. and W. D. Hogg, 1999: Rehabitation and analysis of Canadian daily precipitation time series. Atmosphere-Ocean, **37**, 53-85.

Reeves, J., Chen, J., Wang, X.L., Lund, R., and Lu, Q. (2007). A review and comparison of changepoint detection techniques for climate data. *J. of App. Meteor. Climatol.*. **46**, 900-915.

Seber, G.A.F. (1977). *Linear Regression Analysis.* John Wiley & Sons, New York.

Solow, A. (1987). Testing for climatic change: an application of the two-phase regression model. *J. Climate Appl. Meteorol.*, **26**, 1401-1405.

Trewin, B. C. and A. C. F. Trevitt, 1996: The Development of Composite Temperature Records. *Int. J. Climatol.*, **16**, 1227-1242.

Vincent, L. A., 1998: A technique for the identification of inhomogeneities in Canadian temperature series. *J. Climate*, **11**, 1094-1104.

Wang, X.L., 2008a: Penalized maximal F test for detecting undocumented mean shift without trend change. *J. Atmos. Oceanic Technol.*, **25**, 368-384. DOI: 10.1175/2007/JTECHA982.1

Wang, X.L., 2008b: Accounting for Autocorrelation in Detecting Mean Shifts in Climate Data Series Using the Penalized Maximal t or F test. *J. App. Meteor. Climatol.* **47**, 2423-2444. DOI:10.1175/2008JAMC1741.1.

Wang, X.L., 2006: Climatology and trends in some adverse and fair weather conditions in Canada, 1953-2004. *J. Geophys. Res.*, 111, D09105, doi:10.1029/2005JD006155.

Wang, X.L., and H. Cho, 1997: Spatial-Temporal Structures of Trend and Oscillatory Variabilities of Precipitation over Northern Eurasia. *J. Clim.*, 10, 2285-2298.

Wang, X.L. and Y. Feng, published online 2010: *RHtestsV3 User Manual.* Available online at http://cccma.seos.uvic.ca/ETCCDMI/RHtest/RHtestsV3_UserManual.doc. Climate Research Division, Science and Technology Branch, Environment Canada, Toronto, Ontario, Canada. 26 pp.

Wang, X. L., Q. H. Wen, and Y. Wu, 2007: Penalized maximal t test for detecting undocumented mean change in climate data series. *J. Appl. Meteor. Climatol.*, **46**(No. 6), 916-931. DOI:10.1175/JAM2504.1

Wang, X.L., 2003: Comments on "Detection of Undocumented Changepoints: A Revision of the Two-Phase Regression Model", *J. Clim.*, 16, 3383-3385.

Woolhiser, D. A., 1992: Modelling daily precipitation - progress and problems. In: *Statistics in the Environmental & Earth Sciences* (eds.:A. T. Walden and P. G. Guttorp). Co-published in the Americas by Halsted Press, in imprint of John Wiley & Sons Inc., New York.

Figure captions

Figure 1. Comparison of transPMFred with PMFred in terms of the estimated hit rates HR_U (a); and the HR_U as a function of λ value (b), and of changepoint position K (c). The series length is $N = 600$ for (a) and (b), and $N = 500$ for (c). The size of the shift inserted is expressed in unit of the standard deviation (std) of the homogeneous surrogate series of N non-zero values (generated by block-bootstrapping). A hit is registered when the estimated changepoint time $\hat{c}$ is within the interval $[K - 10, K + 10]$ of the actual changepoint time K. Each value was obtained from 1000 simulations.

Figure 2. The IBC and QM adjustments made to the series of > 0.4 mm daily precipitation recorded at The Pas. The series is adjusted to the latest segment in both cases.

Figure 3. Series of (a) > 0 and (b) > 0.4 mm daily precipitation amounts recorded at The Pas (Manitoba, Canada) in the period from 1 June 1910 to 30 November 2008, (c-d) the corresponding Box-Cox transformed series ($\hat{\lambda} = -0.1$ and $\hat{\lambda} = -0.2$, respectively), and (e-f) IBC and QM adjusted series ($M_q = 4$ for the QM adjustments). The magenta trend lines are the estimated monotonic trend components (or linear trends for the transformed series) with changepoints at statistically identified times.

Figure 4. The distribution of the QM adjustments over empirical cumulative frequency (i.e., the $F(l) \sim A(s, l)$ relationship) for each segment of the AMOS series that needs to be adjusted (step lines) and the corresponding fitted natural cubic splines. The cumulative distribution functions (CDFs) are estimated with the indicated M_q values, using either all available data (all years; panels a-b) or up to 30 or 10 years of data (panel c or d) before and after the respective changepoint.

Figure 5. Annual relative frequency series of measured small precipitation (upper) and reported trace precipitation (lower) at station The Pas.

Table 1. Empirical hit rates (HR; in %) and false alarm rates (FAR; in %) of the (a) transPMFred and (b) PMFred algorithms applied to non-zero daily data of various λ values (representing different distributions). The nominal level of significance is 5%. Each entry is obtained from 1000 simulations of series of length $N = 600$ (i.e., series of 600 non-zero values) for one and no shift cases, and $N = 900$ for the two shifts cases. A hit is registered when the estimated changepoint time $\hat{c}$ is within the interval $[K - 10, K + 10]$ of the actual changepoint time K. The shift size is expressed in units of the standard deviation σ of the homogeneous non-zero daily precipitation series. The subscripts L and U represent the lower and upper bounds of the rate, respectively, which form the 95% range of the rate corresponding to the 95% uncertainty range of the lag-1 autocorrelation estimate (see Wang 2008b for more details).

a. transPMFred

	HR_L, HR_U for series of 1 shift of sizes (unit:σ):				HR_L, HR_U for series of 2 shifts of sizes(unit:σ):				FAR_L, FAR_U
λ	0.5	1.0	1.5	2.5	0.5 & 1.0	1.0 & 1.0	1.0 & 1.5	2.0 & 2.5	(no shift)
-0.2	72.6, 77.2	94.1, 95.8	97.5, 97.9	99.3, 99.6	70.5, 73.9	79.8, 82.3	81.3, 82.6	87.2, 88.3	1.1, 6.1
-0.1	83.9, 86.3	97.7, 98.1	99.1, 99.5	99.9, 99.9	75.5, 78.1	83.9, 86.3	84.7, 86.3	87.1, 87.8	1.5, 10.8
0.0	75.4, 80.0	97.3, 97.9	99.6, 99.6	100.0, 100.0	74.4, 77.1	83.3, 86.2	86.0, 87.3	88.6, 89.4	0.7, 8.6
0.1	63.5, 68.3	94.2, 94.7	99.2, 99.5	100.0, 100.0	69.6, 72.6	82.1, 85.1	85.6, 86.8	86.5, 87.0	2.9, 12.8
0.2	82.0, 84.2	97.8, 98.5	99.6, 99.5	100.0, 100.0	78.1, 79.9	85.4, 88.0	88.4, 89.8	88.2, 88.8	1.4, 8.5

b. PMFred

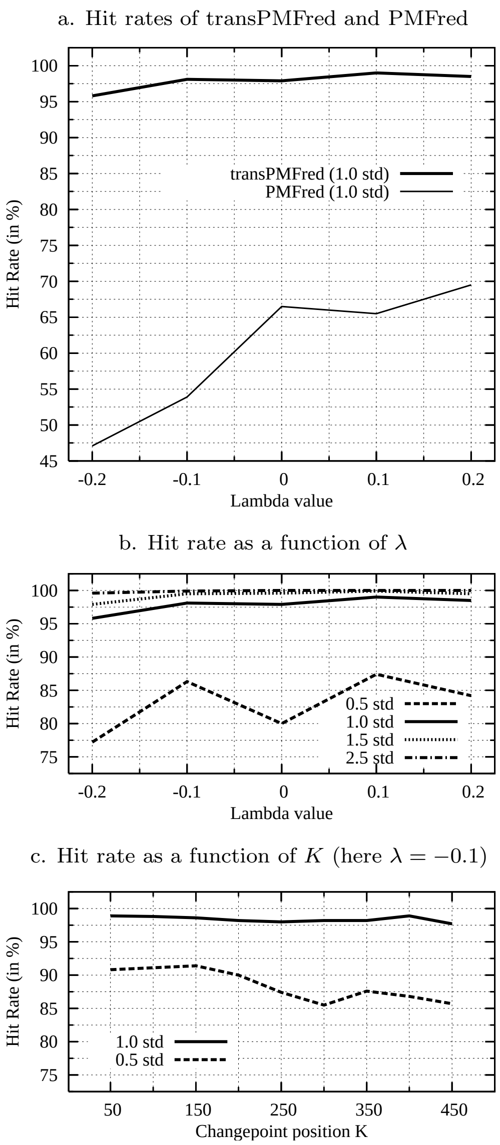
	HR_L, HR_U for series of 1 shift of sizes (unit:σ):				HR_L, HR_U for series of 2 shifts of sizes(unit:σ):				FAR_L, FAR, FAR_U
λ	0.5	1.0	1.5	2.5	0.5 & 1.0	1.0 & 1.0	1.0 & 1.5	2.0 & 2.5	(no shift)
-0.2	5.4, 11.8	39.7, 47.1	57.4, 63.4	74.0, 79.5	29.0, 34.5	43.6, 50.1	57.6, 61.3	72.9, 74.9	1.7, 9.6
-0.1	12.8, 23.3	47.4, 53.9	65.7, 72.7	76.8, 82.5	37.2, 42.2	51.2, 56.7	64.2, 67.6	75.8, 77.4	3.2, 11.1
0.0	13.2, 23.4	57.7, 66.5	79.4, 83.4	92.1, 93.9	41.0, 47.5	58.2, 63.8	70.8, 73.6	79.5, 80.5	0.8, 8.9
0.1	9.4, 17.3	52.5, 58.5	72.7, 78.1	87.1, 89.6	36.3, 40.9	53.4, 58.1	67.4, 70.4	76.5, 78.3	3.8, 12.5
0.2	17.4, 26.2	63.3, 69.5	82.2, 86.2	91.8, 93.1	47.8, 52.4	62.6, 67.7	72.5, 75.1	80.9, 82.7	1.8, 10.1

Table 2. Results of applying the transformed PMFred algorithm to the no-zero ($P > 0$) daily precipitation series for indicated stations, and the related metadata. Date (YYYY.MM.DD) ranges indicate the period in which the change(s) happens (the exact date is unknown). The CV range denotes the critical value range [$\mathrm{PF}_{max,0.05}(\hat{\phi}^L, N_{seg})$, $\mathrm{PF}_{max,0.05}(\hat{\phi}^U, N_{seg})$] (Wang 2008b).

Station (IDs)		Change	PFmax	
(Period tested)	Type	Date	(CV range)	Documented date of change or changes
AMOS,	1	1935.09.27	58.61	mid-1930s: Gauge type change from British gauges
Quebec		(i =2612)	(16.40, 18.36)	to MSC gauge in Canada (Metcalfe et al. 1997).
(7090120)	1	1987.03.27	16.71	1986.05.15-1988.08.11: Gauge type change from MSC
(1914.01.01-		(i =10178)	(16.20, 18.11)	standard to Type B, Nipher rim height change
1999.06.30)				from 153 cm to 161 cm, and graduate replaced.
$\hat{\lambda} = 0.2$				
Dawson A,	1	1939.10.23	39.42	1936.06.30 - 1940.09.16: rim height change from 13.5 inches
Yukon		(i =3427)	(18.79, 20.99)	to 24 inches and exposure change from bad to good.
(2100400,	1	1992.10.03	20.05	1992.07.30 - 1993.05.28: Nipher gauge exposure change
2100402)		(i =9930)	(18.20, 20.30)	from ”does not meet requirement” to good, receiver
(1906.01.01-				leaking and poor snow ruler conditions (worn
2007.02.28)				ends and scale markers) reported on 30 July 1992.
$\hat{\lambda} = 0.0$				
Alert,	1	1964.05.02	56.25	1958.04.13 - 1966.06.15: Change from American
Nunavut		(i =1336)	(18.41, 26.55)	Equipment to MSC standard gauge, Nipher
(2400300,				and snow ruler introduced.
2400306)	1	1997.06.10	53.03	1993.01.30 - 1997.09.14: Fisher Porter weighing gauge
(1950.07.01-		(i =4873)	(18.39, 26.52)	replaced by Belfort weighing gauge.
2007.09.21)				
$\hat{\lambda} = -0.2$				

Table 3. Same as in Table 2 but for the results of applying the transformed PMFred algorithm to the no-zero ($P > 0$ and $P > 0.40$ mm) daily precipitation series for station The Pas, and the related metadata.

Station (IDs) (Period tested)	Type	Change Date	PFmax (CV range)	Documented date of change or changes
The Pas, Manitoba (5052864, 5052880)	1	1938.07.04 (i =2283)	46.26 (16.00, 17.81)	1937.10.09-1938.08.08: Gauge type change (ordinary to MSC gauge), rim height change (12 inches to 15 inches), observation frequency change (twice daily to three times daily), poor gauge condition (no level, no rim circular) reported on 9 October 1937.
(1910.06.01- 2008.11.30) $\hat{\lambda} = -0.1$	1	1946.10.24 (i =3237)	62.36 (16.65, 18.59)	1945.12.31: joining of two nearby stations with different gauge types (Standard vs MSC), rim heights (12 inches vs 13 inches), and observation frequencies (twice daily vs 6-hourly).
$P > 0$	1	1976.10.04 (i =7272)	25.98 (17.85, 20.01)	1975.10.16 - 1977.10.18: Gauge type change (MSC to Type B), rim height change (13 inches to 16 inches); snowfall measuring precision change (0.1 inch (2.5 mm) to 2 mm) (snowfall water equivalent precision change from 0.25 mm to 0.2 mm when the 10:1 ratio was assumed; Environment Canada 1977).
The Pas $P > 0.4$ mm	1	1933.06.06 (i = 1832)	33.86 (14.47, 16.23)	1932.08.01-1937.10.09: poor rain gauge exposure reported on 1 Aug. 1932; rim height change from 19 inches to 12 inches.
	1	1938.07.04 (i = 2264)	57.05 (16.53, 18.78)	1937.10.09-1938.08.08: Gauge type change (ordinary to MSC gauge), rim height change (12 inches to 15 inches), observation frequency change (twice daily to three times daily), poor gauge condition (no level, no rim circular) reported on 9 October 1937.



a. Hit rates of transPMFred and PMFred

b. Hit rate as a function of λ

c. Hit rate as a function of K (here $\lambda = -0.1$)

Figure 1. Comparison of transPMFred with PMFred in terms of the estimated hit rates HR_U (a); and the HR_U as a function of λ value (b), and of changepoint position K (c). The series length is $N = 600$ for (a) and (b), and $N = 500$ for (c). The size of the shift inserted is expressed in unit of the standard deviation (std) of the homogeneous surrogate series of N non-zero values (generated by block-bootstrapping). A hit is registered when the estimated changepoint time $\hat{c}$ is within the interval $[K - 10, K + 10]$ of the actual changepoint time K. Each value was obtained from 1000 simulations.

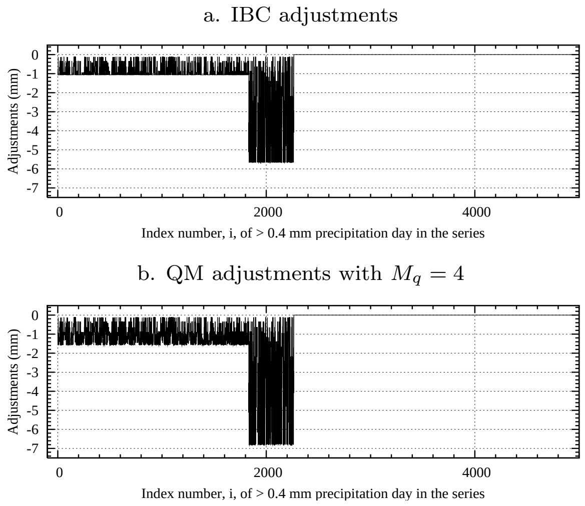
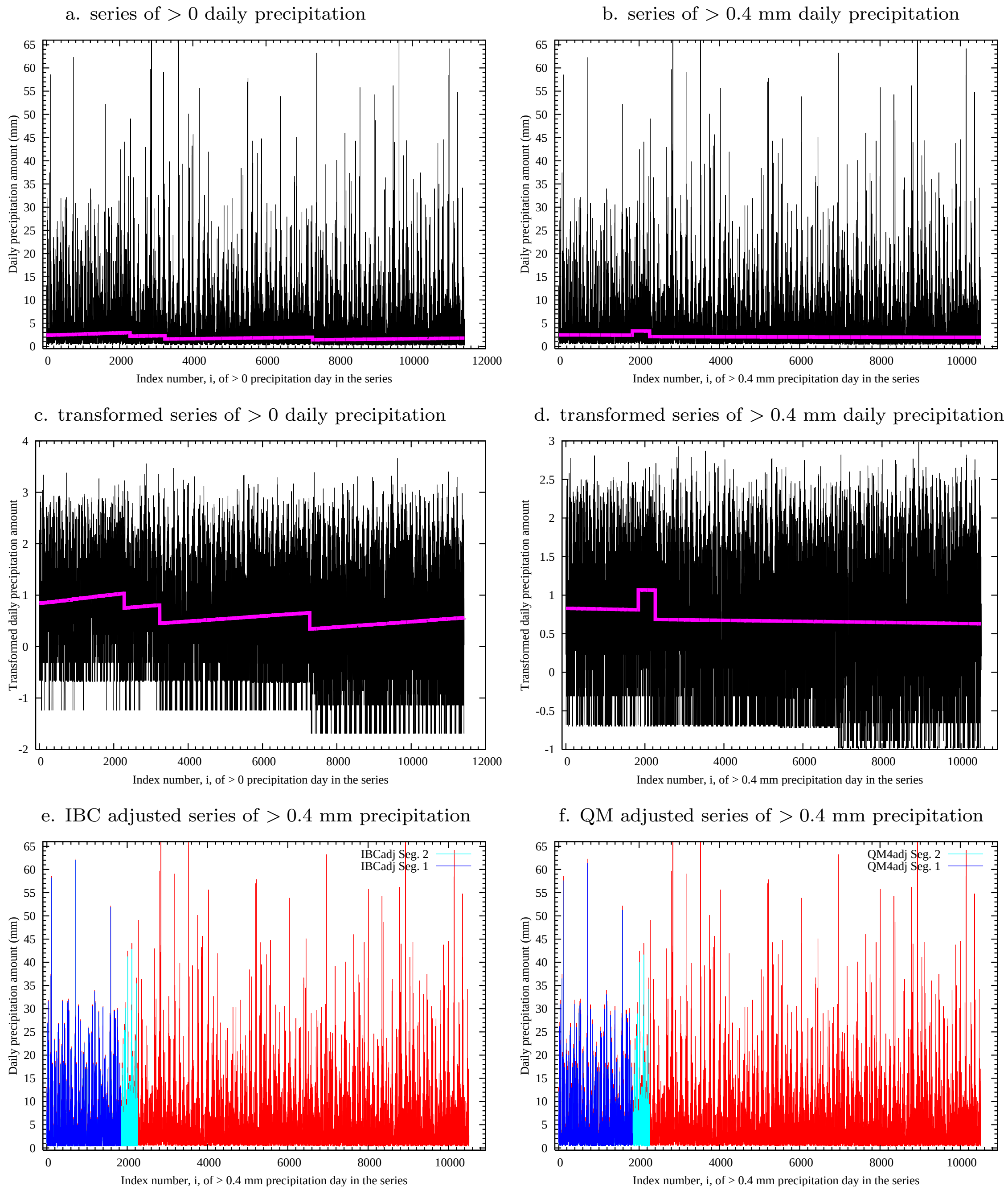


Figure 2. The IBC and QM adjustments made to the series of $>$ 0.4 mm daily precipitation recorded at The Pas. The series is adjusted to the latest segment in both cases.



a. series of > 0 daily precipitation

b. series of > 0.4 mm daily precipitation

c. transformed series of > 0 daily precipitation

d. transformed series of > 0.4 mm daily precipitation

e. IBC adjusted series of > 0.4 mm precipitation

f. QM adjusted series of > 0.4 mm precipitation

Figure 3. Series of (a) > 0 and (b) > 0.4 mm daily precipitation amounts recorded at The Pas (Manitoba, Canada) in the period from 1 June 1910 to 30 November 2008, (c-d) the corresponding Box-Cox transformed series ($\hat{\lambda} = -0.1$ and $\hat{\lambda} = -0.2$, respectively), and (e-f) IBC and QM adjusted series ($M_q = 4$ for the QM adjustments). The magenta trend lines are the estimated monotonic trend components (or linear trends for the transformed series) with changepoints at statistically identified times.

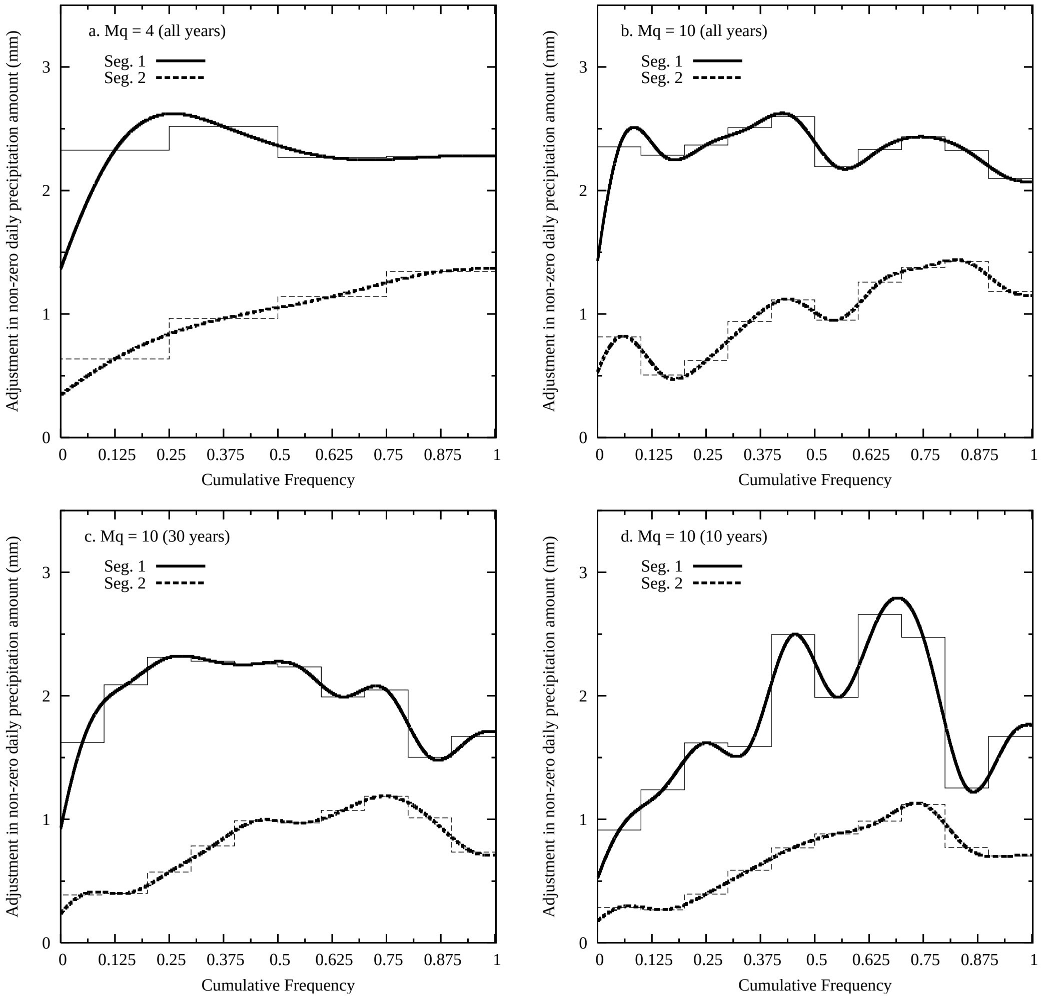


Figure 4. The distribution of the QM adjustments over empirical cumulative frequency (i.e., the $F(l) \sim A(s, l)$ relationship) for each segment of the AMOS series that needs to be adjusted (step lines) and the corresponding fitted natural cubic splines. The cumulative distribution functions (CDFs) are estimated with the indicated M_q values, using either all available data (all years; panels a-b) or up to 30 or 10 years of data (panel c or d) before and after the respective changepoint.

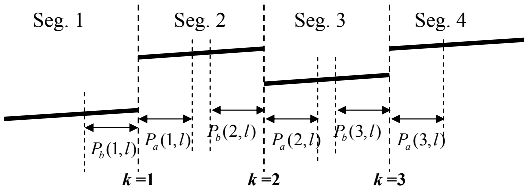


Figure 5. An example of the chosen part of segment for use to estimate the differences in the cumulative distribution (CDF) caused by a shift.

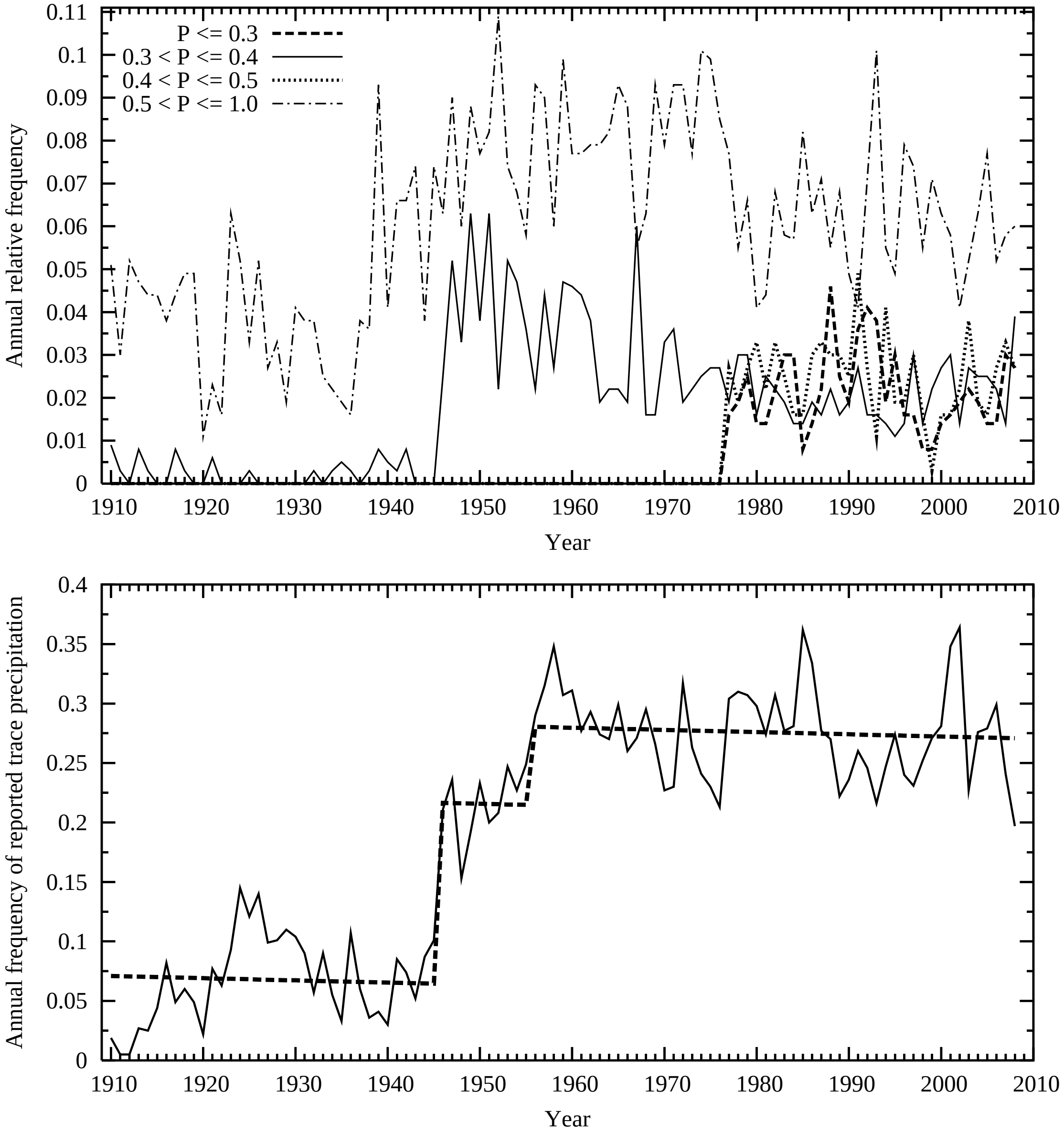


Figure 6. Annual relative frequency series of measured small precipitation events (upper) and reported trace precipitation (lower) at station The Pas.