Penalized Maximal $F$ Test for Detecting Undocumented Mean Shift without Trend Change

XIAOLAN L. WANG

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

(Manuscript received 24 January 2007, in final form 13 June 2007)

ABSTRACT

In this study, a penalized maximal $F$ test (PMFT) is proposed for detecting undocumented mean shifts that are not accompanied by any sudden change in the linear trend of time series. PMFT aims to even out the uneven distribution of false alarm rate and detection power of the corresponding unpenalized maximal $F$ test that is based on a common-trend two-phase regression model (TPR3). The performance of PMFT is compared with that of TPR3 using Monte Carlo simulations and real climate data series.

It is shown that, due to the effect of unequal sample sizes, the false alarm rate of TPR3 has a W-shaped distribution, with much higher than specified values for points near the ends of the series and lower values for points between either of the ends and the middle of the series. Consequently, for a mean shift of certain magnitude, TPR3 would detect it with a lower-than-specified level of confidence and hence more easily when it occurs near the ends of the series than somewhere between either of the ends and the middle of the series; it would mistakenly declare many more changepoints near the ends of a homogeneous series. These undesirable features of TPR3 are diminished in PMFT by using an empirical penalty function to take into account the relative position of each point being tested. As a result, PMFT has a notably higher power of detection; its false alarm rate and effective level of confidence are very close to the nominal level, basically evenly distributed across all possible candidate changepoints. The improvement in hit rate can be more than 10% for detecting small shifts ($\Delta \leq \sigma$, where $\sigma$ is the noise standard deviation).

1. Introduction

In the recent decades, climate change has been a subject of ongoing debate, which is also in part related to the inevitable discontinuities in long-term climate data records that hamper the reliability of the results of historical climate trend assessment, climate change detection, and attribution. The effects of artificial shifts (nonclimatic sudden changes) on the results of climate analysis, especially of historical climate trend assessment, have been illustrated in many studies, for example in Wang (2006), Hanesiak and Wang (2005), Vincent et al. (2002), and Easterling and Peterson (1995), among many others. It has become common knowledge that a realistic and reliable assessment of historical climate trends and variability is hardly possible without a long-term, homogeneous time series of climate data. Thus, more and more effort has been devoted to detecting and adjusting for artificial shifts in climate data series, which are typically due to changes in instrument, in observing location and/or environment, in observing practices/procedure, etc. Some of these changes are documented in the related metadata database, while others are not (due to inaccurate and/or incomplete metadata). Therefore, there exist documented and undocumented mean shifts. With the time of shift documented in metadata (so that one does not need to identify it statistically), documented shifts are much easier to assess; they can be tested using the regular tests of means or variances (e.g., Student’s $t$ test or $F$ test). However, one needs to rely on an appropriate statistical test to detect and assess undocumented shifts (i.e., those that have no metadata support).

Several statistical methods have been proposed for detecting undocumented shifts. Also, there have been several studies that review and compare different methods for testing artificial shifts (e.g., Easterling and Peterson 1995; Peterson et al. 1998; DeGaetano 2006).
Most comprehensively, Reeves et al. (2007) review and compare eight of them, including (but not limited to) the standard normal homogeneity test (SNHT) for mean shifts in time series of zero trend (Alexandersson 1986), the common-trend two-phase regression model–based maximal $F$ test (TPR3) for mean shifts in time series with a constant linear trend throughout the series (Wang 2003), and the two-phase regression model–based maximal $F$ test for a sudden change that occurs simultaneously in both the mean and the linear trend of the series (TPR4; Lund and Reeves 2002). They show that TPR3 seems optimal for most climate data time series, while SNHT is probably best when trend and periodic effects can be diminished by using homogeneous reference series.

However, as noted in Wang et al. (2007, hereafter WW07) the effect of unequal sample sizes results in uneven distributions of false alarm rate (FAR) and detection power of the SNHT type or the maximal $F$ tests. (Why there exists the problem of unequal sample sizes in this type of test is also explained in section 2 below.) WW07 show that the SNHT-type tests have a U-shaped detection power than the SNHT-type tests; its FAR on the location of the shift in the time series (near the ends or the middle), which is undesirable because one wishes to detect a shift of detectable magnitude at the nominal level of confidence/significance regardless of its location. To diminish such undesirable effects, WW07 propose a penalized maximal $t$ test (PMT) for detecting undocumented mean shifts in time series of zero trend, using an empirically constructed penalty function. It is shown that PMT has a significantly higher power of detection than the SNHT-type tests; its FAR is close to the nominal level for all admissible changepoint positions.

Similarly, this study aims to improve TPR3, the common-trend two-phase regression model–based maximal $F$ test for undocumented mean shifts that are not accompanied by any sudden change in the linear trend of time series (the most commonly encountered type of artificial shifts in climate data series; Wang 2003). For this purpose, a penalty function is constructed empirically and imposed on the test statistic of TPR3 to diminish the effect of unequal sample sizes on the FAR and detection power of the test. The focus here is on time series with identically and independently distributed (IID) Gaussian errors that might have at most one changepoint (AMOC). However, it is generally common practice to implement statistical tests that are developed for the AMOC case with an appropriate stepwise testing algorithm to detect multiple changepoints in a climate data time series, as in Wang (2008), Menne and Williams (2005), and Wang and Feng (2007). Nevertheless, readers are referred to Caussinus and Mestre (2004) and Davis et al. (2006) for statistical methods that address directly the issue of detecting multiple changepoints in different settings. Readers are also referred to Wang and Feng (2007), and Wang (2008) for algorithms for testing undocumented and documented artificial mean shifts in tandem.

This article proceeds as follows. The penalized maximal $F$ test (PMFT) and how it is constructed are detailed in section 2 below. Then, the performance of PMFT is compared with that of TPR3 using Monte Carlo simulations in section 3, and using three climate data series from Canadian stations in section 4. Section 5 completes this study with some concluding remarks.

2. Penalized maximal $F$ test

Let $\varepsilon$ denote an IID Gaussian variable of zero mean and unknown variance $\sigma^2$. As proposed by Wang (2003), to test whether or not there exists a mean shift at time $t = k$ in time series $\{X_t\}$ with linear trend $\beta$ is to test the null hypothesis

$$H_0: X_t = \mu + \beta t + \varepsilon_t, \quad t = 1, 2, \ldots, N$$

against the alternative

$$H_a: \begin{cases} X_t = \mu_1 + \beta t + \varepsilon_t, & t \leq k \\ X_t = \mu_2 + \beta t + \varepsilon_t, & k - 1 \leq t \leq N, \end{cases}$$

where $\mu_1 \neq \mu_2$. When $H_a$ is true, the point/time $t = k$ is called a changepoint, and $\Delta = |\mu_1 - \mu_2|$ is called the magnitude of mean shift (or step size alternatively). As shown in Wang (2003), the most probable point to be the changepoint is the one that is associated with

$$F_{\text{max}} = \max_{1 \leq k \leq N-1} F_c(k),$$

where

$$F_c(k) = \frac{(SSE_0 - SSE_A)}{SSE_A/(N - 3)},$$

$$SSE_A = \sum_{i=1}^{k} (X_i - \hat{\mu}_1 - \hat{\beta}t)^2 + \sum_{i=k+1}^{N} (X_i - \hat{\mu}_2 - \hat{\beta}t)^2,$$
and

$$SSE_0 = \sum_{i=1}^{N} (X_i - \hat{\mu}_0 - \hat{\beta}_0 t)^2,$$

where $\hat{\mu}_0$ and $\hat{\beta}_0$ are estimated under the constraint $\mu_1 = \mu_2 = \mu$ (i.e., $X_t = \mu_0 + \beta_0 t + e_t$ for $t = 1, 2, \ldots, N$). Note that under the null hypothesis of no changepoint and assuming IID Gaussian errors $e_t$, the $F_{(k)}$ statistic above has an $F$ distribution of one numerator degree of freedom and $(N - 3)$ denominator degrees of freedom (Wang 2003). Thus, the test using the $F_{(k)}$ statistic above is called the maximal $F$ test (also referred to as TPR3 hereafter). Such hypothesis testing works with a preset or nominal level of significance $\alpha$ (or confidence $p = 1 - \alpha$). A mean shift or changepoint is declared if the $F_{(k)}$ statistic calculated for the time series being tested is greater than the critical value corresponding to the nominal level of significance; otherwise, the time series is declared homogeneous at the nominal level of significance/confidence (Wang 2003). The nominal significance level $\alpha$ represents the rate at which the test mistakenly declares a change point for a homogeneous time series, which is also called the FAR or type-I error rate. That is, $\alpha = \text{FAR}$. The most common choice is $\alpha = 0.05$ (i.e., $p = 0.95$), which is also used in this study.

When $H_a$ is true, the data series can be viewed as two samples: one of size $N_1 = k$ (for $t = 1, 2, \ldots, k$), which has mean response ($\mu_1 + \beta t$); another of size $N_2 = (N - k)$ (for $t = k + 1, k + 2, \ldots, N$), which has mean response ($\mu_2 + \beta t$). The sizes of the two samples are not equal generally; that is, $N_1 \neq N_2$ except when $k = N/2$ (i.e., when the point being tested is the midpoint of the data series). Unequal sample sizes can have huge impact on the results of tests for undocumented mean shifts. For cases in which $\beta = 0$ (i.e., time series of zero trend), the effect of unequal sample sizes makes the FAR of the SNHT-type tests much higher than the nominal level for points near the ends of the series and much lower for the middle points; the FAR as a function of the location $k$ of the point being tested; that is, FAR$(k)$ is U-shaped due to this effect (WW07). Consequently, the resulting level of confidence on the identified changepoints that are near the ends (middle) of the series is much lower (higher) than the nominal level, which is not desirable. Similarly, the effect of unequal sample sizes also exists in the above maximal $F$ test (and in other maximal $F$ tests, such as TPR4).

To diminish such an undesirable effect, the following PMFT is proposed in this study:

$$PF_{\text{max}} = \max_{1 \leq k \leq N-1} [P(k)F_{(k)}],$$

where $P(k)$ is a penalty factor that will be constructed empirically, as described later in this section, and $F_{(k)}$ is as defined above.

The empirical approach here is similar to that of WW07. To estimate FAR$(k)$ values of TPR3, $M_N = (N - 1) \times 100 000$ homogeneous IID Gaussian time series of length $N$, denoted as $Z_j(t)$ ($j = 1, 2, \ldots, M_N$; $t = 1, 2, \ldots, N$), are simulated and a linear trend is then added to these series; that is, $X_j(t) = \beta t + Z_j(t)$ with $\beta = 0.01$. Then, for each series $X_j(t)$, the statistic $F_{(k)}$ for each $k \in \{1, 2, \ldots, N - 1\}$ is calculated, and the maximal value $F_{(j)}(k) = \max_{1 \leq k \leq N-1} F_{(k)}$ and its corresponding position $k_j$ are identified and recorded. The $(1 - \alpha)$th percentile of the $F_{(k)}$ statistic, $F_{\text{max}}(\alpha)$ is then estimated from the $F_{(j)}(k_j)$ values ($j = 1, 2, \ldots, M_N$). Further, let $M_a(k)$ denote the count of cases (out of the $M_N$ cases) where point $k$ is associated with $F_{(j)}(k) > F_{\text{max}}(\alpha)$; that is, point $k$ is mistakenly declared as a changepoint at the $\alpha$ significance level (here $\alpha = 0.05$). Then, the false alarm rate for point $k$ is estimated as

$$\text{FAR} = \frac{M_a(k)}{M},$$

where $M = M_N(N - 1) = 100 000$; $p_a(k) = [1 - \alpha_a(k)] = [1 - \text{FAR}a(k)]$ is called the effective level of confidence of the test, and $\alpha_a(k) = \text{FAR}_a(k)$ the effective level of significance ($k = 1, 2, \ldots, N - 1$). The above calculations are repeated for each of 20 selected values of $N$ (ranging from 10 to 900); the resulting FAR$_a(k)$ curves are shown in Fig. 1. Note that $M = 10 000$ was used for the curve of $N = 900$, while $M = 100 000$ for each of the other 19 $N$ values (i.e., 8 990 000 simulations were done for $N = 900$, instead of 89 900 000 simulations). This is why the curve for $N = 900$ is not as smooth as the others (it is also the reason for the crosses in Fig. 2 to scatter wider for $N = 900$).

Clearly, the FAR$_a(k)$ functions are W-shaped, with the highest values near the ends of the series and the lowest values occurring approximately at the point $W = (11N)/50$ or $W = (21N/100)$ for $N > 500$ from either end of the series, while a secondary peak occurs at the midpoint of the series (see Fig. 1). As the series length increases, the false alarm rate for the midpoint of the series converges to the nominal level (0.05); it approximately equals the nominal level for series of length $N \approx 300$, as shown in Fig. 1. These W-shaped curves indicate that the chance for points near the ends of a homogeneous series to be mistakenly declared as a changepoint is much larger than those around point $W$ from either end of the series; it is also slightly larger for points around the middle of the series when the series length $N < 300$ (see Fig. 1). Consequently, the effective level
of confidence on the results of the above maximal $F$ test is much lower than the nominal level ($p = 0.95$) if the detected changepoints are near the ends of the series (i.e., $p_e < 0.95$ or $\alpha_e > 0.05$), but is much higher if they are near point $W$ from each end of the series (i.e., $p_e > 0.95$ or $\alpha_e < 0.05$). In other words, the above maximal $F$ test (i.e., TPR3) would detect a changepoint of certain magnitude with a lower-than-specified level of confidence and hence more easily when it occurs near the ends of the series than around point $W$ from either end of the series; it would mistakenly declare many more changepoints near the ends of a homogeneous series than around the point $W$. As mentioned before, this problem arises when the two samples (before and after point $k$) are substantially unequal in size (when one sample is very small compared to the other). Similar to the cases shown in WW07, the effect of unequal sample sizes on the statistic $F_c(k)$ is not very obvious either, but it is enough to make a significant difference in the result of searching for the maximum value of $F_c(k)$ across $k \in$...
Fig. 2. The M-shaped penalty term $P_o(k)$ (thin dashed curves) and $P(k)$ (thick curves), as well as their fits to the ratios $R_k$ (crosses, squares, circles, and stars). See section 2 for the definition of these terms.

[1, 2, ..., $N - 1$] (it gets “exaggerated” by this inevitable “maximizing” process).

However, note that the effect of unequal sample sizes on the results of the above TPR3 (shown in Fig. 1) is notably smaller than that on the results of the SNHT-type tests (shown in Fig. 1 of WW07). For example, for points near the ends of a series of length $N = 40$, the FAR values of the SNHT-type tests can reach 0.105 (about 2 times the nominal level 0.05; see WW07), while the FAR values of the TPR3 can only reach about 0.071 (only about 1.4 times the specified level). In general, for a fixed value of $N$, the height of the W-shaped curve is much smaller than that of the U-shaped curve. That is to say, TPR3 is much less severely affected by the inequality of sample sizes than are the SNHT-type tests.

Nevertheless, this type of hypothesis testing is expected and usually assumed to perform effectively at
the nominal level of significance/confidence under all circumstances. It is highly desirable for us to have the same level of confidence on the detected changepoints regardless of their position in the time series, and to have the same false alarm rate (as specified) for all points in a homogeneous series. To construct a maximal $F$ test of such highly desirable features, a penalty factor is empirically constructed in this study using the ratios

$$R_k = \frac{F_{\text{max}}[\alpha_c(k)]}{F_{\text{max}}(0.05)},$$

where $F_{\text{max}}[\alpha_c(k)]$ and $F_{\text{max}}(0.05)$ are the critical values of the $F_{\text{max}}$ in (3) that are corresponding to the $\alpha_c(k)$ and 5% level of significance, respectively; they are also estimated through Monte Carlo simulations. The estimated $R_k$ values are shown in Fig. 2 (see the crosses or squares or circles or stars). Through these simulations, the estimated effect of unequal sample sizes on the $\text{FAR}_c(k)$ values (shown in Fig. 1) is converted into the reverse of the effect on the $F_{\text{max}}$ statistic; thus, a penalty factor on $F_c(k)$ that would even out the W shape of the $\text{FAR}_c(k)$ function should fit the ratios $R_k$ well.

Therefore, in order to construct such a penalty factor, it is attempted in this study to obtain least squares fits to the simulated $R_k$ values for all the 19 selected values of $N$ (the case for $N = 10$ is excluded hereafter because the effect is negligible for $N < 15$). Through trial-and-error Monte Carlo simulation experiments, it is found in this study that the following functions of $k$ and/or $N$ are useful and thus used to obtain the least squares fits mentioned above:

$$P_c(k) = \begin{cases} 
64N^{1/40} + 35 - C & \text{for } k \leq W \text{ or } k \geq N - W, \\
\min\left(1, \frac{98000 + 6N}{100000}\right) & \text{for } W < k < N - W,
\end{cases}$$

where

$$F = \begin{cases} 
1 - Q^{3-C/6} & \text{for } k \leq W \text{ or } k \geq N - W, \\
1 + A^{3-C/6} & \text{for } W < k < N - W
\end{cases}$$

and

$$v = \begin{cases} 
(11.7 - S)B^{1/100} - 11.8 & \text{for } k \leq W \text{ or } k \geq N - W, \\
(C + 353)N^{1/100} - 340 & \text{for } W < k < N - W.
\end{cases}$$

where

$$A = \left\lfloor 1 - \frac{2k}{N}\right\rfloor, \quad B = \log(N), \quad C = \log(B), \quad D = \log(\log(N + 150)),$$

$$Q = \begin{cases} 
\frac{1 - 50k}{11N} & \text{for } k \leq N/2, \\
\frac{1 - 50k - 28N}{11N} & \text{for } k > N/2,
\end{cases}$$

$$S = \min\left(11 \frac{2C^4}{B^3}, \frac{900 - 9K_{\min}}{K_{\min}}\right)$$

with $K_{\min} = \min(k, N - k)$

$$W = \begin{cases} 
\frac{11N}{50} & \text{for } N \leq 500, \\
\frac{21N}{100} & \text{for } N > 500,
\end{cases}$$

where $\lfloor \cdot \rfloor$ means to take the floor value of the integer division. The trial-and-error Monte Carlo simulation results of the present study show that the following penalty function fits the $R_k$ values reasonably well for all the 19 selected values of $N$:

$$P_c(k) = \begin{cases} 
64N^{1/40} + 35 - C & \text{for } k \leq W \text{ or } k \geq N - W, \\
\min\left(1, \frac{98000 + 6N}{100000}\right) & \text{for } W < k < N - W,
\end{cases}$$

where

$$F = \begin{cases} 
1 - Q^{3-C/6} & \text{for } k \leq W \text{ or } k \geq N - W, \\
1 + A^{3-C/6} & \text{for } W < k < N - W
\end{cases}$$

and

$$v = \begin{cases} 
(11.7 - S)B^{1/100} - 11.8 & \text{for } k \leq W \text{ or } k \geq N - W, \\
(C + 353)N^{1/100} - 340 & \text{for } W < k < N - W.
\end{cases}$$

(11)
The thin dashed curves in Fig. 2 show the above penalty function $P_o(k)$, and its fit to the $R_k$ values, for the 19 selected values of $N$ (in each case, most of the thin dashed curve is overlapped with a thick curve, which will be explained below).

Although these fits are reasonably well, it is noticed in this study that, for all time series of $N \geq 40$, the above penalty $P_o(k)$ tends to overpenalize the test statistic for a few extreme end points but underpenalize the rest of about $L$ points at each end of the series, where $L = \text{integer}[1 + (317N^{3/4} - 2867)/1000]$ (take the integer value here; $L$ ranges from 3 to 50 for $40 \leq N \leq 900$; these are also results of the trial-and-error experiments of the current study). In other words, the fit to the $L$ points at each end of the series can be improved.

To obtain such improvement, the approach of trial-and-error experiments is taken again in this study, which leads us to modify the above penalty function to

\[
P(k) = \begin{cases} 
P_o(k) & \text{for all } k \in [1, N-1] \text{ if } N < 40 \\
P_o(L) - (L-k)\eta & \text{for } k = 1, 2, \ldots, L \text{ if } N \geq 40 \\
P_o(k) & \text{for } k = (L + 1), (L + 2), \ldots, (N - L - 1) \text{ if } N \geq 40 \\
P_o(N - L) - (k - N + L)\eta & \text{for } k = (N - L), (N - L + 1), \ldots, (N - 1) \text{ if } N \geq 40, 
\end{cases}
\]

where

\[
\eta = D^{1/3}[P_o(L + 1) - P_o(L)] + \frac{C^3}{5N}
\]

for all time series of $N \in [40, 100]$, and

\[
\eta = \begin{cases} 
\frac{[P_o(L) - P_o(1)]}{L + B - 2C - 1}A^C & \text{for } k = 1, 2, \ldots, L; \\
\frac{[P_o(N - L) - P_o(N - 1)]}{L + B - 2C - 1}A^C & \text{for } k = (N - L), (N - L + 1), \ldots, (N - 1)
\end{cases}
\]

for all time series of length $N > 100$. Figure 2 also shows this modified penalty function $P(k)$ (thick curves), in comparison with the corresponding unmodified penalty $P_o(k)$ (thin dashed curves), and its fit to the $R_k$ values for each of the 19 selected values of $N$. The modified penalty function $P(k)$ is then used in the test statistic in (7); thereby PMFT is constructed.

Figure 3 shows the FAR$_e(k)$ curves of PMFT in comparison with the corresponding curves of TPR3 for 12 selected values of $N$ (it looks similar for other $N$ values, which are thus not shown). Clearly, PMFT has a much more evenly distributed false alarm rate across all points in the series than does TPR3, although it tends to overpenalize slightly the test statistic for the end points of very long time series ($N > 500$; see Fig. 3 for $N = 720$). The penalty does even out, to a great extent, the W shape of the FAR$_e(k)$ curves of TPR3. This is because the new test statistic takes the relative position of each point being tested into account to reduce the distortion of the test statistic due to unequal sample sizes. Observations are treated more equally during the process of searching for the most probable changepoint position/time. Consequently, PMFT has the highly desirable features of great practical importance, that is, a basically uniform false alarm rate across a homogeneous series and, hence, a basically uniform level of confidence on the identified changepoints regardless of their position in the time series.

As in WW07, the approach here is also purely empirical [i.e., both $P_o(k)$ and $P(k)$ are obtained empirically]. A theoretically based penalty term is yet to emerge. This empirical exercise satisfactorily achieves the desired results; it clearly reveals the characteristics of the effect of unequal sample sizes, which may help develop a theoretically based penalty term.

Although it is difficult to derive the theoretical distribution of the test statistic $\text{PF}_{\text{max}}$, its empirical critical values for some selected values of $N$ can be and are obtained by simulating 10 million $\text{PF}_{\text{max}}$ values under the null hypothesis of no changepoint; these values are presented in Table 1 (for other $N$ values not listed here, a linear interpolation between its two neighboring values would be of sufficiently good accuracy). The empirical critical values of $F_{\text{max}}$ in TPR3 are also simulated similarly and listed in Table 1, which will be used for the comparison in section 3.
Fig. 3. Comparison of the FAR$_k(k)$ curves of PMFT and TPR3 applied to homogeneous time series of the indicated lengths at the 5% significance level.
It will be illustrated through the comparison in the next section that, although the effect of unequal sample sizes may not be a big problem for detecting a large mean shift in a very long time series, it cannot be ignored when the magnitude of mean shift is small or medium relative to the noise variance.

3. Comparison of detection power of PMFT with TPR3

Similar to the comparison of PMT with SNHT in WW07, the detection power of PMFT is compared with that of TPR3 through Monte Carlo simulations in this section. The following rates are also used here to measure the power of detection of each method:

1) Position rate (PR), defined as the rate of detecting a changepoint position “correctly” (i.e., within the interval \([K - 3, K + 3]\) of the true position \(K\)), regardless of its significance.

2) Significance rate (SR), defined as the rate of declaring a statistically significant (at the 5% level) changepoint regardless of whether or not its position being identified “correctly.” (This is usually called detection power in statistics.)

3) Hit rate (HR), defined as the rate of detecting a changepoint with both statistical significance (at the 5% level) and “correct” position (as defined for the position rate). For any method, the PR or SR is only a rough measure of its detection ability, while the hit rate is a strict measure of its “accurate” detection power and is of the most practical importance. In practice, identifying a changepoint at a wrong position could be as bad as failing to detect it.
The comparison is carried out as follows. First, for each of the 18 selected values of series length $N$ (ranging from 15 to 720), 10,000 homogeneous time series are generated from an IID Gaussian distribution with zero mean and variance $\sigma^2$ (without lose of generality, $\sigma$ is set to 1 here), and a linear trend component ($\beta = 0.01$) is added to each of these time series. Then a mean-shift $\Delta$ is inserted at point $k$ (between $k$ and $k + 1$) in each of these series, for each $k \in \{2, 3, \ldots, N - 2\}$ and each relative shift size $\delta = \Delta/\sigma \in \{0.25, 0.5, 1, 1.5, 2\}$. Then, PMFT and TPR3 are applied subsequently to each of the 10,000 time series for each combination of $k$ and $\delta$ to detect the mean shift using the same nominal level of significance (5%). The three detection rates are calculated for each combination of $N$, $k$, and $\delta$ values, and are analyzed below.

As an example, Fig. 4 shows the comparison of the three rates as a function of changepoint position $k \in \{2, 3, \ldots, N - 2\}$, where $N = 100$, with three different sizes of shift ($\delta = 0.25, 1.0, 2.0$) at each position. As would be expected, PMFT has higher power of detection (in terms of all the three measures) than does TPR3 for changepoints that occur not too close to the ends of the series; while it has lower power in detecting changepoints that occur within the first or last $N/10$ points of the series. However, the seemingly higher power of TPR3 in detecting changepoints near these end points comes from the fact that it actually applies a lower-than-specified level of confidence so that not all the changepoints it identified are actually significant at the nominal level; while PMFT identifies every changepoint with approximately the same level of confidence as specified, regardless of the changepoint location. That is, PMFT obtains a basically evenly distributed effective level of significance/confidence (and hence power) of the test across all possible changepoint positions $k$. For other selected values of $N$, the situation is similar and hence not shown.

Most importantly, however, when averaged over the possible changepoint positions $k \in \{2, 3, \ldots, N - 2\}$, PMFT has significantly higher power of detection than does TPR3. Tables 2–4 list, respectively, the mean hit rates, mean position rates, and mean significance rates (i.e., those averaged over $k = 2, 3, \ldots, N - 2$) for the 18 selected values of series length $N$. Figure 5 shows these mean rates as a function of series length $N$ and relative shift size $\delta$.

Clearly, in terms of the mean hit rate, the improvement of PMFT over TPR3 is seen for all the five shift sizes regardless of the series length. For small shifts ($\delta = 1$), the improvement increases steadily as the series length increases and reaches its maximal of around 10% for series of length $N = 400 \sim 500$ (Fig. 5a). For
Table 2. The mean hit rates (averaged over $k = 2, \ldots, N - 2$) of PMFT and TPR3 for the indicated different series lengths $N$ and relative shift sizes. The numbers in parentheses are the ratios of the PMFT rate to the TPR3 rate.

<table>
<thead>
<tr>
<th>$N$</th>
<th>$\Delta = 0.25\sigma$</th>
<th>$\Delta = 0.5\sigma$</th>
<th>$\Delta = \sigma$</th>
<th>$\Delta = 1.5\sigma$</th>
<th>$\Delta = 2\sigma$</th>
</tr>
</thead>
<tbody>
<tr>
<td>15</td>
<td>238.7/229.1 (1.042)</td>
<td>240.5/230.3 (1.044)</td>
<td>253.1/243.3 (1.040)</td>
<td>280.3/268.9 (1.042)</td>
<td>322.6/310.7 (1.042)</td>
</tr>
<tr>
<td>20</td>
<td>165.9/165.5 (1.002)</td>
<td>168.6/169.8 (0.993)</td>
<td>188.4/187.9 (1.002)</td>
<td>227.0/226.9 (1.001)</td>
<td>284.5/282.2 (1.008)</td>
</tr>
<tr>
<td>25</td>
<td>143.8/141.2 (1.018)</td>
<td>150.3/150.8 (1.014)</td>
<td>184.6/182.3 (1.012)</td>
<td>233.4/230.1 (1.014)</td>
<td>308.2/305.9 (1.008)</td>
</tr>
<tr>
<td>30</td>
<td>119.0/118.2 (1.007)</td>
<td>126.6/125.3 (1.011)</td>
<td>159.5/156.5 (1.019)</td>
<td>218.2/214.0 (1.020)</td>
<td>314.2/309.0 (1.017)</td>
</tr>
<tr>
<td>40</td>
<td>90.6/89.1 (1.017)</td>
<td>103.1/101.5 (1.026)</td>
<td>146.0/143.2 (1.019)</td>
<td>225.8/220.1 (1.026)</td>
<td>357.7/348.4 (1.027)</td>
</tr>
<tr>
<td>50</td>
<td>73.9/71.3 (1.036)</td>
<td>85.2/83.4 (1.021)</td>
<td>141.2/136.1 (1.037)</td>
<td>245.8/237.9 (1.033)</td>
<td>422.3/408.4 (1.034)</td>
</tr>
<tr>
<td>60</td>
<td>62.4/61.3 (1.019)</td>
<td>75.4/73.7 (1.022)</td>
<td>133.2/129.5 (1.029)</td>
<td>258.9/249.8 (1.036)</td>
<td>474.5/457.2 (1.038)</td>
</tr>
<tr>
<td>70</td>
<td>53.5/52.4 (1.021)</td>
<td>67.8/65.6 (1.036)</td>
<td>136.1/131.0 (1.039)</td>
<td>278.5/267.9 (1.040)</td>
<td>541.5/519.4 (1.043)</td>
</tr>
<tr>
<td>80</td>
<td>49.1/47.0 (1.044)</td>
<td>65.0/62.1 (1.047)</td>
<td>142.4/137.2 (1.037)</td>
<td>305.2/291.5 (1.047)</td>
<td>619.6/593.9 (1.043)</td>
</tr>
<tr>
<td>90</td>
<td>45.7/43.9 (1.041)</td>
<td>63.6/60.9 (1.045)</td>
<td>144.1/137.4 (1.049)</td>
<td>328.4/313.6 (1.047)</td>
<td>682.0/653.0 (1.045)</td>
</tr>
<tr>
<td>100</td>
<td>41.7/39.5 (1.054)</td>
<td>57.9/55.5 (1.043)</td>
<td>144.4/137.3 (1.051)</td>
<td>353.7/334.7 (1.057)</td>
<td>753.7/721.0 (1.045)</td>
</tr>
<tr>
<td>150</td>
<td>30.2/29.6 (1.022)</td>
<td>50.3/48.4 (1.039)</td>
<td>169.2/163.0 (1.058)</td>
<td>479.3/452.1 (1.060)</td>
<td>1121.3/1074.1 (1.044)</td>
</tr>
<tr>
<td>150</td>
<td>24.1/23.2 (1.041)</td>
<td>47.0/43.3 (1.083)</td>
<td>193.1/179.4 (1.076)</td>
<td>620.1/583.2 (1.063)</td>
<td>1513.6/1452.0 (1.042)</td>
</tr>
<tr>
<td>250</td>
<td>21.4/20.7 (1.032)</td>
<td>49.3/46.2 (1.068)</td>
<td>226.8/208.6 (1.087)</td>
<td>773.8/728.6 (1.065)</td>
<td>1900.6/1828.2 (1.039)</td>
</tr>
<tr>
<td>300</td>
<td>20.2/19.2 (1.053)</td>
<td>49.3/45.5 (1.084)</td>
<td>256.3/235.6 (1.088)</td>
<td>918.7/864.7 (1.062)</td>
<td>2266.5/2191.7 (1.034)</td>
</tr>
<tr>
<td>400</td>
<td>17.7/15.9 (1.111)</td>
<td>50.7/45.8 (1.010)</td>
<td>327.2/299.6 (1.094)</td>
<td>1238.0/1170.0 (1.058)</td>
<td>2907.9/2824.5 (1.023)</td>
</tr>
<tr>
<td>500</td>
<td>14.3/13.8 (1.040)</td>
<td>52.0/46.3 (1.122)</td>
<td>401.9/367.0 (1.095)</td>
<td>1544.9/1476.2 (1.049)</td>
<td>3408.3/3360.3 (1.014)</td>
</tr>
<tr>
<td>720</td>
<td>14.4/12.8 (1.121)</td>
<td>61.5/54.1 (1.136)</td>
<td>578.7/533.6 (1.085)</td>
<td>2112.2/2051.4 (1.030)</td>
<td>4040.3/4038.4 (1.000)</td>
</tr>
</tbody>
</table>

medium–large shifts ($\delta > 1$), the improvement is maximal for series of length $100 \leq N \leq 300$. In general, the smaller the $\delta$, the greater the improvement of PMFT over TPR3; and the longer the time series, the greater the improvement of PMFT in detecting shifts in time series of length $N < 100$ (regardless of shift size; Fig. 5a).

In terms of the mean position rate, PMFT is also superior to TPR3. As shown in Fig. 5b, PMFT has up to about 5% higher power in identifying “correctly” the position of small shifts in long time series, while PMFT and TPR3 perform very similarly in detecting the “correct” position of large shifts in very long series ($N > 400$). For series of length $N < 100$, the improvement of PMFT over TPR3 also increases as the series length increases (Fig. 5b).

Note that each of the 10 000 time series tested in each setting here contains a mean shift (i.e., $H_{\text{av}}$ is always true here). Thus, the type-II error rate can be estimated as $(1 - \text{SR})$ roughly [or as $(1 - \text{HR})$ more accurately, which, however, contains the same information as HR], and the higher the significance rate SR, the lower the type-II error rate (and hence the better the test). In this regard, as shown in Fig. 5c, PMFT has 2%–7% improvement over TPR3 in detecting medium–large shifts in medium–long time series, or in detecting small shifts

Table 3. Same as in Table 2, but for the mean position rates of PMFT and TPR3.

<table>
<thead>
<tr>
<th>$N$</th>
<th>$\Delta = 0.25\sigma$</th>
<th>$\Delta = 0.5\sigma$</th>
<th>$\Delta = \sigma$</th>
<th>$\Delta = 1.5\sigma$</th>
<th>$\Delta = 2\sigma$</th>
</tr>
</thead>
<tbody>
<tr>
<td>15</td>
<td>4610.2/4610.9 (1.000)</td>
<td>4622.9/4624.5 (1.000)</td>
<td>4676.4/4680.7 (0.999)</td>
<td>4756.7/4759.6 (0.999)</td>
<td>4862.3/4867.7 (0.999)</td>
</tr>
</tbody>
</table>
PMFT and TPR3 have comparable significance (or changes. However, as mentioned before, TPR4 also suffer from the effect of unequal sample sizes. Thus, a penalized version of TPR4 is currently under development and will be reported in a separate paper.

Serial correlation is another inherent feature of many climate data series, which cannot be diminished by using reference series. Also, long-term climate data series could contain multiple changepoints. Recently, two stepwise algorithms have been developed to empirically account for serial correlation in detecting multiple changepoints using PMT or PMFT (Wang 2008).

In the remainder of this section, examples of application of PMFT and TPR3 to detect undocumented mean shifts in climate data series are presented. The goal is to show which of the two methods performs better in practical use. Thus, we need to be able to verify the results. That is, we need to apply the methods to a time series that contains a documented mean shift (i.e., the time and cause of the shift are known), to see if the methods can detect the shift if it is assumed undocumented. For such purpose, we applied the methods to a three data series: two monthly mean station pressure series, and a monthly mean daily minimum air temperature series.

The first example series is the time series of monthly mean station pressures recorded at Greenwood Airport (Nova Scotia, Canada) for the 52-yr period from 1953 to 2004. The related metadata indicates that this series contains an artificial mean shift (see the thick solid line in Fig. 6a) caused by using the so-called established elevation (instead of station elevation) in the calculation of station pressures from barometer readings in the period prior to January 1977. The shift magnitude is estimated to be 0.35 hPa (which is about 0.12σ by hy-
Fig. 5. Ratios of the mean hit rate, mean position rate, and mean significance rate (PMFTs over TPR3s) as a function of time series length $N$ and the relative step size $\delta = \Delta / \sigma$ (i.e., RSS in the figures).
FIG. 6. Time series of consecutive monthly anomalies (after removing the 12 monthly sample means, one for each calendar month) of station pressure recorded at Greenwood Airport (NS, Canada) for the 52-yr period (1953–2004), and the “undocumented” mean shift detected by applying PMFT to this time series (dashed line). The thick solid line represents the regression fit with the exact time of shift (December 1976).
drostatic model, using hourly pressure and temperature data and correct station elevation (see Wan et al. 2007).

It is also noticed that the roughly deseasonalized pressure time series has a negligible lag-1 autocorrelation of 0.045 (here \(N = 624\)). Thus, PMFT and TPR3 can be, and were, applied to the deseasonalized pressure time series (see the thin solid curve in Fig. 6). At the 5% level of significance, PMFT identified a changepoint in October 1976 (\(k = 286, \text{PF}_{\max} = 17.391\); see the dashed line in Fig. 6), while TPR3 finds the time series homogeneous (it finds \(k = 349\) with \(F_{\max} = 8.699\)). In other words, TPR3 fails to detect the artificial shift between December 1976 and January 1977, while PMFT detects it with good accuracy (only two months earlier than the actual time of change). This is an example showing that PMFT outperforms TPR3 when the changepoint occurs near the middle of the series.

Similarly, the deseasonalized series of station pressures recorded at Daniels Harbor (Newfoundland, Canada) for the 50-yr period from 1953 to 2002 (see Fig. 6b) was also estimated to have a negligible lag-1 autocorrelation (of 0.038). Thus, PMFT and TPR3 were also applied to this series, separately, to see if they can detect the mean shift due to station relocation in the end of July 1955 (as documented in the metadata record). At the 5% level of significance, both tests identified a significant changepoint in August 1955, which is just one month later than the actual time of change (Fig. 6b). This is an example showing that PMFT and TPR3 can perform similarly when the changepoint is located near the end of the series.

The third example series is the series of monthly means of daily minimum air temperature recorded at AMOS (Quebec, Canada) for the 46-yr period from July 1913 to December 1997 (\(N = 1014\)). The station inspection reports indicate that 1) the thermometer screen and stand were replaced on 24 October 1958; and 2) the station elevation changed sometime between August 1948 and 24 October 1958 (from 990 to 1002 ft or, equivalently, from 297 to 300.6 m). The deseasonalized temperature series is estimated to have a lag-1 autocorrelation of 0.173 [its 95% uncertainty range: \((0.112, 0.233)\)], which is highly significant. Taking into account the effect of this autocorrelation in the way proposed in Wang (2008), PMFT identified a changepoint in August 1958 that is significant even without metadata support [\(k = 548, \text{PF}_{\max} = 26.82 > 15.85(14.06 – 17.88);\] the numbers the parentheses are the lower and upper bound of the 95th percentile of the \(\text{PF}_{\max}\) which corresponds to the uncertainty range of autocorrelation estimate above]. This changepoint is obviously due to the replacement of the thermometer screen and stand on 24 October 1958. In addition, PMFT also identified a changepoint in October 1953 that would be significant if it has metadata support [\(k = 490, p = 0.998\)]. Thus, the station elevation change most likely happened around October 1953. The fit with these two changepoints is shown in Fig. 7. TPR3 was not applied to this series because we do not have a version of the TPR3 that can account for the highly significant autocorrelation.

To verify whether it is reasonable to assume that the time series being tested have a common trend throughout, the sum of squared residuals of the common-trend model fit was compared with that of the different-trend model fit to each of the three de-seasonalized series above (via a regular \(F\) test). The results show that the different-trend model (which has one or several more parameters) has no significant improvement over the common-trend model in terms of fit (\(p = 0.744, 0.062,\) and 0.275 for the Greenwood, Daniels Harbor, and AMOS series, respectively). This indicates that there is no significant trend change at the time(s) of the mean shift. The common-trend assumption of PMFT or TPR3 is valid for these series.

Also, the QQ plots (not shown) of the residuals of the PMFT model fit to each of the three series reveal that the error term (prewhitened in the AMOS case) in each of these time series does approximate very well to a Gaussian distribution. Thus, the normality assumption of the PMFT or TPR3 is also valid for these series.

5. Concluding remarks

In this study, the common-trend two-phase regression model–based maximal \(F\) test (TPR3) for detecting undocumented mean shifts that are not accompanied by any sudden change in the linear trend of time series (Wang 2003) is modified to diminish the effect of unequal sample sizes on the false alarm rate and hence on the power of detection, using an empirical penalty function (also constructed in this study).

For time series of length in the wide range \(N \in [15, 720]\), the new penalized maximal \(F\) test (PMFT) has been shown to have basically evenly distributed false alarm (or type-I error) rate over every point in the time series. That is to say, PMFT performs generally at the nominal level of significance/confidence, no matter which point in the time series is being tested (or where the changepoint occurs). This feature is highly desirable in practice, because each point of data should be treated equally, each should have equal chance to be mistakenly declared as a changepoint when the time series is homogeneous. On the other hand, a shift of certain magnitude should have the same probability of
being detected no matter where it occurs. Without this feature, as in the case of TPR3, not all the changepoints identified by the test are actually significant at the nominal level, while some changepoints that are actually significant at the nominal level might go undetected. This is why TPR3, which has W-shaped FAR($k$) curves, seemingly has higher power in detecting changepoints that are near the ends (within the first or last $N/10$ points) of the series than does PMFT, while it has lower power in detecting changepoints that occur near point $W = (0.21 \sim 0.22) \times N$ from either end of the series.

Most importantly, however, when averaged over all possible changepoint positions, PMFT has a higher power of detection than does TPR3. The improvement
in hit rate can be more than 10% for detecting small shifts (\( \Delta \leq \sigma \)). Also, PMFT has up to about 5% higher power in identifying “correctly” the position of small shifts in long time series. In terms of the type-II error rate, it has 2%–7% improvement over TPR3 in detecting medium–large shifts in medium–long time series, or in detecting small shifts (e.g., \( \Delta = 0.5\sigma \)) in long time series \((N \geq 300)\); while PMFT and TPR3 have comparable type-II error rates in detecting small shifts \((\delta < 1)\), especially in time series of short–medium length.

It is also shown that the results of TPR3 are notably less severely affected by the problem of unequal sample sizes than those of the SNHT-type tests. This is because the estimate of the common trend in TPR3 is obtained using the whole series, not affected by the problem of unequal sample sizes. In other words, the whole series is used in the estimate of one of the model parameters and hence in the calculation of the test statistic of TPR3. In contrast, the estimate of all parameters in the SNHT-type tests is affected by the problem of unequal sample sizes. This is also the case for the TPR4 approach (Lund and Reeves 2002), which also has a U-shaped distribution of false alarm rate (similar to what are shown in WW07). An empirically penalized version of TPR4 is also being constructed and will be reported in a separate paper.

Note that PMFT and TPR3, or PMT and the SNHT-type tests, all assume that the errors in the time series being tested are IID Gaussian, which is hardly true for climate data series (even for annual series). Most climate data series exhibit autocorrelation and periodicity. As noted in WW07, the use of a good reference that mate data series exhibit autocorrelation and periodicity in the time series. Therefore, a test of undocumented changepoints that takes into account the effect of autocorrelation needs to be developed. Recently, Lund et al. (2007) addressed this need, proposing a new test for changepoint detection in periodic and autocorrelated time series. However, the effect of unequal sample sizes has yet to be diminished in this new method, which should be the subject for future study. Wang (2008) has developed an algorithm to account for autocorrelation in detecting mean shifts in climate data series using the PMFT.

Finally, a software package RHtestV2 (written in R and FORTRAN languages) for implementing PMFT (or PMT) with a stepwise testing algorithm for detecting multiple mean shifts in a time series has been developed and made freely available (see online at http://ccsma.seos.uvic.ca/ETCCDMI/software.shtml).

Acknowledgments. The author wishes to thank Lucie Vincent and Hui Wan (CRD, Environment Canada) for their helpful comments on an earlier version of this paper. The two anonymous reviewers are also acknowledged.

REFERENCES


