Penalized Maximal *t* Test for Detecting Undocumented Mean Change in Climate Data Series

XIAOLAN L. WANG

Climate Research Division, Atmospheric Science and Technology Directorate, Science and Technology Branch, Environment Canada, and Department of Mathematics and Statistics, York University, Toronto, Ontario, Canada

QIUZI H. WEN AND YUEHUA WU

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

(Manuscript received 24 March 2006, in final form 18 October 2006)

ABSTRACT

In this paper, a penalized maximal *t* test (PMT) is proposed for detecting undocumented mean shifts in climate data series. PMT takes the relative position of each candidate changepoint into account, to diminish the effect of unequal sample sizes on the power of detection. Monte Carlo simulation studies are conducted to evaluate the performance of PMT, in comparison with the most popularly used method, the standard normal homogeneity test (SNHT). An application of the two methods to atmospheric pressure series recorded at a Canadian site is also presented. It is shown that the false-alarm rate of PMT is very close to the specified level of significance and is evenly distributed across all candidate changepoints, whereas that of SNHT can be up to 10 times the specified level for points near the ends of series and much lower for the middle points. In comparison with SNHT, therefore, PMT has higher power for detecting all changepoints that are not too close to the ends of series and lower power for detecting changepoints that are near the ends of series. On average, however, PMT has significantly higher power of detection. The smaller the shift magnitude Δ is relative to the noise standard deviation σ, the greater is the improvement of PMT over SNHT. The improvement in hit rate can be as much as 14%–25% for detecting small shifts ($\Delta < \sigma$) regardless of time series length and up to 5% for detecting medium shifts ($\Delta = \sigma$–1.5σ) in time series of length $N < 100$. For all detectable shift sizes, the largest improvement is always obtained when $N < 100$, which is of great practical importance, because most annual climate data series are of length $N < 100$.

1. Introduction

The accuracy and homogeneity of climate data are indispensable for many aspects of climate research. In particular, a realistic and reliable assessment of historical climate trends and variability is hardly possible without a long-term, homogeneous time series of climate data. Therefore, it is unfortunate that many kinds of changes (instrument/observer changes, station location/exposure changes, observing practices/procedures changes, etc.) that took place during the period of a data record can cause nonclimatic sudden changes (artificial shifts) in the time series. The fast-developing climate-monitoring science and technology (e.g., human observations are being replaced with automated observations or remote sensing) also make discontinuities inevitable in long-term climate data records. Such artificial shifts could have huge impacts on the results of climate analysis, especially those of climate trend analysis, as shown in Wang (2006) and Hanesiak and Wang (2005), among others. The reliability of the results of historical climate trend assessment is often hampered by the existence of discontinuities in instrumental records of climate and is often the subject of the ongoing debate on climate change. Accurate and homogeneous climate data are also indispensable for the calculation of related statistics that are needed and used to define the state of climate and climate extremes. Therefore, artificial shifts should be eliminated, to the extent

Corresponding author address: Dr. Xiaolan L. Wang, Climate Research Division, Atmospheric Science and Technology Directorate, Environment Canada, 4905 Dufferin Street, Toronto, ON M3H 5T4, Canada.
E-mail: xiaolan.wang@ec.gc.ca

DOI: 10.1175/JAM2504.1

possible, from time series prior to their application, especially in climate trend assessment.

There exist two types of artificial shifts: documented and undocumented shifts. Documented shifts refer to those with known position/time of shift (i.e., the time and cause of shift were recorded in the related metadata). Documented shifts are much easier to test/assess, because we do not need to identify the position of the shift statistically (we can find it from metadata); thus, the regular tests of means or variances are applicable. However, metadata often lack accuracy and completeness or in some cases are not available at all. One needs to rely on an appropriate statistical test to detect and assess undocumented shifts (i.e., those that have no metadata support).

In recent decades, several methods have been developed for detection of undocumented changepoints. In the literature of statistics, most of the existing detection methods can be classified into three categories: likelihood-based methods, linear regression–based methods, and nonparametric methods (Csörgő and Horváth 1997). In the climate literature, the most commonly used methods for changepoint detection include the standard normal homogeneity test (SNHT; Alexandersson 1986), the cumulative deviation test (Buishand 1982), two-phase regression-based methods (Solow 1987; Easterling and Peterson 1995; Lund and Reeves 2002; Wang 2003), Bayesian-based methods (Perreault et al. 2000; Chu and Zhao 2004), multiple linear regression (Vincent 1998), and some nonparametric methods such as the Mann–Whitney U test and the Wilcoxon rank test. Readers are referred to Reeves et al. (2007) for a recent, comprehensive review and comparison of these methods and to Peterson et al. (1998) for a review of homogeneity adjustments of in situ atmospheric climate data. DeGaetano (2006) also compared several methods for detecting discontinuities in mean temperature series.

This study attempts to improve a test for detecting undocumented shifts, proposing a new test statistic that treats each candidate changepoint in the time series being tested more equally (such equality is not achieved by the existing methods; see details shown/described later in sections 2 and 5). Although undocumented shifts may take the form of a change in mean, variance, or both, this study only aims at detection of an undocumented shift in the mean. A mean-shift at time $t = k$ refers to the case in which the mean of the data series before this point (i.e., the average over all $t \le k$) is significantly different from that of the data series after this point (i.e., the average over all $t > k$). Here, we consider time series with zero trend and identically and independently distributed (IID) Gaussian errors, and we focus on the case in which the time series being tested contains at most one changepoint (AMOC). However, note that one can implement statistical tests that are developed for the AMOC case with an appropriate recursive testing algorithm to detect multiple changepoints in a time series, as in Wang (2007), Menne and Williams (2005), and Wang and Feng (2004). In the mean time, Davis et al. (2006) and Caussinus and Mestre (2004) are two prominent examples of recent studies that address directly in their statistical model the issue of detecting multiple changepoints. An algorithm for detecting undocumented and documented artificial mean shifts in tandem has also been proposed and implemented (Wang 2007; Wang and Feng 2004).

The rest of the paper is arranged as follows. We describe the proposed new test, a penalized maximal t test, in section 2. Then, we compare this method with SNHT using Monte Carlo simulations in section 3 and using atmospheric pressure data series from a Canadian station in section 4. We give some concluding remarks in section 5.

2. Penalized maximal *t* test

Let $\{X_t\}$ $(t = 1, \ldots, N)$ denote an IID Gaussian series. To detect a changepoint in time series $\{X_t\}$ is to test the null hypothesis

$$H_0: \{X_t\} \sim \text{IID}\mathcal{N}(\mu, \sigma^2) \tag{1}$$

against the alternative

$$H_a: \begin{cases} \{X_t\} \sim \text{IID}\mathcal{N}(\mu_1, \sigma^2), & t = 1, \ldots, k \\ \{X_t\} \sim \text{IID}\mathcal{N}(\mu_2, \sigma^2), & t = k+1, \ldots, N \end{cases}, \tag{2}$$

where $\mu_1 \neq \mu_2$ and "$\{X_t\} \sim \text{IID}\mathcal{N}(\mu, \sigma^2)$" stands for "$\{X_t\}$ follows an IID Gaussian (i.e., normal) distribution of mean μ and variance σ^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). In other words, if there is such a point k, the time series can be viewed as two independent samples from two normal distributions of the same unknown variance σ^2, one for all $t \le k$ and another for all $t > k$. The task of undocumented changepoint detection is to find out the most probable value of k and to test whether the means of these two samples are statistically significantly different from each other. The traditional test for this kind of problem is the so-called likelihood ratio test. The most probable point to be the changepoint is the one that is associated with the maximal value of the following log likelihood ratio (Csörgő and Horváth 1997):

$$l(k) = -(N/2)(\log\tilde{\sigma}^2 - \log\tilde{\sigma}_k^2), \quad (3)$$

where

$$\tilde{\sigma}_k^2 = \frac{1}{N}\left[\sum_{1\le t\le k}(X_t - \overline{X}_1)^2 + \sum_{(k+1)\le t\le N}(X_t - \overline{X}_2)^2\right] \quad \text{and}$$

$$\tilde{\sigma}^2 = \frac{1}{N}\sum_{1\le t\le N}(X_t - \overline{X})^2, \quad \text{with}$$

$$\overline{X}_1 = \frac{1}{k}\sum_{1\le t\le k} X_t, \overline{X}_2 = \frac{1}{N-k}\sum_{(k+1)\le t\le N} X_t, \quad \text{and}$$

$$\overline{X} = \frac{1}{N}\sum_{1\le t\le N} X_t.$$

According to Csörgő and Horváth (1997), maximizing $l(k)$ is equivalent to maximizing

$$T(k) = \frac{1}{\hat{\sigma}_k}\left[\frac{k(N-k)}{N}\right]^{1/2}|\overline{X}_1 - \overline{X}_2|, \quad (4)$$

where

$$\hat{\sigma}_k^2 = \frac{N}{N-2}\tilde{\sigma}_k^2$$

$$= \frac{1}{N-2}\left[\sum_{1\le t\le k}(X_t - \overline{X}_1)^2 + \sum_{(k+1)\le t\le N}(X_t - \overline{X}_2)^2\right],$$

and the test statistic for detecting an undocumented mean shift is

$$T_{\max} = \max_{1\le k\le N-1} T(k), \quad (5)$$

because of the necessary search over all candidate changepoints $k \in \{1, 2, \ldots, N-1\}$ for the most probable position of an undocumented mean shift. This test is called the maximal two-sample t test hereinafter ("two sample" may be suppressed). In comparison with $l(k)$, $T(k)$ is more intuitive in form. Note that without taking the absolute value of $(\overline{X}_1 - \overline{X}_2)$ in $T(k)$ this statistic becomes

$$\frac{1}{\hat{\sigma}_k}\left[\frac{k(N-k)}{N}\right]^{1/2}(\overline{X}_1 - \overline{X}_2),$$

which is just the test statistic of the well-known two-sided two-sample t test (for the equal but unknown variance case) that follows the Student's t distribution with $(N-2)$ degrees of freedom under the null hypothesis (von Storch and Zwiers 1999). Although it is difficult to derive the theoretical distribution of $T_{\max}$, its empirical critical values can be generated by Monte Carlo simulations, as commonly practiced. The popular SNHT is a special case of the above maximal t test [with the test statistic defined in (5)], though SNHT is formulated differently: Alexandersson (1986) assumes a known, unit variance for standardized ratio series. Nevertheless, the above maximal t test and SNHT are equivalent.

Gardner (1975) showed that the power of the t test may decrease considerably when the two samples are of unequal size (relative to the equal-size case). This is because the estimate from the shorter series tends to be less accurate. As a consequence, the maximal t test and SNHT suffer from the disadvantage that points in a homogeneous time series have different probabilities of being mistakenly identified as changepoints. This is illustrated through Monte Carlo simulations below.

To estimate the false-alarm rate as a function of time series length N and changepoint position k ($k = 1, 2, \ldots, N-1$), for various choices of N (see Fig. 1), we simulate $M_N = (N-1) \times 100\,000$ homogeneous IID Gaussian time series, denoted as $X_j(t)$ ($j = 1, 2, \ldots, M_N$; $t = 1, 2, \ldots, N$). Then, for each time series $X_j(t)$, we calculate the statistic $T(k)$ for each $k \in \{1, 2, \ldots, N-1\}$, find the maximal value $T_j(k_j) = \max_{1\le k\le N-1} T(k)$, and record its corresponding position k_j. The $(1-\alpha)$th percentile of the $T_{\max}$ statistic $[T_{max}(\alpha)]$ is then estimated from the $T_j(k_j)$ values ($j = 1, 2, \ldots, M_N$). Further, let $M_\alpha(k)$ denote the count of cases (out of the M_N cases) for which point k is associated with $T_j(k) > T_{\max}(\alpha)$, that is, for which point k is mistakenly identified as a changepoint at the α significance level [which is also often referred to as "at the $p = (1-\alpha)$ level of confidence," although the word "confidence" is somewhat misleading here]. Here we use $\alpha = 0.05$, and thus $p = 0.95$. Then, the false-alarm rate for point k is estimated as

$$\mathrm{FAR}_\alpha(k) = \frac{M_\alpha(k)}{M},$$

where $M = M_N/(N-1) = 100\,000$; $p_e(k) = [1 - \alpha_e(k)] = [1 - \mathrm{FAR}_\alpha(k)]$ is called the effective level of confidence of the test, and $\alpha_e(k) = \mathrm{FAR}_\alpha(k)$ is the effective level of significance ($k = 1, 2, \ldots, N-1$). We repeat the above calculations for each of 18 selected values of N (ranging from 6 to 500); the resulting false-alarm rate as a function of k, $\mathrm{FAR}_\alpha(k)$, is shown in Fig. 1 [the curve for $N = 6$, not shown, is almost flat; note that these curves of $\mathrm{FAR}_\alpha(k)$ values plotted against the corresponding k values are referred to as the $\mathrm{FAR}_\alpha(k) \sim k$ curves in this study]. Note that exactly the same $\mathrm{FAR}_\alpha(k) \sim k$ curves are obtained when the SNHT test statistic is used instead of the $T_{\max}$ above, as expected, because the maximal t test and SNHT are equivalent.

It is clear that the $\mathrm{FAR}_\alpha(k) \sim k$ curves are U shaped and that the larger the series length N is, the flatter is the bottom of the U-shaped curve (Fig. 1). These U-shaped curves indicate that the chance for points near

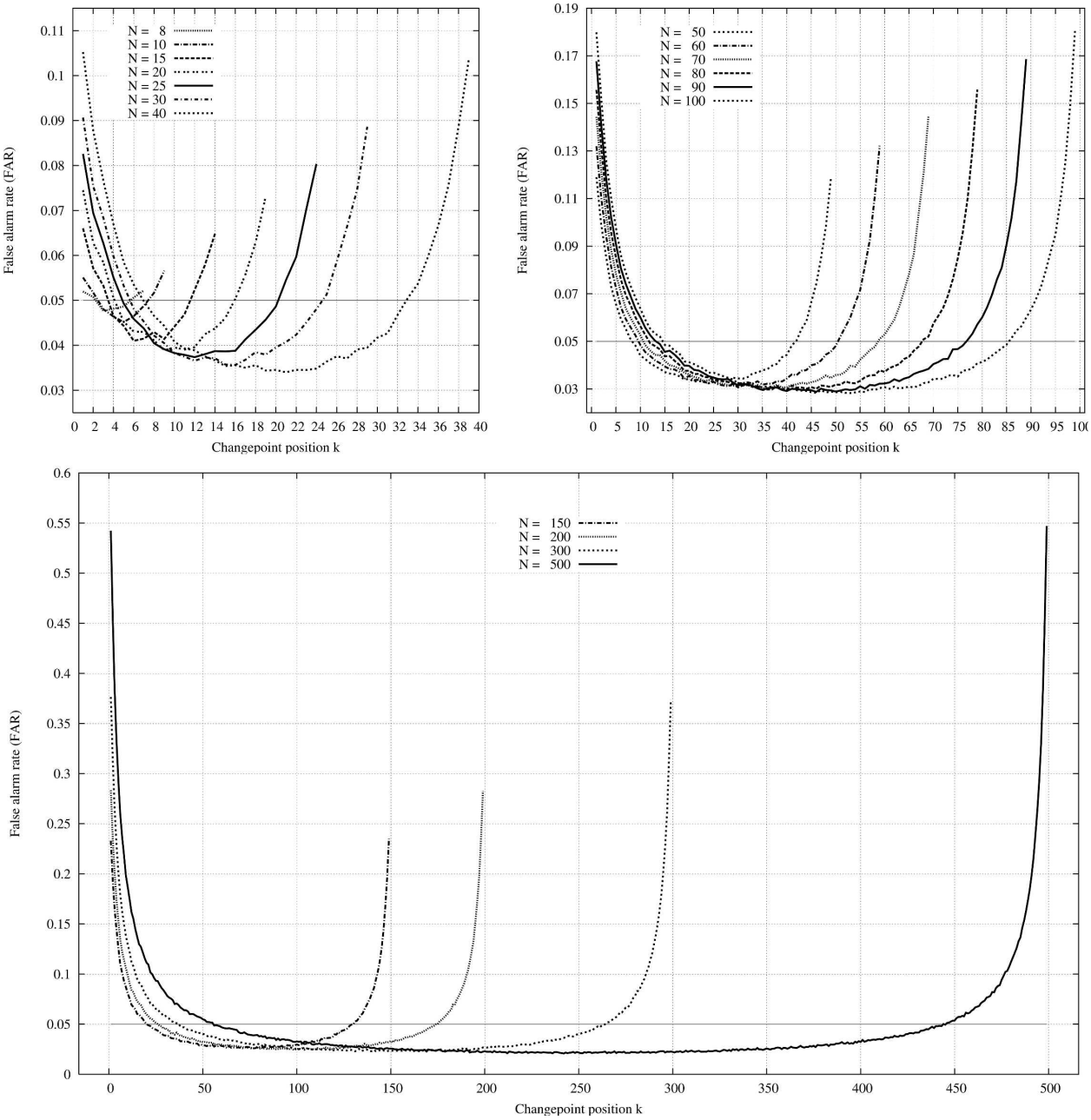

FIG. 1. The $\mathrm{FAR}_{\alpha}(k) \sim k$ curves for series of the indicated lengths N.

the ends of a homogeneous series to be mistakenly identified as a changepoint is much larger than those near the middle of the series. As a consequence, the effective level of confidence on the results of the above maximal t test (and its equivalent SNHT) is much lower than the specified 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 the middle of the series (i.e., $p_e > 0.95$ or $\alpha_e < 0.05$). That is, for a changepoint of certain magnitude, the test 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 when it occurs around the middle, and the test would mistakenly declare many more changepoints near the ends of a homogeneous series than around the middle. This problem arises when the two samples (before and after point k) are substantially unequal in size. The effect of unequal sample sizes on the statistic $T(k)$ is not very obvious, but it is enough to make an important difference in the

result of searching for the maximum value of $T(k)$ across $k \in \{1, 2, \ldots, N-1\}$ (it gets "exaggerated" by this inevitable "maximizing" process).

However, this kind of hypothesis testing works with a specified level of significance/confidence and thus is usually expected/assumed to perform effectively at the specified level under all circumstances. This is because 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 for all points in a homogeneous series. To construct a test of such highly desirable features, we use a penalty factor, which is empirically constructed using the ratios

$$R_k = \frac{T_{\max}[\alpha_e(k)]}{T_{\max}(0.05)}, \tag{6}$$

where $T_{\max}[\alpha_e(k)]$ and $T_{\max}(0.05)$ are the critical values of the $T_{\max}$ in (5) that are corresponding to the $\alpha_e(k) = \mathrm{FAR}_\alpha(k)$ and 5% level of significance, respectively. Similar to the way in which $\alpha_e(k) = \mathrm{FAR}_\alpha(k)$ values are estimated, the $T_{\max}[\alpha_e(k)]$ and $T_{\max}(0.05)$ and, hence, R_k are also estimated through Monte Carlo simulations. The estimated R_k values are shown as crosses, squares, circles, or asterisks in Fig. 2. By conducting these simulations, we convert the effect of unequal sample sizes on the $\mathrm{FAR}_\alpha(k)$ values (shown in Fig. 1) into the reverse of the effect on the $T_{\max}$ statistic, so that a penalty factor on $T(k)$ that would even out the U shape of the $\mathrm{FAR}_\alpha(k) \sim k$ curves should fit the ratios R_k well. Therefore, to construct such a penalty factor, we try to obtain least squares fits to the simulated R_k values, for each of the 18 selected values of N, subsequently. Bearing in mind the characteristics of the effect of unequal sample sizes (they are functions of k and N), we find out through trials that the penalty factor can be constructed using the following functions of k and N:

$$A = \left|1 - \frac{2k}{N}\right|, \quad B = \log(N),$$

$$C = \log(B), \quad \text{and} \quad D = \log[\log(N + 150)]. \tag{7}$$

By trial and error, we find the following penalty function:

$$P_o(k) = \frac{(11C^{9/8} + 195)}{200} F^{\upsilon}, \tag{8}$$

where

$$F = 1 - A^{(7B-2BC)/10} \quad \text{and} \quad \upsilon = \frac{15C^{1/2} - 11}{100} \tag{9}$$

for all time series of length $N \le 100$ and

$$F = 1 - A^{(11BC/50)} \quad \text{and} \quad \upsilon = \frac{2C^2 + 2C - 1}{100} \tag{10}$$

for time series of length $N > 100$. Figure 2 shows the fits of this penalty function (thin dashed curves; most of them are the same as the thick curves that are explained below) to the R_k values for each of the 17 selected values of N (the values for $N = 6$ have little dependence on k and hence are not shown here), which are all very good. Note that for time series of length $N > 100$ the penalty function with $\upsilon = (2C^3 + 2C^2 - 1)/(100C)$ fits the R_k values even better than does the one above, but further checks on the resulting $\mathrm{FAR}\alpha(k) \sim k$ curves indicate that it overpenalizes the test statistic.

In general, our experiments suggest that the penalty function in (8) tends to overpenalize the test statistic for the points that are near the ends of the series [the resulting $\mathrm{FAR}_\alpha(k) \sim k$ curves become M shaped, like the case for $N = 500$ in Fig. 3 but of much bigger amplitude]. Say, for a specific series length N, there are K_1 points at each end that are associated with penalty $P_o(k) \le 1$ [i.e., $P_o(1) < P_o(2) < \cdots < P_o(K_1) \le 1$ (the first K_1 points) and $P_o(N-1) < P_o(N-2) < \cdots < P_o(N-K_1) \le 1$ (the last K_1 points)] while $P_o(k) > 1$ is true for all $k \in [K_1 + 1, N - K_1 - 1]$. The above penalty term overpenalizes approximately the first and last L points, where $L = (\lfloor K_1/2 \rfloor + 3)$ for series of length $10 < N < 50$ and $L = (\lfloor K_1/2 \rfloor + 2)$ otherwise (here $\lfloor K_1/2 \rfloor$ means to take the floor value of the integer division $K_1/2$; L ranges between 1 and 26 for $N \in [6, 500]$). We speculate that this is because the estimates of the false-alarm rate and other statistics for these end points are much more unstable than for all the other points. Also, for very large N (e.g., $N = 500$), the above penalty function is a little too narrow (not quite visible in Fig. 2); the curve is almost vertical at the very ends of series, and so the penalty function is very sensitive to any error in the estimate of false-alarm rate for these points. Besides, we speculate that a penalty factor makes the statistic more sensitive to any estimation error than does an additive penalty term.

To avoid the overpenalization described above, we further modify the above penalty function as follows:

$$P(k) = \begin{cases} P_o(L) - \Theta(L - k) & \text{for} \quad k = 1, 2, \ldots, L \\ P_o(k) & \text{for} \quad k = (L+1), (L+2), \ldots, (N-L-1) \\ P_o(N-L) - \Theta(k - N + L) & \text{for} \quad k = (N-L), (N-L+1), \ldots, (N-1) \end{cases}, \tag{11}$$

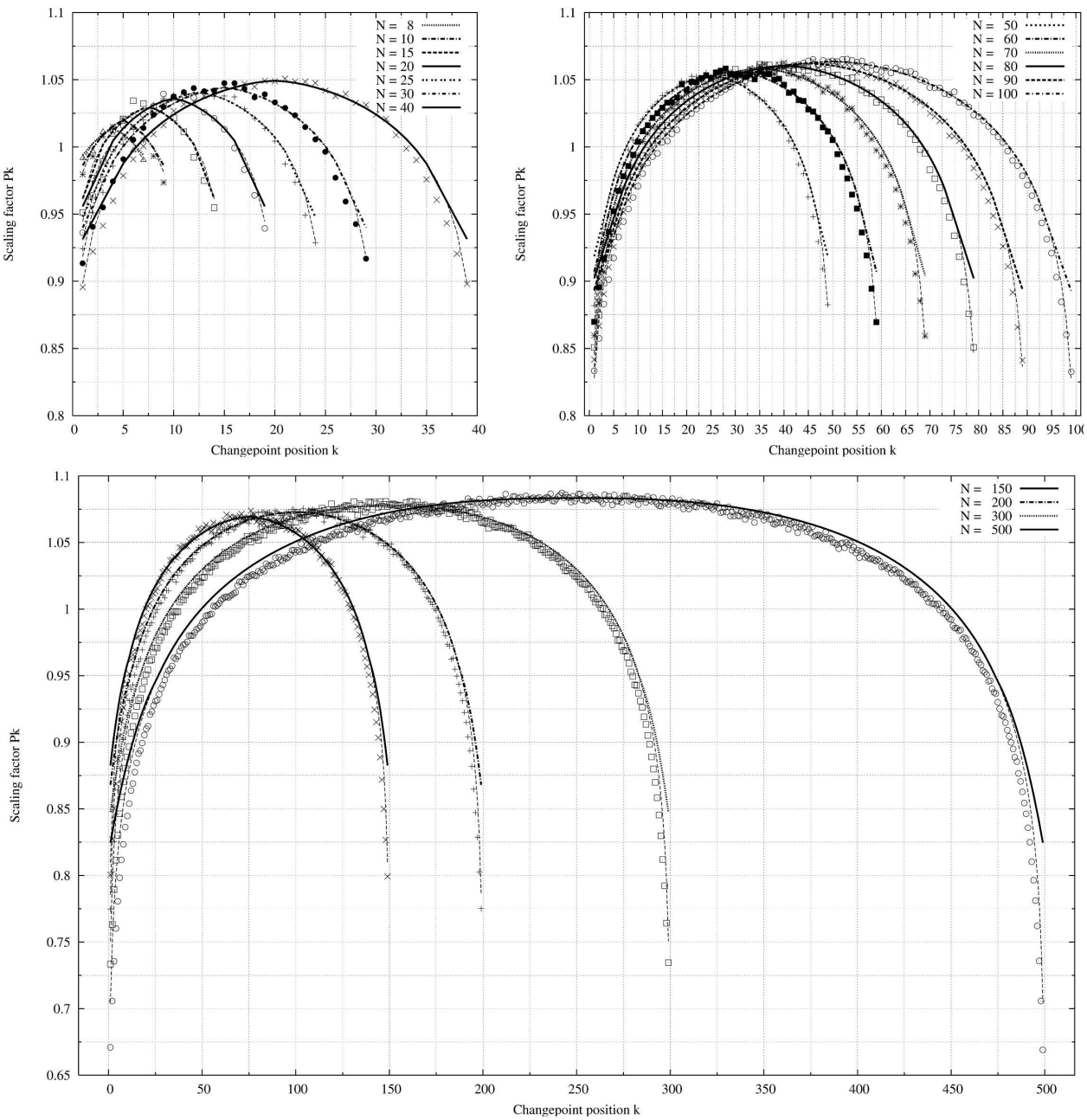

FIG. 2. The penalty term $P_o(k)$ (thin dashed curves) and $P(k)$ (thick curves), as well as their fits to the ratios R_k (crosses, squares, circles, and asterisks). See section 2 for the definition of these terms.

where

$$\Theta = \begin{cases} D^{1/2}\,[P_o(L+1) - P_o(L)] & \text{if} \quad N \le 10 \\ D^{1/3}\,[P_o(L+1) - P_o(L)] + 3/(10N^{4/3}) & \text{if} \quad 10 < N \le 100; \end{cases} \tag{12}$$

for time series of length $N > 100$,

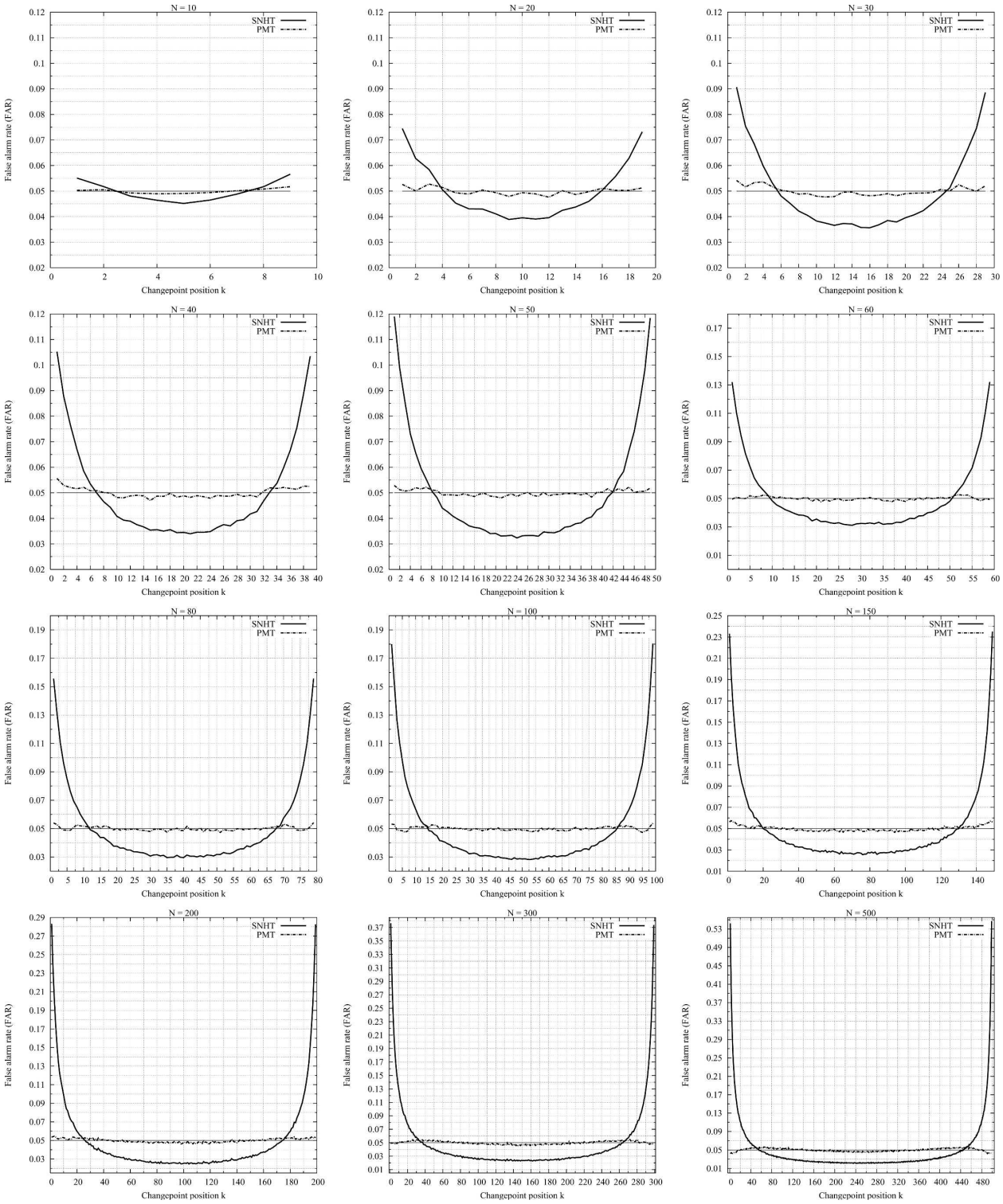

FIG. 3. Comparison of the $FAR_\alpha(k) \sim k$ curves of PMT and SNHT applied to homogeneous time series of the indicated lengths at the 5% significance level.

$$\Theta = \begin{cases} \dfrac{[P_o(L) - P_o(1)]}{2L - 4} A^{C^3} & \text{for} \quad k = 1, 2, \ldots, L \\ \dfrac{[P_o(N - L) - P_o(N - 1)]}{2L - 4} A^{C^3} & \text{for} \quad k = (N - L), (N - L + 1), \ldots, (N - 1). \end{cases} \tag{13}$$

Figure 2 also shows this modified penalty function in comparison with $P_o(k)$ and their fit to the R_k values for each of the 17 selected values of N.

Then, we propose the following penalized maximal t test (PMT):

$$\begin{aligned} &\mathrm{PT}(k) = P(k)T(k) \quad \text{and} \\ &\mathrm{PT}_{\max} = \max_{1 \le k \le N-1} [P(k)T(k)]. \end{aligned} \tag{14}$$

Figure 3 shows the $\mathrm{FAR}_\alpha(k) \sim k$ curves of PMT in comparison with the corresponding curves of SNHT for 12 selected values of N (it looks similar for other N values, and thus they are not shown). It is clear that PMT has a much more evenly distributed false-alarm rate across all points in the series than does SNHT, although it tends to overpenalize slightly the test statistic for the end points of very long time series ($N \ge 500$). The penalty factor does even out, to a great extent, the U shape of the $\mathrm{FAR}_\alpha(k) \sim k$ curves of the unpenalized maximal t test (and SNHT). The new test statistic takes the relative position of each candidate changepoint into account to reduce the distortion of the test statistic that is due to unequal sample sizes. Observations are treated more equally during the process of searching for the most probable changepoint position/time. This is of great practical importance, because it will result in a basically uniform level of confidence on the identified changepoints regardless of their position in the time series, or a uniform false-alarm rate across a homogeneous series.

We acknowledge that our approach here is purely empirical [i.e., both $P_o(k)$ and $P(k)$ are obtained empirically] and that a theoretically based penalty term is yet to be found. This empirical exercise basically achieves the desired results; it clearly reveals the characteristics of the effect of unequal sample sizes, which may help to develop a theoretically based penalty term.

Empirical critical values of $\mathrm{PT}_{\max}$ for some selected values of N are obtained by simulating 10 million $\mathrm{PT}_{\max}$ values under the null hypothesis of no changepoint and 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 $T_{\max}$ in SNHT are also simulated similarly and listed in Table 1, which will be used for the comparison in the next section.

The effect of unequal sample sizes may not be a big problem for detecting a large mean shift, but it cannot be ignored when the magnitude of mean shift is small or medium relative to the noise variance, especially in short time series (such as most annual climate data series). This will be illustrated through the comparison in the next section.

3. Comparison of PMT with SNHT

In this section, the performance of PMT is compared with that of SNHT through Monte Carlo simulations. To evaluate the performance of each method, we use three different measures of detection power. The first measure is position rate (PR), which is the rate of detecting a changepoint position "correctly" (i.e., the detected changepoint position is within the interval $[K - 2, K + 2]$ of the true position K), regardless of its significance (see the discussion below for the choice of this interval). The second measure is significance rate (SR), which is the rate of detecting a statistically significant (at the 5% level) changepoint regardless of whether or not its position is identified correctly. When each of the time series being tested contains a mean shift (i.e., H_a is true), $\hat{\beta} = (1 - \mathrm{SR})$ is a rough estimate of the type-II error rate β (i.e., the rate of mistakenly accepting the null hypothesis of no changepoint when a mean shift does exist in the time series), and $\mathrm{SR} = (1 - \hat{\beta})$ in this case. Note that $(1 - \beta)$ is statistically referred to as the power of test. Thus, SR can also be called the "power" of test when it is equal to $(1 - \hat{\beta})$ (such as in the setting described in the next paragraph). The third measure is the hit rate (HR), which is 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, whereas HR is a strict measure of its "accurate" detection power, and $(1 - \mathrm{HR})$ is a more accurate estimate of the type-II error rate β.

The comparison is carried out as follows. First, for each of the 18 selected values of series length N (ranging from 6 to 500), we generate 1000 homogeneous time series from an IID Gaussian distribution with zero mean and variance σ^2 (without lose of generality, we set $\sigma = 1$ here). We insert a mean-shift Δ at point k (between k and $k + 1$) for each $k \in \{2, 3, \ldots, N - 2\}$ and each mean-shift magnitude $\Delta \in \{0.25\sigma, 0.5\sigma, \sigma, 1.5\sigma, 2\sigma\}$. The ratio Δ/σ is called the relative step/shift size

TABLE 1. Empirical critical values of the $SNHT_{max}$, that is, the T_{max} in Alexandersson (1986) and PT_{max} statistics, obtained through 10 million simulations of each of the two statistics for each N value.

	PT_{max}				$SNHT_{max}$			
N	$\alpha = 0.10$	$\alpha = 0.05$	$\alpha = 0.01$	$\alpha = 0.0001$	$\alpha = 0.10$	$\alpha = 0.05$	$\alpha = 0.01$	$\alpha = 0.0001$
4	5.356	7.674	17.319	178.372	2.803	2.901	2.980	3.000
5	4.185	5.405	9.479	43.869	3.412	3.626	3.870	3.994
6	3.735	4.613	7.188	23.176	3.882	4.206	4.639	4.963
7	3.483	4.211	6.154	16.198	4.242	4.674	5.296	5.887
8	3.324	3.961	5.581	12.838	4.527	5.055	5.862	6.753
9	3.219	3.792	5.206	10.866	4.764	5.369	6.348	7.549
10	3.146	3.673	4.946	9.649	4.964	5.636	6.768	8.280
15	2.965	3.381	4.319	7.210	5.648	6.543	8.230	11.171
20	2.908	3.279	4.093	6.381	6.071	7.090	9.116	13.108
25	2.884	3.230	3.977	5.987	6.368	7.465	9.717	14.494
30	2.871	3.201	3.907	5.751	6.592	7.744	10.157	15.549
35	2.866	3.187	3.866	5.598	6.771	7.969	10.503	16.354
40	2.863	3.175	3.834	5.504	6.920	8.152	10.781	16.985
45	2.861	3.168	3.810	5.402	7.044	8.302	11.000	17.496
50	2.861	3.164	3.793	5.341	7.155	8.435	11.191	17.980
60	2.862	3.158	3.772	5.268	7.331	8.647	11.501	18.681
70	2.866	3.156	3.759	5.214	7.473	8.811	11.739	19.221
80	2.871	3.158	3.751	5.166	7.591	8.950	11.930	19.611
90	2.874	3.159	3.747	5.129	7.695	9.069	12.096	19.897
100	2.878	3.161	3.742	5.121	7.777	9.166	12.230	20.262
150	2.903	3.178	3.742	5.055	8.086	9.518	12.693	21.190
200	2.918	3.191	3.751	5.050	8.283	9.737	12.982	21.599
300	2.941	3.212	3.765	5.034	8.540	10.019	13.324	22.279
400	2.958	3.229	3.779	5.053	8.706	10.201	13.544	22.573
500	2.970	3.240	3.791	5.047	8.827	10.326	13.698	22.898
600	2.980	3.250	3.800	5.058	8.925	10.434	13.812	23.142
700	2.988	3.258	3.808	5.068	8.997	10.512	13.908	23.207
800	2.995	3.265	3.817	5.077	9.063	10.583	13.979	23.313

(RSS), that is, RSS $= \Delta/\sigma$; and we tried with RSS $\in \{0.25, 0.5, 1, 1.5, 2\}$ in this study. Then, we subsequently apply the two tests (PMT and SNHT), at the same 5% level of significance, to each of the 1000 time series for each combination of k and Δ [there are $5 \times (N - 3)$ combinations for each N value with five choices of Δ] to detect the mean shift. We calculate the three detection rates for each combination of N, k, and Δ values. As an example, we show the complete results for the case of $N = 20$ in Tables 2–4. For other selected values of N, only the mean hit rates, mean position rates, and mean significance rates (i.e., those averaged over all the changepoint positions $k \in \{2, 3, \ldots, N - 2\}$) are listed in Tables 5–7, respectively [note that we have tried to insert a changepoint in almost all possible positions (except $k = 1$ and $k = N - 1$; i.e., the case with the first or last datum being an outlier), one at a time, for each selected value of N; thus, it is too much to show the complete results for all 18 selected values of N].

To check the effect of the width of the interval $[K - \delta, K + \delta]$ for defining a correct identification of changepoint position, for the case of $N = 100$ we carry out the above simulations using $\delta = 1, 2, \ldots, 10$, subsequently. The resulting mean position rates are shown in Table 8. It is clear that the results of the comparison of PMT with SNHT have little dependence on the choice of δ value (see the ratios in parentheses in Table 8). We see a small dependence only for the cases of identifying very small shifts ($\Delta = 0.25\sigma$) in which the narrower the width of the interval that is used the more superior PMT is over SNHT. However, there is no notable difference for $\delta = 2$–7. In this study, as described above, we chose $\delta = 2$, so that the interval is also suitable for very short time series (such as $N = 6$, the lowest among the 18 selected values of N we investigate in this study).

As would be expected, PMT has higher power of detection (in terms of all three detection rates) than does SNHT for changepoints that occur not too close to the ends of the series while it has lower power of detection for changepoints that are near the ends of the series (Tables 2–4), because it tries to obtain an evenly distributed effective level of significance/confidence (and hence power) of the test across all possible changepoint positions k, that is, to make the effective level of significance/confidence of the test close to the specified level for every point.

TABLE 2. Hit rates of PMT and SNHT (PMT/SNHT, in counts per 1000) for time series of length $N = 20$. The numbers in parentheses are the ratios of the PMT rate to the SNHT rate. The "mean" row is the average of the column (i.e., over changepoint positions $k \in \{2, 3, \ldots, N - 2\}$).

k	$\Delta = 0.25\sigma$	$\Delta = 0.5\sigma$	$\Delta = \sigma$	$\Delta = 1.5\sigma$	$\Delta = 2\sigma$
2	20/22 (0.91)	29/34 (0.85)	73/83 (0.88)	166/183 (0.91)	339/364 (0.93)
3	23/24 (0.96)	38/42 (0.90)	123/134 (0.92)	258/276 (0.93)	511/526 (0.97)
4	18/19 (0.95)	38/39 (0.97)	157/153 (1.03)	376/371 (1.01)	659/666 (0.99)
5	19/16 (1.19)	44/43 (1.02)	185/183 (1.01)	430/416 (1.03)	733/732 (1.00)
6	19/18 (1.06)	43/39 (1.10)	217/200 (1.09)	526/515 (1.02)	792/783 (1.01)
7	18/18 (1.00)	49/41 (1.20)	245/230 (1.07)	561/540 (1.04)	834/824 (1.01)
8	17/15 (1.13)	55/42 (1.31)	263/234 (1.12)	585/562 (1.04)	866/848 (1.02)
9	20/15 (1.33)	62/51 (1.22)	279/249 (1.12)	600/580 (1.03)	871/852 (1.02)
10	21/18 (1.17)	55/48 (1.15)	282/249 (1.13)	599/563 (1.06)	887/859 (1.03)
11	20/16 (1.25)	60/47 (1.28)	271/248 (1.09)	597/563 (1.06)	863/840 (1.03)
12	22/19 (1.16)	61/50 (1.22)	258/242 (1.07)	580/558 (1.04)	863/836 (1.03)
13	19/19 (1.00)	57/49 (1.16)	254/232 (1.09)	568/544 (1.04)	844/834 (1.01)
14	19/18 (1.06)	51/45 (1.13)	220/212 (1.04)	528/507 (1.04)	809/793 (1.02)
15	15/16 (0.94)	45/43 (1.05)	194/184 (1.05)	473/466 (1.02)	764/752 (1.02)
16	13/12 (1.08)	33/32 (1.03)	157/157 (1.00)	403/399 (1.01)	691/691 (1.00)
17	13/16 (0.81)	24/24 (1.00)	106/112 (0.95)	268/280 (0.96)	540/562 (0.96)
18	7/9 (0.78)	16/18 (0.89)	55/67 (0.82)	165/181 (0.91)	331/361 (0.92)
Mean	17.8/17.1 (1.04)	44.7/40.4 (1.11)	196.4/186.4 (1.05)	451.9/441.4 (1.02)	717.5/713.1 (1.01)

Most important is that, when averaged over all possible changepoint positions $k \in \{2, 3, \ldots, N - 2\}$, PMT significantly outperforms SNHT in detecting small shifts no matter whether in short or long time series and in detecting medium–large shifts in short time series. For large shifts in very long time series, however, SNHT is only slightly better than PMT (Fig. 4 and Tables 5–7).

In terms of the mean hit rate (see Fig. 4a), the improvement of PMT over SNHT can be as much as 14%–25% for detecting small shifts (RSS ≤ 0.5) and up to 5% for detecting medium shifts (e.g., RSS = 1.0 and 1.5) in time series of length $N < 100$, whereas SNHT only has about 1%–2% higher hit rates in detecting medium–large shifts in long time series (of length $N > 100$). Note that the smaller the RSS is, the greater is the improvement of PMT over SNHT (Fig. 4a). Also, the largest improvement occurs when the time series length N is less than 100 (which is true for most annual climate data series), and the larger the RSS is, the shorter is the

TABLE 3. As in Table 2 but for the position rates of PMT and SNHT (PMT/SNHT, in counts per 1000) for time series of length $N = 20$.

k	$\Delta = 0.25\sigma$	$\Delta = 0.5\sigma$	$\Delta = \sigma$	$\Delta = 1.5\sigma$	$\Delta = 2\sigma$
2	300/314 (0.96)	331/352 (0.94)	455/476 (0.96)	631/656 (0.96)	776/797 (0.97)
3	344/367 (0.94)	398/407 (0.98)	557/570 (0.98)	730/747 (0.98)	866/879 (0.99)
4	290/294 (0.99)	349/354 (0.99)	552/558 (0.99)	771/776 (0.99)	897/906 (0.99)
5	265/260 (1.02)	345/335 (1.03)	563/550 (1.02)	777/776 (1.00)	908/911 (1.00)
6	248/239 (1.04)	346/335 (1.03)	602/585 (1.03)	802/791 (1.01)	917/918 (1.00)
7	251/237 (1.06)	356/342 (1.04)	611/593 (1.03)	811/806 (1.01)	913/915 (1.00)
8	246/227 (1.08)	343/326 (1.05)	610/587 (1.04)	811/799 (1.02)	932/926 (1.01)
9	250/234 (1.07)	352/330 (1.07)	613/586 (1.05)	817/808 (1.01)	935/928 (1.01)
10	255/230 (1.11)	341/321 (1.06)	623/600 (1.04)	821/800 (1.03)	941/927 (1.02)
11	250/232 (1.08)	349/325 (1.07)	607/585 (1.04)	829/809 (1.02)	925/919 (1.01)
12	260/240 (1.08)	352/330 (1.07)	601/581 (1.03)	822/808 (1.02)	929/922 (1.01)
13	257/234 (1.10)	334/321 (1.04)	607/587 (1.03)	818/799 (1.02)	922/925 (1.00)
14	284/269 (1.06)	367/354 (1.04)	589/582 (1.01)	802/793 (1.01)	920/915 (1.01)
15	289/280 (1.03)	352/341 (1.03)	584/581 (1.01)	797/792 (1.01)	917/918 (1.00)
16	291/295 (0.99)	358/359 (1.00)	568/565 (1.01)	773/779 (0.99)	911/917 (0.99)
17	322/334 (0.96)	378/397 (0.95)	562/577 (0.97)	731/747 (0.98)	871/889 (0.98)
18	276/286 (0.97)	305/319 (0.96)	460/482 (0.95)	627/660 (0.95)	799/815 (0.98)
Mean	275.2/268.9 (1.02)	350.4/344.0 (1.02)	574.4/567.4 (1.01)	774.7/773.3 (1.00)	898.8/901.6 (1.00)

TABLE 4. As in Table 2 but for the significance rates of PMT and SNHT (PMT/SNHT, in counts per 1000) for time series of length $N = 20$.

k	$\Delta = 0.25\sigma$	$\Delta = 0.5\sigma$	$\Delta = \sigma$	$\Delta = 1.5\sigma$	$\Delta = 2\sigma$
2	57/53 (1.08)	67/66 (1.02)	115/120 (0.96)	214/225 (0.95)	387/407 (0.95)
3	58/58 (1.00)	71/76 (0.93)	165/174 (0.95)	314/327 (0.96)	565/574 (0.98)
4	59/60 (0.98)	80/83 (0.96)	213/210 (1.01)	450/439 (1.03)	722/724 (1.00)
5	66/62 (1.06)	92/89 (1.03)	274/267 (1.03)	535/518 (1.03)	802/795 (1.01)
6	64/64 (1.00)	93/94 (0.99)	315/297 (1.06)	623/611 (1.02)	859/843 (1.02)
7	65/68 (0.96)	99/96 (1.03)	347/328 (1.06)	654/634 (1.03)	897/885 (1.01)
8	66/70 (0.94)	116/101 (1.15)	372/344 (1.08)	680/655 (1.04)	916/899 (1.02)
9	68/64 (1.06)	120/110 (1.09)	384/354 (1.08)	698/679 (1.03)	926/912 (1.02)
10	70/67 (1.04)	120/114 (1.05)	391/361 (1.08)	701/678 (1.03)	935/918 (1.02)
11	69/65 (1.06)	121/108 (1.12)	375/349 (1.07)	702/672 (1.04)	928/910 (1.02)
12	67/65 (1.03)	118/109 (1.08)	362/348 (1.04)	684/660 (1.04)	921/898 (1.03)
13	64/63 (1.02)	109/107 (1.02)	336/318 (1.06)	652/633 (1.03)	902/887 (1.02)
14	64/65 (0.98)	105/99 (1.06)	293/286 (1.02)	599/584 (1.03)	866/850 (1.02)
15	61/63 (0.97)	98/94 (1.04)	265/252 (1.05)	545/535 (1.02)	812/803 (1.01)
16	55/60 (0.92)	82/77 (1.06)	221/221 (1.00)	459/453 (1.01)	734/727 (1.01)
17	55/60 (0.92)	66/64 (1.03)	153/158 (0.97)	326/329 (0.99)	589/602 (0.98)
18	50/55 (0.91)	57/61 (0.93)	93/106 (0.88)	209/220 (0.95)	375/401 (0.94)
Mean	62.2/62.5 (1.00)	94.9/91.1 (1.04)	274.9/264.3 (1.04)	532.1/520.7 (1.02)	772.7/766.8 (1.01)

length of time series in which the largest improvement is obtained (i.e., the peaks of the curves shown in Fig. 4a occur at smaller N for larger RSS).

As shown in Fig. 4b, PMT has up to about 9% higher power in identifying correctly the position of small shifts in all time series (short or long), whereas PMT and SNHT perform very similarly in detecting the correct position of medium–large shifts (with PMT being slightly better when the time series is short and SNHT being slightly better when the time series is very long).

Because each of the 1000 time series tested here contains a mean shift (i.e., H_a is always true here), the type-II error rate can be estimated as (1 − significance rate). Thus, as mentioned before, the higher the significance rate is, the lower is the type-II error rate and hence the better is the test. [In the mean time, a more accurate/strict measure of the type-II error rate can be derived from the hit rate, i.e., the "miss rate," defined as (1 − hit rate), which, however, contains the same information as the hit rate.] In terms of significance rate, as shown in Fig. 4c, the improvement of PMT over SNHT is similar to that in terms of the mean hit rate,

TABLE 5. As in Table 2 but for the mean hit rates of PMT and SNHT for the indicated different series lengths N (i.e., a summary of the results similar to the last row of Table 2 for different N values).

N	$\Delta = 0.25\sigma$	$\Delta = 0.5\sigma$	$\Delta = \sigma$	$\Delta = 1.5\sigma$	$\Delta = 2\sigma$
6	45.0/44.3 (1.02)	56.3/56.3 (1.00)	78.0/77.7 (1.00)	116.0/115.3 (1.01)	173.0/171.0 (1.01)
8	32.0/31.8 (1.01)	42.0/42.2 (1.00)	88.8/86.8 (1.02)	176.6/173.4 (1.02)	295.0/290.8 (1.01)
10	29.6/29.9 (0.99)	45.6/44.9 (1.02)	108.0/105.3 (1.03)	224.6/219.0 (1.03)	399.4/393.7 (1.01)
15	22.8/21.9 (1.04)	42.8/42.0 (1.02)	149.9/142.9 (1.05)	348.0/337.7 (1.03)	610.3/602.2 (1.01)
20	17.8/17.1 (1.04)	44.7/40.4 (1.11)	196.4/186.4 (1.05)	451.9/441.4 (1.02)	717.5/713.1 (1.01)
25	18.6/17.8 (1.05)	49.6/45.5 (1.09)	239.7/227.9 (1.05)	533.7/526.4 (1.01)	780.1/783.4 (1.00)
30	17.6/16.4 (1.07)	54.5/49.3 (1.11)	271.7/258.8 (1.05)	591.3/585.6 (1.01)	814.5/820.4 (0.99)
40	16.7/14.6 (1.15)	61.2/53.9 (1.14)	337.7/323.2 (1.04)	659.6/660.5 (1.00)	846.9/857.6 (0.99)
50	14.7/13.2 (1.11)	65.1/57.3 (1.14)	374.0/362.2 (1.03)	691.2/697.1 (0.99)	861.2/873.5 (0.99)
60	15.9/14.6 (1.08)	75.4/66.8 (1.13)	418.3/407.5 (1.03)	720.2/729.0 (0.99)	877.4/889.3 (0.99)
70	14.4/12.0 (1.20)	82.6/72.5 (1.14)	445.1/437.9 (1.02)	734.0/744.5 (0.99)	882.3/895.3 (0.99)
80	14.6/12.0 (1.22)	90.7/80.2 (1.13)	467.8/464.6 (1.01)	744.3/755.7 (0.98)	887.0/899.2 (0.99)
90	16.7/13.4 (1.25)	98.7/87.9 (1.12)	485.8/484.1 (1.00)	753.8/765.8 (0.98)	891.2/903.4 (0.99)
100	18.6/15.4 (1.20)	107.3/95.6 (1.12)	501.1/503.1 (1.00)	760.8/772.4 (0.98)	894.1/906.7 (0.99)
150	20.0/16.2 (1.24)	147.7/134.6 (1.10)	541.7/548.5 (0.99)	784.9/797.1 (0.98)	909.7/919.2 (0.99)
200	23.9/19.9 (1.20)	177.4/166.8 (1.06)	564.5/573.3 (0.98)	796.8/809.1 (0.98)	915.2/924.0 (0.99)
300	32.3/26.4 (1.23)	210.5/204.0 (1.03)	583.0/594.0 (0.98)	808.1/820.0 (0.99)	921.5/929.6 (0.99)
500	46.2/39.8 (1.16)	241.5/242.2 (1.00)	600.0/611.4 (0.98)	818.9/828.9 (0.99)	927.3/934.7 (0.99)

TABLE 6. As in Table 3 but for the mean position rates of PMT and SNHT for the indicated different series lengths N (i.e., a summary of the results similar to the last row of Table 3 for different N values).

N	$\Delta = 0.25\sigma$	$\Delta = 0.5\sigma$	$\Delta = \sigma$	$\Delta = 1.5\sigma$	$\Delta = 2\sigma$
6	849.7/849.3 (1.00)	855.7/855.3 (1.00)	886.0/885.7 (1.00)	923.7/923.0 (1.00)	956.0/955.3 (1.00)
8	637.2/632.0 (1.01)	655.2/653.4 (1.00)	738.8/738.2 (1.00)	841.8/839.2 (1.00)	918.8/917.6 (1.00)
10	501.6/497.6 (1.01)	538.3/534.3 (1.01)	659.4/656.4 (1.00)	804.1/800.0 (1.01)	904.9/902.7 (1.00)
15	343.5/338.1 (1.02)	405.5/398.9 (1.02)	597.1/589.7 (1.01)	774.4/771.1 (1.00)	891.2/891.0 (1.00)
20	275.2/268.9 (1.02)	350.4/344.0 (1.02)	574.4/567.4 (1.01)	774.7/773.3 (1.00)	898.8/901.6 (1.00)
25	222.7/217.6 (1.02)	315.3/306.1 (1.03)	563.6/556.7 (1.01)	770.5/769.6 (1.00)	896.3/899.5 (1.00)
30	193.4/188.1 (1.03)	298.7/290.4 (1.03)	555.3/548.7 (1.01)	768.0/771.0 (1.00)	895.7/900.8 (0.99)
40	158.0/152.3 (1.04)	270.8/261.3 (1.04)	558.8/555.1 (1.01)	780.1/784.2 (0.99)	901.3/908.0 (0.99)
50	135.8/129.8 (1.05)	258.9/249.6 (1.04)	555.6/552.5 (1.01)	778.7/785.0 (0.99)	903.0/911.9 (0.99)
60	125.8/119.8 (1.05)	254.7/243.6 (1.05)	563.2/561.6 (1.00)	788.5/796.6 (0.99)	911.0/919.2 (0.99)
70	115.0/109.0 (1.05)	250.9/242.6 (1.03)	567.5/567.4 (1.00)	790.8/799.6 (0.99)	911.4/920.6 (0.99)
80	109.6/103.8 (1.06)	246.5/237.0 (1.04)	569.7/571.1 (1.00)	793.6/801.7 (0.99)	912.6/920.8 (0.99)
90	103.7/99.1 (1.05)	244.3/236.7 (1.03)	572.5/575.3 (1.00)	798.0/807.2 (0.99)	914.5/922.3 (0.99)
100	101.5/96.1 (1.06)	243.7/236.8 (1.03)	576.5/579.8 (0.99)	798.2/807.2 (0.99)	914.1/922.9 (0.99)
150	92.1/86.4 (1.07)	250.1/244.1 (1.02)	589.9/595.9 (0.99)	810.3/820.0 (0.99)	923.5/930.3 (0.99)
200	88.8/82.7 (1.07)	254.9/251.8 (1.01)	599.0/606.7 (0.99)	816.3/825.7 (0.99)	925.6/931.6 (0.99)
300	84.1/77.4 (1.09)	258.9/258.2 (1.00)	605.8/615.1 (0.98)	820.9/830.4 (0.99)	928.8/934.7 (0.99)
500	86.7/81.7 (1.06)	267.9/270.6 (0.99)	614.4/623.9 (0.98)	827.2/835.2 (0.99)	932.2/937.7 (0.99)

with up to 16% improvement for small shifts and up to 5% improvement for medium shifts in time series of length $N < 100$, and the two methods perform very similarly in detecting large shifts.

As shown in Fig. 4, the hit rate is much more sensitive to change in the length of time series being tested or in the magnitude of shift than is the position or significance rate. This is because a hit is counted only when the test identifies the changepoint with both statistical significance and correct position.

4. An application of PMT and SNHT to climate data series

In this section, we present an application of PMT and SNHT to detect undocumented mean shifts in climate data series. The purpose is to see 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., we know exactly when the shift

TABLE 7. As in Table 4 but for the mean significance rates of PMT and SNHT for the indicated different series lengths N (i.e., a summary of the results similar to the last row of Table 4 for different N values).

N	$\Delta = 0.25\sigma$	$\Delta = 0.5\sigma$	$\Delta = \sigma$	$\Delta = 1.5\sigma$	$\Delta = 2\sigma$
6	50.7/50.0 (1.01)	63.0/63.0 (1.00)	84.7/84.3 (1.00)	122.7/122.3 (1.00)	175.0/173.3 (1.01)
8	48.0/47.8 (1.00)	56.6/57.0 (0.99)	103.0/100.6 (1.02)	188.0/185.2 (1.02)	302.4/298.4 (1.01)
10	54.0/54.6 (0.99)	71.0/71.3 (1.00)	132.9/130.1 (1.02)	247.9/242.9 (1.02)	416.4/410.9 (1.01)
15	60.3/61.5 (0.98)	83.2/83.1 (1.00)	199.8/194.0 (1.03)	401.1/390.5 (1.03)	650.8/642.4 (1.01)
20	62.2/62.5 (1.00)	94.9/91.1 (1.04)	274.9/264.3 (1.04)	532.1/520.7 (1.02)	772.7/766.8 (1.01)
25	68.7/70.2 (0.98)	116.3/112.6 (1.03)	346.2/333.0 (1.04)	640.1/631.3 (1.01)	847.0/847.7 (1.00)
30	71.7/70.9 (1.01)	127.7/120.4 (1.06)	403.2/387.9 (1.04)	718.3/710.6 (1.01)	887.0/890.7 (1.00)
40	79.6/76.5 (1.04)	158.0/147.2 (1.07)	509.6/491.3 (1.04)	805.4/804.5 (1.00)	923.0/930.8 (0.99)
50	76.7/73.1 (1.05)	175.9/161.9 (1.09)	585.9/569.8 (1.03)	850.7/854.0 (1.00)	938.0/944.9 (0.99)
60	86.3/80.6 (1.07)	210.2/192.4 (1.09)	665.4/652.2 (1.02)	882.2/888.0 (0.99)	950.5/957.8 (0.99)
70	86.2/81.1 (1.06)	237.0/218.6 (1.08)	713.1/703.2 (1.01)	898.7/906.2 (0.99)	955.8/963.4 (0.99)
80	91.4/88.2 (1.04)	269.6/248.8 (1.08)	756.2/751.1 (1.01)	911.2/920.0 (0.99)	960.9/968.3 (0.99)
90	98.3/89.2 (1.10)	299.4/273.8 (1.09)	785.9/784.0 (1.00)	921.0/930.0 (0.99)	965.0/972.4 (0.99)
100	111.1/101.7 (1.09)	336.2/310.5 (1.08)	815.0/817.7 (1.00)	930.3/939.0 (0.99)	968.8/975.8 (0.99)
150	142.2/127.0 (1.12)	481.9/447.0 (1.08)	878.4/884.5 (0.99)	952.0/959.3 (0.99)	978.4/983.6 (0.99)
200	180.5/158.7 (1.14)	594.0/562.6 (1.06)	909.6/917.5 (0.99)	963.2/970.6 (0.99)	983.4/988.2 (1.00)
300	267.2/230.3 (1.16)	732.4/712.9 (1.03)	937.3/945.8 (0.99)	974.0/980.4 (0.99)	987.8/991.9 (1.00)
500	411.9/362.0 (1.14)	846.4/846.6 (1.00)	959.6/968.2 (0.99)	983.1/988.2 (0.99)	991.8/995.2 (1.00)

TABLE 8. As in Table 6 but for the mean position rates of PMT and SNHT (for time series of $N = 100$) as a function of the width of the interval $[K - \delta, K + \delta]$ used to define correct identification of changepoint position.

δ	$\Delta = 0.25\sigma$	$\Delta = 0.5\sigma$	$\Delta = \sigma$	$\Delta = 1.5\sigma$	$\Delta = 2\sigma$
1	67.0/62.9 (1.07)	172.0/167.0 (1.03)	455.8/458.6 (0.99)	687.0/694.4 (0.99)	836.8/845.5 (0.99)
2	101.8/96.1 (1.06)	244.0/236.8 (1.03)	576.5/579.8 (0.99)	798.0/807.2 (0.99)	913.8/922.9 (0.99)
3	133.2/125.9 (1.06)	302.2/293.4 (1.03)	657.3/661.0 (0.99)	858.3/868.2 (0.99)	947.7/956.0 (0.99)
4	161.9/153.2 (1.06)	350.6/340.8 (1.03)	714.0/718.3 (0.99)	893.9/904.0 (0.99)	963.9/971.9 (0.99)
5	188.3/178.3 (1.06)	391.9/381.2 (1.03)	755.3/760.3 (0.99)	916.6/926.4 (0.99)	972.7/980.0 (0.99)
6	212.9/201.7 (1.06)	427.5/416.2 (1.03)	787.0/792.5 (0.99)	931.5/941.3 (0.99)	977.8/984.4 (0.99)
7	236.0/223.6 (1.06)	458.9/446.7 (1.03)	811.3/817.4 (0.99)	941.9/951.2 (0.99)	981.0/987.1 (0.99)
8	257.4/244.2 (1.05)	486.9/474.0 (1.03)	830.8/837.1 (0.99)	949.3/958.4 (0.99)	983.1/988.7 (0.99)
9	277.8/263.6 (1.05)	512.0/498.5 (1.03)	846.8/853.0 (0.99)	954.8/963.5 (0.99)	984.6/989.6 (0.99)
10	297.1/281.8 (1.05)	534.5/520.5 (1.03)	859.7/865.8 (0.99)	958.9/967.3 (0.99)	985.6/990.4 (1.00)

occurred) to see if the methods can detect the mean shift if we did not know about it (i.e., if we treat it as undocumented for the purpose of this application). To this end, we apply PMT and SNHT to time series of monthly mean and annual mean station pressure recorded at Burgeo (Newfoundland, Canada) for the 28-yr period from January 1967 to December 1994 (no data outside this period), because we know that the pressure series contains an artificial mean shift (see Fig. 5a) that is caused by neglecting the station elevation of 10.6 m in the calculation of station pressures from barometer readings in the period prior to January 1977 (a problem of the so-called 50-feet rule, which is to use zero elevation in the calculation of station pressures from barometer readings if the station elevation is less than 50 ft, i.e., 15 m). According to a physically based estimate using a hydrostatic model and hourly pressure and temperature data (Wan et al. 2007), neglecting such an elevation causes a bias of 1.32 hPa on pressure values.

We also know that the station pressure recorded at Yarmouth Airport (Nova Scotia, Canada) for the same 28-yr period is basically homogeneous and is highly correlated with the pressure data recorded at Burgeo. It is the best available reference series for the Burgeo series. Thus, we use the corresponding pressure series from Yarmouth Airport as reference series in this application.

Because both SNHT and PMT assume that the time series being tested is IID Gaussian, we apply the tests to the annual mean and monthly mean pressure series for each calendar month (13 time series in total) separately, to minimize the effect of autocorrelation. As a result, both PMT and SNHT identify a changepoint of at least 5% significance from the annual mean series and from the monthly mean series for July, September, and November (see Table 9 and Figs. 5b–d) and a changepoint (1976) of 5%–10% significance from the December mean pressure series (not shown or listed in

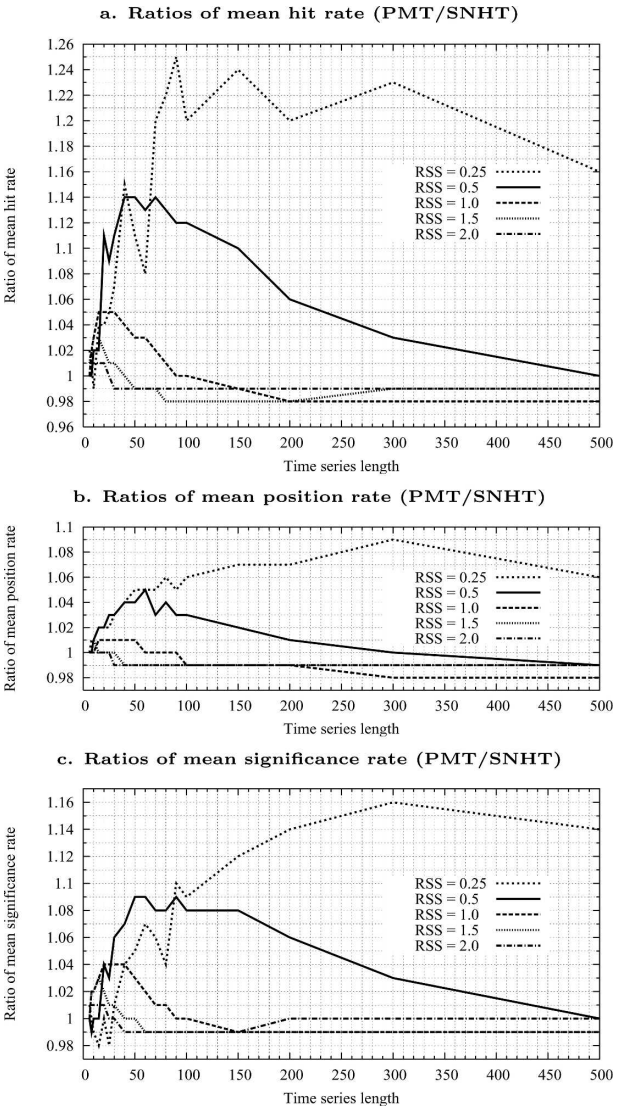

FIG. 4. Ratios of the (a) mean hit rate, (b) mean position rate, and (c) mean significance rate (PMTs over SNHTs) as a function of time series length N and the RSS = Δ/σ.

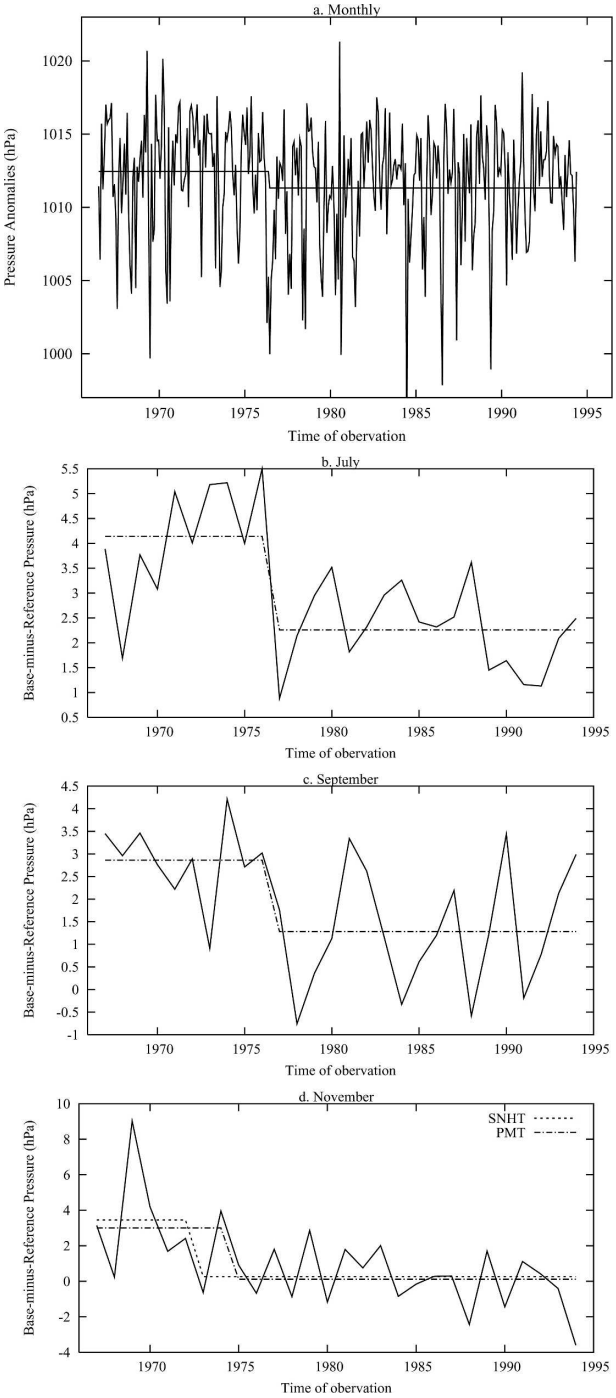

FIG. 5. (a) Time series of consecutive monthly mean station pressure recorded at Burgeo for the 28-yr period 1967–94. (b), (c) Time series of monthly mean pressure differences between Burgeo (base station) and Yarmouth Airport (reference station) for the indicated month of year and the undocumented mean shifts detected by applying PMT and SNHT to the pressure difference series.

Table 9). Both methods are consistently correct in the identification of the mean shift in 1976 from the annual mean series and from the monthly mean series for July and September (Table 9). However, PMT is more accurate in identifying the mean shift from the November mean pressure series; it identified the mean shift to be in 1974 (two intervals/years earlier than the true position), whereas SNHT found it to be in 1972 (four intervals/years earlier; see Table 9 and Fig. 5c).

Both methods found no significant changepoint from the monthly mean pressure series for the other calendar months. One of the reasons for the detection failure is the presence of autocorrelation in the time series. Although all of the base-minus-reference pressure series tested here are annual series (i.e., the interval between two consecutive data is 1 yr), their lag-1 autocorrelation (after taking into account the mean shift in 1976) ranges from -0.303 to 0.117, with more months having a negative autocorrelation. According to Lund et al. (2007), ignoring a positive autocorrelation will increase the false-alarm rate of the test, whereas ignoring a negative one will let real changepoints go undetected, and the larger the autocorrelation is in absolute value, the greater is the effect. The large negative autocorrelations in the base-minus-reference pressure series definitely contributes to the detection failure. In addition, sampling variability is larger for short time series than for long series. The short series length (here $N = 28$) also makes the detection harder.

5. Concluding remarks

Based on empirical methods, a penalized maximal t test is proposed for detecting undocumented mean shifts in climate data series. PMT takes the relative position of each candidate changepoint into account to diminish the effect of unequal sample sizes on the false-alarm rate and hence on the power of detection. It has been shown, for time series of selected lengths $N \in [6, 500]$, that the false-alarm rate of PMT is evenly distributed across all candidate changepoint positions $k \in \{1, 2, \ldots, N - 1\}$; it is very close to the specified level of significance at all candidate changepoints $k \in \{1, 2, \ldots, N - 1\}$. This feature is highly desirable in practice, because each point of data should be treated equally: each should have equal chance to be picked mistakenly as a changepoint when the time series is homogeneous (i.e., the same false-alarm rate); on the other hand, a shift of certain magnitude should have the same probability of being detected no matter where the shift occurs (near the ends or the middle of the series). Without this feature, the resulting level of confidence on the identified changepoints that are near the ends of the series is much lower than the specified level and is much higher

TABLE 9. Significant (at the 5% level) undocumented changepoints detected by applying SNHT and PMT to annual and monthly mean station pressure data recorded at Burgeo during the 28-yr period from January 1967 to December 1994, using as reference series the station pressure data series recorded at Yarmouth Airport for the same period (which is found to be homogeneous).

	PMT results			SNHT results		
Series tested	k	PT_{max}	$PT_{max}(0.05)$	k	$SNHT_{max}$	$SNHT_{max}(0.05)$
Annual means	1976	3.745	3.213	1976	9.058	7.640
Jan means						
Feb means						
Mar means						
Apr means						
May means						
Jun means						
July means	1976	5.177	3.213	1976	13.256	7.640
Aug means						
Sep means	1976	3.503	3.213	1976	8.272	7.640
Oct means						
Nov means	1974	3.432	3.213	1972	8.220	7.640
Dec means						

on those that are near the middle. However, it is also shown that the false-alarm rate of SNHT can be up to 10 times the specified level for points near the ends of the series and much lower for the middle points (Fig. 3); that is, each point of data in the time series is not treated equally.

PMT consequently has higher power in detecting changepoints that are not too close to the ends of series and has lower power for detecting changepoints that are near the ends of series, in comparison with SNHT. However, note that the higher power of SNHT for detecting changepoints near the ends of series arises from the fact that the effect of unequal sample sizes makes the effective level of confidence of SNHT much lower than the specified level for the end points, which is not desirable. What is highly desirable is for a test to perform effectively at the specified level of significance/confidence no matter where the shift occurs, that is, to have the same probability of detecting a shift of certain magnitude regardless of the position of shift (near the ends or the middle of the series). PMT has this highly desirable feature, although it is constructed empirically.

Most important is that, when averaged over all possible changepoint positions, PMT has higher power of detection. In terms of hit rate, the improvement of PMT over SNHT can be as much as 14%–25% for detecting small shifts ($\Delta < \sigma$) regardless of time series length and up to 5% for detecting medium shifts ($\Delta = \sigma$–1.5σ) in time series of length $N < 100$. The smaller the relative shift size RSS $= \Delta/\sigma$ is, the greater is the improvement. The largest improvement is obtained for time series of length $N < 100$, which is of great practical importance, because most annual climate data series are shorter than this length (or the period that contains only one changepoint is shorter than this).

Note that the effect of unequal sample sizes also exists in the two-phase regression model–based tests for undocumented changepoints (Lund and Reeves 2002; Wang 2003), which we are also tackling and will report in a separate paper.

Also, both PMT and SNHT 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 as discussed in section 4). Autocorrelation and periodicity are typically inherent in climate data series. Periodicity and trend can be greatly diminished by using a good reference that has the same trend and periodicity as the base series, but the use of reference series cannot diminish autocorrelation in the time series being tested. Thus, it is of crucial importance for a test of undocumented changepoint to take into account the effect of autocorrelation and periodicity in the time series. Lund et al. (2007) recently proposed a new method for changepoint detection in periodic and autocorrelated time series, although the effect of unequal sample sizes is yet to be taken into account in this new method. The latter should be the subject for our next study.

Acknowledgments. The Climate Research Division of the Atmospheric Science and Technology Directorate of Environment Canada is acknowledged for supporting this research through grants and contributions (KM040–05–0016–IP). The three anonymous reviewers and the editor Dr. Arthur DeGaetano are acknowledged for their helpful review comments.

REFERENCES

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

Buishand, T. A., 1982: Some methods for testing the homogeneity of rainfall records. *J. Hydrol.,* **58,** 11–27.

Caussinus, H., and O. Mestre, 2004: Detection and correction of artificial shifts in climate. *Appl. Stat.,* **53,** 405–425.

Chu, P. S., and X. Zhao, 2004: Bayesian changepoint analysis of tropical cyclone activity: The central North Pacific case. *J. Climate,* **17,** 4893–4902.

Csörgő, M., and L. Horváth, 1997: *Limit Theorems in Change-Point Analysis.* John Wiley and Sons, 414 pp.

Davis, R. A., T. C. M. Lee, and G. A. Rodriguez-Yam, 2006: Structural breaks estimation for non-stationary time series models. *J. Amer. Stat. Assoc.,* **101,** 223–239.

DeGaetano, A. T., 2006: Attributes of several methods for detecting discontinuities in mean temperature series. *J. Climate,* **19,** 838–853.

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

Gardner, P. L., 1975: Scales and statistics. *Rev. Educ. Res.,* **45,** 43–57.

Hanesiak, J. M., and X. L. Wang, 2005: Adverse weather trends in the Canadian Arctic. *J. Climate,* **18,** 3140–3156.

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

——, X. L. Wang, Q. Lu, J. Reeves, C. Gallagher, and Y. Feng, 2007: Changepoint detection in periodic and autocorrelated time series. *J. Climate,* in press.

Menne, M. J., and C. N. Williams Jr., 2005: Detection of undocumented changepoints using multiple test statistics and composite reference series. *J. Climate,* **18,** 4271–4286.

Perreault, L., J. Bernier, and E. Parent, 2000: Bayesian change-point analysis in hydrometeorological time series. Part 1. The normal model revisited. *J. Hydrol.,* **235,** 221–241.

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

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

Solow, A., 1987: Testing for climatic change: An application of the two-phase regression model. *J. Climate Appl. Meteor.,* **26,** 1401–1405.

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

von Storch, H., and F. W. Zwiers, 1999: *Statistical Analysis in Climate Research.* Cambridge University Press, 484 pp.

Wan, H., X. L. Wang, and V. R. Swail, 2007: A quality assurance system for Canadian hourly pressure data. *J. Appl. Meteor. Climatol.,* in press.

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

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

——, 2007: A recursive testing algorithm for detecting and adjusting for multiple artificial changepoints in a time series. *Report of the Fifth Seminar for Homogenization and Quality Control in Climatological Databases,* Budapest, Hungary, World Climate Data and Monitoring Programme, WMO, in press.

——, and Y. Feng, cited 2004: RHTest user manual. [Available online at http://cccma.seos.uvic.ca/ETCCDMI/RHTest/RHTestUserManual.doc.]