Changepoint Detection in Climate Time Series with Long-Term Trends

COLIN GALLAGHER AND ROBERT LUND

Department of Mathematical Sciences, Clemson University, Clemson, South Carolina

MICHAEL ROBBINS

Department of Statistics, The University of Missouri, Columbia, Missouri

ABSTRACT

Climate time series often have artificial shifts induced by instrumentation changes, station relocations, observer changes, etc. Climate time series also often exhibit long-term trends. Much of the recent literature has focused on identifying the structural breakpoint time(s) of climate time series—the so-called changepoint problem. Unfortunately, application of rudimentary mean-shift changepoint tests to scenarios with trends often leads to the erroneous conclusion that a mean shift occurred near the series' center. This paper examines this problem in detail, constructing some simple homogeneity tests for series with trends. The asymptotic distribution of the proposed statistic is derived; en route, an attempt is made to unify the asymptotic properties of the changepoint methods used in today's climate literature. The tests presented here are linked to the ubiquitous t test. Application is made to two temperature records: 1) the continental United States record and 2) a local record from Jacksonville, Illinois.

1. Introduction

Changepoint features can drastically alter inferences made from a climatic time series. For example, Fig. 1 shows how trend estimates change when changepoint information is incorporated or neglected. This example considers annually averaged temperatures recorded at New Bedford, Massachusetts, from 1812 to 1999. The figure reports two statistical regression fits: 1) a line, which has a positive slope of 1.366°C century^{-1}, and 2) a "piecewise line" that allows for four mean shifts at times of known gauge changes and station relocations (these times are taken from the station's metadata and occur in 1888, 1906, 1951, and 1985). The slope in each segment of the piecewise line is constrained to be the same and is estimated at −0.234°C century^{-1}. Obviously, trend inferences are affected by changepoints.

The above data are typical of climate time series in that they have a time trend and multiple mean-shift changepoints that act to influence overall trend inferences.

Corresponding author address: Robert Lund, Department of Mathematical Sciences, Clemson University, Clemson, SC 29634-0975.
E-mail: lund@clemson.edu

The four changepoints here are viewed as artificial as they are induced by changes in data collection (gauge and location changes). Methods that adjust data for artificial shifts of this type are called homogenization techniques. The homogenization changepoint problem is well known in the climate literature; numerous authors have presented changepoint tests for the case of a single mean shift when the series has no trends. A partial list of references for this task includes Page (1955), Hawkins (1977), Potter (1981), Alexandersson (1986), Easterling and Peterson (1995), Vincent (1998), Ducré-Robitaille et al. (2003), and Beaulieu et al. (2012). Some of this literature is reviewed in Reeves et al. (2007).

Homogenizing series with long-term trends is harder. Some methods use information from one or more nearby data collection stations. A station used in this manner is commonly referred to as a reference station. The most common way of using a temperature series reference differences the record with corresponding values from the reference. In many cases, the resulting time series still has a trend. In addition to the data homogenization problem, some climatologists have considered detecting abrupt naturally caused changes in a climate time series. References include Tomé and Miranda (2004), Rodionov (2004), Seidel and Lanzante (2004), and
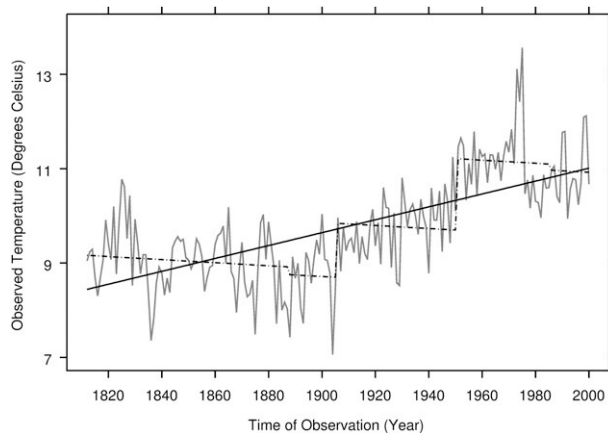
FIG. 1. New Bedford annual temperatures from 1812 to 1999 and two fitted models that ignore (solid) and include (dashed) mean-shift changepoints.

Gallagher et al. (2012). The methods here cannot distinguish between natural and artificial changepoints—such attribution is best done with a metadata record (when available). Whether one is interested in detecting artificial or natural changepoints, it is often prudent to allow for the potential of a trend.

Unfortunately, mean-shift tests may not work well when the series has a trend. If in truth a series has a trend but no mean shift, then a changepoint test designed for zero-trend data often tries to compensate for the mean misspecification by erroneously flagging a changepoint near the record's midpoint. If the series length is subsegmented about found changepoint times and the subsegments are analyzed, additional changepoints are often erroneously declared. It is therefore desirable to develop homogenization methods that are applicable to data with trends.

At most one changepoint (AMOC) tests that allow for trends were considered in Lund and Reeves (2002), Wang (2003, 2008), Wang et al. (2007), and Beaulieu et al. (2012). None of these references identifies an asymptotic statistical distribution for the crucial changepoint existence test. In fact, the current climate literature seems fractured and is often based on case-by-case simulation, despite the fact (which will become apparent here) that tractable asymptotic distributions can be derived.

Of the above papers, Wang et al. (2007) is the most relevant to the current exposition. These authors use empirical methods to improve testing procedures for changepoints in the presence of trends by imposing a "penalization term" that increases as the changepoint time moves away from the series' midpoint. Their penalty for a changepoint occurring at time c, denoted by $P(c)$, is derived through trial and error simulations with normal (Gaussian) distributions. This results in an ad

hoc improvement for the specific models and sample sizes employed, but no general justification is given. The mathematical derivations in this paper explain the necessity for, and provide a specific form for, the penalty term $P(c)$, which is appropriate for all sample sizes. Justification under minimal assumptions is given—we require independent model errors, but no distributional assumptions (such as normality).

This paper seeks to provide a more unified and mathematically justified procedure for changepoint tests for climate time series with possible trends. The good news is that justifiable test statistics are explicit and can be calculated from regression t tests, which are provided by standard statistical software. Furthermore, hypothesis testing decision rules can be based on an asymptotic theory that holds in great generality. Correct quantification of the asymptotic distribution of the AMOC test statistic is needed, which is provided in this paper. For purposes of exposition, attention is restricted to models whose trend slope is constrained to be the same before and after the changepoint time. We provide all technical details so that an interested reader can modify our methods to more complicated models such as those considered in Beaulieu et al. (2012) (that considers a variety of regression structures, moving beyond linear trend paradigms). While several of our asymptotic distributions have appeared in the technical statistics literature, they seem unknown (unappreciated) in climate settings. For example, the asymptotic properties of the classical standard normal homogeneity test (SNHT) of Alexandersson (1986) are presented here for the first time in a climate venue. Also, even in the statistical literature, the connection between AMOC changepoint tests and standard regression t tests has not been made clear.

The rest of this paper proceeds as follows. The next section reviews the simple case of a mean shift only (no trends) and states asymptotic distributions. Section 3 then presents analogous results for regression structures involving a linear trend. Section 4 presents a simulation study that supports the asymptotic theory presented and illuminates some of the properties of the methods. Section 5 shows how the methods can be used to make inferences on two temperature series. Concluding comments are made in section 6. Comparisons of the different tests are made and some words of caution are presented. Technical derivations are collected into an appendix.

2. A mean-shift CUSUM review

First, consider the most basic changepoint problem: detecting a change in the mean of a series at an unknown

time. A regression model describing this scenario for data (X_t) over $t = 1, \ldots, n$ is

$$X_t = \mu + \delta_t + \epsilon_t. \tag{2.1}$$

Here, μ is the mean of the series before the unknown changepoint time c and (ϵ_t) is zero-mean independent random error with unknown variance σ^2. The mean shift occurring at time $c + 1$ has magnitude Δ. Hence, the mean-shift factor (δ_t) obeys

$$\delta_t = \begin{cases} 0, & 1 \le t \le c, \\ \Delta, & c < t \le n. \end{cases}$$

Perhaps the first statistic used to detect mean shifts was the cumulative sum (CUSUM) statistic studied in Page (1955):

$$\text{CUSUM}(c) = \frac{1}{\sqrt{n}} \left(\sum_{t=1}^{c} X_t - \frac{c}{n} \sum_{t=1}^{n} X_t \right). \tag{2.2}$$

If in truth a mean shift occurred in the data at time c (specifically, sometime between times c and $c + 1$, which we call a shift at time c) with magnitude Δ, then taking expectations in (2.2) gives $E[\text{CUSUM}(c)] = n^{-3/2} c(n - c)\Delta$. It follows that if $|\text{CUSUM}(c)|$ is large, then there is statistical evidence that Δ is nonzero and that a changepoint occurred at time c. CUSUM changepoint statistics estimate the changepoint time as the value of c that maximizes $|\text{CUSUM}(c)|$. That is, the changepoint statistic used is

$$C = \frac{\max_{1 \le c \le n} |\text{CUSUM}(c)|}{\hat{\sigma}},$$

where the denominator has been normalized by an estimate of $\hat{\sigma}$ to scale up to a quantifiable asymptotic distribution. The estimate of σ^2 used is typically computed under the null hypothesis of no changepoints and is $\hat{\sigma}^2 = n^{-1} \sum_{t=1}^{n} (X_t - \overline{X}_{1:n})^2$, where $\overline{X}_{1:n} = n^{-1} \sum_{t=1}^{n} X_t$.

Our first result is taken from MacNeill (1974) but does not appear to be appreciated in climate settings. It says that under the null hypothesis of no changepoints, C distributionally converges (as $n \to \infty$) to the supremum (largest value) of a Gaussian process known as the Brownian bridge.

Result 1: As $n \to \infty$,

$$C \xrightarrow{D} \sup_{z \in [0,1]} |B(z)|. \tag{2.3}$$

Here, $B(z) = W(z) - zW(1)$ for $0 \le z \le 1$ and $[W(z)]_{z=0}^{z=1}$ is a standard Brownian motion process. This means that $W(z)$ has a normal distribution with zero mean and

TABLE 1. Critical values of the CUSUM test statistic C. Values are taken from the known expression for the distribution of $\sup_{0<t<1} |B(t)|$ in Resnick (1992).

Quantile	Critical value
90.0	1.224
95.0	1.358
97.5	1.480
99.0	1.628
99.9	1.949

variance z for each $z \in [0, 1]$, the increments of $[W(z)]$ are stationary and independent, and $W(0) = 0$. Brownian motion, Brownian bridges, and other Gaussian processes arise as asymptotic distributions in many limit problems. The interested reader can consult Ross (1996) and problems 6.14 and 6.18 of Resnick (1992) for more on their probabilistic properties. We do not prove this convergence here as it is technical; however, an analogous result for the more complicated case with a linear trend will be derived later.

The immediate implication is that critical values of C can be calculated. Specifically, a known expression [see Resnick (1992) and Robbins et al. (2011a), among others] provides exact tail probabilities of the distribution of $\sup_{z \in [0,1]} |B(z)|$. Here, the resulting critical values are provided in Table 1 for various levels of statistical confidence. Use of these percentiles is generally conservative for a finite n. This will be shown in section 4. Elaborating, should one want 95% confidence and use the 95th quantile in Table 1, then one is slightly more than 95% confident of the declaration of a found changepoint.

A perhaps more intuitive approach for changepoint detection examines the sample means before and after the changepoint. For a changepoint time c, set $\overline{X}_{1:c} = c^{-1} \sum_{t=1}^{c} X_t$ and $\overline{X}_{c+1:n} = (n - c)^{-1} \sum_{t=c+1}^{n} X_t$. If these two sample means greatly differ, a changepoint is statistically suggested at time c. Scaling by a denominator that accounts for the possibly different segment lengths leads to consideration of a statistic $Z(c)$ of the form

$$Z(c) = \frac{\overline{X}_{1:c} - \overline{X}_{c+1:n}}{\text{Var}(\overline{X}_{1:c} - \overline{X}_{c+1:n})^{1/2}}. \tag{2.4}$$

An easy calculation provides the variance margin $\text{Var}(\overline{X}_{1:c} - \overline{X}_{c+1:n}) = [c(n - c)]^{-1} n\sigma^2$. Using this in (2.4) and reworking the numerator into a CUSUM form provides

$$Z(c) = n^{-1/2} \frac{\text{CUSUM}(c)}{\sigma \sqrt{\frac{c}{n} \left(1 - \frac{c}{n} \right)}}. \tag{2.5}$$

Since c is not known a priori and n is constant, our changepoint statistic is taken as a maximum of the squared values multiplied by n:

$$C^* = \max_{1 \le c \le n} \frac{\text{CUSUM}^2(c)}{\hat{\sigma}^2 \frac{c}{n}\left(1 - \frac{c}{n}\right)}. \quad (2.6)$$

The statistic C^* is also a Gaussian likelihood ratio statistic. This means that if one assumes a Gaussian distribution for (ϵ_t) and performs a statistical likelihood analysis, then C^* arises. In general, likelihood ratio statistics can be difficult to compute, but are very powerful (i.e., they detect changepoints with a relatively high probability).

The asymptotic properties of C^* are harder to quantify than those of C. In fact, it can be shown that C^* diverges to infinity as $n \to \infty$. Hence, C^* needs to be scaled in some way. A common scaling practice is to "crop boundaries." One should not expect to accurately detect a changepoint that occurs near the data boundaries as there are too few observations before or after the changepoint to make reliable conclusions. Cropping the first and last $h \times 100\%$ percent of the observations gives a test statistic C_h^* defined by

$$C_h^* = \max_{h \le (c/n) \le 1-h} \frac{\text{CUSUM}^2(c)}{\hat{\sigma}^2 \frac{c}{n}\left(1 - \frac{c}{n}\right)}.$$

The asymptotic distribution of C_h^* is known (Csörgő and Horváth 1997) and again involves the Brownian bridge.

Result 2: As $n \to \infty$,

$$C_h^* \overset{D}{\to} \sup_{h \le z \le 1-h} \frac{B^2(z)}{z(1 - z)}. \quad (2.7)$$

Tail probabilities of the above limit distribution can be approximated using an expression that is provided by Csörgő and Horváth (1997) and Robbins et al. (2011a); simulation is required to obtain more precise values. Thereby, Table 2 lists simulated critical values of C_h^* for various values of h and statistical confidence levels. Up to boundary cropping issues, this result provides the asymptotic distribution of the classical standard normal homogeneity test of Alexandersson (1986). An alternative to using the percentiles in Table 2 would be to simulate critical values for every sample size n.

The similarities and differences between the asymptotic distributions in (2.3) and (2.7) are worth discussing. First, C_h^*'s limit distribution involves the supremum of the square of $B(z)$ rather than $|B(z)|$. This is of little concern: taking the largest absolute value or the largest

TABLE 2. Critical values of the likelihood ratio statistic C_h^*. Values are found by simulating one million samples of a Brownian bridge process.

Quantile	h			
	0.010	0.025	0.050	0.100
90.0	9.209	8.752	8.312	7.728
95.0	10.788	10.338	9.885	9.304
97.5	12.331	11.904	11.409	10.827
99.0	14.364	13.925	13.421	12.827
99.9	19.278	18.890	18.377	17.741

squared value produces the same changepoint time estimator. The C_h^* statistic also has a $z(1 - z)$ factor in the denominator that the C statistic lacks (one views z as c/n). This factor serves to place greater weight on c values near unity or n. As a result, C_h^* should have higher power than C when a changepoint occurs near the record's endpoints. Wang et al. (2007) attempt to reduce this emphasis by imposing a penalization term that increases as the changepoint time moves away from the series midpoint. Specifically, they propose maximizing $|P(c)Z(c)|$, where $P(c)$ is a penalty term derived through trial and error simulation. In the end, their form is unwieldy and statistically inconsistent (does not necessarily get the right answer as $n \to \infty$), but is markedly similar to the $\sqrt{(c/n)(1 - c/n)}$ structure in the denominator in the right-hand side of (2.5).

For practical recommendations, one should prefer C_h^* if only one test can be considered. It turns out, however, that should the changepoint time lie near the center of the data record (around $n/2$), then the CUSUM statistic C is slightly more powerful at detecting the changepoint than the C_h^* statistic [Robbins et al. (2011a) discuss this issue in detail]. Should the changepoint occur closer to the data boundaries (e.g., around 15% or 85% into the record), then the C_h^* statistic is vastly more powerful. In short, it makes sense to compute both statistics but to rely on the likelihood ratio C_h^* statistic should conclusions conflict. As to what value of h one should use, this is not usually important. Authors have truncated 1% and 5% of the data record on both sides with meaningful conclusions; we have yet to see a case where this truncation level greatly matters but do not doubt that such cases exist; especially should a changepoint lie very close to the data boundaries. Finally, one does not need Gaussian data for good performance. Indeed, many Gaussian-based statistics perform well (say consistently as $n \to \infty$) in non-Gaussian settings. In the end, one only requires zero-mean independent errors (ϵ_t) for all of our results. Unfortunately, when the errors are not zero-mean, the performance of the changepoint tests above can be bad (for the reasons

discussed in the introduction). This brings us to our next section.

3. Changepoint statistics in data with trends

Now suppose that a linear trend is allowed; that is, (2.1) is modified to

$$X_t = \mu + \beta t + \delta_t + \epsilon_t, \tag{3.1}$$

where β is a linear trend parameter. In order for the ensuing theory to hold as stated, one should restrict t to the discrete set $1, \ldots, n$ (and not let t denote the raw year, for example). Under a null hypothesis of no changepoints, the ordinary least squares estimators of the mean and trend parameters are the values of μ and β that minimize the sum of squares $\sum_{t=1}^{n}[X_t - (\mu + \beta t)]^2$. A calculus minimization shows these estimators to be

$$\hat{\beta} = \frac{12 \sum_{t=1}^{n} X_t(t - \bar{t})}{n(n+1)(n-1)}, \quad \hat{\mu} = \overline{X}_{1:n} - \hat{\beta}\bar{t}, \tag{3.2}$$

where $\bar{t} = (n + 1)/2$ is the average observation time.

When data are not independent and/or identically distributed, a common approach to changepoint detection is to apply the CUSUM methods of the previous section to a residuals sequence, $(\hat{\epsilon}_t)_{t=1}^{n}$; here, the residuals under the null hypothesis are calculated as $\hat{\epsilon}_t = X_t - (\hat{\mu} + \hat{\beta}t)$ for $t = 1, 2, \ldots, n$. For example, Gallagher et al. (2012) develop a residuals-based statistic for detection of a change in daily precipitation series that incorporates periodicity and that was illustrated to observe the same convergence as the C statistic seen in the previous section. Unfortunately, when a linear trend is involved, the results of the last section will no longer apply. This issue, in fact, confused many early statistics authors (MacNeill 1978; Sen 1982; Kim and Siegmund 1989). Nonetheless (and similar to what was seen in the last section), examining the CUSUM statistic of $(\hat{\epsilon}_t)$ will illuminate better methods.

a. CUSUM-type statistics

To quantify the asymptotic properties of the CUSUM of the residual sequence $(\hat{\epsilon}_t)_{t=1}^{n}$, a separate analysis is needed. For this, set

$$D = \frac{\max\limits_{1 \le c \le n} |D(c)|}{\hat{\sigma}}, \tag{3.3}$$

where

$$D(c) = \frac{1}{\sqrt{n}\hat{\sigma}}\left(\sum_{t=1}^{c} \hat{\epsilon}_t - \frac{c}{n}\sum_{t=1}^{n} \hat{\epsilon}_t\right) = \frac{1}{\hat{\sigma}\sqrt{n}}\sum_{t=1}^{c} \hat{\epsilon}_t. \tag{3.4}$$

TABLE 3. Critical values of the CUSUM test statistic D. Values are found by simulating one million samples of $[\tilde{B}(t)]$.

Quantile	Critical value
90.0	0.836
95.0	0.906
97.5	0.970
99.0	1.047
99.9	1.222

Here, $\hat{\sigma}^2 = \sum_{t=1}^{n} \hat{\epsilon}_t^2/(n-2)$ is the null hypothesis estimate of σ^2 and the second equality in (3.4) follows from the fact that ordinary least squares residuals sum to zero: $\sum_{t=1}^{n} \hat{\epsilon}_t = 0$. What can be shown (following MacNeill 1978) is that the asymptotic distribution of the trend adjusted CUSUM statistic in (3.3) has the following limit law. Our argument for this is presented in the appendix.

Result 3: In the linear trend case, as $n \to \infty$,

$$D \xrightarrow{D} \sup_{z \in [0,1]} |\tilde{B}(z)|. \tag{3.5}$$

Here, $[\tilde{B}(z)]_{z=0}^{1}$ is a process related to the Brownian bridge $[B(z)]_{z=0}^{1}$ via

$$\tilde{B}(z) = B(z) - 6z(1-z)\int_0^1 B(x)\,dx. \tag{3.6}$$

From Gaussian process theory [see Ross (2010) for more on Gaussian processes], it is known that $[\tilde{B}(z)]$ is another Gaussian process with $\tilde{B}(0) = 0$ and variance

$$\mathrm{Var}[\tilde{B}(t)] = t(1-t)[1 - 3t(1-t)], \quad 0 \le t \le 1. \tag{3.7}$$

The variance in (3.7) is the key nuance between no-trend and trend cases—the analogous variance for the no-trend model is $\mathrm{Var}[B(t)] = t(1-t)$.

Table 3 provides asymptotic percentiles for D. The percentiles are again conservative when applied to scenarios with a finite n.

b. Relationship to t statistics

Here, we connect the $D(c)$ statistic, regression t tests, and likelihood ratio statistics for series with linear trends. Suppose that a changepoint is known to occur at time c. An involved calculation shows that the estimates that minimize the sum of squares $\sum_{t=1}^{n}\{X_t - [\mu + \beta t + \Delta 1_{(t>c)}]\}^2$ are

$$\hat{\beta}(c) = \frac{\sum_{t=1}^{c} X_t\left[t - \frac{c(c+1)}{2}\right] + \sum_{t=c+1}^{n} X_t\left[t - \frac{n(n+1) - c(c+1)}{2(n-c)}\right]}{\frac{c(c+1)(c-1)}{12} + \frac{(n-c)(n-c+1)(n-c-1)}{12}},$$

$$\hat{\mu}(c) = \frac{1}{c} \sum_{t=1}^{c} [X_t - \hat{\beta}(c)t], \quad \text{and}$$

$$\hat{\Delta}(c) = \frac{1}{n-c} \sum_{t=c+1}^{n} [X_t - \hat{\beta}(c)t] - \hat{\mu}(c). \tag{3.8}$$

Here, the notation $\hat{\beta}(c)$, for example, signifies that the trend is estimated assuming a changepoint at time c. A changepoint is suggested at time c when $\hat{\Delta}(c)$ is significantly nonzero. A t statistic for such a test is simply

$$T(c) = \frac{\hat{\Delta}(c)}{\widehat{\text{Var}}[\hat{\Delta}(c)]^{1/2}}. \tag{3.9}$$

To explicitly identify the numerator and denominator in (3.9), the appendix derives the two identities

$$\hat{\Delta}(c) = \frac{-\sum_{t=1}^{c} \hat{\epsilon}_t}{c\left(1 - \frac{c}{n}\right)\left[1 - 3\frac{c(n-c)}{n^2 - 1}\right]} \tag{3.10}$$

and

$$\widehat{\text{Var}}[\hat{\Delta}(c)] = \frac{\hat{\sigma}^2}{n\frac{c}{n}\left(1 - \frac{c}{n}\right)\left[1 - 3\frac{c(n-c)}{n^2 - 1}\right]}. \tag{3.11}$$

Using (3.10) and (3.11) in (3.9) gives the following. Let $r = c/n$. When n is large, division by n^2 is approximately division by $n^2 - 1$, $c(n-c)/(n^2 - 1) \approx (c/n)(1 - c/n)$, and under the null hypothesis, the approximation

$$T(c) \approx \frac{-n^{-1/2}\sum_{t=1}^{c} \hat{\epsilon}_t}{\hat{\sigma}\sqrt{r(1-r)[1-3r(1-r)]}} \tag{3.12}$$

emerges. To link $T(c)$ and $D(c)$ statistics, we show in the appendix that

$$D(c) = \frac{1}{\hat{\sigma}\sqrt{n}}\left(\sum_{t=1}^{c} \hat{\epsilon}_t\right)$$
$$= \frac{1}{\hat{\sigma}\sqrt{n}}\left[\sum_{t=1}^{c} \epsilon_t - \frac{c}{n}\sum_{t=1}^{n} \epsilon_t - (\hat{\beta} - \beta)\frac{c}{2}(n-c)\right], \tag{3.13}$$

where $\hat{\beta}$ is the slope estimate under the null hypothesis. Substituting (3.13) for the numerator of (3.12) shows that, for large n,

$$|T(c)| \approx \frac{|D(c)|}{\sqrt{r(1-r)[1-3r(1-r)]}}. \tag{3.14}$$

Equation (3.14) has important ramifications. This is because any standard statistical software will calculate $T(c)$ for each admissible changepoint time c. A t-based test for an unknown changepoint then simply looks at the maximum of $|T(c)|$ [or $T(c)^2$] over all possible c. Such a statistic is also the Gaussian likelihood ratio statistic.

As in the simple mean-shift case, one cannot extract a meaningful limit law without first cropping. Specifically, it can be shown that $\max_{1 \leq c \leq n} |T(c)| \to \infty$ as $n \to \infty$ [also, $\max_{1 \leq c \leq n} T(c)^2 \to \infty$ as $n \to \infty$]. This is a consequence of the fact that $r(1-r)$ converges to zero as $r \to 1$ and/or as $r \to 0$ and a deeper analysis of the behavior of $\tilde{B}(z)$ near $z = 0$ and $z = 1$.

To quantify a limit law for a cropped statistic, let

$$D_h^* = \max_{h \leq (c/n) \leq 1-h} \frac{D(c)^2}{\hat{\sigma}^2 r(1-r)[1-3r(1-r)]} \quad \text{and} \tag{3.15}$$

$$\tilde{D}_h^* = \max_{h \leq (c/n) \leq 1-h} T(c)^2. \tag{3.16}$$

Our last result is the following.
Result 4: In the linear trend case, as $n \to \infty$,

$$\tilde{D}_h^* \xrightarrow{D} \sup_{h \leq z \leq 1-h} \frac{\tilde{B}^2(z)}{z(1-z)[1-3z(1-z)]}. \tag{3.17}$$

The limit law for $\tilde{D}_h^*$ is the same as that for D_h^*.

Asymptotic percentiles for $\tilde{D}_h^*$ are given in Table 4. As the next section shows, percentiles are again conservative when applied to scenarios with a finite n (see also Robbins et al. 2011a).

Notice that D_h^* requires less computation than $\tilde{D}_h^*$ as one does not need to compute parameter estimators for each and every c. However, both have the same limiting distribution and hence the same asymptotic detection power. This said, $\tilde{D}_h^*$ should be slightly more powerful

TABLE 4. Critical values of the likelihood ratio statistic $\tilde{D}_h^*$. Values were found by simulating one million samples of $[\tilde{B}(t)]$.

Quantile	h			
	0.010	0.025	0.050	0.100
90.0	10.341	10.061	9.790	9.459
95.0	11.956	11.684	11.415	11.077
97.5	13.532	13.271	12.989	12.662
99.0	15.540	15.301	15.011	14.692
99.9	20.600	20.218	20.114	19.758

for finite n as it incorporates alternative hypothesis estimators.

It is easy to design a changepoint test statistic that has the advantages of $\tilde{D}_h^*$ (when compared to D_h^*) but also has the same asymptotic distribution as D. To do this, set

$$\tilde{D} = \max_{1 \leq c \leq n} \sqrt{r(1-r)[1-3r(1-r)]}|T(c)|, \quad (3.18)$$

where $r = c/n$. Then the limit law is

$$\tilde{D} \xrightarrow{D} \sup_{z \in [0,1]} |\tilde{B}(z)|,$$

where $\tilde{B}(z)$ is as in (3.6).

Extending the rationale of Wang et al. (2007) to models with a linear trend, Wang (2008) empirically adds a penalty term to the above t-type tests. Elaborating, Wang et al. (2007) suggest maximizing $P^*(c)T(c)^2$, where $P^*(c)$ is a penalty that is "M-shaped" when graphed as a function of c. The author develops piecewise expressions for $P^*(c)$ through simulation. The resulting penalty term is unwieldy, but markedly similar to $r(1-r)[1-3r(1-r)]$ with $r = c/n$, the scaling term in (3.18) here (which is M-shaped when graphed). In effect, Wang (2008) is trying to introduce a test statistic with the behavior of $\tilde{D}$. Unfortunately, Wang's (2008) statistic cannot be easily calculated or quantified asymptotically, whereas the statistics introduced here are easily calculated and have quantifiable limiting distributions.

For practical recommendations, we suggest using $\tilde{D}_h^*$ in favor of $\tilde{D}$ should conclusions conflict for reasons illustrated in the next section.

It remains to derive results 3 and 4, which is done in the appendix. This task is important since the asymptotic distribution depends on the presence/absence of a trend. One would have to perform yet another derivation for regression structures employing sinusoidal, quadratic, or exponential terms. Sinusoidal terms, for example, could arise in the analysis of daily or monthly series. This said, the derivations are similar in spirit and we offer the linear case as perhaps useful guidance.
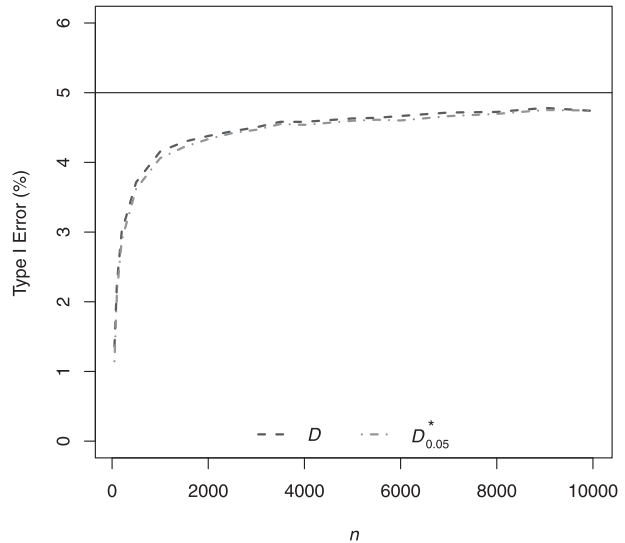

FIG. 2. Simulated type I error probabilities (false detection) of the D and $D_{0.05}^*$ tests as a function of the sample size n. Simulations are based on one million replications for each n.

4. A simulation study

This section uses simulation to assess the efficacy of the methods in section 3. All series are simulated from (3.1) with $\mu = 1$, $\beta = 0.5$, and $\sigma^2 = 1$. The performance of all methods seemed invariant of the regression parameters chosen; hence, we report results for the above choices only. The significance level $\alpha = 0.05$ is used throughout.

To begin, the no changepoint null hypothesis performance of the methods is studied. For each series length n, one million series were generated from (3.1) with $\delta_t = 0$ for all t and Gaussian errors $\{\epsilon_t\}$. Because of the large number of replications (one million), little error is attributable to simulation. For each generated series, the statistics D and $D_{0.05}^*$, as specified in (3.3) and (3.16), are computed (tests using $\tilde{D}$ and $\tilde{D}_{0.05}^*$ are too computationally cumbersome to be evaluated here, but their asymptotic performance is the same as the corresponding statistics considered). For both tests, the empirical proportion of the one million series that falsely declare a changepoint is graphed in Fig. 2. The individual series are changepoint assessed using the critical values in Tables 3 and 4. Figure 2 reveals two features: 1) as the sample size n becomes larger, the probability of erroneously declaring that a changepoint exists converges to 0.05 (as the asymptotic theory demands), and 2) the tests are conservative in that p values will not overestimate the probability of changepoint presence. This last property is convenient for those wishing to avoid overhomogenizing a record for too many changepoints.
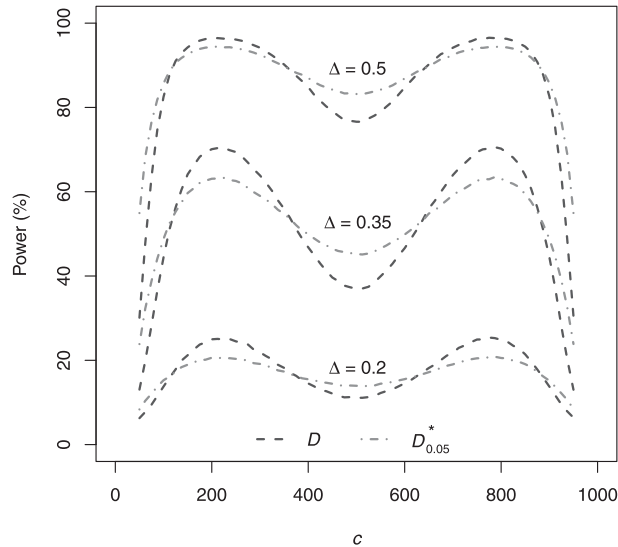
FIG. 3. Simulated power of detection for the D and $D^*_{0.05}$ tests for three values of Δ as a function of the changepoint time c with $n = 1000$. Simulations are based on 100 000 replications for each combination of c and Δ.

The detection power of the methods in the presence of a changepoint (i.e., $\delta_t \neq 0$ for some t) is also of interest. While it would be instructive to compare to the climate-realistic benchmarks developed in Venema et al. (2012), these results were developed for multiple changepoint features. It is important to emphasize that our methods are single changepoint tests. While one can turn a single changepoint test into a multiple changepoint procedure via subsegmenting, Li and Lund (2012) show that some undesirable issues arise in so doing. In particular, mean shifts that occur closely in time become difficult to distinguish, or mean-shift magnitudes Δ that alternate in sign can fool segmentation methods.

Here, the series length $n = 1000$ is fixed for convenience (this large value is taken so that the asymptotics hold, but the scenario is not overly different for shorter series lengths). Several values of the changepoint time c and three nonzero values of the mean-shift magnitude Δ are studied: 0.2, 0.35, and 0.50. These levels are slightly less than the mean-shift magnitude variations simulated in the Venema et al. (2012) benchmarks. One hundred thousand replicated series are simulated for each c and Δ. The detection powers on the vertical axis in Fig. 3 show the proportion of series where the null hypothesis is correctly rejected, while the horizontal axis indicates the location of the changepoint time. The detection powers seem reasonable, with the tests performing relatively poorer when the changepoint time is located near either endpoint.

Figure 3 shows that the detection power of both tests increases as the mean-shift magnitude Δ increases. Both

tests show M-shaped powers as a function of the changepoint time c—the test using $D^*_{0.05}$ is more powerful than D when the changepoint occurs near the center or endpoints of the dataset, and the D test is more powerful elsewhere. This said, the power in D is never drastically greater than that for $D^*_{0.05}$—the greatest power discrepancy in favor of D is approximately 7%, which is seen when $\Delta = 0.35$ and $c = 220$ or 780. However, $D^*_{0.05}$ may have a markedly higher power than D when the changepoint occurs near the endpoints of the dataset. For instance, the discrepancy in power is approximately 25% in favor of $D^*_{0.05}$ when $\Delta = 0.5$ and $c = 50$ or 950. As previously mentioned, a similar phenomenon was noted by Robbins et al. (2011a) for the no-trend case. Hence, we recommend use of D^*_h in practice. Of course, it is not difficult to compute both D and D^*_h.

5. Examples

This section uses the results of the previous sections to study two temperature series. The first series contains annual temperatures for the contiguous United States (CONUS)—the so-called CONUS data discussed in Menne et al. (2010). This is a situation in which a reference series is not available; moreover, since the data are averaged over many stations, no good metadata exist. Our second series moves to a more local scenario: monthly high temperatures at Jacksonville, Illinois. Here, we construct a reference series to help make conclusions.

All tests calculate the following four statistics: 1) D in (3.5), 2) D^*_h in (3.15), 3) $\tilde{D}$ in (3.18), and 4) $\tilde{D}^*_h$ in (3.16). Tests 1 and 3 use the critical values from Table 3 and tests 2 and 4 use critical values from Table 4. All conclusions are made at the 95% significance level; $h = 0.05$ is used in any test requiring cropping. We remind the reader that the methodology presented in this article is designed for the AMOC alternative, and climate series often have multiple changepoints. Here, the possibility of multiple changepoints is handled through subsegmenting. Specifically, the whole series is tested via AMOC methods; if any changepoints are detected, the methods are subsequently applied to the resulting subsegments. This process is repeated until all segments test as homogeneous. While this procedure is simple and has demonstrated utility in climate settings (Robbins et al. 2011b), methods that handle multiple changepoint aspects could be devised with further analyses (see, e.g., Li and Lund 2012).

Our first application studies annual temperatures from the CONUS record from 1895 to 2011 ($n = 117$). The D^* and $\tilde{D}^*_h$ statistics flag a changepoint in 1997, whereas the D and $\tilde{D}$ statistics both flag changepoints in 1985. As our preferred methods require some truncation,

TABLE 5. Changepoint tests for the annual continental U.S. temperature data.

1895–2011			
	Stat.	$\hat{c}$ (year)	p value
D	0.994	91 (1985)	0.0189
$D^*_{0.05}$	13.35	103 (1997)	0.0217
$\tilde{D}$	1.047	91 (1985)	0.0102
$\tilde{D}^*_{0.05}$	14.97	103 (1997)	0.0104
1895–1997			
	Stat.	$\hat{c}$ (year)	p value
D	0.796	26 (1920)	0.1435
$D^*_{0.05}$	8.831	63 (1957)	0.1489
$\tilde{D}$	0.824	26 (1920)	0.1115
$\tilde{D}^*_{0.05}$	9.584	63 (1957)	0.1093
1998–2011			
	Stat.	$\hat{c}$ (year)	p value
D	0.528	10 (2007)	0.7683
$D^*_{0.05}$	3.519	10 (2007)	0.8489
$\tilde{D}$	0.604	10 (2007)	0.5534
$\tilde{D}^*_{0.05}$	4.617	10 (2007)	0.6612

FIG. 4. Changepoint test statistics for the annual CONUS temperature data.

we support the 1997 changepoint if forced to choose one conclusion. This changepoint shifts temperatures upward. Phrased another way, we conclude that recent temperatures are increasing at a rate that is faster than can be explained by a linear trend. All subsegments test as homogeneous with 95% confidence. The p values for all tests are reported in Table 5 (definitive conclusions are made at 95% confidence). Figure 4 plots the $T^2(c)$ values for this and all subsegments; the fit of the resulting changepoint model is plotted in Fig. 5. We comment that there are secondary peaks in the $T(c)^2$ statistics for the 1895–1997 subsegment in Fig. 4 circa 1920 and 1957. While these times are ''suspect,'' they are not significant at level 95% (the Table 5 p values report them as significant with approximately 85%–90% confidence). Note that one cannot take all the times that exceed the 95% quantile in Fig. 4 and declare them as changepoints.

We next examine monthly maximum temperatures from Jacksonville, Illinois, from January 1928 to December 2010 ($n = 996$). A reference series was constructed by averaging the 40 most correlated neighboring stations to Jacksonville. (The Jacksonville minus reference series is plotted in Fig. 7.) Our changepoint methods are applied to the difference of the two series. Differencing acts to reduce or remove the effect of many complicating data characteristics (e.g., autocorrelation, periodicity, trend, etc.), thereby allowing the analyst to isolate non-climatic changes (such as a gauge change). Nonetheless, a differenced series may still exhibit a linear trend, caused perhaps by external agents (e.g., urbanization) that influence the series under study. We have chosen
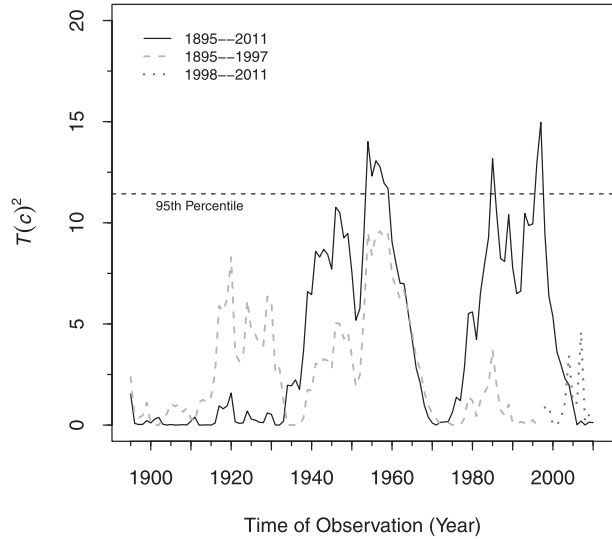
to make a reference by averaging 40 adjacent stations to illustrate the methods, but avoid some of the complications of pairwise analyses that arise should one consider all $\binom{40}{2} = 780$ differenced series. Menne and Williams (2005, 2009) admirably discuss pairwise comparison issues.

Table 6 displays the results from all changepoint tests. All methods flag a changepoint at January 1957. After subsegmentation, no further changepoints were detected at the 5% significance level. This declaration is again somewhat perilous as the $T(c)^2$ statistics for the subsegments almost exceed the 95th percentile circa the early 1960s and mid-1980s. The 1957 changepoint acts to lower temperatures relative to the reference.
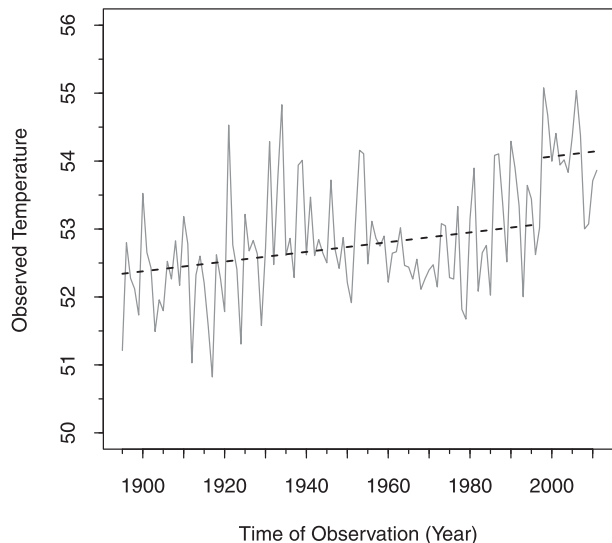

FIG. 5. Estimated model fit for the CONUS temperature data.

TABLE 6. Changepoint tests for the monthly Jacksonville, IL, temperatures.

		January 1928–December 2010	
	Stat.	$\hat{c}$ (month–year)	p value
D	1.75	349 (Jan 1957)	0.000
$D^*_{0.05}$	42.26	349 (Jan 1957)	0.000
$\tilde{D}$	1.78	349 (Jan 1957)	0.000
$\tilde{D}^*_{0.05}$	44.09	349 (Jan 1957)	0.000
		January 1928–January 1957	
	Stat.	$\hat{c}$ (month–year)	p value
D	0.846	251 (Nov 1948)	0.092
$D^*_{0.05}$	10.54	18 (Jun 1929)	0.073
$\tilde{D}$	0.856	251 (Nov 1948)	0.083
$\tilde{D}^*_{0.05}$	10.84	18 (Jun 1929)	0.064
		February 1957–December 2010	
	Stat.	$\hat{c}$ (month–year)	p value
D	0.822	62 (Mar 1962)	0.114
$D^*_{0.05}$	10.70	331 (Aug 1984)	0.069
$\tilde{D}$	0.829	62 (Mar 1962)	0.107
$\tilde{D}^*_{0.05}$	10.859	331 (Aug 1984)	0.064

FIG. 6. Jacksonville, IL, changepoint test statistics for the differenced data.

Notice that the common trend slope in the two segments is positive, meaning that Jacksonville maximums are warming more rapidly than its reference. Graphical support for these conclusions are provided in Fig. 6, which displays the fitted model (changepoint and linear trend) and Fig. 7, which plots the $T(c)^2$ values for the series and all subsegments.

In concluding these examples, we caution the reader that mean-shift models with a constant slope may not adequately describe all climate time series. The above conclusions, while supplying good rudimentary guidance, are not set in stone. Indeed, climatic forcings are unlikely to be linear in time; moreover, localized considerations, such as moving a station into or out of an urban heat island zone, may merit consideration of more complicated regression structures (perhaps varying slopes for the different segments).

6. Comments

In cases for which it is not clear whether the series has a linear trend, it is prudent to apply the methods that allow for a trend. Conclusions will then be conservative and changepoint conclusions will be more justifiable than if the trend were ignored. This said, if it is known concretely that a series does not have a trend, then one should not apply the methods for trends. In general, detection power is sacrificed when one examines models with more model parameters than needed.

As eluded to in the last section, another issue involves whether the trend slope should change at the mean-shift
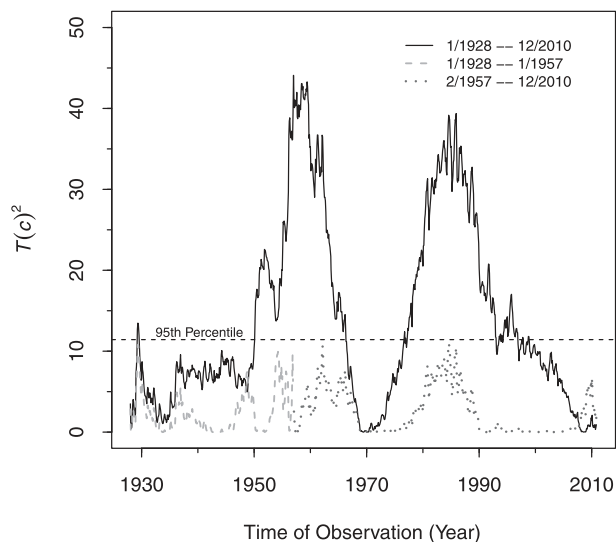
time. Two-segment models that allow for different slopes before and after the changepoint time are called two-phase regressions and are studied in Lund and Reeves (2002). Although the F_{max} statistic examined there has similarities to our statistics, quantification of the F_{max} limit distribution is too onerous to pursue here. In the case of multiple segments (more than two) with distinct trends, multiple changepoint techniques designed to handle such regression structures will likely be more reliable than segmentation approaches with AMOC tests. Our belief is that series mean misspecification is the primary unresolved issue in multiple changepoint problems. Li and Lund (2012) discuss multiple changepoint problems in more depth. In general, changepoints that occur close together in time or act to move the series in oscillating manners can fool subsegmenting methods.
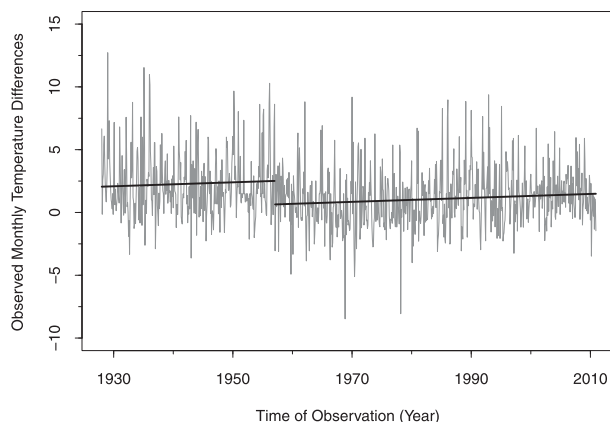

FIG. 7. Jacksonville, IL, differences with the fitted changepoint model.

Two other Medusas of changepoint methods merit mention. First, even moderate levels of error autocorrelation are common in climate series and will influence changepoint conclusions. The analyses here ignore autocorrelation. Naturally caused (Rodionov) mean shifts can be attributed to autocorrelation in part. Second, seasonality is often present in nonannualized series. For example, it is not clear that the Jacksonville minus reference series of the last section does not have any seasonal features. Both of these issues are discussed in detail in Lund et al. (2007).

APPENDIX

Derivation of Relevant Formulas in Section 3

Proof of result 3: We begin by justifying the convergence in result 3. Suppose that the null hypothesis model with $\delta_t = 0$ is in force. Substituting the right-hand side of (3.1) into (3.2) and simplifying yields

$$\hat{\beta} - \beta = \frac{12 \sum_{t=1}^{n} (t - \bar{t})\epsilon_t}{n^3 - n} \qquad (A1)$$

and

$$\hat{\mu} = \mu - (\hat{\beta} - \beta)\bar{t} + \bar{\epsilon}.$$

[Recall that $\bar{t} = (n + 1)/2$ is the average observation time.]

Using these formulas to calculate the residuals under the null hypothesis gives

$$\hat{\epsilon}_t = \epsilon_t - \bar{\epsilon} - (\hat{\beta} - \beta)(t - \bar{t}),$$

where $\bar{\epsilon} = n^{-1}\sum_{t=1}^{n}\epsilon_t$. Using this expression for the residuals, (A1), and $\sum_{t=1}^{c}(t - \bar{t}) = c(c - n)/2$ proves (3.13):

$$\sum_{t=1}^{c} \hat{\epsilon}_t = \left(\sum_{t=1}^{c} \epsilon_t - c\bar{\epsilon} \right) + \frac{12(c/2)(n - c)}{n^3 - n} \sum_{t=1}^{n} (t - \bar{t})\epsilon_t. \qquad (A2)$$

The advantage of the representation in (A2), which expresses the sum of the residuals as a sum of the (unobservable) independent variables $\{\epsilon_t\}$, is twofold. First, well-known results for weighted sums of independent variables can be used to justify the Gaussian process limit (see Billingsley 1999). Second, the independence of the summands can be used to calculate the covariance of the Gaussian process. Since the $[\epsilon_t]$ are independent, $\mathrm{Cov}(\epsilon_t, \epsilon_\ell) = 0$ when $t \neq \ell$. From this and linearity of covariances, when $k \geq c$, one has

$$\mathrm{Cov}\left(\sum_{t=1}^{c} \epsilon_t - c\bar{\epsilon}, \sum_{t=1}^{k} \epsilon_t - k\bar{\epsilon} \right) = \sigma^2(n - k)c/n,$$

$$\mathrm{Cov}\left[\sum_{t=1}^{c} \epsilon_t - c\bar{\epsilon}, k(n - k)(\hat{\beta} - \beta)/2 \right]$$

$$= \frac{-3\sigma^2 c(n - c)k(n - k)}{n^3 - n}$$

and

$$\mathrm{Cov}[c(n - c)(\hat{\beta} - \beta)/2, k(n - k)(\hat{\beta} - \beta)/2]$$

$$= \frac{\sigma^2 3c(n - c)k(n - k)}{n^3 - n}.$$

Using these in (A1) and properties of covariance gives our key equation:

$$n^{-1}\mathrm{Cov}\left(\sum_{t=1}^{c} \hat{\epsilon}_t, \sum_{t=1}^{k} \hat{\epsilon}_t \right) = \mathrm{Cov}\left(\sum_{t=1}^{c} \epsilon_t - c\bar{\epsilon}, \sum_{t=1}^{k} \epsilon_t - k\bar{\epsilon} \right) + 2\mathrm{Cov}\left[\sum_{t=1}^{c} \epsilon_t - c\bar{\epsilon}, k(n - k)(\hat{\beta} - \beta)/2 \right]$$

$$+ \mathrm{Cov}[c(n - c)(\hat{\beta} - \beta)/2, k(n - k)(\hat{\beta} - \beta)/2]$$

$$= \frac{\sigma^2(n - k)c}{n^2}[1 - 3k(n - c)/(n^2 - 1)] \approx \sigma^2(1 - s)r[1 - 3s(1 - r)],$$

where $s = k/n$ and $r = c/n$. This equation identifies the limiting covariance structure of the Gaussian process. In particular, when $s = r$

$$\mathrm{Var}\left(\sum_{t=1}^{c} \hat{\epsilon}_t \Big/ \sqrt{n} \right) \approx \sigma^2 r(1 - r)[1 - 3r(1 - r)].$$

The convergence claimed in result 3 now follows from process convergence results found in Billingsley (1999).

Proof of result 4: While (3.8) gives the ordinary least squares estimators of $\mu(c)$, $\beta(c)$, and $\Delta(c)$ for each changepoint time c, the computations simplify if we work with a linear models orthogonal decomposition on the

design matrix. Stapleton (2009) gives a comprehensive discussion of these techniques. For $t > c$, denote the predicted values under the alternative model as $\hat{x}_t = \hat{\mu} + \hat{\beta}(c)t + \hat{\Delta}(c)$. Let $\mathbf{v}_1$ be an $n \times 1$ vector whose entries are all ones, let the vector $\mathbf{v}_2$ have tth component $t - \bar{t}$, and let the vector $\mathbf{v}_3$ have tth component satisfying

$$v_3(t) = \begin{cases} \dfrac{c-n}{n^3-n}[n^2 - 1 + 6c(t - \bar{t})], & t \le c \\[2ex] 1 + \dfrac{c-n}{n^3-n}[n^2 - 1 + 6c(t - \bar{t})], & t > c \end{cases}. \quad (A3)$$

The vector of predicted values has representation

$$\hat{\mathbf{X}} = \overline{X}_{1:n}\mathbf{v}_1 + \hat{\beta}\mathbf{v}_2 + \hat{\Delta}(c)\mathbf{v}_3$$

and linear models theory (Stapleton 2009, section 3.7) provides

$$\hat{\Delta}(c) = \frac{\mathbf{v}_3'\mathbf{X}}{\|\mathbf{v}_3\|^2} \quad \text{and} \quad \mathrm{Var}[\hat{\Delta}(c)] = \frac{\sigma^2}{\|\mathbf{v}_3\|^2},$$

with $\mathbf{X} = (X_1, \ldots, X_n)'$. Here, the norm is the sum of squares $\|\mathbf{y}\|^2 = \sum_{t=1}^{n} y_t^2$ and $\mathbf{y} = (y_1, \ldots, y_n)'$.

A direct calculation with (A3) shows

$$\|\mathbf{v}_3\|^2/n = \frac{n-c}{n}\left(1 - \frac{n-c}{n-3c^2}\frac{n-c}{n^3-n}\right)$$
$$= \frac{c}{n}\left(1 - \frac{c}{n}\right)\left[1 - 3\frac{c(n-c)}{n^2-1}\right],$$

which establishes (3.11), and

$$\mathbf{v}_3'\mathbf{X} = c(n-c)(\beta - \hat{\beta})/2 - \sum_{t=1}^{c} \epsilon_t - c\bar{\epsilon} = -\sum_{t=1}^{c} \hat{\epsilon}_t,$$

where $\hat{\epsilon}_t$ is the residual calculated under the null hypothesis. This establishes (3.10).

The squared t statistic becomes

$$T(c)^2 = \frac{\hat{\Delta}(c)^2}{\widehat{\mathrm{Var}}[\hat{\Delta}(c)]} = \frac{(\mathbf{v}_3'\mathbf{X})^2/n}{\sigma^2\|\mathbf{v}_3\|^2/n},$$

where, as $n \to \infty$,

$$n^{-1}\|\mathbf{v}_3\|^2 \to r(1-r)[1 - 3r(1-r)],$$

with $r = c/n$. Under the null hypothesis of no change-points, as $n \to \infty$,

$$T(c)^2 = \frac{\left(\sum_{t=1}^{c} \hat{\epsilon}_t\right)^2 \big/ n}{\hat{\sigma}^2\|\mathbf{v}_3\|^2/n} \approx \frac{\left(\sum_{t=1}^{c} \hat{\epsilon}_t \big/ \sqrt{n}\right)^2}{\sigma^2 r(1-r)[1 - 3r(1-r)]}.$$

The above derivations yield (3.12) and the convergence in result 4 now follows from (3.5), (3.12), and the continuous mapping theorem of Billingsley (1999).

REFERENCES

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

Beaulieu, C., J. Chen, and J. L. Sarmiento, 2012: Change-point analysis as a tool to detect abrupt climate variations. *Philos. Trans. Roy. Soc.,* **370A,** 1228–1249.

Billingsley, P., 1999: *Convergence of Probability Measures.* 2nd ed. Wiley, 285 pp.

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

Ducré-Robitaille, J.-F., L. A. Vincent, and G. Boulet, 2003: Comparison of techniques for detection of discontinuities in temperature series. *Int. J. Climatol.,* **23,** 1087–1101.

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

Gallagher, C., R. Lund, and M. Robbins, 2012: Changepoint detection in daily precipitation data. *Environmetrics,* **23,** 407–419.

Hawkins, D. M., 1977: Testing a sequence of observations for a shift in location. *J. Amer. Stat. Assoc.,* **72,** 180–186.

Kim, H.-J., and D. Siegmund, 1989: The likelihood ratio test for a change-point in simple linear regression. *Biometrika,* **76,** 409–423.

Li, S., and R. B. Lund, 2012: Multiple changepoint detection via genetic algorithms. *J. Climate,* **25,** 674–686.

Lund, R. B., 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,* **20,** 5178–5190.

MacNeill, I., 1974: Tests for change of parameter at unknown times and distributions of some related functionals on Brownian motion. *Ann. Stat.,* **2,** 950–962.

——, 1978: Limit processes for sequences of partial sums of regression residuals. *Ann. Probab.,* **6,** 695–698.

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

——, and ——, 2009: Homogenization of temperature series via pairwise comparisons. *J. Climate,* **22,** 1700–1717.

——, ——, and M. A. Palecki, 2010: On the reliability of the U.S. surface temperature record. *J. Geophys. Res.,* **115,** D11108, doi:10.1029/2009JD013094.

Page, E. S., 1955: A test for a change in a parameter occurring at an unknown point. *Biometrika,* **42,** 523–527.

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

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

Resnick, S. I., 1992: *Adventures in Stochastic Processes.* Birkhäuser, 626 pp.

Robbins, M. W., C. M. Gallagher, R. B. Lund, and A. Aue, 2011a: Mean shift testing in correlated data. *J. Time Ser. Anal.,* **32,** 498–511.

——, R. B. Lund, C. M. Gallagher, and Q. Lu, 2011b: Changepoints in the North Atlantic tropical cyclone record. *J. Amer. Stat. Assoc.,* **106,** 89–99.

Rodionov, S. N., 2004: A sequential algorithm for testing climate regime shifts. *Geophys. Res. Lett.,* **31,** L09204, doi:10.1029/2004GL019448.

Ross, S. M., 1996: *Stochastic Processes.* Wiley, 510 pp.

——, 2010: *Introduction to Probability Models.* 10th ed. Elsevier, 784 pp.

Seidel, D. J., and J. R. Lanzante, 2004: As assessment of three alternatives to linear trends for characterizing global atmospheric temperature changes. *J. Geophys. Res.,* **109,** D14108, doi:10.1029/2003JD004414.

Sen, P. K., 1982: Invariance principles for recursive residuals. *Ann. Stat.,* **10,** 307–312.

Stapleton, J., 2009: *Linear Statistical Models.* 2nd ed. Wiley, 474 pp.

Tomé, A. R., and P. M. A. Miranda, 2004: Piecewise linear fitting and trend changing points of climate parameters. *Geophys. Res. Lett.,* **31,** L02207, doi:10.1029/2003GL019100.

Venema, V. K. C., and Coauthors, 2012: Benchmarking homogenization algorithms for monthly data. *Climate Past,* **8,** 89–115.

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

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

——, 2008: Penalized maximal *F* test for detecting undocumented mean shift without trend change. *J. Atmos. Oceanic Technol.,* **25,** 368–384.

——, Q. H. Wen, and Y. Wu, 2007: Penalized maximal *t* test for detecting undocumented mean change in climate data series. *J. Appl. Meteor. Climatol.,* **46,** 916–931.