Changepoint Detection in Periodic and Autocorrelated Time Series

ROBERT LUND,* XIAOLAN L. WANG,+ QIQI LU,# JAXK REEVES,@ COLIN GALLAGHER,* AND YANG FENG+

*Department of Mathematical Sciences, Clemson University, Clemson, South Carolina
+ Climate Research Division, Atmospheric Science and Technology Directorate, Science and Technology Branch, Environment Canada, Toronto, Ontario, Canada
# Department of Mathematics and Statistics, Mississippi State University, Mississippi State, Mississippi
@ Department of Statistics, The University of Georgia, Athens, Georgia

(Manuscript received 24 March 2006, in final form 1 February 2007)

ABSTRACT

Undocumented changepoints (inhomogeneities) are ubiquitous features of climatic time series. Level shifts in time series caused by changepoints confound many inference problems and are very important data features. Tests for undocumented changepoints from models that have independent and identically distributed errors are by now well understood. However, most climate series exhibit serial autocorrelation. Monthly, daily, or hourly series may also have periodic mean structures. This article develops a test for undocumented changepoints for periodic and autocorrelated time series. Classical changepoint tests based on sums of squared errors are modified to take into account series autocorrelations and periodicities. The methods are applied in the analyses of two climate series.

1. Introduction

A changepoint is a time at which the structural pattern of a time series changes. Instrumentation/observer changes, station location changes, and changes in observation practices are frequent culprits behind changepoints. In many cases, the changepoint time and cause are documented and it is reasonably straightforward to statistically adjust (homogenize) the series for the effects of the changepoint. Unfortunately, many changepoint times are undocumented.

While undocumented changepoints are sometimes evident in a plot of the series, debatable cases also abound. Visual detection of a changepoint in a series with a prominent seasonal mean can be difficult. Moreover, the statistical methods used to identify undocumented changepoints are known to be important [Reeves et al. (2007) reviews the topic for models with independent and identically distributed (IID) errors]. Undocumented changepoint detection methods can greatly facilitate metadata investigations by identifying times around which the investigation should focus. Hence, the development of statistically sound tests for undocumented changepoints is desirable. Undocumented changepoint detection in climate settings has been previously explored by Potter (1981), Alexandersson (1986), Solow (1987), Gullet et al. (1991), Rhoades and Salinger (1993), Easterling and Peterson (1995), Vincent (1998), Lund and Reeves (2002), Wang (2003), and Ducré-Robitaille et al. (2003) (among others). The statistical side of the subject is also vast, with Page (1955), Kander and Zacks (1966), Hinkley (1971), Hawkins (1977), Caussinus and Mestre (2004), and Davis et al. (2006) composing a prominent sample. Neither of these lists is complete.

In this paper, we develop a method for undocumented changepoint detection for series with autocorrelated and periodic features. The periodic and autocorrelation aspects are modeled in tandem rather than separately. The results enable one to test for undocumented changepoints in a wide variety of realistic climate settings. The methods can be used with or without a reference series. This paper is perhaps the first detailed investigation of changepoint detection in climate settings involving autocorrelation; changepoint detec-

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

DOI: 10.1175/JCLI4291.1

tion techniques for periodic series were previously considered in Gullett et al. (1991).

The rest of this paper proceeds as follows. Section 2 introduces a two-phase time series regression model with autocorrelated and periodic features. A test statistic weighing overall series homogeneity against the alternative of one undocumented changepoint is developed in section 3. Section 4 shows why autocorrelation and periodicities are important in changepoint detection problems. Section 5 examines power aspects of the proposed test. Applications of the methods to two series are made in section 6. Section 7 closes with several remarks.

2. The model

In the time-homogeneous (nonperiodic) setting, a simple but useful model allowing for *one* changepoint in a time series $\{X_t\}$ is the following regression:

$$X_t = \mu + \beta t + \Delta 1_{[t>c]} + \epsilon_t, \quad 1 \le t \le N, \qquad (2.1)$$

where c is the unknown time of change, the magnitude of the changepoint effect (step size) is Δ, and $\{\epsilon_t\}$ is a zero-mean random sequence that may be autocorrelated (a time series). The factor βt allows for a simple linear trend in series values. Following Wang (2003), the linear trend β is constrained to be the same before and after the changepoint time c. The focus here is on mean changes at the changepoint time and not on autocovariance, trend, or other types of changes. The simple linear structure in (2.1) may require modification in some settings. For example, Lund and Reeves (2002) study a carbon dioxide series where a quadratic trend is apparent; the existence of a good reference series may render the inclusion of a trend component unnecessary.

This paper works in the at-most-one changepoint (AMOC) setting. Multiple undocumented changepoints are indeed a frequent problem in climate series; however, as current multiple changepoint detection algorithms will need to assess the presence (or lack thereof) of a single changepoint over various subsegments of the series, AMOC methods essentially represent ground zero. Further discussion on this point is presented in section 7.

Equation (2.1) is a simple linear regression model with two phases; such models and their variants have been studied in Hinkley (1969), Solow (1987), Easterling and Peterson (1995), Vincent (1998), Lund and Reeves (2002), and Wang (2003). A periodic variant of (2.1) merely allows the location parameter μ to vary periodically with known period T: $\mu_{t+T} = \mu_t$ for each t. Using n as a cycle index and ν as a phase (called a season here) index permits (2.1) to be rewritten in seasonal form, which is

$$X_{nT+\nu} = \mu_\nu + \beta(nT + \nu) + \Delta 1_{(nT+\nu>c)} + \epsilon_{nT+\nu}, \quad 1 \le nT + \nu \le N. \qquad (2.2)$$

In (2.2), $X_{nT+\nu}$ is the series value during the νth season, $1 \le \nu \le T$, of the nth cycle of data. Our bookkeeping takes d complete cycles of data and labels these as $0, \ldots, d-1$, respectively; this makes X_1 the observation for season 1 of cycle 0 and X_N the observation for season T of cycle $d-1$. The total number of observations is $N = dT$. The setup here assumes a time-homogeneous (nonperiodic) linear trend β and a time-homogeneous mean shift Δ; this is emphasized notationally in that neither β nor Δ are subscripted with ν. Changepoints that induce different mean shifts during different seasons could be described by allowing Δ to depend on ν.

The mean series response at time $nT + \nu$ in (2.2) is

$$E(X_{nT+\nu}) = \mu_\nu + \beta(nT + \nu) + \Delta 1_{(nT+\nu>c)};$$

seasonality in the first moment arises through μ_ν, $1 \le \nu \le T$. In addition to seasonality in mean, many climatic series also display seasonality in variance and autocorrelations. For example, nontropical temperature series have larger variabilities (lag zero autocovariances) during winter seasons and many western United States precipitation series have minimal variability during late summer and early fall seasons. To allow for autocorrelation and periodicities, the errors $\{\epsilon_t\}$ are modeled as a periodically stationary time series (periodic series). A general overview of periodic series and their applications in climate modeling is presented in Lund et al. (1995).

For simplicity of computation, presentation, and overall flexibility, we will work with perhaps the simplest periodic time series model for $\{\epsilon_t\}$: a first-order periodic autoregression [PAR(1)]. Such an $\{\epsilon_t\}$ is governed by the difference equation

$$\epsilon_{nT+\nu} = \phi_\nu \epsilon_{nT+\nu-1} + Z_{nT+\nu}, \qquad (2.3)$$

where $\{Z_t\}$ is zero-mean periodic white noise; that is, Z_t and Z_s are uncorrelated when $t \ne s$, Z_t has zero mean for every t, and the variance of Z_t is periodic with $\mathrm{Var}(Z_{nT+\nu}) = \sigma_\nu^2$. The model in (2.3) has $2T$ parameters [in addition to the $T + 2$ regression parameters in (2.2)], which may be a large number if the series is observed frequently. For example, a daily PAR(1) model ($T = 365$) has 365 autoregressive parameters and 365 white noise variance parameters. Parsimony issues for periodic series are discussed in Lund et al. (2006).

3. The test statistic

An undocumented changepoint test statistic weighs the null hypothesis that $\Delta = 0$ (termed a null or H_0 model) against the alternative that $\Delta \neq 0$ (termed a full or H_A model). The changepoint time c is an unknown parameter of the full model. The general form of our test statistic coincides with that in Lund and Reeves (2002) and Wang (2003):

$$F_{\max} = \max_{1 \le c \le N-1} F_c, \tag{3.1}$$

where F_c defined by

$$F_c = \frac{\mathrm{SSE}_0 - \mathrm{SSE}_A(c)}{\mathrm{SSE}_A(c)/(N-p)} \tag{3.2}$$

is a regression F-type statistic measuring closeness of the null model and a full model with a single changepoint at time c. In (3.2), SSE_0 is the sum of squared errors under H_0 and $\mathrm{SSE}_A(c)$ is the H_A sum of squared errors when a changepoint exists at time c. SSE_0 does not depend on the value of c but $\mathrm{SSE}_A(c)$ does. In (3.2), p is the number of regression parameters involved in the full model with a changepoint at time c. In our section 6 applications with monthly series, $p = 14$.

In classical regression settings with IID Gaussian $\{\epsilon_t\}$, F_c has an F distribution (exactly) with 1 numerator and N–p denominator degrees of freedom. The larger F_c is, the more evidence points to an undocumented changepoint at time c. Intuitively, the $F_{\max}$ statistic selects the time of largest discrepancy in the two phases of the model as the estimator of c; H_0 is accepted when $F_{\max}$ is small enough to be explained by chance variation and rejected when $F_{\max}$ is excessively large.

For IID $\{\epsilon_t\}$, Alexandersson (1986) and Lund and Reeves (2002) connect the $F_{\max}$ statistic to Gaussian likelihood ratios and maxima of correlated t and F random variates. Since the F_cs are correlated in c, $F_{\max}$ does not behave statistically as the maximum of independent F statistics.

The key methodological innovation put forth here involves modifying sums of squares in autocorrelated and periodic settings. Here, sums of squared errors are best linear predictors. Because we work with seasonal series, these sums of squares should be weighted for seasonal variabilities. Specifically, for known time series parameters ϕ_ν and σ_ν^2 for $1 \le \nu \le T$,

$$\mathrm{SSE}_0 = \sum_{n=0}^{d-1} \sum_{\nu=1}^{T} \frac{(X_{nT+\nu} - \hat{X}^0_{nT+\nu})^2}{\sigma_\nu^2};$$

$$\mathrm{SSE}_A(c) = \sum_{n=0}^{d-1} \sum_{\nu=1}^{T} \frac{[X_{nT+\nu} - \hat{X}^A_{nT+\nu}(c)]^2}{\sigma_\nu^2}, \tag{3.3}$$

where the predictions $\hat{X}^0_{nT+\nu}$ and $\hat{X}^A_{nT+\nu}(c)$ are best one-step-ahead linear predictions from the observed past:

$$\hat{X}^0_t = P_0[X_t | X_1, \ldots, X_{t-1}, 1],$$

$$\hat{X}^A_t(c) = P_A[X_t | X_1, \ldots, X_{t-1}, 1]. \tag{3.4}$$

The notation $P[Y|X_1, \ldots, X_m, 1]$ denotes the best (minimum mean square error) linear prediction of Y from linear combinations of $X_1, \ldots, X_m$ and a constant. The subscript under P (or the superscript on $\hat{X}_t$) indicates the model (H_0 or H_A) under which the linear prediction is to be computed. Brockwell and Davis (1991, chapter 8) provide the theory for sums of squared errors in time series settings. The assumed PAR(1) structure renders $\mathrm{Var}(X_{nT+\nu} - \hat{X}_{nT+\nu}) = \sigma_\nu^2$; for more "complicated time series models," one merely replaces σ_ν^2 in the denominators in (3.3) by the appropriate expression for the residual variance $\mathrm{Var}(X_{nT+\nu} - \hat{X}_{nT+\nu})$. The $F_{\max}$ statistic proposed here reduces to the $F_{\max}$ statistic for IID errors when autocorrelation and seasonality are not present.

The computation of $F_{\max}$ requires SSE_0 and $\mathrm{SSE}_A(c)$ for each c. We first tackle SSE_0. For the H_0 model, the PAR(1) structure gives

$$\hat{X}^0_{nT+\nu} = \mu_\nu + \beta(nT+\nu) + \phi_\nu[X_{nT+\nu-1} - \mu_{\nu-1} - \beta(nT+\nu-1)] \tag{3.5}$$

for $2 \le nT + \nu \le N$, where the start-up convention $\hat{X}^0_1 = \mu_1 + \beta$ is made. Estimated values of μ_ν, $1 \le \nu \le T$, and β are computed by weighted least squares methods (see Fuller 1996 for background), with the weights set optimally according to the covariance matrix of the observations $X_1, \ldots, X_N$, and used in (3.5). The equations governing the PAR(1) covariance structure are (5.4)–(5.7) in Lund and Basawa (2000). The PAR(1) covariance matrix depends only on the ϕ_νs and σ_ν^2s, which are viewed here as nuisance parameters. In practice, one needs only a rough idea of their values since small perturbations in the autocovariance parameters do not induce radical changes in the sum of squares, and this is easily accomplished by several methods, one of which is presented in section 6a.

To compute $\mathrm{SSE}_A(c)$ for a fixed c, we proceed as with the null model except that (3.5) is modified to account for the changepoint at time c:

$$\hat{X}^A_{nT+\nu} = \mu_\nu + \beta(nT+\nu) + \Delta 1_{(nT+\nu>c)} + \phi_\nu[X_{nT+\nu-1} - \mu_{\nu-1} - \beta(nT+\nu-1) - \Delta 1_{(nT+\nu-1>c)}] \tag{3.6}$$

for $2 \le nT + \nu \le N$, where the start-up convention $\hat{X}^A_1 = \mu_1 + \beta$ is made.

One rejects the null hypothesis of series homogeneity if the $F_{\max}$ statistic exceeds a critical value calibrated to a preset level of statistical confidence; in our section 6 applications, these critical values (thresholds) are set for 95% confidence and are computed by simulation. These critical values, in theory, depend slightly on the time series parameters ϕ_ν and σ_ν for $1 \le \nu \le T$. This dependence decays as N increases; accounting for this dependence by simulation for each N and set of time series parameters allows one to attach exact margins of error. The applications in section 6 provide some feel for these critical values.

When N is small, it may be wise to investigate parsimony constraints for the regression and time series parameters. Physical logic may dictate the form of the constraints. For example, when a good reference series is available and the differenced series is tested for an undocumented changepoint, it might be sufficient to set $\mu_\nu \equiv \mu$ (a constant) and $\beta = 0$ (zero trend). Constrained periodic time series modeling is considered in Lund et al. (2006).

4. Effects of autocorrelation and periodicities

This section shows how autocorrelations and periodicities degrade changepoint detection procedures. Here, we present simulations that quantify how an $F_{\max}$ statistic performs when autocorrelations and periodicities are ignored.

We first investigate the effects of autocorrelation. To study the effects of autocorrelation only, we examine a time-homogeneous setup: (2.1) and $\{\epsilon_t\}$ satisfying the first-order autoregression [AR(1)]

$$\epsilon_t = \phi\epsilon_{t-1} + Z_t, \tag{4.1}$$

where $\{Z_t\}$ is IID zero-mean Gaussian noise with variance σ^2 and $|\phi| < 1$. We consider series of length $N = 100$; other series lengths behave similarly. As the autoregressive coefficient $\phi > 0$ increases, the degree of serial autocorrelation in the errors increases. When $\phi = 0$, the errors are independent. The white noise variance σ^2 is selected to make the variance of ϵ_t unity in all cases; this entails setting $\sigma^2 = 1 - \phi^2$ and allows for meaningful comparisons across different table entries.

Table 1 reports empirical probabilities of erroneously rejecting H_0 (false alarm rates), at level 5%, for various values of the autocorrelation coefficient ϕ. Each empirical probability was aggregated from 100 000 independent simulations; hence, simulation error is minimal. In each simulation run, a Gaussian AR(1) error series $\{\epsilon_t\}$ with specified ϕ is first generated. Two $F_{\max}$ statistics are then computed, one that takes into account autocorrelations (the "new" column) and one that ignores autocorrelations (the "old" column). In computing the $F_{\max}$ statistic that takes autocorrelations into account, we have used the true values of ϕ and σ^2 to compute one-step-ahead predictions and their mean-squared errors. This allows us to exclude the effects of time series parameter estimation. Results for estimated time series parameters are presented later in this section. The mean of $\{X_t\}$ is taken as zero ($\beta = \mu = 0$). In this case, the hypothesis of no changepoint is rejected whenever an $F_{\max}$ statistic exceeds 11.054 [the percentiles of Wang (2003) apply in this setting]. The actual (up to simulation error) 95th percentiles of the $F_{\max}$ statistic, denoted by $F_{\max,0.95,}$ that ignore autocorrelation are included for comparison's sake; notice that they are much larger than the 11.054 threshold applicable to $\phi = 0$. In these cases, an error (type I) is made any time the test declares an undocumented changepoint to exist (H_A), as, in truth, no changepoints are present (H_0).

TABLE 1. Effects of autocorrelation on changepoint detection.

AR(1) ϕ	Old probability	New probability	$F_{\max,0.95}$
0.95	0.997	0.0515	176.753
0.75	0.924	0.0499	64.185
0.50	0.601	0.0509	29.547
0.25	0.238	0.0515	17.250
0.15	0.138	0.0497	14.350
0.00	0.0508	0.0508	11.054
−0.25	0.006 34	0.0499	7.460
−0.50	0.002 63	0.0515	5.323
−0.75	0.002 71	0.0502	4.211
−0.95	0.002 55	0.0507	4.531

The Table 1 rejection probabilities are extremely high for the old $F_{\max}$ column, dramatically so for values of ϕ slightly less than unity. The rejection probabilities should be close to 0.05 in a well-functioning test. This agrees with the findings of Percival and Rothrock (2005) and is not surprising geometrically: as $\phi > 0$ becomes larger, the series makes longer sojourns above and below its mean response levels, which effectively imitate the effects of a mean shift due to a changepoint. This said, we are somewhat surprised with the drastic performance degradations in autocorrelated settings; even the $\phi = 0.15$ case, which entails minor autocorrelation, shows empirical rejection rates (type-I errors) that are some 3 times too large. When $\phi = 0.5$, the rejection rate is a whopping 60%. When $\phi < 0$ (which is encountered less frequently in climate modeling), consecutive observations tend to split the mean response level (one above and one below) and make changepoints easier to detect. Notice that the $F_{\max}$ statistic accounting for autocorrelations, however, is performing as it should in all cases, with empirical rejection

TABLE 2. Effects of estimating AR(1) parameters.

	$n = 100$	$n = 500$	$n = 1000$
$\phi = 0$	0.032 25	0.048 11	0.053 89
$\phi = 0.1$	0.031 68	0.047 53	0.052 59
$\phi = 0.25$	0.030 12	0.046 82	0.054 39
$\phi = 0.50$	0.027 80	0.045 91	0.051 71
$\phi = 0.75$	0.030 77	0.042 98	0.050 25

rates very close to 5%. The deficiencies of the test should not be blamed on use of $F_{\max}$ procedures; other types of changepoint tests that ignore autocorrelations will see the same performance degradations. In fact, the $F_{\max}$ test is statistically optimal (a likelihood ratio test) in the case of IID Gaussian errors. Overall, the theme is clear: one should be extremely careful about the modeling methods in changepoint settings with positively autocorrelated series. Ignoring positive autocorrelations can lead to a higher than specified false alarm rate, while ignoring negative autocorrelations may let a true changepoint go undetected.

To show that the effects of estimating time series parameters are negligible, Table 2 reports sample type-I errors for our new $F_{\max}$ statistic when the AR(1) time series parameters are estimated, as opposed to assuming them to be known as was done in Table 1. The simulations consider various values of ϕ and the sample sizes $n = 100$, 500, and 1000; $\sigma^2 = 1$ was taken in all cases and a level 95% test was again used. Each table entry is based on 100 000 independent simulations. The results show that estimating ϕ and σ^2 does not overly impact the type-I errors, with lessening effects as the sample size increases. The Table 2 values merely reflect that the estimators of ϕ and σ^2 become more accurate as the sample size increases.

Next, we consider how periodicities in the white noise variances influence changepoint detection. Here, our model is (2.2) and $\{\epsilon_t\}$ is periodic Gaussian white noise. To study the effects of periodicities in the error variances only, our parameter choices are $\mu_\nu \equiv 0$, $\phi_\nu \equiv 0$, and $\beta = 0$. The white noise variances were assumed sinusoidal in season for simplicity:

$$\sigma_\nu^2 = C_0 + C_1 \cos\left[\frac{2\pi(\nu - \xi)}{T}\right]. \tag{4.2}$$

Akin to Table 1, Table 3 reports empirical type-I errors. The "old" column here refers to an $F_{\max}$ statistic that ignores seasonality in the variances of $\{\epsilon_t\}$; the "new" column refers to an $F_{\max}$ statistic that takes these seasonal variances into account (by using exact values of σ_ν^2). Again, each table entry is based on 100 000 independent simulations. The table varies the values of

TABLE 3. Effects of seasonal variances on changepoint detection.

C_0	C_1	Old probability	New probability	True 95th percentile
1.00	0.00	0.0502	0.0502	11.107
1.00	0.50	0.0676	0.0501	11.926
1.00	0.95	0.0984	0.0500	13.253
10.00	0.00	0.0500	0.0500	11.093
10.00	5.00	0.0677	0.0501	11.926
10.00	9.50	0.0970	0.0501	13.254

C_0 and C_1 but sets $\xi = 0$ in all cases. If $C_1 = 0$, the error variances are nonseasonal and the setting reduces to that studied in Wang (2003). The larger C_1 is relative to C_0 the more seasonality there is in white noise variances (across varying seasons ν). Of course, we need $C_1 < C_0$ or σ_ν^2 could become negative. This table employs $N = 120$ ($d = 10$, $T = 12$), which corresponds to a decade of monthly data. For series of this length, an $F_{\max}$ statistic must be 11.105 or greater to declare H_a [the percentiles of Wang (2003) are applicable when $C_1 = 0$].

The Table 3 results show that the rejection probabilities increase slightly with increasing seasonal variability. This is as expected: when σ_ν^2 varies greatly with the season ν, there is a larger chance for an outlying ϵ_t to pull the least squares regression fit away from its true zero-mean level, hence mimicking a mean shift caused by a changepoint. Note, however, that the effects of periodic variances are nowhere near as drastic as those of autocorrelation. Also, observe that the percentiles for the first three rows are approximately those for the last three rows and that those for $C_1 = 0$ coincide with those reported in Wang (2003), up to simulation error.

5. Power aspects

The last section demonstrated that the type-I error rate of the test was as advertised in autocorrelated and seasonal settings. This section studies the power of detection attributes. Specifically, we will examine how frequently the test detects changepoints in settings when, in truth, changepoints exist.

The methods we compare here are 1) the test proposed in section 3; 2) the $F_{\max}$ statistic of Wang (2003) applied to the seasonally mean adjusted series $\{X_{nT+\nu} - \overline{S}_{nT+\nu}\}$, where $\overline{S}_\nu = d^{-1}\Sigma_{n=0}^{d-1} X_{nT+\nu}$ and $\overline{S}_{nT+\nu} = \overline{S}_\nu$ (seasonal-mean adjustment); and 3) the $F_{\max}$ statistic of Wang (2003) applied to the annual averages $\{\overline{X}_n\} = \{T^{-1}\Sigma_{\nu=1}^{T} X_{nT+\nu}\}$ (annual averaging). Methods 2 and 3 entail crude ways of producing stationary series from series with periodic characteristics; they do not take autocorrelations into account.

TABLE 4. Detection powers.

Method with $\phi = 0.15$	$\kappa = 0.0$	$\kappa = 0.5$	$\kappa = 1.0$
New method	0.0505	0.2607	0.8902
Season-mean adjustment	0.1751	0.3922	0.9058
Annual averaging	0.0546	0.2586	0.8545
Method with $\phi = 0.25$	$\kappa = 0.0$	$\kappa = 0.5$	$\kappa = 1.0$
New method	0.0514	0.1803	0.8058
Season-mean adjustment	0.3107	0.3505	0.8621
Annual averaging	0.0569	0.1884	0.7835
Method with $\phi = 0.50$	$\kappa = 0.0$	$\kappa = 0.5$	$\kappa = 1.0$
New method	0.0518	0.0579	0.3948
Season-mean adjustment	0.7536	0.2193	0.6130
Annual averaging	0.0735	0.0698	0.4329

We will use simulation to study the detection powers. Each simulation run generates an error series $\{\epsilon_t\}$ with a time-constant autoregressive coefficient ϕ (for simplicity and interpretability) and $N = 600$, which represents 50 yr of monthly data. For realism, the means μ_ν, seasonal variances σ_ν^2, and trend β are set to those estimated for the Longmire series in section 6b under the alternative hypothesis of a changepoint. In each simulation run, a synthetic mean shift is introduced at a changepoint time chosen randomly in $\{1, \ldots, N - 1\}$. The magnitude of the mean shift at the changepoint time is reported in terms of the parameter $\kappa = \Delta/\overline{\sigma}$. Here, Δ is the actual magnitude of the mean shift at the changepoint time and $\overline{\sigma} = T^{-1}\Sigma_{\nu=1}^{T}\sigma_\nu$ is the average error standard deviation over a full cycle; κ should be a good measure of power in that larger κs will make the changepoint time easier to detect. The ϕ values of 0.15, 0.25, and 0.5 and the κ values of 0.5 and 1.0 are studied. Despite generating the error series with a constant ϕ_ν, the full algorithm in section 3 is put to the test; this entails seasonal estimation of the ϕ_νs.

Table 4 reports empirical powers of the three methods aggregated from 10 000 independent simulations in a level 5% test. Our conventions call a simulation a success if the correct changepoint time is found within ± 18 months of the true changepoint time in the cases where $\kappa > 0$ and merely makes the correct conclusion of no changepoint when $\kappa = 0$. In the annual tests, the above specifications translate into getting the changepoint time correct to within one year. For example, when $\phi = 0.25$ and $\kappa = 0.5$, the seasonal-mean adjustment procedure is signaling for a changepoint within ± 18 months of the true changepoint time 35.05% of the time. Here, $\overline{\sigma} = 1.54$. The 95th percentile for the new test, as estimated from 10 000 independent simulations, is 14.0 when $\phi = 0.15$, 14.3 when $\phi = 0.25$, and 15.6 when $\phi = 0.50$. The 95th percentile for the seasonal-mean adjusted procedure is 11.55, while that for annual averaging is 11.07.

The type-I error rate (the $\kappa = 0$ column) for the seasonal-mean adjustment procedure is unacceptably high for a level 5% test, making this procedure unusable in our work. Seasonal-mean adjustment procedures ignore all correlations in the series, which, as the last section showed, is not wise when there is indeed correlation. Having a higher power, as the seasonal-mean adjusted procedure possesses, does not necessarily yield a superior test: a test that always rejects the null hypothesis of no changepoints will always make the correct conclusions when in truth a changepoint exists and the wrong ones when changepoints do not exist. The method of annual averaging merits some consideration, however, as its powers are similar to those of the new method and its type-I error rate is only slightly higher than the nominal 0.05 level. Annual averaging reduces correlation magnitudes in positively autocorrelated series; that is, year-to-year autocorrelations are significantly less than month-to-month autocorrelations. The simulations here involve only moderate autocorrelation; the type-I error of annual averaging methods will degrade as the correlation becomes larger (ϕ becomes larger). Also worth mentioning is the issue of parsimony: tests with many parameters in general have lower powers than more parsimoniously devised tests. Annual averaging methods involve only three parameters: an annual mean, trend, and error variance. This parameter count is substantially less than the 37 needed to describe general periodic monthly data. If one is willing to take the time to parsimonize the parameters in the periodic model, improvements in the power of detection should result. As the sample size becomes larger, the new method will perform better and better in comparison to annual averaging. Also, as we will see in the first example of the next section, the new test is able to detect changepoints when the data record is relatively short; in this case, annually averaged series are not long enough to have good detection power.

With the type-I error and the power aspects of the new test clearly performing well, it remains to see how the test works on actual data. Before doing so, we must address a matter which was glossed over earlier. In the above work, we estimated the time series parameters ϕ_ν and σ_ν^2 at the c that is associated with an ordinary least squares version of $F_{\max}$, then regarded these time series parameters as fixed and computed optimal estimates of the regression parameters μ_ν, β, and Δ under the null and alternative hypotheses. This allows us to calculate the F_c and $F_{\max}$ statistics of section 3. The next section illustrates the procedure with explicit examples. A challenge lies with the need to determine the critical values of the $F_{\max}$ statistic for each individual series being

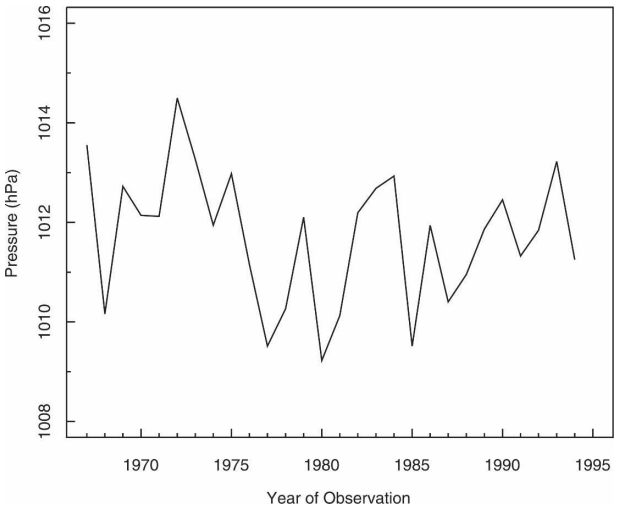

FIG. 1. Burgeo annual mean station pressures.

tested. For the results presented in this section (and section 6), the critical values are obtained from 10 000 independent simulations, using a computer program (written by the third author in FORTRAN) that computes all the parameter estimates for each simulation run in less than 1 s. Roughly, at least 100 000 simulations are required if one wants to estimate the 95th percentiles of the $F_{\max}$ to one decimal place, which, for a 50-yr monthly series ($N = 600$), would take about one day to finish on a standard computer with the FORTRAN code mentioned above (available from the authors). Quicker methods, such as a parametric bootstrap procedure, are currently being investigated.

6. Examples

a. A monthly mean atmospheric pressure series

The above methods were applied to a series of monthly mean atmospheric pressures recorded at Burgeo, Newfoundland (Canada), from 1967 to 1994 ($d = 28$ yr; $N = 336$). The annual averages of this series are plotted in Fig. 1; the monthly series, both raw and adjusted for a seasonal mean, can be viewed (along with a mean fit explained below) in Fig. 2. Mean series pressures are relatively lower in winter (December–March) and peak in summer. Winter pressure variabilities are higher than those in summer, a property also shared by temperatures. These structures are reflected in the rela-

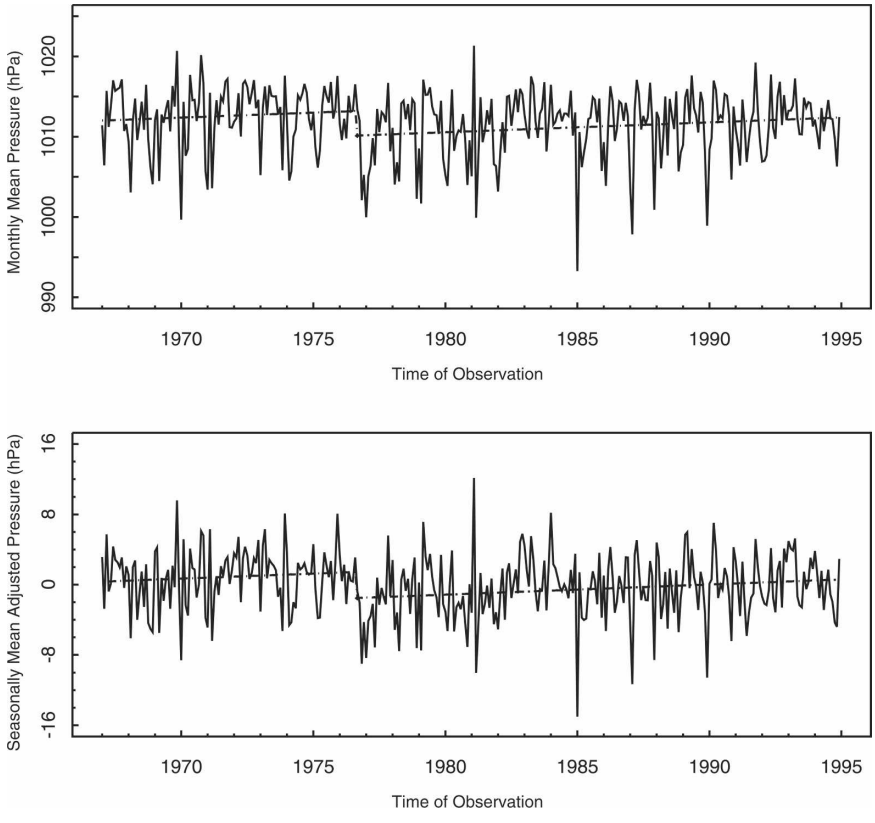

FIG. 2. (top) Burgeo monthly mean station pressures and (bottom) seasonally mean adjusted series with model fits.

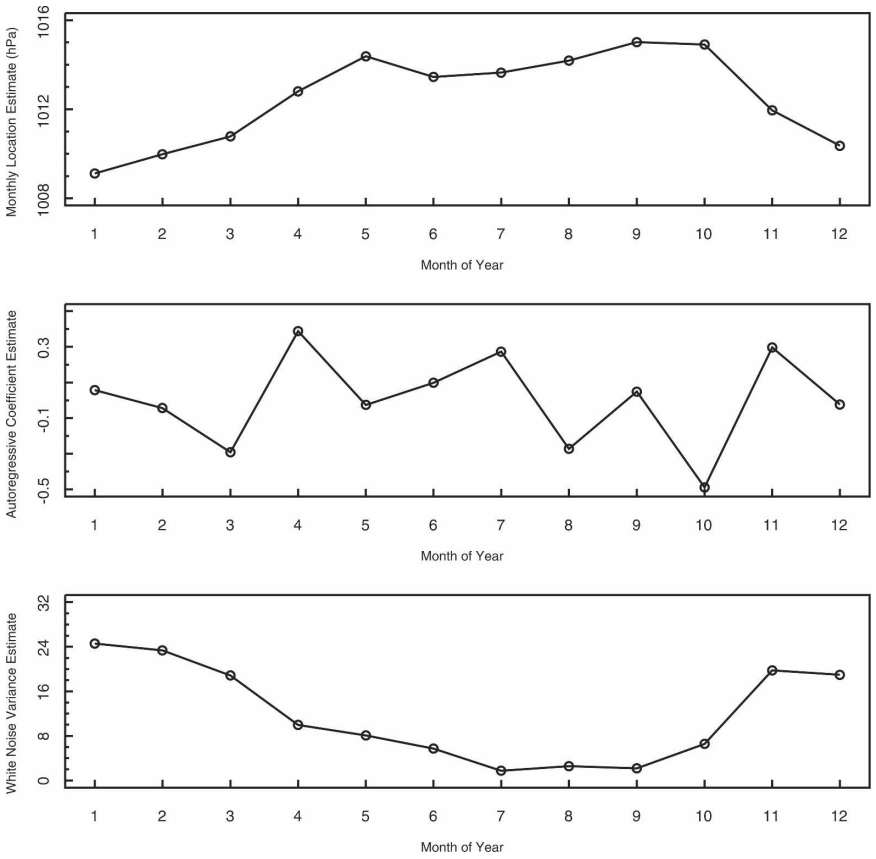

FIG. 3. Burgeo seasonal parameter estimates: (top) $\hat{\mu}_\nu$, (middle) $\hat{\phi}_\nu$, and (bottom) $\hat{\sigma}_\nu^2$.

tive magnitudes of the parameter estimates displayed in Fig. 3.

The $F_{\max}$ test of Wang (2003) for undocumented changepoints was applied to annually averaged Burgeo pressure values. No changepoint was found at the 95% level. This is not surprising given the paucity of data (28 points). The $F_{\max}$ test of Wang (2003) was also applied to the seasonally mean adjusted series $\{X_{nT+\nu} - \overline{S}_{nT+\nu}\}$. This test signals a changepoint at $c = 116$ (August 1976) at the 95% level. However, as noted in the previous section, this test has an extremely high false alarm rate when autocorrelation is present, since it does not account for serial autocorrelation. Hence, we will scrutinize the results by applying the section 3 methods.

To compute the section 3 $F_{\max}$ test statistic, a full model for each admissible changepoint time c was first fitted to the series by numerically minimizing the sum of squared errors in (3.3) with the autocorrelation aspects component ignored [$\{\epsilon_t\}$ is taken as time-homogeneous white noise]. The sum of squared errors at $c = 116$ is the smallest and hence serves as our preliminary estimate of a changepoint time. The ϕ_ν and σ_ν^2 parameters are then estimated from simple moment equations involving the residuals $\{R_t\} = \{X_t - \hat{X}_t^A(c)\}$ computed from a full model fit with $c = 116$:

$$\hat{\phi}_\nu = \frac{\hat{\gamma}_\nu(1)}{\hat{\gamma}_{\nu-1}(0)}, \quad \hat{\sigma}_\nu^2 = \hat{\gamma}_\nu(0) - \hat{\phi}_\nu\hat{\gamma}_\nu \tag{6.1}$$

and

$$\hat{\gamma}_\nu(h) = d^{-1}\sum_{n=0}^{d-1} R_{nT+\nu}R_{nT+\nu-h},$$

with $R_0 = 0$. Estimates of Δ, β, and μ_ν for each season ν are now computed via weighted least squares methods, with the weights set according to the covariance matrix of $X_1, \ldots, X_N$ (which depends only on the ϕ_ν's and σ_ν^2s). A new set of residuals was then computed and refits of the parameters in (6.1) were calculated. This procedure was iterated 5 times in the spirit of Cochrane and Orcutt (1949) to obtain good estimates of the time series parameters ϕ_ν and σ_ν^2. These values are displayed in the bottom two panels of Fig. 3.

The next step is to compute F_c for each admissible c, accounting for the effects of autocorrelations and peri-

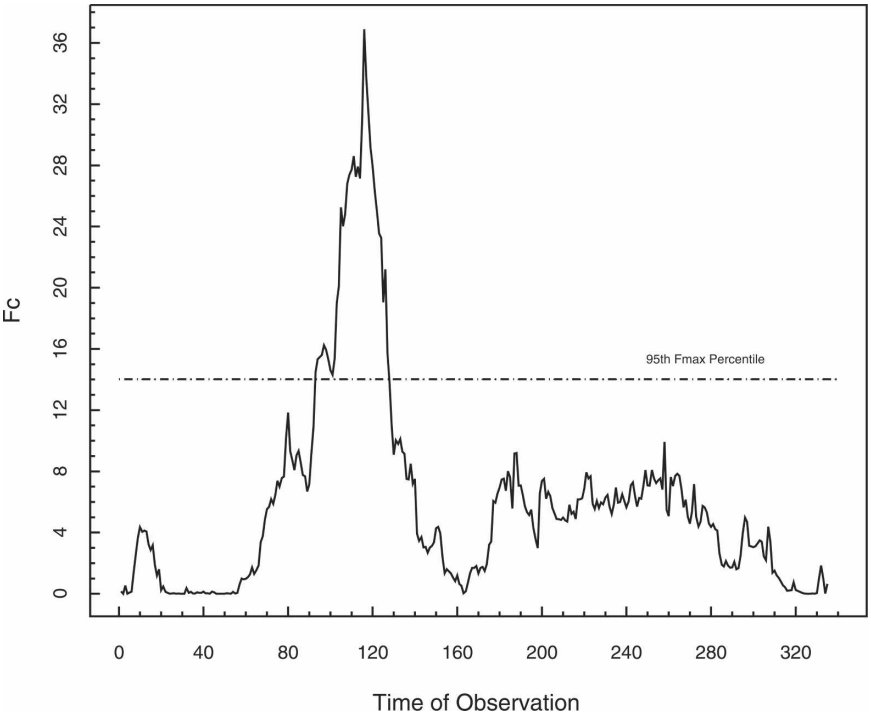

FIG. 4. Burgeo F_c statistics.

odicities. In this computation, the ϕ_ν and σ_ν^2 parameters are held fixed as estimated above and the sums of squares are computed to minimize (3.3) over values of β, Δ, and μ_ν, $1 \le \nu \le T$. Figure 4 plots the F_c statistics. For this series, the 95th percentile is 14.0. The largest F_c is 36.90, which occurred again at $c = 116$ (August 1976) and greatly exceeds the 95% threshold. Hence, evidence suggests a significant changepoint in 1976. The mean shift at the changepoint time is estimated as $\Delta = -2.893$ hPa by the full model and the trend estimate is $\hat{\beta} = 0.006\,36$ hPa month^{-1}. Figure 2 also plots the mean structure of this series, less the seasonal cycle contained in the μ_νs (this is for visual clarity), against the data (dashed curve). This fit appears to describe the series well. The estimates of μ_ν are plotted in the top panel of Fig. 3.

Investigation of the related metadata suggests that the changepoint was caused by neglecting the 10.6-m elevation in the calculation of station pressures from barometer readings prior to 1977 (i.e., an elevation of 0 m was used instead of 10.6 m). According to a physically based estimate using a hydrostatic model and hourly pressure and temperature data (see Wan et al. 2007), neglecting such an elevation causes a bias of 1.32 hPa on pressure values. The estimated changepoint time is very close to its plausible value. Additional changes that happened between 1976 and 1977, such as the use of computer-produced pressure reduction tables and the addition of a plateau correction, may have also contributed to the magnitude of the mean shift.

The autocorrelations in this series are not overly large (many are not significantly different from zero, perhaps due in part to the short length of the series). In fact, the average $\hat{\phi}_\nu$ (over all 12 months) is zero to two decimal places. In retrospect, a more efficient estimation strategy might have been to model ϕ_ν as constant in the season ν, a sound general approach. Other model parsimonizing steps, such as parameterizing the σ_ν^2 as a simple cosine wave, could also be pursued if desired.

b. A monthly temperature series

Figure 5 displays 50 yr of average monthly temperatures recorded at Longmire, Washington (inside Mount Rainier National Park), from 1951 to 2000 ($d = 50$ yr; $N = 600$); both the raw and the seasonally adjusted date are plotted. The seasonal cycle in the data is clear, with winter temperatures being colder and slightly more variable than summer temperatures. Estimates of the μ_νs are displayed in the top panel of Fig. 6.

The method of Wang (2003) was applied to the annually averaged series and does not find a changepoint at the 95% level. However, applying Wang's (2003)

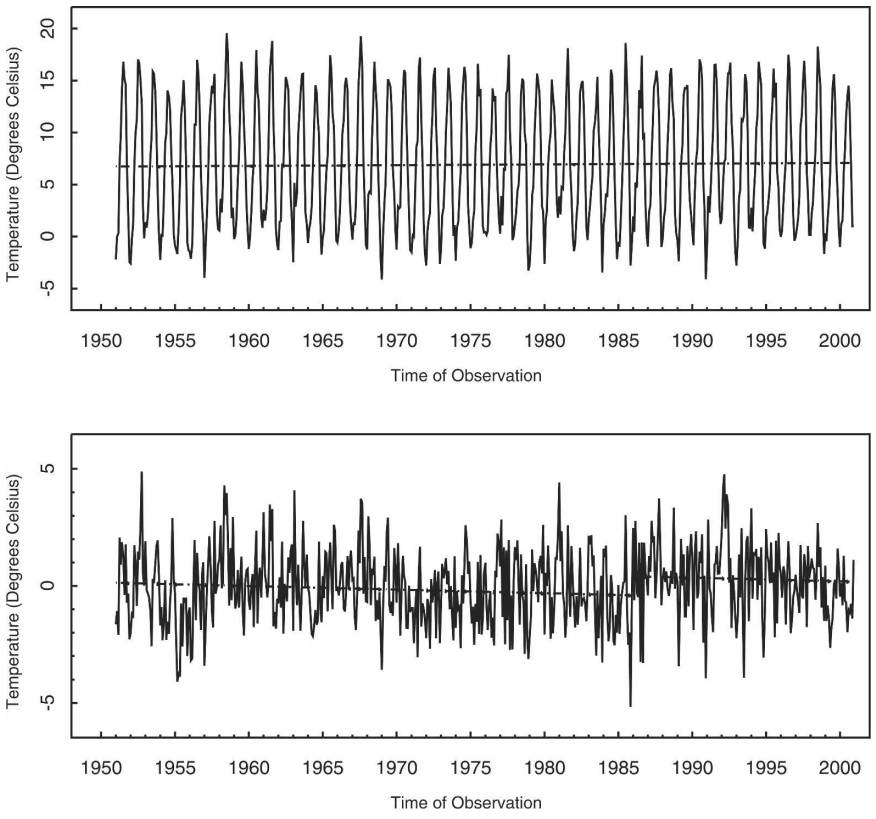

FIG. 5. (top) Longmire monthly mean temperatures and (bottom) seasonally mean adjusted series with model fits.

methods to the seasonally mean adjusted series reveals a changepoint at $c = 420$ (December 1985) with $F_{\max} = 12.18$, which slightly exceeds the 95% threshold of 11.55 for that test.

Examination of the related metadata does not reveal a reason for the changepoint. The Longmire series shows stronger autocorrelations than the Burgeo series (the lag-one sample autocorrelation of the seasonally mean adjusted series is approximately 0.245); hence, one is concerned that the changepoint declaration is due to the effects of autocorrelation.

Proceeding as with the Burgeo series, the methods of section 3 were fitted for each candidate changepoint time c. The best-fitting model had $c = 420$; values of the PAR(1) parameter estimates for this c are plotted in the middle and lower panels of Fig. 6. From these PAR(1) parameters, F_c statistics were then computed for each admissible c; these are plotted in Fig. 7 against a 95% confidence threshold. The largest F statistic was 12.29, occurred at $c = 419$ (November 1985), and does *not* exceed the 95% confidence threshold of 14.2 for this series. The null hypothesis model fit (no changepoint) has $\hat{\beta} = 0.000\,234$°C month^{-1}.

We believe that the Longmire series is typical in that an analysis that accounts for periodicities and autocorrelations does not find a changepoint in this series, but methods that ignore these features do. Given the results in Table 1, we expect that such conclusions will arise frequently. Of course, the methods of section 3 are preferable as they account for the series autocorrelations and periodicities; mistakes can be made when autocorrelations are ignored. In fact, as observed in the two examples of this section, the method of section 3 obtained what we believe to be the "correct" answer in both cases (detecting a changepoint at the correct time in the Burgeo series but not finding any changepoints in the Longmire series), whereas the annual-averaged method failed to detect a changepoint in either series and the monthly adjusted method detected "changepoints" in both series. This is consistent with the behavior of these procedures noted in the power study of section 5.

7. Remarks

We conclude with several comments. First, the setting examined here is the classical "at-most-one

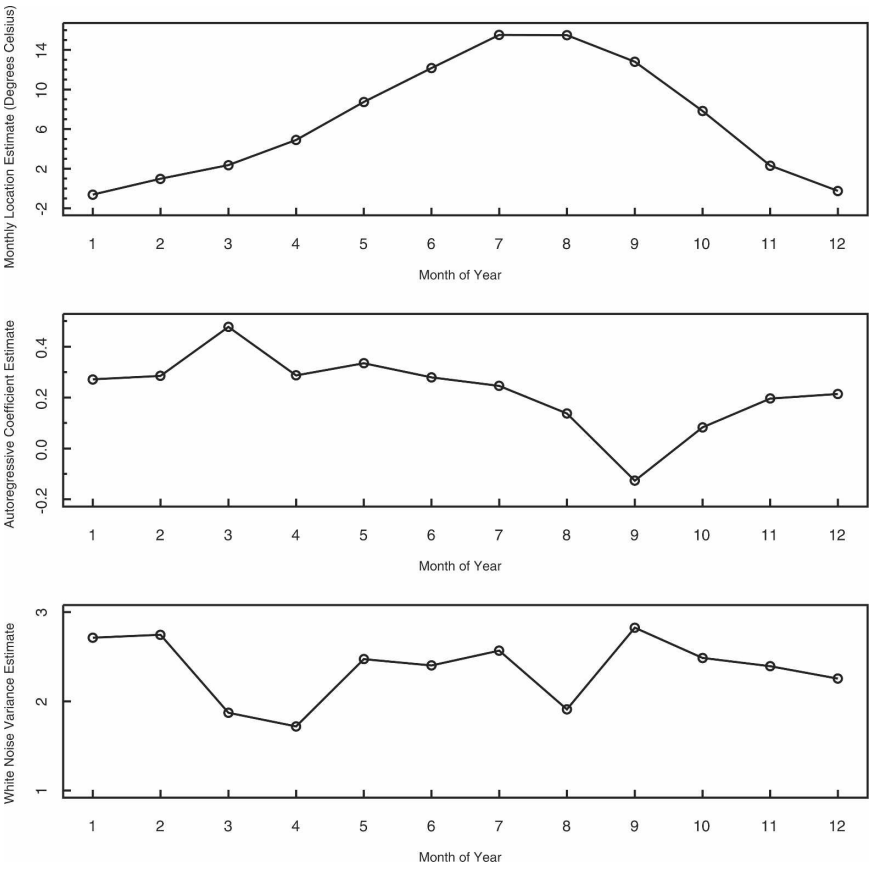

FIG. 6. Longmire seasonal parameter estimates: (top) $\hat{\mu}_\nu$, (middle) $\hat{\phi}_\nu$, and (bottom) $\hat{\sigma}^2_\nu$.

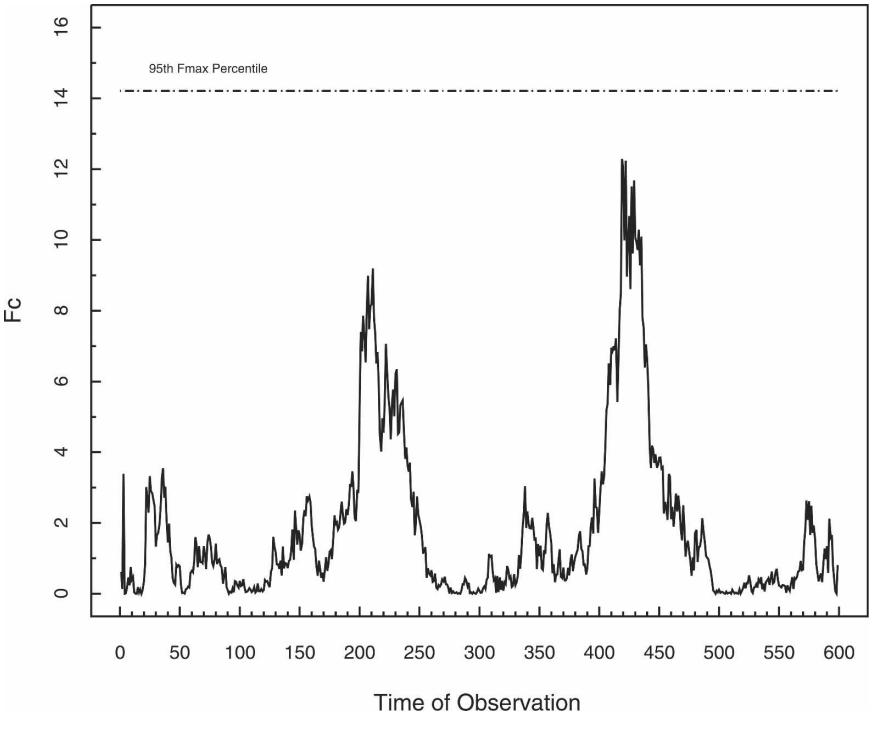

FIG. 7. Longmire F_c statistics.

changepoint" scenario. This may not be realistic for longer series. In practice, climatologists typically apply AMOC methods recursively to series subsegments to detect multiple changepoints (see Wang 2006; Menne and Williams 2005; Wang and Feng 2007). Recent work by Caussinus and Mestre (2004) and Davis et al. (2006) are rooted in efficient algorithms to find the locations of multiple changepoints. Davis et al. (2006) do allow for autocorrelated model errors but focus on applications where the covariance structure of the model errors shifts at the changepoint times, but the mean remains constant (in contrast to keeping the time series parameters fixed across varying series segments and studying mean shifts). Their methods should be adaptable to our setting and hence show promise for multiple changepoint problems.

One should be cautious when seasonally adjusting a series before checking for undocumented changepoints. This is because the estimates of the seasonal parameters will be biased, sometimes quite heavily, if the mean shift(s) induced by the changepoint is large and ignored. A remedy to the problem is as we have done above: fit the changepoint and seasonal structures simultaneously.

The PAR(1) model used here for $\{\epsilon_t\}$ seems to work well for localized series (series that have not been spatially averaged over a large region). Localized series typically display short memory (Handcock and Wallis 1994). Series that have been aggregated over a large area can possess a longer memory. For such series, the model for $\{\epsilon_t\}$ may need to be modified. Covariate effects, such as El Niño and the North Atlantic Oscillation, could also be incorporated into the regression model if needed.

The changepoint detection techniques here are based on (2.2) and the methods should not be applied to series that grossly violate this regression structure. For instance, these methods may not work well for series that display quadratic trends, such as those considered in Lund and Reeves (2002). Reeves et al. (2007) provide a recent overview of regression response function selection issues in changepoint settings.

We envision a plethora of climate series where previously declared undocumented changepoints might be erroneously diagnosed, or at least need to be reassessed. The seasonal resolution of the model will also prevent a changepoint occurring in the middle of a year from spreading its effects over two adjacent cycles.

Acknowledgments. The comments made by three referees stimulated much of the work presented here, as well as clarified our own ideas on the topic.

REFERENCES

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

Brockwell, R., and R. A. Davis, 1991: *Time Series: Theory and Methods.* 2d ed. Springer-Verlag, 577 pp.

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

Cochrane, D., and G. H. Orcutt, 1949: Application of least squares regression to relationships containing autocorrelated error terms. *J. Amer. Stat. Assoc.,* **44,** 32–61.

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

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

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.

Fuller, W. A., 1996: *Introduction to Statistical Time Series.* 2d ed. John Wiley and Sons, 698 pp.

Gullet, D. W., L. Vincent, and L. Malone, 1991: Homogeneity testing of monthly temperature series: Application of multiphase regression models with mathematical changepoints. Canadian Climate Centre Rep. 91-10, Atmospheric Environment Service, Downsview, ON, Canada, 47 pp.

Handcock, M. S., and J. R. Wallis, 1994: An approach to statistical spatial-temporal modeling of meteorological fields (with discussion). *J. Amer. Stat. Assoc.,* **89,** 368–390.

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

Hinkley, D. V., 1969: Inference about the intersection in two-phase regression. *Biometrika,* **56,** 495–504.

——, 1971: Inference in two-phase regression. *J. Amer. Stat. Assoc.,* **66,** 736–743.

Kander, Z., and S. Zacks, 1966: Test procedures for possible changes in parameters of statistical distributions occurring at unknown time points. *Ann. Math. Stat.,* **37,** 1196–1210.

Lund, R. B., and I. V. Basawa, 2000: Recursive prediction and likelihood evaluation for periodic ARMA models. *J. Time Ser. Anal.,* **21,** 75–93.

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

——, H. Hurd, P. Bloomfield, and R. L. Smith, 1995: Climatological time series with periodic autocorrelations. *J. Climate,* **8,** 2787–2809.

——, Q. Shao, and I. V. Basawa, 2006: Parsimonious periodic time series modeling. *Aust. N. Z. J. Stat.,* **48,** 33–47.

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.

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

Percival, D. B., and D. A. Rothrock, 2005: "Eyeballing" trends in climate time series: A cautionary note. *J. Climate,* **18,** 886–891.

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.

Rhoades, D. A., and M. J. Salinger, 1993: Adjustment of temperature and rainfall records for site changes. *Int. J. Climatol.,* **13,** 899–913.

Solow, A. R., 1987: Testing for climate 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.

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: 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 2007: RHTest user manual. [Available online at http://cccma.seos.uvic.ca/ETCCDMI/RHTest/RHTestUserManual.doc.]