A Review and Comparison of Changepoint Detection Techniques for Climate Data

JAXK REEVES AND JIEN CHEN

Department of Statistics, The University of Georgia, Athens, Georgia

XIAOLAN L. WANG

Climate Research Division, Atmospheric Science and Technology Directorate, Science and Technology Branch, Environment Canada, Toronto, Ontario, Canada

ROBERT LUND

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

QIQI LU

Department of Mathematics and Statistics, Mississippi State University, Mississippi State, Mississippi

(Manuscript received 29 March 2006, in final form 25 September 2006)

ABSTRACT

This review article enumerates, categorizes, and compares many of the methods that have been proposed to detect undocumented changepoints in climate data series. The methods examined include the standard normal homogeneity (SNH) test, Wilcoxon's nonparametric test, two-phase regression (TPR) procedures, inhomogeneity tests, information criteria procedures, and various variants thereof. All of these methods have been proposed in the climate literature to detect undocumented changepoints, but heretofore there has been little formal comparison of the techniques on either real or simulated climate series. This study seeks to unify the topic, showing clearly the fundamental differences among the assumptions made by each procedure and providing guidelines for which procedures work best in different situations. It is shown that the common trend TPR and Sawa's Bayes criteria procedures seem optimal for most climate time series, whereas the SNH procedure and its nonparametric variant are probably best when trend and periodic effects can be diminished by using homogeneous reference series. Two applications to annual mean temperature series are given. Directions for future research are discussed.

1. Introduction

Changepoints are times of discontinuities in a time series that can be induced from changes in observation locations, equipment, measurement techniques, environmental changes, and so on. Inferences drawn from climatic series frequently depend on "continuity of the measurement process," that is, the lack of changepoints. For example, Easterling and Peterson (1995), Chen and Gupta (2000), Lu et al. (2005), Hanesiak and Wang (2005), and Wang (2006) note that linear trend estimates are trustworthy only when the series are homogeneous in time. A changepoint-free record is difficult to ensure; moreover, many changepoints occur without documentation. Before one studies trends, the relative homogeneity of the series should be assessed.

The World Meteorological Organization (WMO) Climate Program guidelines on climate metadata and homogenization (Llanso 2003) list at least 14 data homogenization assessment techniques, and many more methods have been suggested. Different homogenization techniques (methods/models) may be required for different climate elements or the same climate element on different time scales. With so many procedures, many of which yield conflicting conclusions when applied to the same series, a need has arisen for a careful discussion and comparison of these methods, along with some recommendations concerning which procedure(s) are best to use in commonly encountered situations. This paper attempts to survey, contrast/com-

Corresponding author address: Dr. Jaxk Reeves, Department of Statistics, The University of Georgia, Athens, GA 30602.
E-mail: jaxk@stat.uga.edu

DOI: 10.1175/JAM2493.1

TABLE 1. Changepoint detection procedures.

Name	Section
Standard normal homogeneity test	2a(1)
Nonparametric SNH test	2a(2)
Two-phase regression of Wang (2003)	2a(4)
TPR of Lund and Reeves (2002)	2a(3)
New generalized method of this study	3b
Method of Vincent (1998)	2b(1)
Akaike's information criteria	2b(2)
Sawa's Bayes criteria	2b(2)

pare, and modify eight prominent changepoint detection methods, as displayed in Table 1. Although these eight do not cover all methods proposed in the climate literature, they are very representative.

We are not the first to compare different changepoint procedures for climatologists. Easterling and Peterson (1995) conducted the first major attempt, although their introduced statistic did not always produce a clear conclusion. More recent review attempts have been made by Peterson et al. (1998), Ducré-Robitaille et al. (2003), Rodionov (2004), and DeGaetano (2006). A difficulty in attempting such a review lies in placing the methods on a common footing—they were often devised for different situations. In this article, we start with the simplest assumptions, so that the fundamental characteristics of each procedure can be understood and compared, before introducing various complications. We believe this will enable practitioners to understand better the similarities and differences among the different procedures.

To begin, we impose the following assumptions:

1) Under the null hypothesis of a homogeneous series (no changepoints), the series of interest $\{Y_t\}$ can be adequately described by a regression equation (also called a linear model) with error terms that are independent and identically distributed (IID) Gaussian (also called normal) random variables.
2) Over the period examined, $\{Y_t\}$ experiences at most one changepoint.
3) Except where noted, the procedures examined are being applied directly to $\{Y_t\}$. Discussion of reference series is contained in section 6a.

Although these assumptions are commonly made, they might be somewhat unrealistic in some climate applications. Relaxation of these assumptions and their effects on conclusions are discussed in section 6.

The remainder of this article proceeds as follows. Section 2 reviews many of the existing homogeneity assessment procedures, with section 3 constructing some modifications of these. Section 4 compares these procedures under various scenarios, and section 5 presents applications to two climate series. Section 6 discusses the effects (or lack thereof) of the assumptions on conclusions.

2. Review of existing methods

a. Simple changepoint methods

The methods described in this section are designed for cases in which the underlying regression response form is known (e.g., linear, quadratic, sinusoidal), aside from whether a changepoint exists. One should not expect a single method to perform optimally over all regression response forms; indeed, the tests presented here are most powerful when the assumed regression structure and normality of errors hold but may have less power under departures from these assumptions. Regression response form uncertainties are considered in sections 2b and 3.

1) SNH AND ITS VARIANTS

The standard normal homogeneity (SNH) test, which has roots in Hawkins (1977), was first applied to climatic data by Alexandersson (1986) and then was used by many others such as DeGaetano (1996) and Rosenbluth et al. (1997). Alexandersson (1986) originally scaled a target series by a reference series to create a series $\{q_t\}$ whose components were assumed to be normal. He standardized $\{q_t\}$ into a series of standardized anomalies $\{Z_t\}$ through $Z_t = (q_t - \overline{q})/s$, where $\overline{q}$ and s are the sample mean and sample standard deviation of $\{q_t\}$. The elements in $\{Z_t\}$ were treated as normally distributed, with the following null (H_0) and alternative (H_A) hypotheses:

$$H_0\colon Z_t \sim N(0, 1), \quad 1 \le t \le n, \quad \text{and} \tag{2.1}$$

$$H_A\colon \begin{cases} Z_t \sim N(\mu_1, 1), & 1 \le t \le c \\ Z_t \sim N(\mu_2, 1), & c+1 \le t \le n\,, \end{cases} \tag{2.2}$$

where $\mu_1 \ne \mu_2$, $1 \le c < n$, and the parameters μ_1, μ_2, and c are unknown. The symbol "$\sim N(\mu, \sigma^2)$" represents a normal distribution with mean μ and variance σ^2. Under these assumptions, Alexandersson (1986) derived the likelihood ratio statistic to assess H_0 versus H_A, that is, to determine the existence of a changepoint c. His test statistic was

$$T_0 = \max_{1 \le c < n} \{\tilde{T}_c\}, \quad \text{with} \quad \tilde{T}_c = c\overline{Z}_1^2 + (n-c)\overline{Z}_2^2, \tag{2.3}$$

where

$$\overline{Z}_1 = \frac{1}{c}\sum_{i=1}^{c} Z_i \quad \text{and} \quad \overline{Z}_2 = \frac{1}{n-c}\sum_{i=c+1}^{n} Z_i$$

denote sample means before and after time c.

A major assumption behind the procedure of Alexandersson (1986) is that the standardization [i.e., $Z_t = (q_t - \overline{q})/\sigma_q$, with the true standard deviation σ_q replaced by its estimate s] produces normal variables with a unit variance. This assumption is reasonable if the data are homogeneous (in which case s is a consistent estimator of σ_q). In the presence of a changepoint c, the (overall) sample standard deviation s is a biased and inconsistent estimator of σ_q, in which case σ_q should be estimated by the pooled sample standard deviation

$$s_p = \left[\frac{(c-1)s_1^2 + (n-c-1)s_2^2}{n-2}\right]^{1/2},$$

where s_1 and s_2 are the sample standard deviations of the two samples $\{q_1, \ldots, q_c\}$ and $\{q_{c+1}, \ldots, q_n\}$, respectively). However, because the existence and location of c are unknown, correctly estimating σ_q (by s or s_p) is not feasible, and therefore there is no guarantee that the variance of Z_t is close to unity, which invalidates the assumptions in (2.1) and (2.2).

To fix this inaccuracy, we propose a more precise variant of the SNH test, which can be applied to $\{Y_t\}$ when a good reference series is not available, provided that $\{Y_t\}$ is IID and Gaussian. In particular, the hypotheses in (2.1) and (2.2) become

$$H_0\colon Y_t \sim N(\mu, \sigma^2), \quad 1 \le t \le n, \quad \text{and} \tag{2.4}$$

$$H_A\colon \begin{cases} Y_t \sim N(\mu_1, \sigma^2), & 1 \le t \le c \\ Y_t \sim N(\mu_2, \sigma^2), & c+1 \le t \le n, \end{cases} \tag{2.5}$$

where c, μ, μ_1, μ_2, and σ^2 are all unknown. To assess the existence of a changepoint $1 \le c < n$, we derive the likelihood ratio statistic as in Alexandersson (1986) and obtain our test statistic $T_{\max}$, defined as

$$T_{\max} = \max_{1\le c<n} |T_c|, \quad \text{with} \quad T_c = \frac{\overline{Y}_1 - \overline{Y}_2}{s_p\sqrt{c^{-1} + (n-c)^{-1}}}, \tag{2.6}$$

where $\overline{Y}_1$ and $\overline{Y}_2$ denote the sample means of $\{Y_t\}$ before and after c and s_p is the pooled estimate of the standard deviation of $\{Y_t\}$.

Both T_0 in Alexandersson (1986) and our $T_{\max}$ are likelihood ratio statistics, which affords one statistical optimality. If c is known, $\tilde{T}_c$ in (2.3) is the likelihood ratio statistic for testing whether $\mu_1 \ne \mu_2$ when the variance of the data is unity and will be a χ^2 variable with 2 degrees of freedom under H_0 in (2.1). In a similar way, $|T_c|$ in (2.6) is the standard two-sample t-test statistic for equality of means when the variance σ^2 is unknown; this statistic has a t distribution with $n - 2$ degrees of freedom under H_0 in (2.4). In both cases, the value of c that maximizes $\tilde{T}_c$ or $|T_c|$ is declared the most probable changepoint position. Because $|T_c|$ depends on c, $T_{\max}$ is the maximum of t statistics over all "admissible" changepoint positions c, and similarly for T_0. Our version of the SNH procedure avoids the inaccurate standardization used in Alexandersson's procedure, and hence the statistic $T_{\max}$ will perform better. Henceforth, we consider only the modified version of the SNH procedure when referring to SNH.

TABLE 2. The 95% critical values for four simple changepoint models.

Method		SNH	NPW	XLW	LR
Statistic		$T^2_{\max}$	$W_{\max}$	$F_{\max}$	$F_{\max}$
n	25	10.36	7.08	11.67	7.37
	50	9.83	7.93	11.07	6.92
	75	9.94	8.38	11.06	6.88
	100	10.10	8.77	11.09	6.91
	200	10.17	9.25	11.21	7.01
	500	10.26	9.86	11.54	7.24
	1000	10.72	10.27	11.75	7.42
	2500	11.19	10.75	12.06	7.65

Because T_c^2 and $|T_c|$ provide the same changepoint information,

$$T^2_{\max} = \max_{1\le c<n} T_c^2$$

is equivalent to $T_{\max}$ in (2.6). For comparability with statistics to be introduced later, we henceforth use $T^2_{\max}$. The critical values of $T^2_{\max}$ for some values of n and levels of type-I error have been simulated under H_0 in (2.4) and are reported in Table 2. If $T^2_{\max}$ exceeds a rejection threshold set to a specified tolerance level (frequently 5%), then the test concludes that a changepoint exists, and the c that maximizes T_c^2 (and $|T_c|$) is the estimate of the changepoint time.

2) NONPARAMETRIC VARIANTS OF THE SNH PROCEDURE

The SNH procedure is also a likelihood ratio test when the model errors are IID and Gaussian. Normality of errors is a debatable assumption for many climatological series. To guard against a spurious "false changepoint" on the basis of one or two outliers, especially near the record boundaries ($t = 1$ or $t = n$), one might prefer a more robust procedure. Statisticians typically make parametric test procedures more robust (less sensitive to distributional departures from normality) by applying parametric procedures to the relative ranks of the data rather than to the observed values. If the sample size is large, such nonparametric tests may be only slightly less powerful than parametric tests and will typically provide better false-detection rates (type-I errors) and powers when the parametric assumptions are violated.

Because the $T_{\max}$ statistic is the maximum of $n - 1$ two-sample t statistics, the obvious nonparametric analog is the maximum of $n - 1$ Wilcoxon rank-sum statistics (or, equivalent, Mann–Whitney statistics). Thus, a nonparametric SNH procedure (hereinafter referred to as NPW) is one that detects a changepoint at time c when $W_{\max}$ is sufficiently large, where

$$W_{\max} = \max_{1 \le c < n} W_c \tag{2.7}$$

and W_c is the square of a normalized Wilcoxon rank-sum statistic for each fixed c:

$$W_c = 12 \frac{\left[\sum_{t=1}^{c} r_t - c(n+1)/2\right]^2}{c(n-c)(n+1)}, \tag{2.8}$$

where r_t is the rank of the tth element in the series (e.g., if X_{10} is the 32d largest value, then $r_{10} = 32$). Critical values for the $W_{\max}$ statistic under the null hypothesis of no changepoint could be obtained, as with $T_{\max}$, by simulation. Examples of these critical values are presented in Table 2. The time c at which W_c attains its maximum is the nonparametric estimator of the changepoint time.

The $W_{\max}$ statistic or variants thereof have been proposed by several climate authors, most notably Karl and Williams (1987), with a refinement by Ducré-Robitaille et al. (2003). Lanzante (1996) also bases his procedure on the Wilcoxon statistic, and Yonetani and McCabe (1994) use the Lepage modification of the Wilcoxon statistic. Pettit (1979) uses a Mann–Whitney statistic, which is equivalent to the Wilcoxon rank-sum statistic and should hence perform equivalently. Of these references, only Pettit gives a clear procedure for determining critical values. The others either assume that the location of the changepoint is approximately known from metadata or downplay multiple testing aspects (the many candidate times at which a changepoint could occur). If one ignores multiple testing aspects, any detection method will yield too many false changepoints.

Parameter estimation generally poses more difficulty in nonparametric cases. For example, if the $T_{\max}$ statistic suggests a changepoint at time c, the Gaussian maximum likelihood estimator of Δ would be

$$\hat{\Delta} = \frac{1}{n-c} \sum_{t=c+1}^{n} Y_t - \frac{1}{c} \sum_{t=1}^{c} Y_t. \tag{2.9}$$

The typical nonparametric estimator of the shift is

$$\hat{\Delta} = \text{median}\,\{Y_{t_2} - Y_{t_1}\},$$

where the median is taken over all pairs of indices (t_1, t_2) that satisfy $1 \le t_1 \le c$ and $c + 1 \le t_2 \le n$: a total of $c(n - c)$ pairs of differences. Although this testing procedure yields a relatively simple estimator for the shift Δ, the methods break down when trend features are included in the model (see also section 6b).

3) The two-phase regression model and recent revision

Hinkley (1969, 1971) proposed a two-phase regression (TPR) model with a changepoint at time c:

$$Yt = \begin{cases} \mu_1 + \beta_1 x_t + \varepsilon_t, & 1 \le t \le c \\ \mu_1 + \beta_2 x_t + \varepsilon_t, & c + 1 \le t \le n, \end{cases} \tag{2.10}$$

where predictor values $x_1 \le x_2 \le \ldots \le x_n$ are ordered and known; the errors $\{\varepsilon_t\}$ are zero mean, IID, and Gaussian with variance σ^2; and μ_1, μ_2, β_1, β_2, and c are unknown. Hinkley assumed continuity of the regression response at the changepoint c, which translates to the constraint $\mu_2 = \mu_1 + (\beta_1 - \beta_2)x_c$. Solow (1987) used (2.10) to test the homogeneity of a temperature series. Solow took $x_t = t$ to allow for a time trend [the continuity constraint is $\mu_2 = \mu_1 + (\beta_1 - \beta_2)c$]. To study changes in the trend, the null and alternative hypotheses are H_0: $\beta_1 = \beta_2$ and HA: $\beta_1 \neq \beta_2$.

A drawback of Solow's (1987) application lies with the continuity constraint. Although slow continuous trends can be caused by increasing urbanization (the so-called urban heat island effect), a deterioration of the instruments, a gradual change in the station environment and/or station instrumentation, or location changes typically induce a mean shift discontinuity into a series. In such cases, a continuity constraint is undesirable.

Instead of imposing continuity constraints, Lund and Reeves (2002) revised the TPR model to

$$Y_t = \begin{cases} \mu_1 + \beta_1 t + \varepsilon_t, & 1 \le t \le c \\ \mu_2 + \beta_2 t + \varepsilon_t, & c + 1 \le t \le n, \end{cases} \tag{2.11}$$

which allows both step-type ($\mu_1 \neq \mu_2$) and trend-type ($\beta_1 \neq \beta_2$) changepoints. The null and alternative hypotheses are

$$\begin{aligned} &H_0\colon \mu_1 = \mu_2 \quad \text{and} \quad \beta_1 = \beta_2 \\ &H_A\colon \mu_1 \neq \mu_2 \quad \text{and/or} \quad \beta_1 \neq \beta_2. \end{aligned} \tag{2.12}$$

If c were fixed and known, then H_0 could be tested by simply using the standard F test for model reduction:

$$F_c = \frac{(\text{SSE}_0 - \text{SSE}_A)/2}{\text{SSE}_A/(n-4)} \sim F_{2,n-4}, \tag{2.13}$$

where SSE_0 and SSE_A are the sum of squared errors computed under H_0 and H_A (with a changepoint at c), respectively. As suggested by the term $\sim F_{2,n-4}$, this statistic follows the F distribution with (2, $n - 4$) degrees of freedom under H_0. Large F_c values suggest H_A with a changepoint at time c.

When c is unknown, we use

$$F_{\max} = \max_{1 \leq c < n} F_c \tag{2.14}$$

as the test statistic. If $F_{\max}$ is too large to be attributed to chance variation, one concludes H_A, and the c maximizing F_c is taken as the estimate of the changepoint time. The true distribution of $F_{\max}$ under H_0 does not follow any well-known distribution type; this is in part due to serial correlation in the F_cs (across different values of c). Lund and Reeves (2002) presented simulated critical values of $F_{\max}$ that enable one to reach statistically valid conclusions about the existence of a changepoint. A drawback of the original method of Hinkley (1971) (subsequently used by others) is the inaccuracy of the percentiles of the $F_{\max}$ statistic under H_0. In particular, Lund and Reeves show that the $F_{3,n-4}$ null hypothesis distribution reported by Hinkley gives erroneously low critical values. Use of $F_{3,n-4}$ critical values in lieu of the correct values in Lund and Reeves (2002) will result in acceptance of many spurious changepoints. Turner et al. (2006) provide an application of the TPR model of Lund and Reeves (2002), hereinafter referred to as the LR method.

4) Two-phase regressions with a common trend

Wang (2003) noted that the LR model, while correcting the existing testing flaws, may be unrealistic in climate settings. This occurs because the most typical changepoint effect would shift mean series levels rather than affecting both the mean and the trend. Thus, a more realistic trend model is

$$Y_t = \begin{cases} \mu_1 + \beta t + \varepsilon_t, & 1 \leq t \leq c \\ \mu_2 + \beta t + \varepsilon_t, & c + 1 \leq t \leq n, \end{cases} \tag{2.15}$$

where the terms are as defined previously. The hypotheses of interest are H_0: $\mu_1 = \mu_2$ and H_A: $\mu_1 \neq \mu_2$.

If c were fixed and known, then H_0 could be tested by simply using

$$F_c = \frac{(\text{SSE}_0 - \text{SSE}_A)/1}{\text{SSE}_A/(n - 3)} \sim F_{1,n-3}, \tag{2.16}$$

where the terms are defined analogous to those above; when c is unknown, $F_{\max}$ in (2.14) is again used as the test statistic. Wang (2003) simulated critical values of this $F_{\max}$ test statistic for some common values of n, as in Table 2. Her method has been applied to time series of several climate variables (e.g., Vincent et al. 2005; Wang 2006).

This model (henceforth XLW) differs from the SNH model in that SNH assumes no trend and differs from the LR model in that it does not allow trend shifts at the changepoint time. Each procedure is based on an $F_{\max}$ statistic ($T^2_{\max}$ is also an $F_{\max}$ statistic because $T^2_c \sim F_{1,n-2}$ for fixed c) and is statistically most powerful if the hypothesized structural form is correct and the errors are Gaussian. All three procedures are very powerful at correctly identifying the changepoint if the relative shift size RSS = Δ/σ is large. However, the power to detect a changepoint decreases as RSS decreases, and use of an incorrectly specified model increases variability in the estimates of both c and Δ.

A generalization of the XLW method would replace the time factor t by a known "covariate series" $\{x_t\}$, possibly a reference series. The use of a homogeneous reference series with the same climate signal (i.e., trends and periodicity) as the target series has the potential to reduce greatly the model error variance and hence to increase power. Maronna and Yohai (1978) develop a bivariate changepoint detection procedure in this case, and Potter (1981) furthers the work. In cases in which $\{x_t\}$ is deterministic and known, the analysis proceeds as before with t replaced by x_t; in cases in which $\{x_t\}$ is random (such as a reference series for $\{Y_t\}$), the critical rejection percentiles are not purely a function of the series length n (as they are under the XLW procedure) but are also a function of the covariates $\{x_t\}$. Buishand (1984) discusses various modifications to these statistics that make critical values depend almost only on n. It is typically easier to incorporate reference series by modifying the response variable [e.g., using $\{D_t = Y_t - x_t\}$ or $\{Q_t = \ln(Y_t/x_t)\}$] than it is to include the $\{x_t\}$ directly as an explanatory factor (see section 6a).

b. *Hierarchical changepoint methods*

The methods in section 2a involve simple alternatives in that the regression response form of the alternative model is mathematically specified. The tests presented are optimal (most powerful) for detecting changepoints when the underlying regression response structure is parameterized correctly. In some cases, a specific functional regression response form is not known, and it could be worthwhile to consider a hierarchy of possible forms.

Table 3 presents a hierarchy of regression models. The SNH and NPW methods test reduction from model 3 to model 1, and the XLW and LR methods test reduction from model 4 to model 2 and from model 5 to model 2, respectively. Solow's model (Solow 1987) is a

TABLE 3. A hierarchy of models. Here, the changepoint indicator $I(t > c)$ is unity for $t > c$ and zero for $t \le c$.

Model 1 (M1)	$Y_t = \mu$	$+ \varepsilon_t$
Model 2 (M2)	$Y_t = \mu + \beta_1 t$	$+ \varepsilon_t$
Model 3 (M3)	$Y_t = \mu + \Delta I(t > c)$	$+ \varepsilon_t$
Model 4 (M4)	$Y_t = \mu + \beta_1 t + \Delta I(t > c)$	$+ \varepsilon_t$
Model 5 (M5)	$Y_t = \mu + \beta_1 t + \Delta I(t > c) + \beta_2 t I(t > c)$	$+ \varepsilon_t$

special case of model 5 with the constraint $\Delta = -\beta_2 c$ imposed. The modified Vincent's method, described below, and the penalized likelihood methods of section 2b(2) allow one to choose the best-fitting model in Table 3 while simultaneously assessing whether an undocumented changepoint exists.

1) MODIFIED VINCENT'S METHOD

The first climate changepoint detection methods to consider regression response form adequacy and changepoints simultaneously were introduced in Vincent (1998) and refined in Vincent and Gullett (1999). An application to Canadian temperatures can be found in Bonsal et al. (2001). Neglecting covariate terms, Vincent's procedure is a type of "forward regression" algorithm in that the significance of the nonchangepoint parameters in the regression model is assessed before (and after) a possible changepoint is introduced. In the end, the most parsimonious model is used to describe the data. Vincent includes models 1, 2, 3, and 5 of Table 3 in her hierarchy; we add model 4 to this hierarchy and call the procedure MLV.

Vincent's method is not easy to interpret or program. Figure 1 presents a flow diagram showing our interpretation of the sequence of tests needed for the MLV method. At each node of Fig. 1, one asks if the residuals from the fitted model are acceptable (white noise), with the scheme stopping if the answer is "yes," and continuing otherwise. Residual adequacy is assessed with the Durbin–Watson test and, if that is inconclusive, by sample autocorrelation assessment. If the MLV procedure opts for a changepoint model (models 3, 4, or 5), the changepoint location is estimated by the c that minimizes the SSE. However, the decision to proceed to the next node is based on residual adequacy tests and not on an F statistic. Although Vincent's procedure is less powerful than likelihood-based procedures, Vincent's idea of searching for an appropriate model within a regression hierarchy is of considerable merit and is the basis for other more recently developed detection procedures.

2) PENALIZED LIKELIHOOD CRITERIA

Choosing the best model within a hierarchy is frequently resolved by minimizing penalized likelihood statistics. Two common penalized likelihoods are Akaike's information criteria (AIC) and Sawa's Bayes criteria (SBC; sometimes also called BIC):

$$\mathrm{AIC}(p) = -2\ln(L) + 2p \quad \text{and} \tag{2.17}$$

$$\mathrm{SBC}(p) = -2\ln(L) + \ln(n)p, \tag{2.18}$$

where p is the number of parameters in the model under consideration, n is the series length, and L is the likelihood of the model being evaluated at the estimated model parameters. For linear models in the hierarchy of Table 3,

$$-2\ln(L) = n\ln(\mathrm{SSE}/n), \tag{2.19}$$

where SSE is the sum of squared errors for the model being fitted.

An AIC or SBC selector chooses the model with the minimum AIC or SBC statistic. The driving idea is to penalize for an excessive number of model parameters. Note that SBC penalizes more heavily than AIC and hence tends to yield simpler models. Both AIC and SBC were originally developed for assessing the significance of regression parameters. There is continuing debate in the statistical community about whether a changepoint parameter should be treated the same as other parameters; whether it should be penalized more heavily is still not clear.

For a known changepoint time c, it would be a simple matter to evaluate AIC or SBC for each of the five models in the hierarchy. However, in applying AIC or SBC criteria, $-2\ln(L)$ would need to be computed for models 1 and 2, as well as for each $c \in \{1, \ldots, n-1\}$ for models 3, 4, and 5. In parameter enumeration, models 3–5 are penalized one extra parameter for the maximization over c. Thus, if one searches for the maxima over models 1–5, there are $p = 1, 2, 3, 4,$ and 5 parameters, respectively, in the AIC and SBC statistics. These criteria are appealing in that they circumvent the simulations required to obtain critical values for the $T_{\max}$, $F_{\max}$, and $W_{\max}$ statistics, although they do not provide fixed levels (α) of statistical significance.

3. Method modifications

In cases in which the true regression response form is not clear, hierarchical alternative methods that consider a variety of regression forms have merit. We first modify the LR and XLW methods to allow for model selection.

a. *Modified LR and XLW procedures*

Lund and Reeves (2002) assess changepoint existence by testing the reduction of model 5 to model 2; in

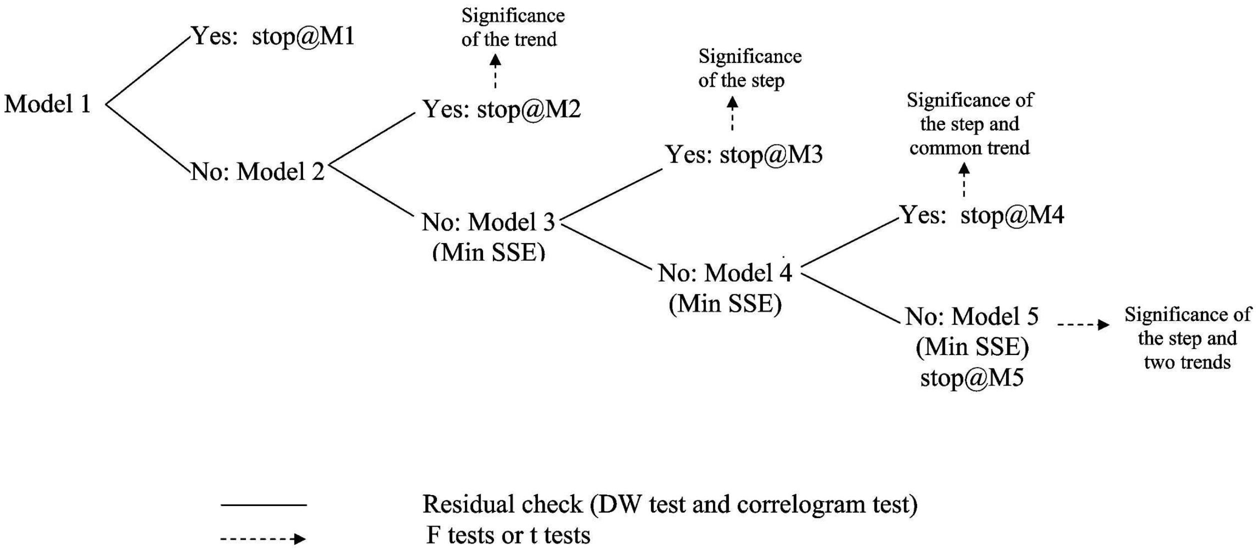
Model 1
Yes: stop@M1
No: Model 2
Significance of the trend
Yes: stop@M2
No: Model 3 (Min SSE)
Significance of the step
Yes: stop@M3
No: Model 4 (Min SSE)
Significance of the step and common trend
Yes: stop@M4
No: Model 5 (Min SSE) stop@M5
Significance of the step and two trends
Residual check (DW test and correlogram test)
F tests or t tests

FIG. 1. MLV method: "Yes" and "No" indicate whether the residuals are acceptable using the Durbin–Watson test.

a similar way, Wang (2003) tests the reduction of model 4 to model 2. If no changepoint is detected, both the LR and XLW procedures would select the trend model (model 2). However, if model 2 is deemed appropriate, one could further test whether the trend term β_1 should be included (i.e., testing H_0: $\beta_1 = 0$ against H_A: $\beta_1 \neq 0$). Thus, the best model might be model 1.

In a similar way, the original LR and XLW procedures can be modified to assess the need of other model parameters after c is estimated. For example, if model 4 is selected after applying XLW, one may still test whether to reduce to model 3. This is done with a conventional regression F test because c is now fixed at the position estimated by model 4. Such parsimony modifications are unrelated to our focus of changepoint detection, although these modifications can affect the power of detecting the "true" model given that the presence of a changepoint is correctly gauged.

b. A new generalized algorithm

Figure 2 shows a new generalized algorithm (GNL) for detecting a changepoint under regression response form uncertainty. This algorithm builds from the hierarchical methods in section 2b and the modified LR procedure. The modified LR and XLW procedures estimate the changepoint location and assess the changepoint's significance only on the first model fit. It is conceivable that estimation of the changepoint time could be confounded with model choice. That is, if the null and full models used for testing homogeneity are incorrect, the estimate of the changepoint time may not be accurate. For the modified LR and XLW procedures, one cannot revise the changepoint time estimate in model fits at later stages if the initial models are incorrectly specified. However, the new generalized algorithm allows the changepoint time to be reestimated as the procedure evolves.

In Fig. 2, a "C" beneath an arrow denotes a stage at which a test is performed between a changepoint model (models 5, 4, or 3) and a nonchangepoint model (models 2 or 1) over all possible changepoint times. An "x" beneath an arrow indicates testing a regression structure reduction between two changepoint models or two nonchangepoint models, with the changepoint position fixed. Each node in Fig. 2 asks whether reduction from a higher-level model to a lower-level model is statistically permissible, with the next fit (or termination of the algorithm) dependent on the answer. For example, if the test for reduction from model 5 to model 2 (denoted by "M5 → M2" with subscript C in Fig. 2) rejects model 2 and therefore an estimate, say, $\hat{c}_1$ of the changepoint is obtained under model 5, then the next question would be whether model 4, a simpler changepoint model, is more appropriate. This is addressed by testing reduction from model 5 to model 4 (denoted by "M5 → M4" with subscript x) with the changepoint fixed at $\hat{c}_1$. If reduction from model 5 to model 4 is permissible, the next node asks whether there is still a significant changepoint and what is its best estimate under model 4. The second question is then answered by testing reduction from model 4 to model 2 over all possible changepoint positions. Here, a new estimate, say, $\hat{c}_2$ of the changepoint time is obtained if the last test also rejects model 2.

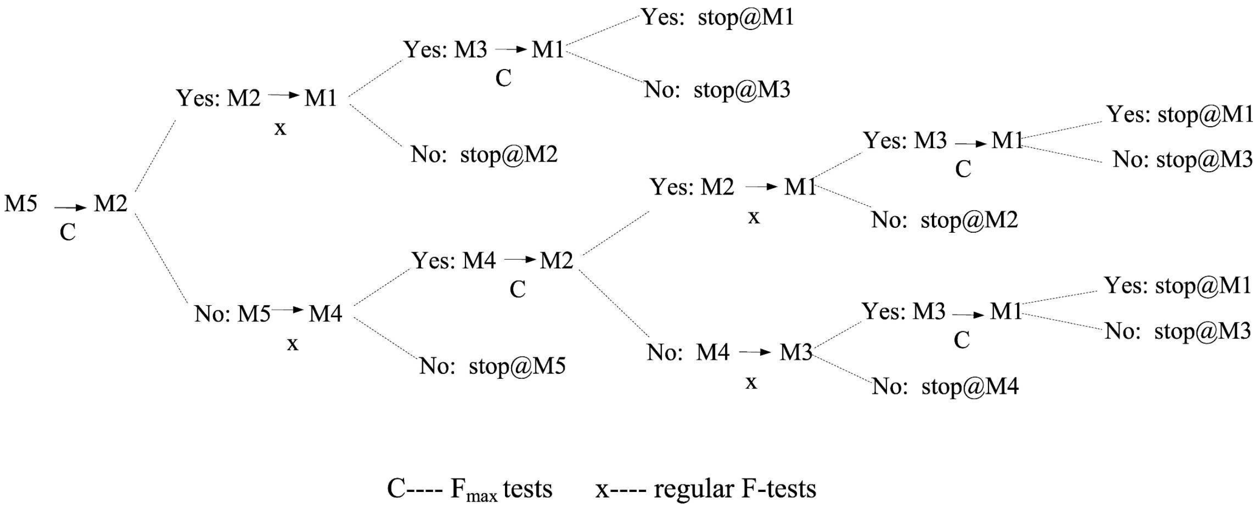

FIG. 2. GNL: "Yes" and "No" indicate whether the reduction from a higher-level model to a lower-level model is acceptable through the test.

Similar to backward-regression methods, each test is conducted at the desired α level. There are several opportunities in the GNL algorithm for the changepoint time to be reestimated, and this differs fundamentally from the modified XLW and LR procedures. The GNL and MLV procedures have similar merits, although MLV is not as comprehensive as GNL, and MLV is a "forward regression" procedure, whereas GNL uses "backward regression." The GNL procedure requires the type-I α-level critical values for the SNH, XLW, and LR tests, as well as conventional $F_{1,n-2}$, $F_{1,n-3}$, and $F_{2,n-4}$ critical values, for each sample size n.

4. Comparisons of methods

a. Simulation setup

Simulations were conducted to compare the eight methods shown in Table 4. All simulations were governed by one of the five models in Table 3. In all cases, without loss of generality, μ was taken as zero and the errors were generated as IID $N(0, 1)$ noise. When the changepoint models 3–5 were used, the parameters c, Δ, β_1, and β_2 were varied as $c = 50$, 65, or 80; $\Delta = 0.5$, 1, or 2 and β_1 or $\beta_2 = 0.005$, 0.01, or 0.02, as explained later. Each table entry is based on $M = 10\,000$ runs. For each run, a series $\{Y_t\}$ of length $n = 100$ was generated and subjected to the eight methods, with a type-I error rate $\alpha = 0.05$ used for the first six methods. The variation in c represents changepoints near the center, slightly off center, and far from the center, given that $n = 100$. No attempt was made to simulate "more extreme" values of c, because it is well known that most of these procedures have a higher-than-specified false-alarm rate for detecting changepoints near the boundary. The three choices for each of Δ, β_1, and β_2 represent low, moderate, and high variation effects, given the series length ($n = 100$) and the error standard deviation ($\sigma = 1$).

The XLW and LR critical values are taken from Wang (2003) and Lund and Reeves (2002), respectively. The critical values needed for the SNH and NPW methods were estimated from 1 million simulations under the null hypothesis. These four critical values, for $n = 100$ and $\alpha = 0.05$, are displayed in Table 2. Critical values for conventional F, Durbin–Watson, and correlogram tests were taken from standard tables. The AIC and SBC procedures have no critical values.

b. Power and fit statistics

The traditional way to compare changepoint detection algorithms (after ensuring a common type-I error rate) examines their power of detection given that a changepoint actually exists. This study considers four

TABLE 4. Changepoint detection procedures and associated statistics.

Code	Name	Statistic
SNH	Modified SNH test	$T^2_{\max}$
NPW	Nonparametric SNH test	$W_{\max}$
XLW	Modified Wang's TPR method	$F_{\max}$
LR	Modified Lund and Reeves TPR method	$F_{\max}$
GNL	New generalized method (Fig. 2)	(Multiple)
MLV	Modified Vincent's method (Fig. 1)	(Multiple)
AIC	Akaike's information criteria	AIC
SBC	Sawa's Bayes criteria	SBC

TABLE 5. Results of applying the procedures to 10 000 simulations under model 1 (no trend; no changepoint).

	General simulation results					Fit and power statistics			
Procedure	M1	M2	M3	M4	M5	RMP	CRM	CRC	CRB
SNH	9502	0	498	0	0	0.0927	9502	9502	9502
NPW	9509	0	491	0	0	0.1107	9509	9509	9509
XLW	8999	490	153	358	0	0.1029	8999	9489	8999
LR	9013	494	30	153	310	0.1040	9013	9507	9013
GNL	8858	495	190	147	310	0.1081	8858	9353	8858
MLV	9528	177	131	85	79	0.0873	9528	9705	9528
AIC	2899	126	2633	2084	2258	0.2427	2899	3025	2899
SBC	9104	229	476	143	48	0.1028	9104	9333	9104

measures of model adequacy—three power measures (CRM, CRC, and CRB) and one measure of fit (RMP). In particular, CRM is the probability of selecting the correct model, irrespective of estimated parameters, CRC is the probability of "closely estimating the changepoint time *c*," irrespective of models (by closely estimating *c* we mean that model 1 or model 2 is selected for series with no changepoint or that the estimated *c* is within ±3 of the true *c* value when a changepoint actually exists), CRB is the probability of *both* identifying the correct model *and* locating the changepoint (if one exists) within ±3 of the true location, and RMP is the average root-mean-squared-prediction error for procedure *j*:

$$\mathrm{RMP}(j) = \frac{1}{M}\sum_{m=1}^{M}\left\{\frac{\sum_{t=1}^{n}[\hat{Y}_{m,t}(j) - E(Y_t)]^2}{n - p_m(j)}\right\}^{1/2}, \quad (4.1)$$

where $\hat{Y}_{m,t}(j)$ is the predicted series value at time *t* using procedure *j* on the *m*th simulation run, $E(Y_t)$ is the true expected value of the series at time *t* under the simulated conditions, and $p_m(j)$ is the number of estimated parameters (including *c*) in the model selected by procedure *j* in the *m*th simulation (here, $M = 10\,000$ and $n = 100$). The RMP statistic quantifies separation between procedures even when the procedures themselves yield different models. In contrast, the three power measures are all dichotomous success/failure judgments of each simulation's correctness.

c. Simulation results for nonchangepoint models

Tables 5 and 6 summarize simulations for models 1 and 2. Here, the integer displayed in row *i* and column *j* of the left-hand part of the table displays the number of simulations for which procedure *i* chose model *j* as the best model over the 10 000 runs. The right-hand portion of the tables shows RMP and empirical counts of CRM, CRC, and CRB.

The Table 5 results are as expected when model 1 (no trend or shift) holds. For the SNH, NPW, and MLV methods, the correct model 1 is chosen in 95% of the simulations. For XLW and LR, there is a 95% chance that a nonchangepoint model (model 1 or model 2) is found and a 90% (≈95%^{2}) chance that model 1 is chosen after the subsequent test of model 2 versus model 1. The GNL method chooses model 1 with an 89% probability. The AIC and SBC procedures both overparameterize relative to the other six procedures. The penalty term used by SBC appears reasonable, with only 6.7% of the simulations yielding an incorrect changepoint model (models 3, 4, or 5). However, AIC is extremely underpenalized, because an incorrect model is selected in about 70% of the simulations. Among all eight methods, the MLV method has the best RMP of 0.087. From a practical viewpoint, all except AIC perform well under this null model (model 1) simulation.

Table 6 summarizes simulations in which model 2 holds, that is, with no shift but with the trend increasing from $\beta_1 = 0.005$ to $\beta_1 = 0.01$ to $\beta_1 = 0.02$, respectively. As the trend increases, the procedures diverge. The XLW, LR, and GNL methods still report a 5% chance of falsely choosing models 3, 4, or 5. However, whether the 95% of non-changepoint-detected simulations are assigned correctly to model 2 depends on the magnitude of β_1, with power increasing for all three methods as β_1 increases, and little difference between the three. The SNH and NPW methods decide between models 1 and 3. If the true trend is small, they typically choose the null model 1, whereas, as the trend increases, they opt for model 3, making their best step-function approximation to a linear increase. As β_1 increases, MLV eventually selects the correct model 2 but is much less powerful than XLW, LR, or GNL. The two penalized procedures again overparameterize, with AIC being much worse than SBC. As β_1 increases from 0 ($\beta_1 = 0$ is model 1), type-I errors of AIC and SBC increase from 70% to about 91% and from 7% to 32%, respectively (when β_1 is slightly larger than 1/*N*; not shown in Table 6), and then slowly drop to 0 as β_1 increases

Table 6. Results of applying the procedures to 10 000 simulations under model 2 [nonzero trend (β_1) but no changepoint].

		General simulation results					Fit and power statistics			
β_1	Procedure	M1	M2	M3	M4	M5	RMP	CRM	CRC	CRB
0.005	SNH	7954	0	2046	0	0	0.1970	0	7954	0
	NPW	7793	0	2207	0	0	0.2085	0	7793	0
	XLW	6592	2878	193	337	0	0.1774	2878	9470	2878
	LR	6617	2893	50	145	295	0.1782	2893	9510	2893
	GNL	6417	2895	252	141	295	0.1808	2895	9312	2895
	MLV	9224	491	128	83	74	0.1762	491	9715	491
	AIC	1447	413	3645	2054	2441	0.2689	413	1860	413
	SBC	6768	1612	1309	243	68	0.1944	1612	8380	1612
0.01	SNH	3410	0	6590	0	0	0.2833	0	3410	0
	NPW	3159	0	6841	0	0	0.2873	0	3159	0
	XLW	1771	7736	196	297	0	0.1620	7736	9507	7736
	LR	1760	7757	72	105	306	0.1631	7757	9517	7757
	GNL	1639	7759	191	105	306	0.1630	7759	9398	7759
	MLV	7991	1741	129	75	64	0.2778	1741	9732	1741
	AIC	155	982	3910	2162	2791	0.2783	982	1137	982
	SBC	1928	5206	2402	344	120	0.2065	5206	7134	5206
0.02	SNH	21	0	9979	0	0	0.3710	0	21	0
	NPW	18	0	9982	0	0	0.3731	0	18	0
	XLW	3	9489	133	375	0	0.1399	9489	9492	9489
	LR	3	9502	58	123	314	0.1406	9502	9505	9502
	GNL	3	9507	59	117	314	0.1405	9507	9510	9507
	MLV	1989	7717	76	149	69	0.2217	7717	9706	7717
	AIC	0	1829	1257	3416	3498	0.2840	1829	1829	1829
	SBC	4	8064	1142	633	157	0.1736	8064	8068	8064

further. Their rates of convergence to the true model are much slower than those of XLW, LR, and GNL, with SBC yielding a power of 80% (and AIC 18%) at $\beta_1 = 0.02$ (see column CRM). The RMP statistic reveals more variety. For a small trend ($\beta_1 = 0.005$), the MLV method is similar to XLW, LR, and GNL, with SNH, NPW, and SBC all displaying slightly greater RMP, and AIC again being much worse. For a moderate trend ($\beta_1 = 0.01$), XLW, LR, and GNL are about equally good, followed by SBC, with SNH, NPW, MLV, and AIC all considerably worse. Last, with the large trend ($\beta_1 = 0.02$), XLW, LR, and GNL continue to perform best, followed by SBC, MLV, and AIC, with SNH and NPW bringing up the rear.

d. *Simulation results for changepoint models*

1) Simulation results for model 3

Table 7 displays fit and power statistics for the eight procedures for the changepoint locations $c = 50$, 65, and 80 and shift $\Delta = 0.5$, 1.0, and 2.0 under model 3. SNH and NPW uniformly yield the largest power. For any procedure, as Δ increases, the power increases, RMP decreases, and performance differences of the procedures become evident. Grouping procedures with respect to the power statistics gives the ranking

$$(\text{SNH}, \text{NPW}) > \text{SBC} > \text{XLW} > (\text{LR}, \text{GNL}) > \text{MLV}, \tag{4.2}$$

where procedures in parentheses are deemed roughly equivalent. AIC is not listed here, because its relative performance varies considerably, being relatively good for small Δ but worsening as Δ increases. The RMP rankings are similar, with the relative differences better discriminated for moderate ($\Delta = 1.0$) than for small or large shifts. For fixed shifts Δ and procedures, the effect of the location of the changepoint c is mixed. For SNH and NPW, the power of detection decreases as the changepoint moves away from the center. The AIC and SBC powers are approximately constant, with a slight drop at $c = 80$. For XLW, LR, GNL, and MLV, the power of detecting the changepoint (for fixed Δ) increases slightly as c moves away from the center. For all procedures, RMP generally becomes slightly larger as c moves away from the center for a moderate shift $\Delta = 1.0$, but very little variation in RMP over c is noted for small or large shifts.

2) Simulation results for model 4

Table 8 displays simulation results under model 4. For these simulations, c was fixed at 50; β_1 and Δ were

TABLE 7. Fit and power statistics obtained from applying the procedures (PROC) to 10 000 simulations under model 3, for each combination of changepoint c and shift Δ.

		$\Delta = 0.5$				$\Delta = 1.0$				$\Delta = 2.0$			
c	PROC	RMP	CRM	CRC	CRB	RMP	CRM	CRC	CRB	RMP	CRM	CRC	CRB
50	SNH	0.2572	4512	1638	1638	0.2132	9841	6874	6874	0.1858	10 000	9715	9715
	NPW	0.2632	4822	1767	1767	0.2249	9855	6829	6829	0.1994	10 000	9730	9730
	XLW	0.2296	335	360	142	0.2729	2642	2482	2137	0.2036	9210	9315	8968
	LR	0.2319	118	251	59	0.2857	1638	1867	1353	0.2245	8269	8927	8049
	GNL	0.2316	257	280	93	0.2856	1627	1859	1341	0.2221	8263	8975	8094
	MLV	0.2629	227	131	83	0.3677	775	711	622	0.3534	4834	4989	4735
	AIC	0.2839	3937	2078	1272	0.2642	6024	5676	4365	0.2130	7663	9565	7488
	SBC	0.2474	2629	1136	1002	0.2345	7243	5583	5324	0.1944	9548	9595	9293
65	SNH	0.2532	4105	1544	1544	0.2148	9723	6820	6820	0.1864	10 000	9713	9713
	NPW	0.2601	4333	1657	1657	0.2269	9753	6833	6833	0.2045	10 000	9658	9658
	XLW	0.2374	420	403	192	0.2819	3122	2830	2495	0.2019	9282	9428	9030
	LR	0.2398	167	280	81	0.2988	1966	2092	1579	0.2221	8381	9039	8125
	GNL	0.2393	454	365	165	0.2971	2001	2146	1629	0.2183	8394	9122	8211
	MLV	0.2554	264	142	98	0.3774	943	826	722	0.3355	5796	5921	5662
	AIC	0.2836	3851	2031	1262	0.2633	5950	5633	4391	0.2159	7609	9481	7432
	SBC	0.2489	2699	1198	1066	0.2351	7512	5829	5575	0.1946	9551	9617	9295
80	SNH	0.2340	2798	1132	1132	0.2319	8846	6151	6151	0.1857	10 000	9625	9625
	NPW	0.2427	2986	1238	1238	0.2435	8844	6113	6113	0.2083	10 000	9449	9449
	XLW	0.2350	659	465	332	0.2929	3924	3341	3036	0.1978	9371	9457	9038
	LR	0.2377	233	293	137	0.3184	2144	2224	1709	0.2258	8008	8579	7745
	GNL	0.2383	744	524	366	0.3067	2595	2622	2100	0.2235	8013	8636	7800
	MLV	0.2261	302	163	129	0.3673	1155	998	893	0.3141	6707	6761	6511
	AIC	0.2833	3807	1979	1299	0.2705	5650	5269	4048	0.2190	7448	9131	7235
	SBC	0.2384	2338	1101	1017	0.2459	7440	5611	5382	0.1942	9527	9515	9189

varied as {0.005, 0.01, 0.02} and {0.5, 1.0, 2.0}, respectively. The SNH and NPW methods cannot select the correct model (model 4), and so they have CRM = CRB = 0. They do a reasonable job at detecting the correct location of c, as seen from the CRC statistic, although this is partially because the changepoint occurs at the midpoint ($c = 50$) in these simulations. For the other six procedures, the CRMs and CRCs increase as the trend β_1 or the shift Δ increases, that is, as one moves from upper left to lower right in Table 8. Some exceptions occur for AIC, LR, and GNL. Overall, there is no unique "best" procedure when model 4 holds. Using RMP or any of the power statistics, the following ordering seems to hold:

$$\text{XLW} > (\text{LR, GNL}) > \text{MLV}. \tag{4.3}$$

Neither SNH nor NPW can model trends and hence are typically worse than the above four procedures. The XLW method is generally the "winner," as one would expect under model 4. By the RMP criterion, SBC is generally comparable to XLW and better than AIC, whereas AIC is better than either by CRM. By the CRC criteria, SBC is always better than XLW, with the difference in power being fairly dramatic under some conditions (53% vs 24% for $\beta_1 = 0.005$ and $\Delta = 1.0$) and less so as the parameters increase (95% vs 94% in the lower-right panel of Table 8).

3) SIMULATION RESULTS FOR MODEL 5

Table 9 displays simulation results under model 5 (i.e., with trend β_1 before c and $\beta_1 + \beta_2$ afterward, along with a mean shift Δ at c). For these simulations, $\beta_1 = 0.01$ and $\Delta = 1.0$, and c and β_2 were varied as {50, 65, 80} and {.005, 0.01, 0.02}, respectively. The SNH, NPW, and XLW procedures cannot choose the correct model 5. For these three procedures, for fixed c, CRC increases as β_2 increases. For a fixed β_2, as c moves further from the center, the XLW method performs better in terms of RMP and CRC, but neither SNH nor NPW improves significantly. Among the five methods that could select model 5, MLV has a very disappointing performance, with its largest CRM being 3%; LR and GNL are very similar, even identical under some scenarios; AIC and SBC behave similarly under model 5, with AIC always better than SBC (which is usually slightly better than GNL/LR) by any criteria. Overall, AIC is the best procedure to use if one "knows" that model 5 is correct. Because AIC tends to overparameterize, it is not surprising that it wins in cases in which

Table 8. Fit and power statistics obtained from applying the procedures to 10 000 simulations under model 4 with $c = 50$, for each combination of trend β_1 and shift Δ.

		$\Delta = 0.5$				$\Delta = 2.0$				$\Delta = 1.0$			
β_1	PROC	RMP	CRM	CRC	CRB	RMP	CRM	CRC	CRB	RMP	CRM	CRC	CRB
0.005	SNH	0.2586	0	3296	0	0.2336	0	7127	0	0.2086	0	9691	0
	NPW	0.2661	0	3375	0	0.2444	0	7140	0	0.2217	0	9706	0
	XLW	0.2039	450	332	54	0.2750	517	2441	70	0.2223	764	9346	642
	LR	0.2067	192	231	43	0.2845	185	1870	63	0.2414	532	8965	483
	GNL	0.2062	191	239	43	0.2846	194	1849	66	0.2390	528	8990	481
	MLV	0.3301	91	138	9	0.3275	92	731	31	0.3625	142	5000	123
	AIC	0.2803	2108	2231	217	0.2764	1664	5466	377	0.2277	1955	9497	1797
	SBC	0.2273	452	1763	33	0.2499	398	5332	49	0.2180	588	9504	491
0.010	SNH	0.2818	0	4184	0	0.2756	0	7410	0	0.2554	0	9697	0
	NPW	0.2884	0	4197	0	0.2837	0	7465	0	0.2645	0	9708	0
	XLW	0.1984	387	358	4	0.2872	475	2519	28	0.2570	2458	9283	2249
	LR	0.2001	127	228	3	0.2934	134	1845	16	0.2735	1964	8883	1829
	GNL	0.2000	125	230	3	0.2930	143	1853	16	0.2715	1948	8935	1835
	MLV	0.3014	114	128	4	0.2983	159	686	52	0.3881	278	4909	242
	AIC	0.2930	2280	2187	82	0.2966	2713	5283	1039	0.2466	4007	9468	3793
	SBC	0.2278	453	1830	3	0.2809	472	4842	30	0.2565	2047	9452	1880
0.020	SNH	0.3919	0	4871	0	0.3896	0	7585	0	0.3697	0	9743	0
	NPW	0.3945	0	5074	0	0.3947	0	7710	0	0.3768	0	9769	0
	XLW	0.2043	421	380	24	0.3023	1762	2557	1002	0.2463	7701	9339	7371
	LR	0.2029	118	260	9	0.3051	842	1931	555	0.2624	6782	8970	6553
	GNL	0.2027	108	255	9	0.3052	814	1925	539	0.2608	6774	9002	6576
	MLV	0.1972	209	142	35	0.2943	308	676	143	0.4286	931	4968	848
	AIC	0.3014	4406	1723	950	0.2972	6128	5019	3616	0.2364	7554	9509	7301
	SBC	0.2321	697	1049	37	0.3099	2263	3494	1261	0.2525	7117	9477	6811

the simulations are run at the upper limit of the model hierarchy.

e. Conclusions

In summary, there is no unique best procedure by any criteria. Under the null hypothesis of no changepoints (models 1 or 2), all six of the α-level procedures do yield 5% chance of falsely detecting a changepoint. AIC behaves very poorly with respect to this criterion, with a 70% type-I error under model 1 and up to 91% type-I error under certain model-2 scenarios. SBC performs acceptably under model 1 (6.7% type-I error) but could have type-I error as high as 32% under certain model-2 scenarios.

For model 3, which many consider to be the most realistic when good reference series are available (so that trend and periodic effects can be diminished), the SNH and NPW procedures are best and similar if the errors are truly normally distributed. If gross outliers are present, NPW will be more powerful than SNH. If the changepoint occurs near the boundaries of the series, neither SNH nor NPW will be very powerful at detecting the changepoint. For models of more complexity than model 3, SNH and NPW are reasonable at detecting changepoints but are useless in estimating other parameters.

The XLW, LR, and GNL procedures behave similarly, especially if the true model complexity is model 3 or below. In such cases, XLW, because it is simpler, is less prone to overparameterization and thus performs slightly better. If model 4 truly holds, XLW is better than LR or GNL. If model 5 is correct, LR and GNL will eventually outperform XLW, but not until the trend difference parameter β_2 is very large. However, large trend changes (i.e., large β_2 values) do not appear to be very realistic for most climate series. The MLV procedure does not appear to perform very well, except under models 1 and 2, where it has the correct type-I error. The main deficiency of MLV is that it uses an inefficient test to detect changepoints.

Overall, one would not want to use AIC or MLV because of the high type-I error rate and low power, respectively. It also seems that LR and GNL are more complex than necessary in most situations. Thus, XLW, SNH, NPW, and SBC would be reasonable alternatives. Of these, SBC has a simplicity advantage in that no critical values are needed, although its high type-I error under model 2 is a drawback. Perhaps a modified SBC procedure that penalizes the changepoint parameter more heavily than the other parameters merits further study. If good reference series (see section 6 for definition) are available, the SNH and NPW procedures

TABLE 9. Fit and power statistics obtained from applying the procedures to 10 000 simulations under model 5 with $\beta_1 = 0.01$ and $\Delta = 1.0$, for each combination of changepoint c and trend change β_2.

		$\beta_2 = 0.005$				$\beta_2 = 0.010$				$\beta_2 = 0.020$			
c	PROC	RMP	CRM	CRC	CRB	RMP	CRM	CRC	CRB	RMP	CRM	CRC	CRB
50	SNH	0.2998	0	8360	0	0.3289	0	8942	0	0.4033	0	9494	0
	NPW	0.3073	0	8445	0	0.3353	0	9026	0	0.4071	0	9609	0
	XLW	0.3174	0	4662	0	0.3213	0	6559	0	0.3068	0	8865	0
	LR	0.3281	944	3769	435	0.3338	1391	5941	919	0.2908	3287	9055	2998
	GNL	0.3279	944	3780	435	0.3335	1391	5934	919	0.2915	3287	9024	2998
	MLV	0.3387	146	1352	92	0.3859	209	2331	151	0.4465	291	5384	269
	AIC	0.2899	2213	6810	918	0.2780	2609	8002	1762	0.2642	5009	9334	4705
	SBC	0.3103	222	5965	79	0.3141	452	7309	295	0.2993	2109	9176	1975
65	SNH	0.2922	0	8616	0	0.2978	0	9314	0	0.3166	0	9896	0
	NPW	0.3032	0	8517	0	0.3110	0	9211	0	0.3391	0	9772	0
	XLW	0.3130	0	6135	0	0.2897	0	8356	0	0.2827	0	9684	0
	LR	0.3316	1153	5194	398	0.3028	1290	7855	641	0.2789	2135	9621	1883
	GNL	0.3304	1153	5238	398	0.3024	1290	7848	641	0.2786	2135	9616	1883
	MLV	0.3729	145	2212	87	0.4039	147	4378	117	0.3573	261	8684	233
	AIC	0.2844	1834	7160	782	0.2706	1980	8444	1296	0.2579	3536	9634	3292
	SBC	0.3028	243	7010	74	0.2903	325	8538	153	0.2869	1019	9683	888
80	SNH	0.3201	0	8327	0	0.3108	0	9350	0	0.2978	0	9938	0
	NPW	0.3484	0	7387	0	0.3531	0	8262	0	0.3759	0	9030	0
	XLW	0.2986	0	7261	0	0.2657	0	9214	0	0.2454	0	9925	0
	LR	0.3252	2223	5561	340	0.2903	1852	7946	435	0.2574	1237	9579	847
	GNL	0.3236	2223	5613	340	0.2879	1852	7983	435	0.2563	1237	9577	847
	MLV	0.3996	147	3061	80	0.3966	170	6093	117	0.3042	192	9685	163
	AIC	0.2817	2495	7047	831	0.2624	2133	8482	1039	0.2384	2181	9638	1842
	SBC	0.2996	439	7450	82	0.2773	484	8970	126	0.2576	529	9788	363

are probably best, but for the many cases in which such reference series are unavailable or are of uncertain quality, the XLW and SBC procedures appear to be optimal.

5. Applications

The eight methods above are now applied to two annual average temperature series of interest. The first example comes from Tuscaloosa, Alabama, for which the complete series record extends over the 100-yr period of 1901–2000. This is a well-behaved series for which documentation is believed to be fairly good. There are three documented changepoints resulting from equipment changes or station relocations: June 1939, November 1956, and June 1987. Here, we will examine the 47-yr segment from 1940 to 1986 (Fig. 3a). Assuming that there are no other undocumented changepoints during this period, the procedures should detect only the known changepoint in 1956.

For the 1940–86 segment, all eight procedures detect the changepoint at $c = 1957$, one year later than actuality. All methods except AIC select model 3, with means of 17.789°C from 1940 to 1957 and 16.960°C from 1958 to 1986; that is, $\Delta = -0.829$°C. AIC also detects the changepoint in 1957 but prefers model 4. In this example, all eight methods give approximately the same answers and jibe with historical records. Of course, this is the best answer in retrospect, given that we knew approximately where all changepoints should occur and were able to segment so as to reduce consideration to at most one changepoint. In practice, if one is not sure that the series being examined has at most one changepoint, as demonstrated next, the methods can easily produce conflicting models, none of which may actually be correct.

Our second example examines a 90-yr (1911–2000) annual series of temperatures from Libby, Montana. Again, the documentation is believed to be very good and indicates only one changepoint, in 1938 (resulting from a change in latitude and elevation). A plot of this series is shown in Fig. 3b. All eight detection methods were applied to the series. The two simplest methods, SNH and NPW, use model 3 to detect a changepoint in 1930, with a shift of $\Delta = +0.818$°C between 1930 and 1931. The SBC, XLW, and MLV methods detect a changepoint in 1947, preferring model 4 with $\mu = 6.008$°C, $\beta_1 = +0.032$°C yr^{-1}, and a shift of $\Delta = -1.101$°C between 1947 and 1948. Both model fits are superimposed on the original series in Fig. 3b. The AIC, LR, and GNL methods, not shown in Fig. 3b, all detect the changepoint in 1945, using model 5. Thus, all

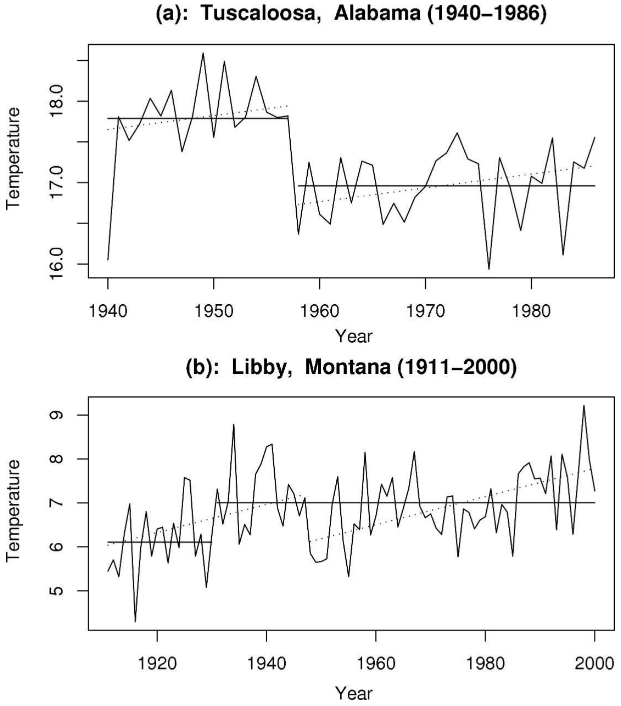

FIG. 3. Annual mean temperature series for (a) Tuscaloosa and (b) Libby, along with the regression fits.

eight procedures detect changepoints in the general neighborhood of the documented changepoint (c = 1938), but location estimates and the preferred model vary. Perhaps this discrepancy is attributable to the effects of other undocumented (but small in magnitude) changepoints elsewhere in the series, or perhaps none of the models examined sufficiently describes this series' behavior. With respect to the last explanation, the estimated lag-one autocorrelation for the series, after correcting for the changepoint, is 0.11, perhaps too seriously violating the IID error assumptions under which all eight procedures are predicated.

6. Other considerations and future directions

a. Reference series

As seen in section 4, the power of detection may be low, even for moderately large shifts, especially when the changepoint occurs near the record boundaries. Incorporating a good reference series, when available, can boost changepoint detection power. A good reference series $\{R_t\}$ should be homogeneous and highly correlated with the target series $\{Y_t\}$. Gauging $\{Y_t\}$ relative to $\{R_t\}$ in some fashion results in a new series [e.g., $\{D_t = Y_t - R_t\}$ or $\{L_t = \ln(Y_t/R_t)\}$] with much less variability than $\{Y_t\}$ itself and with features such as trends and periodicities removed or simplified. Then the previously mentioned procedures should yield markedly improved results when applied to the new series.

One must be certain that the reference series $\{R_t\}$ is changepoint free over the period of comparison. The use of a reference series that is not homogeneous and/or has different climate signals (trends and periodicities) would complicate the problem of changepoint detection/adjustment. For the Libby series in section 5, an attempt was made to find a reference series. Temperature series from 34 stations in Montana were examined, three of which were very highly correlated with the Libby series after subtracting monthly means: Fortine (r = 0.903), Kalispell (r = 0.906), and Saint Ignatius (r = 0.902). The average of these three series was used as a reference, and the eight detection procedures were applied to the annual differences: $D_t = Y_t - R_t$, for t = 1911–2000. None of these methods came close to detecting the known changepoint in 1938. Results were not appreciably better when any of the three individual series was used as a reference. This example illustrates a well-known deficiency of reference series; they do not work well unless the reference has no changepoints.

b. Departures from normality

Section 2a(2) shows how a parametric changepoint detection procedure can be made nonparametric by applying parametric methods to the relative ranks of the series. Nonparametric procedures are less sensitive to outlying and skewed data, making them good for hypothesis testing but sometimes inconvenient for parameter estimation. It is fortunate that, as noted in section 2a(2), the NPW procedure will yield simple parameter estimates for both c and Δ when no trend is present; that is, under the conditions of model 3. Nonparametric generalizations to trend-inclusive procedures such as XLW or LR could be easily devised, but estimating model parameters after a changepoint is detected by such methods is problematic. There is discussion of using Theil's estimator (a nonparametric slope estimator) in Lanzante (1996), but this method has not yet been fully developed. Nevertheless, if one suspects outliers, then it is wise to examine nonparametric and parametric changepoint procedures in tandem.

c. Periodicities and autocorrelated errors

The models considered here were developed for annual series. Many climate series, such as monthly and daily series, have periodic features. The procedures examined in this article should not be expected to work well for periodic data. In theory, correcting for a periodic mean is not too difficult, because one would effectively replace the mean in the previous models by a periodic mean with a specified (known) period.

Autocorrelation is another concern. The methods developed previously assume independent error terms. This assumption, while tenable for some annual climate series, is not realistic for daily or monthly series, where there is much empirical evidence of autocorrelation. Although the difficulties posed by periodicity and autocorrelation are distinct, it is pragmatic to attempt to account for them simultaneously. An article that does exactly this is Lund et al. (2007). One conclusion therein is that changepoint detection procedures developed for independent error series when positive autocorrelation is truly present will result in the detection of too many false changepoints.

d. Multiple changepoints

The procedures presented control the type-I error rate and are heavily dependent on the “at most one changepoint” assumption. If the series is long and undocumented changepoints are plausible, then more than one undocumented changepoint may be present. Many practitioners merely search for the most obvious changepoint, correct for that (if one is found), and then reapply the method to the corrected series. This can lead to erroneous adjustments, because the effects of the first changepoint are heavily biased when other unaccounted changepoints are present (Wang and Feng 2004). In the ideal case, all possible changepoint times should be identified jointly before their mean shift magnitudes are estimated.

Multiple changepoints are an active current area of statistical research. In climate settings, Wang and Feng (2004) introduce a semihierarchical splitting algorithm to identify multiple changepoints [Menne and Williams (2005) introduced a similar algorithm]. In this procedure, “prior” changepoint times are reassessed after additional changepoint times are located. Methods that impose a Bayesian prior on the number of possible changepoints and then fit the model conditional on these changepoint times show promise. Among the methods discussed in this article, the method most amenable to this approach is the SBC procedure, because it could easily be evaluated and penalized for any model and number of changepoints. In fact, the multiple-changepoint procedure suggested by Caussinus and Mestre (2004) is very similar to an SBC approach generalized to multiple changepoints.

Acknowledgments. The Climate Research Division of the Atmospheric Science and Technology Directorate of Environment Canada is acknowledged for supporting Jaxk Reeves and Jien Chen through Contracts KM040-04-5200 and KM040-05-0121. Robert Lund and QiQi Lu were supported in part by National Science Foundation Grant DMS 0304407.

REFERENCES

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

Bonsal, B. R., X. Zhang, L. A. Vincent, and W. D. Hogg, 2001: Characteristics of daily and extreme temperatures over Canada. *J. Climate,* **14,** 1959–1976.

Buishand, T. A., 1984: Tests for detecting a shift in the mean of a hydrological time series. *J. Hydrol.,* **73,** 51–69.

Caussinus, H., and O. Mestre, 2004: Detection and correction of artificial shifts in climate. *J. Roy. Stat. Soc.,* **C53,** 405–425.

Chen, J., and A. K. Gupta, 2000: *Parametric Statistical Change Point Analysis.* Birkhäuser, 240 pp.

DeGaetano, A. T., 1996: Recent trends in maximum and minimum temperature threshold exceedences in the northeastern United States. *J. Climate,* **9,** 1646–1660.

——, 2006: Attributes of several methods for detecting discontinuities in temperature series: Prospects for a hybrid homogenization procedure. *J. Climate,* **19,** 838–853.

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

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

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.

Karl, T. R., and C. N. Williams Jr., 1987: An approach to adjusting climatological time series for discontinuous inhomogeneities. *J. Climate Appl. Meteor.,* **26,** 1744–1763.

Lanzante, J. R., 1996: Resistant, robust, and non-parametric techniques for the analysis of climate data: Theory and examples, including applications to historical radiosonde station data. *Int. J. Climatol.,* **16,** 1197–1226.

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

Lu, Q., R. Lund, and L. Seymour, 2005: An update of U.S. temperature trends. *J. Climate,* **18,** 4906–4914.

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

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

Maronna, R., and V. J. Yohai, 1978: A bivariate test for the detection of a systematic changepoint in mean. *J. Amer. Stat. Assoc.,* **73,** 640–645.

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

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

Pettit, A. N., 1979: A non-parametric approach to the change-point problem. *Appl. Stat.,* **28,** 126–135.

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

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

Rosenbluth, B., H. A. Fuenzalida, and P. Aceituno, 1997: Recent temperature variations in southern South America. *Int. J. Climatol.,* **17,** 67–85.

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

Turner, J., T. A. Lachlan-Cope, S. Colwell, G. J. Marshell, and W. M. Connolley, 2006: Significant warming of the Antarctic winter troposphere. *Science,* **311,** 1914–1917.

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

——, and D. W. Gullett, 1999: Canadian historical and homogeneous temperature datasets for climate change analysis. *Int. J. Climatol.,* **19,** 1375–1388.

——, and Coauthors, 2005: Observed trends in indices of daily temperature extremes in South America 1960–2000. *J. Climate,* **18,** 5011–5023.

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

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

——, and Y. Feng, cited 2004: RHTest (0.95) user manual. [Available online at http://cccma.seos.uvic.ca/ETCCDMI/RHTest/RHTestUserManual.doc.]

Yonetani, T., and G. S. McCabe, 1994: Abrupt changes in regional temperature in the conterminous United States. *Climate Res.,* **4,** 12–23.