Detection of Undocumented Changepoints Using Multiple Test Statistics and Composite Reference Series

MATTHEW J. MENNE AND CLAUDE N. WILLIAMS JR.

NOAA/NESDIS/National Climatic Data Center, Asheville, North Carolina

(Manuscript received 27 May 2004, in final form 14 April 2005)

ABSTRACT

An evaluation of three hypothesis test statistics that are commonly used in the detection of undocumented changepoints is described. The goal of the evaluation was to determine whether the use of multiple tests could improve undocumented, artificial changepoint detection skill in climate series. The use of successive hypothesis testing is compared to optimal approaches, both of which are designed for situations in which multiple undocumented changepoints may be present. In addition, the importance of the form of the composite climate reference series is evaluated, particularly with regard to the impact of undocumented changepoints in the various component series that are used to calculate the composite.

In a comparison of single test changepoint detection skill, the composite reference series formulation is shown to be less important than the choice of the hypothesis test statistic, provided that the composite is calculated from the serially complete and homogeneous component series. However, each of the evaluated composite series is not equally susceptible to the presence of changepoints in its components, which may be erroneously attributed to the target series. Moreover, a reference formulation that is based on the averaging of the first-difference component series is susceptible to random walks when the composition of the component series changes through time (e.g., values are missing), and its use is, therefore, not recommended. When more than one test is required to reject the null hypothesis of no changepoint, the number of detected changepoints is reduced proportionately less than the number of false alarms in a wide variety of Monte Carlo simulations. Consequently, a consensus of hypothesis tests appears to improve undocumented changepoint detection skill, especially when reference series homogeneity is violated. A consensus of successive hypothesis tests using a semihierarchic splitting algorithm also compares favorably to optimal solutions, even when changepoints are not hierarchic.

1. Introduction

Climatic time series that are free of artificial changepoints are indispensable to the study of observed climate variability and change, especially at local and regional scales (Easterling et al. 1996). Unfortunately, few climate series of even modest historic length are characterized only by variations in weather and climate. Even minor changes in a meteorological station's environment or in observation practices can artificially alter the mean level of measurements and/or introduce a local trend (Conrad and Pollack 1962). In situ observation practice changes include instrument relocations or replacement, sensor drift from calibration, changes in land use/land cover surrounding the observing site, and changes to the daily observation schedule. The challenge of artificial changepoint detection and adjustment in climate series is reflected by the expansive literature on the subject. Peterson et al. (1998a) provide a review of many of the techniques that have been used or proposed in the climate literature. Techniques to evaluate documented risks of changepoints have been used (e.g., Karl and Williams 1987), in addition to those applied in the detection of unknown (undocumented) changepoints (e.g., Solow 1987; Easterling and Peterson 1995; Alexandersson and Moberg 1997; Vincent 1998; Lund and Reeves 2002). Archives or other knowledge of observational practice can be used to test for artificial shifts at the instant of known observation practice changes. Unfortunately, station histories (metadata) are often incomplete, and climate series may contain undocumented changepoints, even when relatively extensive metadata exist.

Corresponding author address: Matthew J. Menne, NOAA/National Climatic Data Center, 151 Patton Avenue, Asheville, NC 28801.
E-mail: Matthew.Menne@noaa.gov

In the absence of corroborating metadata, however, questions regarding the veracity of apparent undocumented changepoints can arise. This is especially true when the interest lies in the continuity of a single or small number of time series versus a collection of series used *en masse* to calculate, for example, the spatial mean across a large region (Easterling et al. 1996). Some questions are probably inevitable because a certain background rate of type I and type II errors is always present. Nevertheless, determining the appropriate sensitivity of an undocumented changepoint test can be an iterative process, and many "false" changepoints may be revealed if an inappropriate sensitivity level or test statistic is used (Lavielle 1998, Lund and Reeves 2002). Visual inspection of a time series can provide insight into possible changepoints, but such inspection becomes impractical when a large number of time series requires evaluation. Moreover, even with visual inspection, the presence of a nonclimatic changepoint may still be debatable (Lund and Reeves 2002), and the analyst has little recourse other than to speculate on its cause or lack thereof.

Given the necessity of testing for undocumented changepoints and requirements for automated detection in some circumstances (e.g., the reprocessing and/or update of large datasets), a comparison of the characteristics of some commonly used test statistics is described below. Rather than compare, for example, the percentage of simulated changepoints that are identified by various tests (see Ducré-Robitaille et al. 2003 for a recent comparison of eight methods), this comparison was undertaken to ascertain whether multiple tests can be combined to improve overall confidence in undocumented changepoint detection. Specifically, the goal was to evaluate to what degree various test statistics provide independent assessments of the presence of undocumented changepoints and their position in a series. The comparison between tests was likewise motivated by the desire to evaluate undocumented changepoint detection as a function of the method that was used to formulate a composite reference series against which a target (candidate) series is compared. Frequently, a difference or ratio series is formed between the target and reference series in order to differentiate artificial changepoints from those rooted in true climate change and variability. Changepoint detection skill was, therefore, evaluated using different formulations of composite reference series. In addition, because the test statistics that are commonly applied to climate series are strictly relevant to determining the likelihood of a single changepoint, the skill of detection was evaluated for series that contain multiple changepoints, including in the component series that are used to form a composite reference. In practice, multiple changepoints are commonly present in both the target climate series and in series from nearby locations used to estimate the background climate signal. Situations in which multiple undocumented changepoints occur in all series are particularly challenging. Therefore, the skill of successive hypothesis testing using multiple tests is compared to an alternative approach, which optimizes a statistic based on an exhaustive comparison of all possible changepoint number and position combinations.

A description of the test statistics used in the comparison is given in section 2. Methods used to detect multiple undocumented changepoints and the framework for evaluating changepoint detection skill are also described in section 2. Three alternative formations of composite reference series are discussed in section 3, as well as the simulation of groups of cross-correlated climate series. Changepoint detection results are presented in section 4. A discussion and concluding remarks are provided in section 5.

2. Changepoint tests and quantification of detection skill

Three test statistics that are commonly applied to climate series were used in the comparison. Ducré-Robitaille et al. (2003) found these statistics to be among the highest performing in terms of the combination of changepoints that are correctly identified and the number "falsely detected" in the series with multiple step changes. Thorough descriptions of the test statistics can be found in Alexandersson (1986), Vincent (1998), and Lund and Reeves (2002), so only brief descriptions are provided below. These tests can be used with or without comparison to observations from nearby stations. In practice, however, a reference series is commonly used and the exposed changepoints are relative nonhomogeneities (Conrad and Pollack 1962; Alexandersson and Moberg 1997). While each test statistic may be used to detect a change in slope (trend) as well as a change in mean, here changes in mean level only were considered. Lund and Reeves (2002) note that step- and trend-type changes are difficult to unconfound in general. Wang (2003) discusses the potential of confounding artificial changepoints and those that are associated with true periodic variations in a climate series. That risk should be alleviated with the use of a reference series, provided that it and the target series are characterized by similar true variations. Nevertheless, the automated, skillful detection of local climate trends remains a difficult problem.

The likelihood ratio test for a shift in mean described by Hawkins (1977) and Alexandersson (1986) involves the comparison of the means of adjacent segments that

form a series $\{Y_t\}$ of length n in its standardized form. The series $\{Y_t\}$ may be either a raw climate series or a sequence of differences or ratios formed with a reference series. Assuming that $\{Y_t\}$ is normally distributed, a single shift in the level of the standardized series $\{z_t\}$ is determined using the null hypothesis H_o and alternative hypothesis H_a, given by

$$H_o: \quad z_t \rightarrow N(0, 1), \qquad t = 1, n$$
$$H_a: \left\langle \begin{matrix} z_t \rightarrow N(\mu_1, 1), & t = 1, c \\ z_t \rightarrow N(\mu_2, 1), & t = c + 1, n \end{matrix} \right\rangle.$$

If H_o is rejected in favor of H_a, the implication is that there has been a shift in the level of the z series. With the sample means that are used as the maximum likelihood estimators for the means before ($\bar{z}_1$) and after ($\bar{z}_2$) all possible instances of shift, the test statistic can be calculated as (Hawkins 1977; Alexandersson 1986)

$$T_c = c\bar{z}_1^2 + (n - c)\bar{z}_2^2. \tag{1}$$

Percentiles of T_c are generated via Monte Carlo simulations of z under the null hypothesis, recording the maximum T_c value for each realization as

$$T_{\max} = \max_{1 \le c < n} T_c = \max_{1 \le c < n} [c\bar{z}_1^2 + (n - c)c\bar{z}_2^2]. \tag{2}$$

Here H_o is rejected when T_c in a series exceeds the chosen percentile of $T_{\max}$ for one or more values of c, the instant of the change (defined here as the last value at the former level). Alexandersson and Moberg (1997) discuss how the likelihood of a change in trend can be similarly obtained using a likelihood ratio test. Potter (1981) describes a different version of the likelihood ratio test in which comparison to a reference is implicit to the test statistic (see also Maronna and Yohai 1978).

The formulation for a simple two-phase regression describing a series $\{Y_t\}$ is given by (Lund and Reeves 2002)

$$Y_t = \begin{Bmatrix} \mu_1 + \alpha_1 t + \varepsilon_t, & 1 \le t \le c \\ \mu_2 + \alpha_2 t + \varepsilon_t, & c < t \le n \end{Bmatrix}. \tag{3}$$

Under the null hypothesis of no changepoint, the two phases of the regression should be statistically equivalent. In that case, both the difference in means ($\mu_1 - \mu_2$) and slope ($\alpha_1 - \alpha_2$) should be close to zero for each $c \in \{1, \ldots, n\}$, and a single phase of the regression would be justified because $\mu_1 \approx \mu_2 \approx \mu_{\text{RED}}$ and $\alpha_1 \approx \alpha_2 \approx \alpha_{\text{RED}}$. The subscript "RED" refers to a single-phase or "reduced" model. To evaluate the null hypothesis of no changepoint versus the alternative hypothesis of an undocumented changepoint, an F statistic is calculated at each position c in the time series as

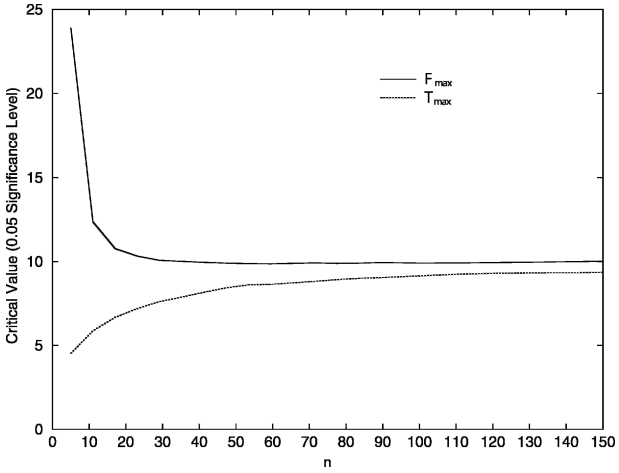

FIG. 1. The $F_{\max}$ and $T_{\max}$ critical values (0.05 significance level); $F_{\max}$ critical values based on the two-phase model given as in (6).

$$F_c = \frac{(\text{SSE}_{\text{RED}} - \text{SSE}_{\text{FULL}})/2}{(\text{SSE}_{\text{FULL}}/(n - 4))}, \tag{4}$$

where SSE_{FULL} refers to the sum of the squared errors about each of the two phases (the "full" model). As with a $T_{\max}$ test statistic, percentiles of the F statistic are obtained via simulations under the null hypothesis (Lund and Reeves 2002), in this case, recording the maximum F_c value in each series of F_cs,

$$F_{\max} = \max_{1 \le c \le n} F_c. \tag{5}$$

The null hypothesis of a "one phase" time series is rejected when the magnitude of $F_{\max}$ is greater than the chosen percentile (significance level).

If there is reasonable confidence that there is no trend in a time series, then estimation of the slope parameter α can be eliminated and the two-phase model for $\{Y_t\}$ becomes (Lund and Reeves 2002)

$$Y_t = \begin{Bmatrix} \mu_1 + \varepsilon_t, & 1 \le t \le c \\ \mu_2 + \varepsilon_t, & c < t \le n \end{Bmatrix}, \tag{6}$$

and the F statistic will have 1 numerator degree of freedom and $n - 2$ denominator degrees of freedom. If (6) is used, the two-phase regression test is equivalent to the likelihood ratio test; however, while a series of F_cs based on (6) will be similar to a series of T_cs using (1), critical values depend on which form of test statistic is used, as shown in Fig. 1. Unlike the $F_{1,n-2}$ and t statistic, which can be appropriate for evaluating the likelihood of a shift at the instant of a known risk of changepoint (Lund and Reeves 2002), $F_{\max}$ for model (6) is not the

square of a $T_{\max}$ statistic, and large differences in critical values exist for smaller sample sizes (n).

a. Detection of multiple changepoints

The presence of multiple breaks in a series can complicate the interpretation of these test statistics. When there are K segments to a series (or $K - 1$ changepoints), the time series may be treated as

$$Y_t = \mu_k + \varepsilon_t, \qquad c_{k-1} + 1 \le t \le c_k, \qquad k = 1, K, \quad (7)$$

assuming, as in (6), that it is *piecewise* stationary and $c_0 = 1$ and $C_K = n$. The solution to (7) frequently has been based on successive hypothesis testing using a hierarchic binary segmentation of the series (Hawkins 2001). In this approach, a series is split at the location where the hypothesis test statistic reaches a maximum, provided that its critical value is exceeded. Subsequences on either side of the split are likewise evaluated, and the process is repeated recursively until either the magnitude of the statistic does not exceed the chosen significance level in the remaining subsequences or the sample size in a segment is too small to test. This kind of solution is called "greedy" because changepoints are selected to maximize the separation between segments at each split, as opposed to evaluating all possible changepoint combinations iteratively to identify the optimal multiway split. The solution is hierarchic because it will reliably converge to the optimal solution only when the true changepoints are hierarchic, which may not be the case (Hawkins 2001).

An optimization algorithm also may be used to solve (7) by, for example, minimizing a penalized contrast statistic, the pooled residual sum of squares about each kth segment. The penalized contrast function in Lavielle's (1998) approach takes the form

$$U = \sum_{k=1}^{K} \sum_{t=c_{k-1}+1}^{c_k} [Y_t - \hat{\mu}_k]^2 + \beta(K - 1), \quad (8)$$

where $\beta = 2\alpha\sigma_\varepsilon^2$. The first term on the right-hand side of (8) measures the fidelity of the model to the observations $\{Y\}$, while the second term, the penalty function, is proportional to the number of changepoints. The estimated number of segments will be, in this case, the greatest K with a p value that is larger than α, the configuration of which is determined by minimizing U. An optimal global solution, therefore, requires evaluation of the large number of possible changepoint number and position combinations (a total of 2^{n-1}), for which dynamic programming can be used to reduce computational complexity (Lavielle 1998; Hawkins 2001).

Because the number of artificial shifts in a climate series is generally unknown, an optimal global solution will likely require calibration in order to avoid revealing too many "unimportant" or "false" changepoints (Lavielle 1998). The nature of the jumps that are identified in a series is calibrated via a penalty function like β (see also Akaike 1974; Schwarz 1978; Caussinus and Mestre 2004). Ideally, the penalty function should set the desired balance between the probability (power) of detection and probability of false detection. A solution using a relatively large penalty function will expose only the more important "jumps," but will overlook others. On the other hand, a small penalty function may reveal too many changes that are caused only by chance variation in the time series. Consequently, the best choice of the penalty function may not be obvious, but could be selected by a specialist with experience using the data. We used a very small p value (0.000 01) to solve (8) because too many changes are detected with a larger value (M. Lavielle 2005, personal communication). Nevertheless, the choice will likely require some level of intervention, ideally for each series tested (Lavielle 1998; Caussinus and Mestre 2004).

For successive hypothesis testing, we used a semihierarchic splitting algorithm to compare hypothesis testing to optimal solutions. In the semihierarchic algorithm, each splitting step is followed by a merging step to test whether a split chosen at an earlier stage has lost its importance after subsequent break points are identified (Hawkins 1976). At each splitting step, H_o is evaluated separately for all subsequences that occur between the apparent changepoints identified up to that stage. The subsequences are defined as 1 to c_1, $c_1 + 1$ to c_2, etc., up to $c_{K-1} + 1$ to n. If H_o is rejected in any subsequence, that segment is split and K is incremented. In the merging step that follows each splitting step, H_o is evaluated for all subsequences that include only one of the $K - 1$ apparent changepoints. In this case, the segments are defined from 1 to c_2, from $c_1 + 1$ to c_3, up to $c_{K-2} + 1$ to n. If H_o is not rejected in one of these subsequences, the apparent changepoint that is contained therein is removed and K is decremented. The process ends when no subsequence is split and no subsequences are merged on a pass through the full sequence. Although an improvement over strictly hierarchic solutions, this algorithm may not always converge to an optimal solution when K is greater than two and the changepoints are not hierarchic (Hawkins 2001) and/or occur close in time. In such circumstances, an optimal approach should have a higher power of detection.

b. Quantification of detection skill

The general framework in which forecast skill is quantified in light of the joint probability distributions

TABLE 1. Contingency table for the detection of undocumented changepoints. The null hypothesis for each test is series homogeneity (no changepoint).

	Changepoint occurred		
Changepoint detected	Yes	No	Total
Yes	a (f_1, o_1) (hit/correct rejection of H_o)	b (f_1, o_0) (false alarm/false positive/type I)	$a + b$
No	c (f_0, o_1) (miss/false negative/type II)	d (f_0, o_0) (correct acceptance of H_o)	$c + d$
Total	$a + c$	$b + d$	$(a + b + c + d) = n$

of forecasts f and observations o (e.g., Murphy and Winkler 1987) also may be applied to hypothesis testing (Stephenson 2000). In this case, "forecast" refers to the rejection or acceptance of the null hypothesis of homogeneity at each position in a series. "Observation" refers to the true, known occurrence or nonoccurrence of a simulated changepoint. The possible joint outcomes of changepoint detection (f) and occurrence (o) are represented as

$$f = \begin{matrix} 1, H_o \text{ rejected} \\ 0, H_o \text{ accepted} \end{matrix} \qquad o = \begin{matrix} 1, \text{change point present} \\ 0, \text{change point not present} \end{matrix}.$$

Measures that quantify various aspects of the joint frequency distribution of f and o then can be calculated using a 2×2 contingency table, containing counts of the four possible outcomes as shown in Table 1. The rate of type I (reject null hypothesis when it is true: a "false alarm" or "false positive") and type II (fail to reject null hypothesis when it is false: a "miss" or "false negative") errors can be calculated for each test statistic individually and for the "consensus" of multiple tests. The hit rate H measures the ratio of correctly classified changepoints to the total number of changepoints and is known as the *sensitivity*. Here, H and its counterpart, the false alarm rate F, are calculated as

$$H = \frac{a}{a + c} = \text{"the probability (power) of detection"} \quad (9)$$

$$F = \frac{b}{b + d} = \text{"the probability of false detection"}, \quad (10)$$

where $a = f_1, o_1$ (hit), $b = f_1, o_0$ (false alarm), $c = f_0, o_1$ (miss), and $d = f_0, o_1$ (correct acceptance of H_o). The term hit rate is sometimes used to refer to the quantity $(a + d)/n$ (the "percent correct") while the false alarm rate or ratio (FAR), or false positive rate, will sometimes (e.g., Wilks 1995) refer to the quantity

$$\text{FAR} = \frac{b}{(a + b)}. \quad (11)$$

A third quantity, bias, is calculated as

$$B = \frac{(a + b)}{(a + c)}, \quad (12)$$

which represents the ratio of the number of H_o rejections to the number of simulated changepoints. When the base rate of event occurrence is much lower than the rate of nonoccurrence, skill scores like the Heidke Skill Score (HSS) are commonly used to adjust for the large number of correctly predicted nonevents. Because changepoints do not occur in a majority of years (or months), that is, the quantity d in Table 1 is much larger than $a + c$, a changepoint reasonably can be treated as a rare event. The HSS compares the proportion that are correct to a random no-skill forecast with the same base rate of event occurrence (Doswell et al. 1990; Stephenson 2000), and can be calculated as

$$\text{HSS} = \frac{2(ad - bc)}{(a + c)(c + d) + (a + b)(b + d)}. \quad (13)$$

The most likely position of a changepoint is where the test statistic reaches a maximum (or minimum in the case of the simple sum of squares). A hit is tallied when this maximum (or minimum) coincides with the true position of a simulated changepoint. If the test statistic exceeds the critical value, but the maximum (minimum) is not coincident with a simulated changepoint, it is counted as a false alarm. When the null hypothesis is not rejected in a sequence that contains a changepoint, a miss is recorded. As shown in Fig. 2, the time series of a test statistic may exceed the critical value across a range of locations around the true position of the changepoint, and the highest value is subject to some chance variation. Consequently, it may be desirable to qualify a rejection of the null hypothesis as a "hit" when the maximum in the test statistic occurs within one to a few time steps of its true position. Here, coincidence between tests was defined as ± 2 time steps.

3. Reference series formulation and simulation of climate series

A good choice of reference series should capture the background climate signal that is common to the target and surrounding station series. Under the assumption that the composite reference series is at least approximately homogeneous, when a changepoint is revealed in a difference or a ratio series is formed between the target and its reference, the conclusion is that

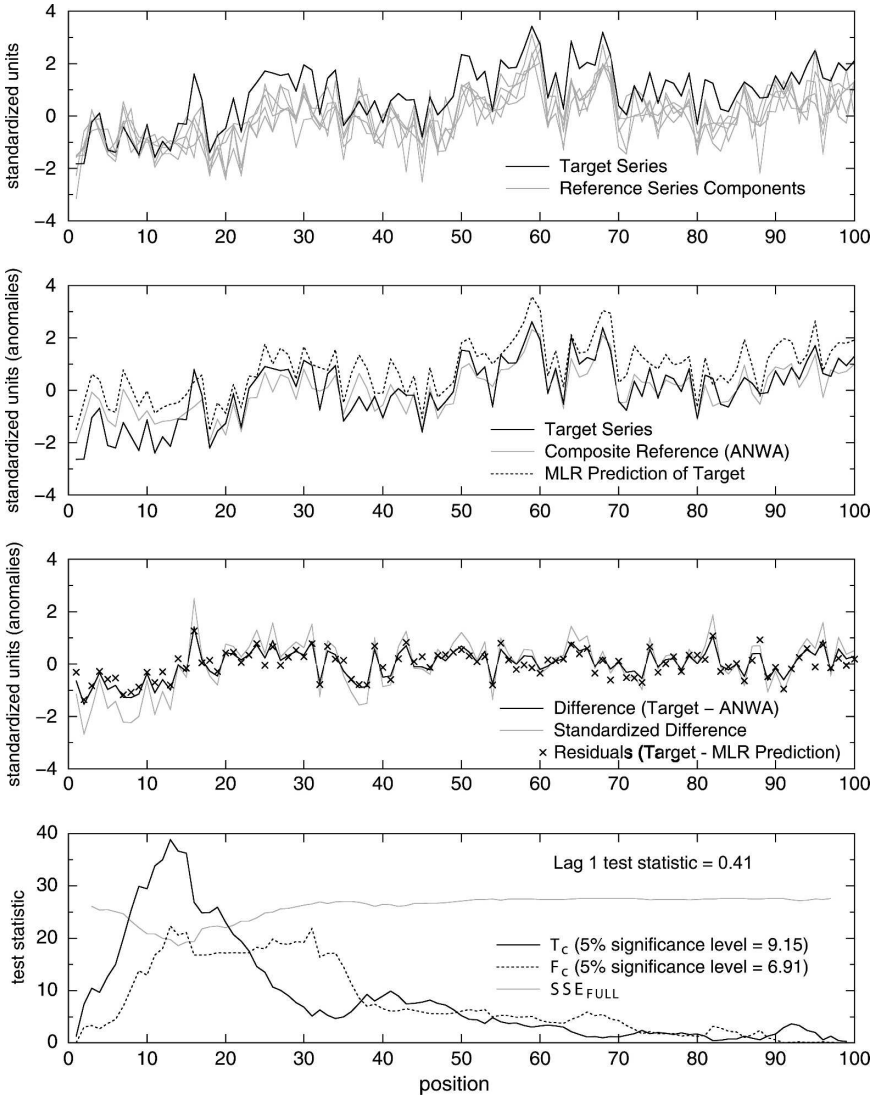

FIG. 2. Example of Monte Carlo simulation with one changepoint in target series (in bold) and none in the composite reference series components (case 2a). (a) Target (candidate) series and five correlated composite reference component ("neighboring") series; (b) target and ANWA composite reference series (correlation-weighed average of five reference component series); (c) difference between candidate and composite reference series and residuals from multiple linear regression prediction of target series; (d) series of T_c and F_c statistics and residual sum of squares (SSE_{FULL}) for two-phase model.

an artificial shift is present in the target series. However, an artificial shift in one or more of the m nearby stations series that is used to form the reference may carry through to the series of differences or ratios and the assumption of reference series homogeneity can be invalid. In that case, a changepoint in the reference may be erroneously attributed to the target series (see also Szentimrey 1999).

To quantify the potential for such impacts, three different composite reference series were calculated using component series with and without simulated changepoints. In one case, changepoints were identified in a difference series $\{Y_t\}$, formed between observations at the target station and the average from nearby stations, calculated according to Alexandersson and Moberg (1997) as

$$Y_t = (y_t - \bar{y}) - \frac{\sum_{j=1}^{m} \rho_j^2 (x_{jt} - \bar{x}_j)}{\sum_{j=1}^{m} \rho_j^2}, \qquad t = 1, n, \quad (14)$$

TABLE 2. Matrix of possible composite reference series formulation and test statistic pairings.

Test statistic	Reference series formulation		
	MLR	ANWA	FDWA
Lag 1 test (lag 1)	MLR–lag 1	ANWA–lag 1	FDWA–lag 1
Likelihood ratio test (T_{max})	MLR–T_{max}	ANWA–T_{max}	FDWA–T_{max}
Two-phase regression test (F_{max})	MLR–F_{max}	ANWA–F_{max}	FDWA–F_{max}

where y_t and x_{jt} are monthly or annual temperatures for the candidate and each of m neighboring stations, respectively, and ρ_j represents the correlation coefficients between observations at the candidate station and the jth instance of m surrounding stations. The quantities with an overbar may be calendar monthly (e.g., Menne and Duchon 2002) or annual means over a series of length n. We refer to the term to the right of the minus sign in (14) as the anomaly-weighted average (ANWA) composite reference series.

Peterson and Easterling (1994) suggest using first-difference-filtered values of each series (i.e., where $Y'_t = y_t - y_{t-1}$) to calculate each ρ_j to reduce the chance of making poor estimates of the magnitude of correlation between the candidate and neighboring series when one or both series contain a shift or trend. In their method, the first-difference correlation coefficients are used as weights to form an *average* first-difference series from the m nearby station series. We refer to this formulation as the first-difference-weighted average (FDWA) composite reference series. Using common weights and serially complete (i.e., no missing values) reference series components, the ANWA and FDWA composite reference series are exactly correlated and differ only by the offset that is used to convert the FDWA series back to raw averages.

Vincent (1998) does not use a reference series per se. Rather, the residuals e_t from a multiple linear regression (MLR) equation, using observations from neighboring stations to estimate values at the candidate station, are examined for evidence of changepoints using either the Durbin–Watson or lag-1 test for serial correlation, e_t ($e_t = Y_t - \hat{Y}_t$). In the case of identifying a step change or artificial trend, the null hypothesis of serial independence in the residuals is evaluated against the alternative hypothesis that they are consistent with a first-order autoregressive process (Wilks 1995; Durbin and Watson 1950, 1951, 1971). A step or trend in the target series will tend to cause serial correlation in the regression residuals. When the value of the test statistic is sufficient to reject the null hypothesis of uncorrelated residuals, a binary variable is introduced iteratively at each series position to separate the multiple linear regression estimates into all combinations of two phases. The changepoint position, which minimizes the pooled residual sum of squares (SSE_{FULL}) about the two phases, is considered to be the most likely break point. The relative performance of these three formulations of reference series was evaluated by controlling for the test statistic whereby each reference was paired with each test as shown in Table 2.

a. Simulation of climate series

Simulations of temperature anomaly series were produced by generating large numbers (1000 in each of several cases) of Box–Jenkins first-order autoregressive [AR(1)] model realizations, given by

$$x_{t+1} - \mu = \phi(x_t - \mu) + \varepsilon_{t+1}, \qquad \varepsilon \sim N(0, 1), \quad (15)$$

where μ is the mean of the time series (in this case 0), ϕ is the autoregressive parameter, and ε is a random error component (Wilks 1995). For each realization ($n = 100$), the autoregressive parameter ϕ was randomly selected from a sample distribution of observed lag-1 (1 yr) autocorrelation coefficients that are calculated using the time series of mean annual temperatures from stations in the United State Historical Climatology Network (USHCN; Karl et al. 1990). The values in each AR(1) series, though approximately standard normal, then were restandardized. To create groups of cross-correlated series, a constant of 2.0 times a random cross-correlation coefficient, also drawn from observed values, was added a total of ($m + 1$) times to each of the original 1000 series ($m = 5$ "neighboring" series plus the target). Each of the ($m + 1$) series in a group is formed, therefore, from the same "parent" series, which is not used. Because the target (candidate) and reference component (neighbor) series are all "sibling" series, each has approximately the same degree of cross correlation, on average, with every other series in its group.

b. Addition of random changepoints

Detection results are based on time series that contain zero, one, two, or a variable number changepoints in the combinations shown in Table 3. The amplitude of each simulated changepoint was selected at random from the standard normal distribution with no restric-

tions. As shown in Fig. 3, the standard normal distribution is a reasonable proxy for the distribution of known changepoints in the USHCN (expressed in standardized form). The simulated changepoint position was allowed to vary randomly. It should be noted, however, that when two changepoints are separated in time by no more than a few time steps, a changepoint detection algorithm may identify only one changepoint that is, in effect, an amalgam of the two nearby changepoints. If the two are of a comparable amplitude but opposite in sign, neither changepoint may be detected. On the other hand, if the changepoints are of disparate amplitudes, the larger shift may eclipse the smaller. To avoid sorting out the impact of these confounding scenarios on measures of detection skill, the results presented below are based on simulated shifts separated in time by no fewer than five positions in a sequence. In practice, however, nearby changepoints are a distinct possibility, especially in the analysis of annual values. In Fig. 2, a realization of a target/neighbor series from case 2a was shown that includes the values of each reference series formulation and test statistic at the first splitting step.

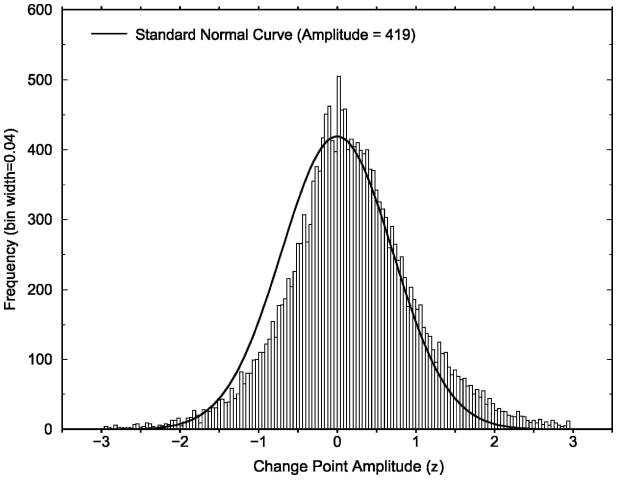

FIG. 3. Distribution of the estimated amplitude of step changes at known risks of artificial changepoints in the USHCN. Step-change amplitude is expressed in standardized form.

4. Results

In practice, the number of true changepoints in a climate series is unknown. Moreover, the presence of multiple shifts can sometimes suppress the magnitude of the test statistics near each true break point to such a degree that none exceeds its critical value. In these situations, the first split can be made at the position where the test statistic reaches a maximum without regard to the significance level (0.05 is used here) to avoid the possibility that the series will be overlooked completely when there is a complicated multi-break-point configuration. Of course, it is generally not known in advance that such a situation exists so the first split must be made in all series, which necessitates the merging steps in the semihierarchic method. It is worth pointing out, however, that the number of H_o rejections using hierarchic binary splitting in series where the null hypothesis is true is greater relative to the expected value in a test for a single changepoint (no splitting). This is because when the critical value is exceeded somewhere in a sequence, a split is made at the point where the test statistic reaches a maximum. At that point, the subsequences on either side of the split are evaluated separately, each having some probability of type I error as in the case of testing the full series. Inclusion of the merging steps will rejoin some of the "false splits" and reduce the number of type I errors, but the number may, nevertheless, be larger than the expected value.

TABLE 3. Number of added changepoints in each target–reference component series groups used in five Monte Carlo case studies. Each case comprised of 1000 simulated series groups.

	Number of simulated changepoints ($K - 1$)	
	Candidate series	Each reference series component
Case 1 (null case)	0	0
Case 2	(a) 1 (b) 2	(a) 0 (b) 0
Case 3	(a) 0 (b) 0	(a) 1 (b) 2
Case 4	(a) 1 (b) 2	(a) 1 (b) 2
Case 5 (with missing values)	(a) 0 (b) 1	(a) 0 (b) 1
Case 6	Between 0 and 6*	Between 0 and 6*
Case 7	6**	0

* The number of changepoints in each series is approximately normally distributed about an average of 3.

** Changepoint position and amplitude are fixed as in Caussinus and Mestre (2004). See Table 9 for details.

The consequence of the larger number of hypothesis tests can be seen in Table 4, where changepoint detection is summarized for case 1 (null hypothesis always true). For most of the nine test statistic/reference pairs, the number of false alarms is higher than the expected value of 5% (not all combinations are shown). On the other hand, when H_o is true, the null hypothesis is rarely rejected at the same position in a series (± 2 time steps) by more than one test so the false alarm rate when agreement between tests is required is less than the expected value for one test, suggesting some independence between tests.

TABLE 4. Skill scores for simulations with no changepoints in the candidate or in the reference component series (case 1).

Case 1 Reference series–statistic	Hit	False alarm	Miss	*B*	*H*	*F*	FAR	HSS
(1) MLR–lag 1	0.0%	4.9%	0.0%	—	—	0.0004	1.00	0.00
(2) FDWA–F_{max}	0.0%	8.2%	0.0%	—	—	0.0008	1.00	0.00
(3) ANWA–T_{max}	0.0%	11.2%	0.0%	—	—	0.0011	1.00	0.00
Consensus								
(1) and (2)	0.0%	1.0%	0.0%	—	—	0.0000	1.00	0.00
(2) and (3)	0.0%	1.9%	0.0%	—	—	0.0002	1.00	0.00
(1) and (3)	0.0%	0.8%	0.0%	—	—	0.0001	1.00	0.00
(1) and/or (2) and/or (3)*	0.0%	2.7%	0.0%	—	—	0.0003	1.00	0.00
MLR–T_{max} and/or ANWA–F_{max} and/or MLR–lag 1**	0.0%	2.3%	0.0%	—	—	0.0002	1.00	0.00
Optimal	0.0%	125.4%	0.0%	—	—	0.0125	1.00	0.00

* Any two of three.
** Best 2 of 3 of all 84 possible reference series–test statistic triplets.

The detection results summarized for cases 2 through 6 (Tables 5–9), indicate that, apart from the optimal algorithm, the likelihood ratio statistic is generally the most sensitive of the three calculated statistics (cf. the single test hit rates *H*, for cases 2, 4, 5b, and 6). The superior sensitivity, however, comes at the price of a larger number of false alarms, especially when changepoints are present in the reference series components (cases 3, 4, 5b, and 6), but even when the reference series is truly homogeneous (case 2). In the case 2 simulations, as in case 1, a large reduction in type I errors occurs when a consensus is required that includes at least two different test statistics. Likewise, in case 2 changepoint detection is essentially the same whichever composite reference series formulation is paired with the likelihood ratio test. In fact, results from pairing all three reference formulations with a common statistic (all combinations not shown) suggest that the choice of

TABLE 5. Skill scores for simulations with one (case 2a) or two (case 2b) changepoints in the candidate and none in the reference series components.

Case 2a Reference series–statistic	Hit	False alarm	Miss	*B*	*H*	*F*	FAR	HSS
(1) MLR–lag 1	42.9%	16.0%	57.0%	0.59	0.43	0.0016	0.27	0.54
(2) FDWA–F_{max}	44.9%	20.0%	55.5%	0.65	0.45	0.0020	0.31	0.54
(3) ANWA–T_{max}	58.9%	21.7%	41.0%	0.81	0.59	0.0022	0.27	0.65
Consensus								
(1) and (2)	38.9%	4.8%	61.0%	0.44	0.39	0.0005	0.11	0.54
(2) and (3)	44.2%	4.0%	55.7%	0.48	0.44	0.0004	0.08	0.59
(1) and (3)	41.6%	3.7%	58.3%	0.45	0.42	0.0004	0.08	0.57
(1) and/or (2) and/or (3)*	48.3%	9.2%	51.6%	0.58	0.48	0.0009	0.16	0.61
ANWA–F_{max} and/or ANWA–T_{max} and/or MLR–T_{max}**	57.1%	16.9%	42.8%	0.74	0.57	0.0017	0.23	0.65
Optimal	60.1%	132.3%	39.9%	1.92	0.60	0.0134	0.69	0.40
Case 2b								
(1) MLR–lag 1	40.9%	14.0%	59.1%	0.55	0.41	0.0028	0.25	0.52
(2) FDWA–F_{max}	41.4%	12.8%	58.7%	0.54	0.41	0.0026	0.24	0.53
(3) ANWA–T_{max}	55.4%	14.6%	44.6%	0.70	0.55	0.0030	0.21	0.65
Consensus								
(1) and (2)	35.4%	3.6%	64.7%	0.39	0.35	0.0007	0.09	0.50
(2) and (3)	40.0%	2.6%	60.1%	0.43	0.40	0.0005	0.06	0.56
(1) and (3)	39.6%	2.7%	60.5%	0.42	0.40	0.0005	0.06	0.55
(1) and/or (2) and/or (3)*	45.6%	6.4%	54.5%	0.52	0.46	0.0013	0.12	0.59
MLR–T_{max} and/or ANWA–lag 1 and/or ANWA–T_{max}**	53.8%	11.1%	46.3%	0.65	0.54	0.0023	0.17	0.65
Optimal	58.5%	62.6%	41.6%	1.2105	0.58	0.0128	0.52	0.52

* Any two of three.
** Best 2 of 3 of all 84 possible reference series–test statistic triplets.

TABLE 6. Skill scores for simulations with no changepoints in the candidate and one (case 3a) or two (case 3b) changepoints in each reference series component

Case 3a Reference series–statistic	Hit	False alarm	Miss	*B*	*H*	*F*	FAR	HSS
(1) MLR–lag 1	0.0%	9.9%	0.0%	—	—	0.0010	1.00	0.00
(2) FDWA–F_{max}	0.0%	30.0%	0.0%	—	—	0.0030	1.00	0.00
(3) ANWA–T_{max}	0.0%	71.4%	0.0%	—	—	0.0071	1.00	0.00
Consensus								
(1) and (2)	0.0%	2.6%	0.0%	—	—	0.0003	1.00	0.00
(2) and (3)	0.0%	14.1%	0.0%	—	—	0.0014	1.00	0.00
(1) and (3)	0.0%	2.6%	0.0%	—	—	0.0003	1.00	0.00
(1) and/or (2) and/or (3)*	0.0%	16.5%	0.0%	—	—	0.0016	1.00	0.00
MLR–T_{max} and/or MLR–F_{max} and/or MLR–lag 1**	0.0%	5.6%	0.0%	—	—	0.0006	1.00	0.00
Optimal	0.0%	202.3%	0.0%	—	—	0.0202	1.00	0.00
Case 3b								
(1) MLR–lag 1	0.0%	14.3%	0.0%	—	—	0.0014	1.00	0.00
(2) FDWA–F_{max}	0.0%	57.2%	0.0%	—	—	0.0057	1.00	0.00
(3) ANWA–T_{max}	0.0%	106.4%	0.0%	—	—	0.0106	1.00	0.00
Consensus								
(1) and (2)	0.0%	4.2%	0.0%	—	—	0.0004	1.00	0.00
(2) and (3)	0.0%	23.9%	0.0%	—	—	0.0024	1.00	0.00
(1) and (3)	0.0%	4.1%	0.0%	—	—	0.0004	1.00	0.00
(1) and/or (2) and/or (3)*	0.0%	28.5%	0.0%	—	—	0.0029	1.00	0.00
MLR–T_{max} and/or MLR–F_{max} and/or MLR–lag 1**	0.0%	10.2%	0.0%	—	—	0.0010	1.00	0.00
Optimal	0.0%	229.9%	0.0%	—	—	0.0230	1.00	0.00

* Any two of three.
** Best 2 of 3 of all 84 possible reference series–test statistic triplets.

test statistic is more important than choice of reference series formulation when the reference series components are serially complete and homogeneous. In that case, the ANWA and the FDWA reference series are identical because they differ only by an offset.

The impact of changepoints in reference component series is illustrated in Table 6 (case 3). Because in these realizations each reference component contains either 1 (case 3a) or 2 (case 3b) changepoints, a composite reference will incorporate 5 (case 3a) or 10 (case 3b) changepoints of various amplitudes and locations. The number of false alarms using the likelihood ratio test and the ANWA or FDWA composite reference increases from just over 100 in the null case (case 1) to over 700 in case 3a and over 1000 in case 3b. Similarly, the ANWA and FDWA composite reference paired with the two-phase regression test statistic show a fourfold or better increase in the number of false alarms. In contrast, paired with the MLR reference series, the likelihood ratio and two-phase regression tests have less than half the number false alarms (not shown) and the increase over the sample of "ideal" reference series (case 1, 2, or 5a) is, therefore, much smaller. The MLR–lag 1 combination produced the smallest number of false alarms for a single reference series–test statistic pair.

It appears that a step change in a reference component series of anomalies or raw values will reduce the magnitude of the series coefficient in the MLR equation, and, therefore, its weight, effectively filtering the impact of the step changes in the composite. On the other hand, using first-difference-filtered series to calculate truer correlation-based weights when artificial break points may be present helps to ensure that step changes in the component series will carry through to the composite by minimizing the impact of a step change on the correlation coefficients. Nevertheless, the value of a consensus result is especially evident in case 3 from the large reduction in false alarms linked to nonhomogeneities in the composite reference series.

In case 4, when all series contain one or two changepoints, the number of false alarms can approach, or, in the case of ANWA–T_{max} even exceed the number of hits. As in other cases, the advantage to using a consensus result is apparent by the large reduction in false alarms relative to most single tests. Unfortunately, no consensus combination of reference series–test statistic pairs clearly stands out as the more skillful because many appear to optimize test sensitivity while others produce the fewest false alarms. Nevertheless, the pairing of MLR–T_{max} forms a good combination with many other reference series–test statistic pairs because this

TABLE 7. Skill scores for simulations with one (case 4a) or two (case 4b) changepoints in the candidate and one (case 4a) or two (case 4b) changepoints in each reference series component.

Case 4a Reference series–statistic	Hit	False alarm	Miss	*B*	*H*	*F*	FAR	HSS
(1) MLR–lag 1	35.3%	28.8%	64.7%	0.64	0.35	0.0029	0.45	0.43
(2) FDWA–F_{max}	45.4%	38.2%	54.7%	0.84	0.45	0.0039	0.46	0.49
(3) ANWA–T_{max}	57.8%	63.2%	42.2%	1.21	0.58	0.0064	0.52	0.52
Consensus								
(1) and (2)	32.2%	5.4%	67.8%	0.38	0.32	0.0005	0.14	0.47
(2) and (3)	43.9%	11.5%	56.1%	0.55	0.44	0.0012	0.21	0.56
(1) and (3)	34.2%	5.5%	65.8%	0.40	0.34	0.0006	0.14	0.49
(1) and/or (2) and/or (3)*	46.9%	19.1%	53.1%	0.66	0.47	0.0019	0.29	0.56
MLR–T_{max} and/or ANWA–T_{max} and/or ANWA–F_{max}**	53.0%	26.5%	47.0%	0.80	0.53	0.0027	0.33	0.59
Optimal	61.7%	181.1%	38.3%	2.43	0.62	0.0183	0.75	0.35
Case 4b								
(1) MLR–lag 1	29.7%	53.6%	70.3%	0.56	0.30	0.0055	0.47	0.37
(2) FDWA–F_{max}	38.0%	50.5%	62.1%	0.63	0.38	0.0052	0.40	0.46
(3) ANWA–T_{max}	50.3%	85.0%	49.7%	0.93	0.50	0.0087	0.46	0.51
Consensus								
(1) and (2)	24.2%	11.0%	75.8%	0.30	0.24	0.0011	0.19	0.37
(2) and (3)	34.3%	14.6%	65.7%	0.42	0.34	0.0015	0.18	0.48
(1) and (3)	27.4%	9.4%	72.7%	0.32	0.27	0.0010	0.15	0.41
(1) and/or (2) and/or (3)*	39.2%	27.7%	60.8%	0.53	0.39	0.0028	0.26	0.51
MLR–T_{max} and/or ANWA–T_{max} and/or ANWA–lag 1**	45.1%	42.3%	54.9%	0.66	0.45	0.0043	0.32	0.54
Optimal	55.7%	189.6%	88.6%	1.51	0.56	0.0193	0.63	0.43

* Any two of three.
** Best 2 of 3 of all 84 possible reference series–test statistic triplets.

pairing filters the impact of changepoints in the reference series components while retaining much of the test sensitivity. However, skillful consensus detection with this reference series–test statistic pair is possible only when none of the other pairings includes the MLR reference series because its use has a large impact on test sensitivity with all of the test statistics.

Based on the HSS, a consensus of any two of three tests is generally more skillful than agreement between a single pair of reference series–test statistic combinations. In fact, a consensus of any two or three reference series–test statistic pairs is more skillful than the use of either any two of four, or three of five pairs, etc. This is because a consensus of a large number of reference–test combinations will maximize both the number of coincident hits and the number of coincident false alarms. Because there is probably more independence between tests in terms of the position of false alarms, the small gain in hits using an agreement between, say, any two of four over any two of three tests is more than offset by the gain in the number of consensus false alarms. In case 2, the highest skill scores for any 2 of 3 of the 84 reference series–test statistic pairings are those paired combinations that include all three test statistics. In case 3, it is for pairings that include only the MLR reference.

In case 5a and 5b, which are comprised of simulations with randomly censored values, results are similar to the analogous cases 1 and 4a, respectively, with one exception: all test statistics that are paired with the FDWA show a large increase in false alarm numbers relative to the serially complete counterpart scenarios (cf. e.g., the false alarm column in Tables 4 and 8). This large increase in pairings that include the FDWA appears to be caused by random walks introduced into the FDWA series that are a consequence of biased estimates of the average first difference when one or more of a component's series values are censored (missing) at various positions. Such biased estimates are unavoidable when values are missing, and they also impact the ANWA reference series, but in that case cause only a small increase in false alarms. In the case of the FDWA, however, a biased estimate at one position in a series will cause all subsequent composite averages (or, in this case, working backward, all earlier averages) to exhibit the same bias. If there are missing values in the various reference series at different positions scattered throughout the summary period, the combination can cause a random walk, rather than a simple step change, the range of which may be large (e.g., one standard deviation), as shown in Fig. 4. When the FDWA composite reference series is used to form a difference (or

TABLE 8. Skill scores for simulations with one, two, or five missing values in a row at random positions but no changepoints in the candidate and or reference series components (case 5a) or with missing values and one changepoint in the candidate and one changepoint in each reference series component (case 5b).

Case 5a Reference series–statistic	Hit	False alarm	Miss	*B*	*H*	*F*	FAR	HSS
(1) MLR–lag 1	0.0%	5.3%	0.0%	—	—	0.0005	1.00	0.00
(2) FDWA–F_{max}	0.0%	27.5%	0.0%	—	—	0.0027	1.00	0.00
(3) ANWA–T_{max}	0.0%	13.8%	0.0%	—	—	0.0014	1.00	0.00
Consensus								
(1) and (2)	0.0%	1.4%	0.0%	—	—	0.0001	1.00	0.00
(2) and (3)	0.0%	3.9%	0.0%	—	—	0.0004	1.00	0.00
(1) and (3)	0.0%	1.3%	0.0%	—	—	0.0001	1.00	0.00
(1) and/or (2) and/or (3)*	0.0%	5.1%	0.0%	—	—	0.0005	1.00	0.00
MLR–T_{max} and/or MLR–F_{max} and/or MLR–lag 1**	0.0%	2.1%	0.0%	—	—	0.0002	1.00	0.00
Optimal	0.0%	115.9%	0.0%	—	—	0.0116	1.00	0.00
Case 5b								
(1) MLR–lag 1	27.6%	25.6%	72.4%	0.53	0.28	0.0026	0.48	0.36
(2) FDWA–F_{max}	39.3%	51.7%	60.0%	0.91	0.39	0.0052	0.57	0.41
(3) ANWA–T_{max}	54.7%	64.8%	45.3%	1.20	0.55	0.0065	0.54	0.49
Consensus								
(1) and (2)	23.2%	4.3%	77.0%	0.27	0.23	0.0004	0.16	0.36
(2) and (3)	37.4%	12.1%	62.6%	0.50	0.37	0.0012	0.24	0.50
(1) and (3)	26.0%	3.5%	74.0%	0.29	0.26	0.0004	0.12	0.40
(1) and (2) and (3)*	41.5%	17.8%	58.5%	0.59	0.41	0.0018	0.30	0.52
MLR–T_{max} and/or ANWA–T_{max} and/or ANWA–lag 1**	49.2%	28.1%	50.8%	0.77	0.49	0.0028	0.36	0.55
Optimal	58.0%	185.7%	42.0%	2.44	0.58	0.0188	0.76	0.33

* Any two of three.
** Best 2 of 3 of all 84 possible reference series–test statistic triplets.

ratio) with the target series, the difference series will incorporate the characteristics of a step change or random walk in the reference and lead to a large increase in false alarms relative to that based on serially complete data or other reference series formulation. Thus, the averaging of first difference series should be avoided when serially incomplete values or a changing station mix must be used. In addition, the potential for a biased estimate using the ANWA or FDWA formulation will differ according to the relative magnitude of the field variance of anomalies versus the field variance of first differences (interannual variability).

Not surprisingly, a comparison of detection results for simulations containing one shift to those with two suggests that there is a general reduction in the hit rate when more than one undocumented changepoint occurs in a series. However, in the simulated scenarios with a maximum of two changepoints, the number of false alarms increases proportionately less than the number of hits when there are two shifts in the series versus one, so the skill of detection (the HSS) is not necessarily greatly reduced. The disproportionate change in the number of false alarms relative to hits is reflected by the reduction in bias (*B*) when there are two changepoints instead of one. Compare, for example, case 2a to 2b (Table 5) or case 4a to 4b (Table 7). In some reference series–test statistic pairings, the HSS is essentially equivalent in scenarios with one and two breaks, especially in reference series–test statistic pairings that include the likelihood ratio test.

In general, the optimal solution using the method defined in (8) is more sensitive than successive hypothesis tests, especially in case 7. The case 7 scenario is precisely the kind of situation in which the optimal solution should be superior because the imposed changepoints are not hierarchic and they have equal amplitudes but opposite signs at positions 70 and 75. Nevertheless, while the optimal solution has a higher hit rate than any single hypothesis test, the best of an agreement between any two of three hypothesis test statistic–reference series pairings also has a very high hit rate. Given that a consensus of successive hypothesis tests has many fewer false alarms than the optimal solution, the skill of a consensus of successive hypothesis testing is nearly identical to the optimal skill in case 7 and is higher than the optimal solution in the other cases.

Because the number of false alarms in the optimal solution, expressed in Tables 4–9 as the total number of false alarms over the number of target series, is high, a different penalty function might be used to reduce this

TABLE 9. Skill scores for simulations with zero to six changepoints of random amplitude and position (case 6) and with six changepoints of fixed amplitude and position (case 7). In case 7, changepoints with an amplitude of 2.0 were added or subtracted as in Caussinus and Mestre (2004), i.e., +2.0 at position 20, +2.0 at position 40, −2.0 at position 50, −2.0 at position 70, +2.0 at position 75, and +2.0 at position 85.

Case 6 Reference series–statistic	Hit	False alarm	Miss	*B*	*H*	*F*	FAR	HSS
(1) MLR–lag 1	28.1%	63.8%	72.0%	0.49	0.28	0.0066	0.43	0.36
(2) FDWA–F_{max}	33.8%	58.3%	66.2%	0.53	0.34	0.0060	0.36	0.43
(3) ANWA–T_{max}	45.4%	96.9%	54.7%	0.78	0.45	0.0100	0.41	0.50
Consensus								
(1) and (2)	20.9%	10.5%	79.1%	0.24	0.21	0.0011	0.14	0.33
(2) and (3)	28.3%	15.0%	71.7%	0.33	0.28	0.0015	0.15	0.42
(1) and (3)	23.5%	11.9%	76.5%	0.27	0.24	0.0012	0.14	0.36
(1) and/or (2) and/or (3)*	35.1%	31.9%	64.9%	0.46	0.35	0.0033	0.23	0.47
ANWA–T_{max} and/or MLR–T_{max} and/or ANWA–lag 1**	41.5%	49.4%	58.5%	0.58	0.41	0.0051	0.28	0.51
Optimal	44.3%	200.8%	55.7%	1.14	0.44	0.0215	0.61	0.40
Case 7								
(1) MLR–lag 1	86.3%	47.1%	13.9%	0.94	0.86	0.0050	0.08	0.88
(2) FDWA–F_{max}	81.8%	31.7%	18.4%	0.87	0.82	0.0034	0.06	0.87
(3) ANWA–T_{max}	70.4%	26.3%	29.8%	0.75	0.70	0.0028	0.06	0.79
Consensus								
(1) and (2)	73.7%	7.1%	26.4%	0.75	0.74	0.0008	0.02	0.83
(2) and (3)	60.5%	2.1%	39.7%	0.61	0.60	0.0002	0.01	0.74
(1) and (3)	65.3%	2.9%	34.8%	0.66	0.65	0.0003	0.01	0.78
(1) and (2) and (3)*	86.2%	10.1%	13.8%	0.88	0.86	0.0011	0.02	0.91
ANWA–F_{max} and/or MLR–lag 1 and/or ANWA–lag 1**	94.3%	15.8%	5.7%	0.97	0.94	0.0017	0.03	0.95
Optimal	99.9%	46.8%	0.1%	1.08	0.999	0.0050	0.07	0.96

* Any two of three.
** Best 2 of 3 of all 84 possible reference series–test statistic triplets.

total. Caussinus and Mestre (2004), for example, specified a penalty function for the same type of optimal solution, which, in contrast to methods by Akaike (1974) and Schwarz (1978), did not produce an excessive number of changepoints. In Fig. 5, a histogram of the number of detected changepoints by position is shown using the solution that is provided by the best two of three test statistic–reference series pairings. A comparison of Fig. 5 and a similar histogram provided in Table 1 of Caussinus and Mestre (2004) indicates that the consensus result of successive hypothesis tests is more sensitive than their approach to a penalty function, while at the same time it limits the number of false alarms. Thus, successive hypothesis testing using multiple tests might be a reasonable alternative to optimal solutions, even when complicated multi-break-point scenarios occur. Moreover, in the most realistic of changepoint scenarios, case 6, the optimal hit rate is not as high as the ANWA–T_{max} combination when a semiempirical splitting algorithm is used for the hypothesis test.

5. Discussion and conclusions

The quantification of detection skill using Monte Carlo case studies indicates that the likelihood ratio test is the most sensitive of the three successive hypothesis test statistics in all but one of the simulated scenarios. As a result, it is also the most sensitive to changepoints in reference series components and, thus, has a higher probability of detection and a higher probability of false detection. The higher sensitivity of the likelihood ratio test is not surprising given that the assumption of no slope in the form of the test used here was met perfectly by the Monte Carlo simulations. The assumption that there is no local trend may be reasonable in many situations, but nevertheless should be evaluated in practice. Wang (2003), arguing from the standpoint of sampling variability, noted that the sensitivity of the two-phase regression test can be increased, especially in short segments, by using a common slope parameter between the two phases. By eliminating the slope parameter altogether, the sensitivity of the two-phase regression test is equivalent to that of the zero-slope version of likelihood ratio test, and there is no benefit to including both zero-slope test models in multiple testing.

Even when no phase break points or trends are anticipated, there are step-change configurations in which allowance for trend changes may vastly increase step-change detection sensitivity. This was illustrated by the

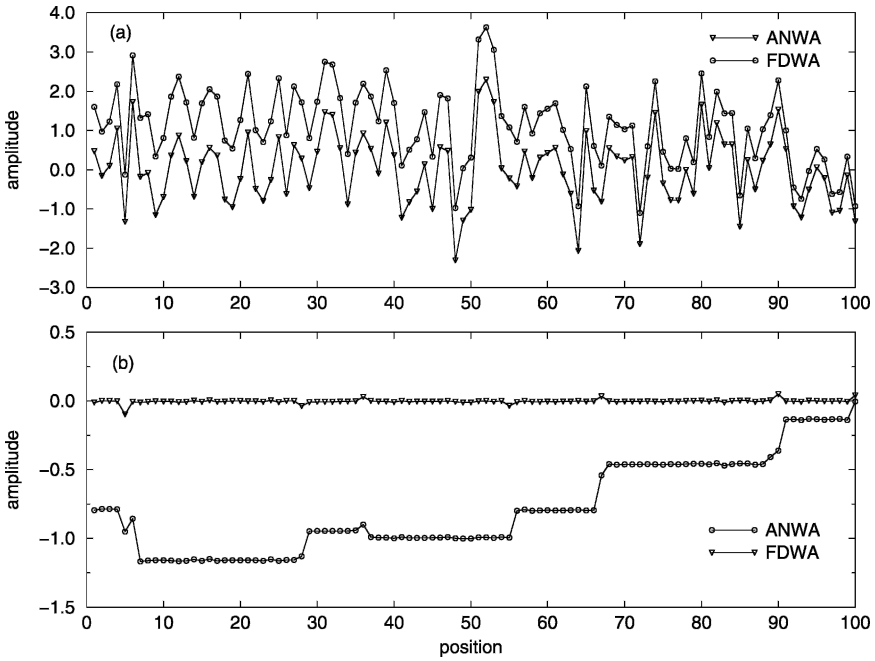

FIG. 4. (a) Example of FDWA and ANWA composite reference series with randomly censored values in the five component series; (b) difference between the true value of the ANWA and FDWA composite reference series and its estimate using component series with censored values (from case 5a).

case 7 detection results and in a simulation by Easterling and Peterson (1995) who imposed simulated changepoints with equal amplitudes but opposite signs 10 positions apart. Inclusion of the two-phase regression model with a slope parameter greatly increases the likelihood of finding the "temporary" jumps relative to a zero-slope test model because the apparent change in trend near the step is frequently sufficient to reject the null hypothesis of a one-phase segment. In case 7, where the step changes are not hierarchic, the semihierarchic splitting algorithm that seeks to resolve only changes in mean often failed to converge on the optimal solution. Even in case 7, however, detection using successive hypothesis testing is improved with the use of multiple tests and the consensus skill is comparable to an optimal solution. In arguably the most realistic of the simulations, case 6, the consensus of successive hypothesis testing can be more skillful at undocumented changepoint detection than an optimal solution by limiting the number of false alarms without reducing sensitivity too much. Thus, successive hypothesis testing may be preferable in situations where intervention in the result of an optimal algorithm is impractical.

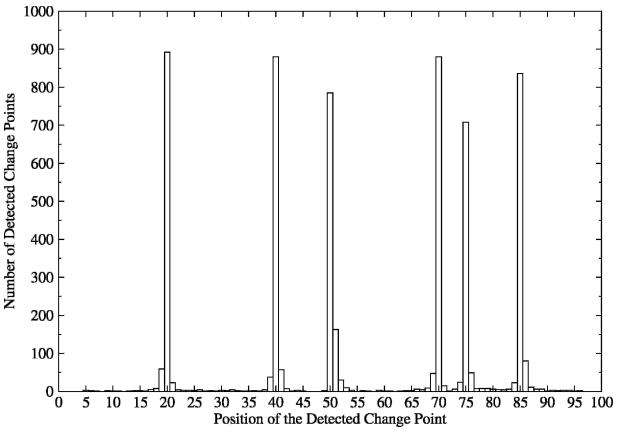

FIG. 5. Histogram of the number of detected changepoints by position for case 7. Detection is from the best consensus result using three test statistic–reference series pairs (agreement between any two of three) as indicated in Table 9.

The comparison of various combinations of test statistics and composite reference series formulations suggests that for reasonably well correlated time series, the choice of reference series formulation has relatively little impact on target series changepoint detection skill, provided that the reference component series are homogeneous. Though probably rare in practice, under such circumstances the choice of the test statistic has a greater impact. In the case where reference series components contain changepoints, and/or values are missing, the choice of reference series formulation has more important implications in changepoint detection. Step

changes in the various component series are more readily transferred to the composite reference when first-difference-filtered climate series are used to calculate truer correlation-based weights and increase the likelihood that heterogeneities in the composite reference will be erroneously identified as changepoints in the target series. On the other hand, a multiple linear regression or non-first-difference correlation-based-weighted reference series will tend to reduce the impact step changes on the composite reference. To confound the problem, an analyst risks weighting most heavily those station series that contain similar artificial breaks when anomaly or raw value correlation weights are used, reducing changepoint detection sensitivity. This is a pervasive problem when coincident or nearly coincident network-wide practice changes are imposed.

The first-difference composite reference has been advocated to facilitate changepoint detection in shorter, incomplete series for which anomaly calculation using a common base period is problematic (Peterson et al. 1998b). Moreover, the removal of spurious first differences in reference series components, presumably caused by step changes, also has been recommended prior to computing the average (Peterson and Easterling 1994). The results of this analysis suggest that composite first differences should be avoided if values are missing or removed from one or more component series or, more generally, when the composition of component series changes through time. In such circumstances, the averaging of first differences introduces step changes or random walks when the series is converted back to a raw value average. Random walks and spurious steps increase the number of false alarms if this form of composite reference is subtracted from a target series and may lead to erroneous conclusions about the nature of the background climate signal in a region when only a small number of reference component series is available. To avoid such artifacts in first-difference-based reference calculations, only serially complete segments should be used. In that case, the average first-difference series that is converted back to raw observations is exactly correlated with a similarly weighted average anomaly or raw value series, and there is no advantage to the use of first differences. If missing values are estimated, the estimate error still will cascade throughout the average first difference series and potentially lead to the same type of random walks.

A principal benefit of a multitest consensus, in addition to improved detection when changepoints are not hierarchic, occurs in situations where the composite reference series is not homogeneous because there appears to be greater independence between tests in the occurrence of false alarms than in detected changepoints. A consensus of a large number of test statistic–reference series pairs, however, maximizes the number of false alarms, while the consensus of only two test statistics–reference series pairs limits detection to the least sensitive pairing. Consequently, the most skillful consensus appears to be any two of three test statistic–composite reference series pairs. Using the agreement between any two of three tests, detection skill in simulations where reference series components contain changepoints (case 4) are comparable to the perfectly homogeneous reference series case and the use of a single test (case 2). The reference series–test statistic combinations that are most appropriate for a particular evaluation of nonclimatic changepoints may depend, in large part, on the relative priority of reducing false alarms or avoiding misses. If a climate series is to be adjusted for undocumented changepoints, then the reduction of false alarms may be critical and multiple linear regression or non-first-difference-based correlation weights should be used in a least one of the three reference series formulations. If sensitivity is critical, then multiple linear regression or non-first-difference-based weighting should be used in, at most, one of the reference series formulations.

Even in the use of a consensus approach, however, the proportion of detected changepoints that are false (the FAR) remains substantial, over 25%, for example, in the most realistic simulations (case 6). Alternative strategies, therefore, are likely required to reduce the false alarm rate in real world applications. Such strategies may include increasing the number of references series components that are used to form the average (Li et al. 2005, manuscript submitted to *J. Geophys. Res.*), or an iterative recalculation of the composite reference where the target series is the only series common to all calculated difference series (Szentimrey 1999). False alarms that are linked to changepoints in the composite reference series also may be avoided through a pairwise comparison of climate series (Jones et al. 1986; Menne and Duchon 2001; Caussinus and Mestre 2004). In a pairwise approach, the concept of target and reference lose their meaning and the offending series may be more readily identified. Detection skill based on this approach to undocumented changepoint analysis will be discussed in a forthcoming paper.

Acknowledgments. The authors wish to thank Dr. Thomas C. Peterson for bringing recent climate changepoint detection work, summarized in the WMO publication "Guidance on metadata and homogenization" (WMO/TD 1186), to our attention. Special thanks also to Dr. Xiaolan Wang, Tressa Fowler, and two anonymous reviewers whose thoughtful and construc-

tive comments greatly improved this manuscript. Algorithms for solving the penalized contrast function were provided by Dr. Marc Lavielle (http://www.math.u-psud.fr/~lavielle/programs/index.html). Partial support for this work was provided by the Office of Biological and Environmental Research, U.S. Department of Energy.

REFERENCES

Akaike, H., 1974: A new look at the statistical identification model. *IEEE Trans. Autom. Control,* **19,** 716–723.

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

——, and A. Moberg, 1997: Homogenization of Swedish temperature data. Part I: Homogeneity test for linear trends. *Int. J. Climatol.,* **17,** 25–34.

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

Conrad, V., and C. Pollack, 1962: *Methods in Climatology*. Harvard University Press, 459 pp.

Doswell, C. A., III, R. Davies-Jones, and D. L. Keller, 1990: On summary measures of skill in rare event forecasting based on contingency tables. *Wea. Forecasting,* **5,** 576–585.

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.

Durbin, J., and G. S. Watson, 1950: Testing for serial correlation in least squares regression. I. *Biometrika,* **37,** 409–428.

——, and ——, 1951: Testing for serial correlation in least squares regression. II. *Biometrika,* **38,** 159–178.

——, and ——, 1971: Testing for serial correlation in least squares regression. III. *Biometrika,* **58,** 1–19.

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.

——, ——, and T. R. Karl, 1996: On the development and use of homogenized climate datasets. *J. Climate,* **9,** 1429–1434.

Hawkins, D. M., 1976: Point estimation of the parameters of a piecewise regression model. *Appl. Stat.,* **25,** 51–57.

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

——, 2001: Fitting multiple change-points to data. *Comput. Stat. Data Anal.,* **37,** 323–341.

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.

——, ——, F. T. Quinlan, and T. A. Boden, 1990: United States Historical Climatology Network (HCN) serial temperature and precipitation data. Oak Ridge National Laboratory, Carbon Dioxide Information and Analysis Center, Environmental Science Division Publication No. 3404, 389 pp.

Jones, P. D., S. C. B. Raper, R. S. Bradley, H. F. Diaz, P. M. Kellyo, and T. M. L. Wigley, 1986: Northern Hemisphere surface air temperature variations: 1851–1984. *J. Climate Appl. Meteor.,* **25,** 161–179.

Lavielle, M., 1998: Optimal segmentation of random processes. *IEEE Trans. Signal Process.,* **46,** 1365–1373.

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

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

Menne, M. J., and C. E. Duchon, 2001: A method for monthly detection of inhomogeneities and errors in daily maximum and minimum temperatures. *J. Atmos. Oceanic Technol.,* **18,** 1136–1149.

——, and ——, 2002: Quality assurance of monthly temperature data at the National Climatic Data Center. Preprints, *13th Conf. on Applied Climatology*, Portland, OR, Amer. Meteor. Soc., 18–21.

Murphy, A. H., and R. L. Winkler, 1987: A general framework for forecast verification. *Mon. Wea. Rev.,* **115,** 1330–1338.

Peterson, T. C., and D. R. Easterling, 1994: Creation of homogeneous composite climatological reference series. *Int. J. Climatol.,* **14,** 671–679.

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

——, T. R. Karl, P. F. Jamason, R. Knight, and D. R. Easterling, 1998b: First difference method: Maximizing station density for the calculation of the long-term global temperature change. *J. Geophys. Res.,* **103** (D20), 25 967–25 974.

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

Schwarz, G., 1978: Estimating the dimension of a model. *Ann. Stat.,* **6,** 461–464.

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

Stephenson, D. B., 2000: Use of the "odds ratio" for diagnosing forecast skill. *Wea. Forecasting,* **15,** 221–232.

Szentimrey, T., 1999: Multiple analyses of series for homogenization (MASH). *Proc. of the Second Seminar for Homogenization of Surface Climatological Data*, WMO-TD-962, Budapest, Hungary, WMO, 27–46.

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

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

Wilks, D. S., 1995: *Statistical Methods in the Atmospheric Sciences*. Academic Press, 467 pp.