# Homogenization of Temperature Series via Pairwise Comparisons

MATTHEW J. MENNE AND CLAUDE N. WILLIAMS JR.

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

(Manuscript received 2 October 2007, in final form 2 September 2008)

ABSTRACT

An automated homogenization algorithm based on the pairwise comparison of monthly temperature series is described. The algorithm works by forming pairwise difference series between serial monthly temperature values from a network of observing stations. Each difference series is then evaluated for undocumented shifts, and the station series responsible for such breaks is identified automatically. The algorithm also makes use of station history information, when available, to improve the identification of artificial shifts in temperature data. In addition, an evaluation is carried out to distinguish trend inhomogeneities from abrupt shifts. When the magnitude of an apparent shift attributed to a particular station can be reliably estimated, an adjustment is made for the target series. The pairwise algorithm is shown to be robust and efficient at detecting undocumented step changes under a variety of simulated scenarios with step- and trend-type inhomogeneities. Moreover, the approach is shown to yield a lower false-alarm rate for undocumented changepoint detection relative to the more common use of a reference series. Results from the algorithm are used to assess evidence for trend inhomogeneities in U.S. monthly temperature data.

## 1. Introduction

Discontinuities in a climate series can be induced by virtually any change in instrumentation or observation practice. The relocation, replacement, or recalibration of an instrument, for example, can lead to an abrupt shift in time-ordered observations that is unrelated to any real change in climate. Likewise, alterations to the land use or land cover surrounding a measurement site might induce a sudden or ''creeping'' change (Carretero et al. 1998; Karl et al. 1988) that could limit the degree to which observations are representative of a particular region. Such artifacts in the climate record ultimately confound attempts to quantify climate variability and change (Thorne et al. 2005). Unfortunately, changes to the circumstances behind a series of climate observations are practically inevitable at some point during the period of record. For this reason, testing for artificial discontinuities or ''inhomogeneities'' is an essential component of climate analysis. Often, the test results can then be used to adjust a series so that it more closely reflects only variations in weather and climate.

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

DOI: 10.1175/2008JCLI2263.1

Numerous approaches have been employed to detect discontinuities in climate series (Peterson et al. 1998a), and comparison studies have recently proliferated (e.g., Ducré-Robitaille et al. 2003; DeGaetano 2006; Reeves et al. 2007, hereafter R07). The goal of this work is to describe an automated homogenization algorithm for monthly data that builds on the most efficient changepoint detection techniques using a holistic design approach. For example, the algorithm relies upon a pairwise comparison of temperature series in order to reliably distinguish artificial changes from true climate variability, even when the changes are undocumented (Caussinus and Mestre 2004). Consequently, the procedure detects inhomogeneities regardless of whether there is a priori knowledge of the date or circumstances of a change in the status of observations (Lund and Reeves 2002). In addition, the algorithm employs a recursive testing strategy to resolve multiple undocumented changepoints within a single time series (Menne and Williams 2005, hereafter MW05). Last, the procedure explicitly looks for both abrupt ''jumps'' as well as local, unrepresentative trends in the temperature series (DeGaetano 2006).

The organization of the paper is as follows: additional background on the design considerations for constructing this ''pairwise'' homogenization algorithm is provided in section 2. In section 3, the specific components of the algorithm are described. In section 4, an assessment of the

algorithm's skill at changepoint detection and how this skill compares to previous studies is provided by means of simulated temperature series. Because of recent interest in land use change and its impact on the temperature record (e.g., Peterson and Owen 2005; Kalnay et al. 2006; Parker 2006; Pielke et al. 2007), the algorithm was also applied to historical temperature data from the U.S. Cooperative Observer (Coop) Network to assess the frequency of local, nonrepresentative trends as discussed in section 5. Some concluding remarks are offered in section 6.

## 2. Design considerations for the pairwise algorithm

### a. Relative changepoint testing

Conrad and Pollak (1962) state that "*a climatological series is relatively homogeneous with respect to a synchronous series at another place if the temperature differences (or precipitation ratios) of pairs of homologous averages constitute a series of random numbers*" (i.e., white noise). The assumption is that similar variations in climate occur at nearby locations because of the spatial correlation inherent to meteorological fields (Livezey and Chen 1983). A statistically significant and persistent violation of relative homogeneity is presumed to be artificial or, at least, to have origins other than the background variations in weather and climate. Relative homogeneity testing is therefore conducted primarily to distinguish artificial breaks from real climate variability, although it may also improve the power of detecting artificial shifts. The reason is that when two temperature series $\{X_t\}$ and $\{Y_t\}$ are highly correlated [i.e., Corr $(X_t, Y_t) = \rho > ½$], the variance of their differences will be lower relative to the original series.

To carry out relative homogeneity testing, a reference series is commonly constructed by averaging values from locations near the target site whose observations are in question (Karl and Williams 1987; Alexandersson and Moberg 1997; Vincent 1998). Unfortunately, the homogeneity of the reference series cannot be taken for granted because undocumented changepoints may be present in any one of the averaged series (Hanssen-Bauer and Førland 1994; MW05). Strategies for reducing changepoint attribution errors have included assessing the homogeneity of the reference series itself (McCarthy et al. 2008) and building a reference from previously adjusted series (González-Rouco et al. 2001). Unfortunately, conducting a separate assessment of reference series homogeneity fails to exploit the enhanced sensitivity of relative homogeneity testing, and many small-amplitude changepoints may go undetected in the reference series only to be later attributed to the target series. Similar problems may arise when adjusted data are used to build a reference series because artifacts from the original imperfect reference series can be transferred to the adjusted data themselves.

Alternatively, relative homogeneity testing can be implemented via a pairwise comparison of individual climate series (Jones et al. 1986; Slonosky et al. 1999; Menne and Duchon 2001; Caussinus and Mestre 2004). In pairwise testing, the cause of undocumented changepoints can be traced more directly, that is, without first testing the reference series or assuming it is homogeneous. Unfortunately, implementing pairwise testing has usually required a manual review of the results. For example, Jones et al. (1986) conducted an arduous station-by-station homogenization by manually determining the cause of changepoints in paired difference series. Caussinus and Mestre (2004) computed the locations of changepoints in difference series automatically, but still deferred to an analyst to attribute the cause. In contrast, an automated approach was developed for the pairwise algorithm, as described in section 3.

### b. Distinction between documented and undocumented changepoints

In the absence of station history records, the date of inhomogeneity must be treated as an unknown parameter. In such cases, a systematic search through all values in a series is required to identify the dates of statistically significant discontinuities. The systematic nature of the search necessitates the use of a more conservative set of critical values relative to the standard values that are appropriate for testing the significance of known changes to observation practice (Lund and Reeves 2002). This means that tests for undocumented changepoints are less sensitive than comparable tests for documented changes. It follows that to maximize the power of changepoint detection, station histories should be exploited whenever possible.

The strategy used by the pairwise algorithm is to first identify all evidence of changepoints using the less sensitive tests for undocumented changepoints. Where possible, the results are then combined with information about documented changes whose impact may go undetected by these less sensitive tests. An important benefit of this approach is that all possible changepoints are identified before estimates of their magnitude are made.

### c. Resolving multiple undocumented changepoints

While the issue of accurately resolving multiple undocumented changepoints remains an active area of statistical research (R07), two approaches are in operative use. The first, more common approach uses a recursive testing procedure (e.g., Vincent 1998) to overcome the "at most one changepoint" assumption behind most

hypothesis tests for undocumented changepoints. The second approach relies on a penalty function to constrain the number of changepoints resolved through an optimization routine used to maximize the contrast between sequential mean levels of a series (e.g., Caussinus and Mestre 2004).

A recursive testing approach is used in the pairwise algorithm for the following two reasons: First, the approach is associated with a low probability of false changepoint detection without requiring an analyst to interpret the results (cf. Caussinus and Mestre 2004). Second, MW05 noted that when the recursive hypothesis test method is carried out using a semihierarchical splitting algorithm (Hawkins 1976), the power of changepoint detection can be comparable to that of optimal algorithms.

Recursive testing is based on a hierarchic, binary segmentation of the test series whereby a series is split at the location where the test statistic reaches a maximum, that is, the point at which the separation between the mean before and after the breakpoint is greatest. Then, the subsequences on either side of the first split are likewise evaluated, and the process is repeated recursively until the magnitude of the statistic does not exceed the chosen significance level in any remaining subsequences (or the sample size in a segment is too small to test). A semihierarchic implementation of this method means that 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 breakpoints are identified, thereby more closely approximating an optimal solution.

### *d. Impact of local, unrepresentative trends*

Ideally, a changepoint detection method would differentiate trend changes from step changes. In practice, however, many of the commonly used tests for undocumented changepoints are not robust to the presence of trends in the test data because they are based solely on comparing the means of two sequential intervals. Use of such tests in the presence of trends can lead to falsely detected step changes as well as to inaccurate estimates of the magnitude of a shift when it occurs within a general trend (DeGaetano 2006; Pielke et al. 2007). Conversely, methods that directly account for both step changes and trend changes (e.g., Vincent 1998; Lund and Reeves 2002; Wang 2003) are characterized by much lower powers of detection than the simpler difference in means tests.

While no one test clearly outperforms others under all circumstances, the standard normal homogeneity test (SNHT; Alexandersson 1986) has been shown to have superior accuracy in identifying the position of a step change under a wide variety of step and trend inhomogeneity scenarios relative to other commonly used methods (DeGaetano 2006; R07). For this reason, the pairwise algorithm uses the SNHT along with a verification process that identifies the form of the apparent changepoint (e.g., step change, step change within a trend, etc.). In fact, the pairwise testing procedure is similar to the Vincent (1998) and R07 forward and backward regression methods, respectively, but is more easily adaptable to a recursive testing approach for resolving multiple undocumented changepoints, and at the same time retains the higher power of detection of the SNHT.

## 3. Description of the pairwise algorithm

The pairwise algorithm is executed according to the following six steps:

(i) Select a set of ''neighbors'' for each ''target'' series in the network, and form pairwise difference series between the target and its neighbors.
(ii) Identify the timing of shifts in all target-minus-neighbor difference series using SNHT.
(iii) Verify each apparent shift identified by SNHT in the pairwise differences (e.g., does the apparent shift look more like a trend?).
(iv) Attribute the cause of shifts in the set of target-minus-neighbor difference series to the various ''culprit'' series.
(v) Quantify the uncertainty in the timing of shifts attributed to each culprit series.
(vi) Estimate the magnitude of the identified shifts for use in adjusting the temperature series to reflect the true background climate signal.

Each of these steps is described in some detail below.

### *a. Selection of neighbors and formulation of difference series*

The pairwise algorithm starts by finding the 100 nearest neighbors for each temperature station within a network of stations. These neighboring stations are then ranked according to their correlation with the target. The first differences of the monthly anomalies are used to calculate the correlation coefficients [i.e., Corr $(X_t - X_{t-1}, Y_t - Y_{t-1})$] in order to minimize the impact of artificial shifts in determining the correlation (Peterson et al. 1998b). A series must simply be positively correlated with the target series to be eligible as a neighbor. Eligible neighbors could also be restricted to those series for whom $\rho \geq$ ½. This restriction effectively occurs in practice for monthly temperature values from the U.S. Cooperative Observer Network where more than 99.5% of the monthly temperature series from the 100 nearest neighbors are correlated at this level (and $\rho \geq 0.8$ in 90% of cases).

From all eligible neighbors, the set used for the pairwise analysis is selected using a two-step process. First, an account is made of the years and months for which both the target and its 40 most highly correlated neighbors report monthly mean maximum and minimum temperature data. Then, beginning with the 41st most highly correlated neighbor, the algorithm assesses whether an additional neighbor adds any data for the years and months that have fewer than seven viable neighbors. If the neighbor in question provides records for such data-sparse periods, it replaces the least correlated of the original 40 with the new neighbor provided that the addition does not remove data for other data-sparse periods. This process ensures that, whenever possible, at least seven neighbors are available at all times during the target station's period of record (the rationale for attempting to make at least seven target–neighbor comparisons is provided in section 4).

Next, time series of differences $\{D_t\}$ are formed between all target–neighbor monthly temperature series. To illustrate this, take two monthly series $\{X_t\}$ and $\{Y_t\}$, that is, a target and one of its correlated neighbors. Following Lund et al. (2007), these two series can be represented as

$$X_{mT+\nu} = \mu_\nu^X + \beta^X(mT+\nu) + \delta_{mT+\nu}^X + \varepsilon_{mT+\nu}^X \quad (1)$$

and

$$Y_{mT+\nu} = \mu_\nu^Y + \beta^Y(mT+\nu) + \delta_{mT+\nu}^Y + \varepsilon_{mT+\nu}^Y, \quad (2)$$

where $\mu$ represents the monthly mean anomaly at the specific series, $T = 12$ represents the months in the annual cycle, $\nu \in \{1, \ldots, 12\}$ is the monthly index, $m =$ the year (or annual cycle) number, and the $\varepsilon_t$ terms denote mean zero error terms at time $t$ for the two series. The $\delta_t$ terms represent shift factors cause by station changes, which are thought to be step functions. Following Lu and Lund (2007), these shift factors are of the form

$$\delta_{nT+\nu}^X = \begin{Bmatrix} \Delta_1^X, 1 \le t < c_1^X \\ \Delta_2^X, c_1^X \le t < c_2^X \\ \vdots \\ \Delta_k^X, c_{k-1}^X \le t < n^X \end{Bmatrix} \quad \text{and}$$

$$\delta_{nT+\nu}^Y = \begin{Bmatrix} \Delta_1^Y, 1 \le t < c_1^Y \\ \Delta_2^Y, c_1^Y \le t < c_2^Y \\ \vdots \\ \Delta_k^Y, c_{k-1}^Y \le t < n^Y \end{Bmatrix}, \quad (3)$$

where $n =$ the total number of values common to $\{X_t\}$ and $\{Y_t\}$, and $\Delta$ and $c$ represent the size and time of a shift, respectively. Because the timing of the level shifts is often unknown in climate networks, the goal of the pairwise algorithm is to reveal the shift times $\{c_1, c_2, \ldots, c_{k-1}\}$ not only for $\{X_t\}$ and $\{Y_t\}$, but for all of the series in the network no matter how complete the station metadata. Once the timing of the shifts is known, their magnitudes $\{\Delta_1, \Delta_2, \ldots, \Delta_{k-1}\}$ can be estimated.

Differencing the $\{X_t\}$ and $\{Y_t\}$ yields the $\{D_t\}$ series, which has the form

$$D_{mT+\nu} = (\mu_\nu^X - \mu_\nu^Y) + (\beta^X - \beta^Y)(mT+\upsilon) + (\delta_{mT+\nu}^X - \delta_{mT+\nu}^Y) + \varepsilon_{mT+\nu}^X - \varepsilon_{mT+\nu}^Y. \quad (4)$$

In reality, it is unrealistic to assume that $\beta^X$ and $\beta^Y$ are stationary in $t$ given the nature of multidecadal variations in climate series; however, it may be that $\beta^X \approx \beta^Y$ in general. This assumption is evaluated further in subsequent steps because if $\beta^X \neq \beta^Y$, then a local, unrepresentative trend (i.e., creeping inhomogeneity) is present in $\{X_t\}$ and/or $\{Y_t\}$. At present, the periodicity in (4) is considered to be negligible, especially since $\{X_t\}$ and $\{Y_t\}$ are first deseasonalized.

Figure 1 provides an example $\{D_t\}$ series formed between mean monthly maximum temperature anomalies from Chula Vista, California, and nine highly correlated neighbor series. The reduction in the variance of the $\{D_t\}$ series relative to the original target series is clearly evident. The variety of overlapping periods and relative shifts between the records from Chula Vista and its neighbors is common in surface temperature records.

### b. Identification of undocumented changepoints

After all difference series have been formed, the SNHT is used to identify undocumented changepoints in each $\{D_t\}$ using the semihierarchical splitting algorithm and a 5% significance level ($\alpha = 0.05$). The SNHT evaluates the null hypothesis ($H_0$) that a $\{D_t\}$ series has a constant mean against the alternative hypothesis ($H_A$) that there is an undocumented step change on date $c$. To account for the possibility of multiple changepoints, the difference series is assumed to consist of $K$ segments, each bounded by two changepoints ($c_{k-1}$ and $c_k$). In the pairwise algorithm, SNHT takes the form

$$H_0: \{D_t\} \rightarrow N(\mu_k, \sigma^2),\ c_{k-1}+1 \le t \le c_k, \quad (5)$$

$$H_A: \begin{cases} \{D_t\} \rightarrow N(\mu_1, \sigma^2),\ c_{k-1}+1 \le t \le c \\ \{D_t\} \rightarrow N(\mu_2, \sigma^2),\ c+1 \le t \le c_k \end{cases}, \quad (6)$$

where $N(\mu, \sigma^2)$ refers to a random normal variable with a mean $\mu$ and variance $\sigma^2$, and $\mu_1 \neq \mu_2$. For convenience

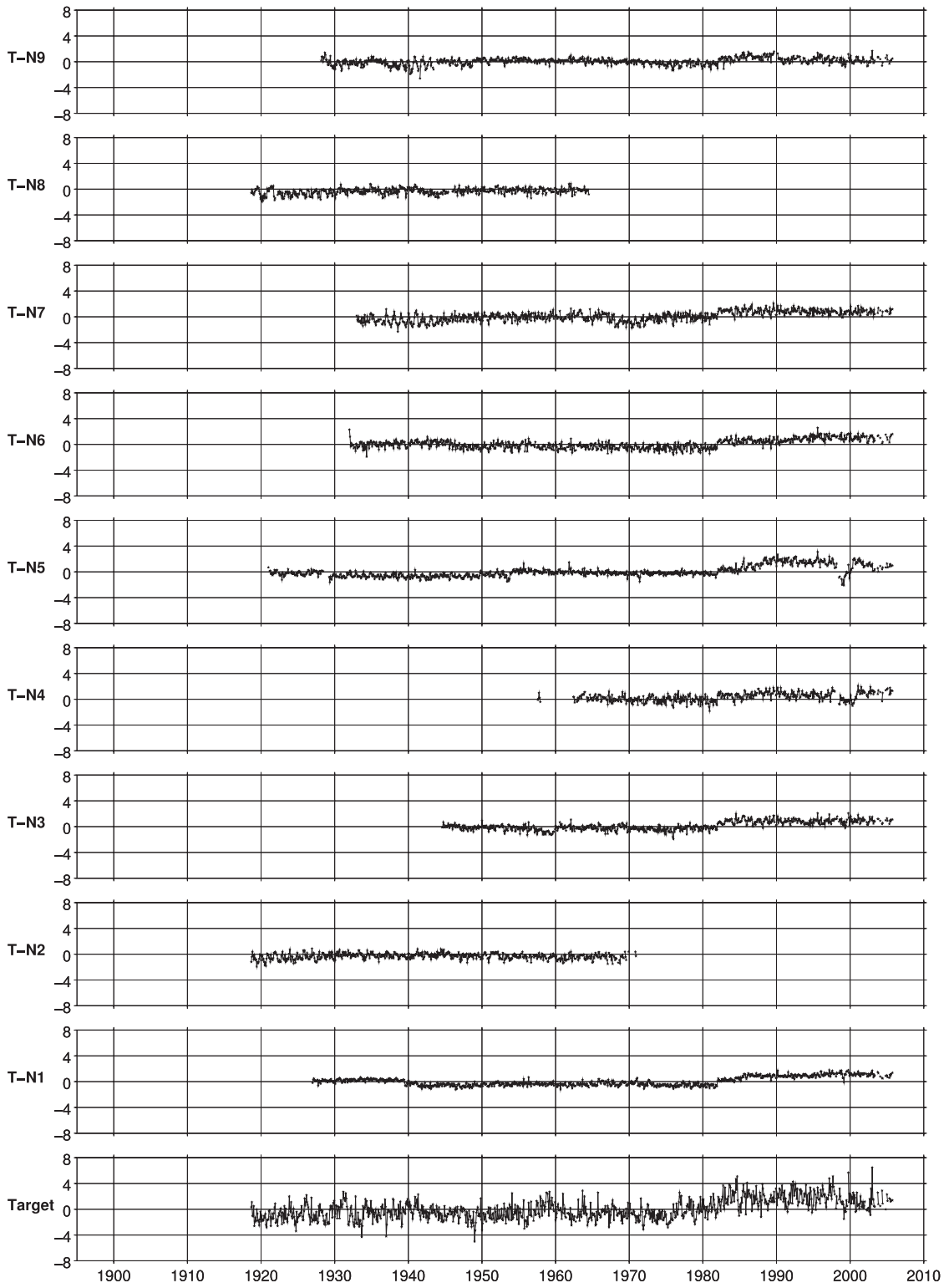


FIG. 1. Mean monthly maximum temperature anomalies for Chula Vista (target) and differences between monthly temperature anomalies at Chula Vista and nine neighboring series (T-N1 through T-N9).

TABLE 1. Hierarchy of changepoint models for a temperature difference series $\{D_t\}$, where the subscript $t$ refers to the time step of the series (e.g., 1 month), $\mu$ refers to the mean, $\beta$ refers to the trend, and $\varepsilon_t$ represents a random error term.

| Model | Description | Schematic of model | Number of parameters $p$ required to fit model |
|---|---|---|---|
| M1 | $D_t = \mu + \varepsilon_t$ |  | 1 |
| M2 | $D_t = \mu + \beta t + \varepsilon_t$ |  | 2 |
| M3 | $D_t = \begin{cases} \mu_1 + \varepsilon_t, t \le c \\ \mu_2 + \varepsilon_t, t > c \end{cases}$ |  | 3 |
| M4 | $D_t = \begin{cases} \mu_1 + \beta t + \varepsilon_t, t \le c \\ \mu_2 + \beta t + \varepsilon_t, t > c \end{cases}$ |  | 4 |
| M5 | $D_t = \begin{cases} \mu_1 + \beta_1 t + \varepsilon_t, t \le c \\ \mu_2 + \beta_2 t + \varepsilon_t, t > c \end{cases}$ |  | 5 |

we define $c_0 = 1$ and $c_K = n$, the total number of values in the $\{D_t\}$ series. The unsubscripted $c$ in (6) refers to the assignment of an undocumented changepoint between two previously established changepoints ($c_{k-1}$ and $c_k$) as the semihierarchal splitting algorithm iterates through the succession of splitting and merging steps, ultimately converging on a solution of $K$ segments bounded by $K - 1$ shifts.

### c. Classification of breakpoints identified by the SNHT test

The result of step b is a set of $K - 1$ apparent changepoints for each $\{D_t\}$ series. Because the SNHT assumes that each series is of the form

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

the next step determines whether this piecewise stationary model is justified for each changepoint. The determination is made by fitting a hierarchy of potential models for all segments centered on each $k$th breakpoint. The five models (M1–M5) are described in Table 1 (after R07). The model that minimizes the Bayesian information criterion (BIC; Schwarz 1978) is selected as the best representation for each changepoint.

Procedurally, the BIC is calculated by fitting M1–M5 to every segment $c_{k-1} + 1$ to $c_{k+1}$ for all $k = 1, K$. The BIC is defined as

$$\text{BIC}(p) = -2\log(L) + \log(n')p, \quad (8)$$

where $p$ is the number of parameters required to fit the model, $n'$ is the number of data points in the segment from $c_{k-1} + 1$ to $c_{k+1}$, and $L$ is the likelihood of the model in question. For the models listed in Table 1,

$$-2\log(L) = n'\log(\text{SSE}/n'), \quad (9)$$

where SSE refers to the sum of squared errors for the particular model fit.

In some cases, one or more of the original $K - 1$ changepoints may be eliminated from the solution for a particular $\{D_t\}$ series. For example, if the true model between the values of $c_{k-1} + 1$ and $c_{k+1}$ is a constant increasing trend (M2), the SNHT may have identified an apparent jump in the middle of the trend interval, whereas the BIC is likely to be lower for M2 than for any of the other four models. In such a case, the false changepoint time is removed from $\{c_1^D, c_2^D, \ldots, c_{k-1}^D\}$ and $K$ is decremented. Alternatively, the use of the BIC may determine that the $\{D_t\}$ segment between $c_{k-1} + 1$ and $c_{k+1}$ more appropriately follows M4 (step change within a constant trend) or M5 (a step change separated by different trends). If so, there is evidence of a relative trend between the two series, and the magnitude of the step change $\Delta$ required in subsequent steps e and f should be calculated using the higher dimension models (M4 and M5) to avoid calculating a biased estimate of the step.

### d. Attribution of shifts in the difference series

Given that breaks in a difference series will be induced by discontinuities in either $\{X_t\}$ or $\{Y_t\}$, the next step is to identify the series responsible for a particular discontinuity. To begin, an array of change dates by station is formed, and all of the changepoint dates detected in the $\{D_t\}$ series are temporarily assigned to both $\{X_t\}$ and $\{Y_t\}$. Specifically, a count is incremented for the date of shift each time a station is implicated by a break in one of its difference series. The resulting array of change dates by station is then ''unconfounded'' by systematically identifying those stations that are common to numerous difference series with the same date of change. More specifically, the station/date with the highest overall changepoint count is identified. This station is then tagged as the ''culprit'' or ''perpetrator,'' that is, as the cause of the breaks on the date with the highest breakpoint count. The corresponding count on that particular change date is then decremented for all of the perpetrator's neighbors, and the process is repeated using the updated shift–date tallies. The procedure continues recursively until no station/shift date count is greater than one for any station/date in the period of record.

### e. Assignment of undocumented changepoint dates

Although undocumented shifts are assigned to a perpetrating series in step d, the date of an undocumented changepoint returned by the SNHT is subject to sampling variability. As illustrated in Fig. 2, the degree of this sampling variability is a function of the magnitude of changepoint, with larger changepoints associated with more precise estimates of the date of change.

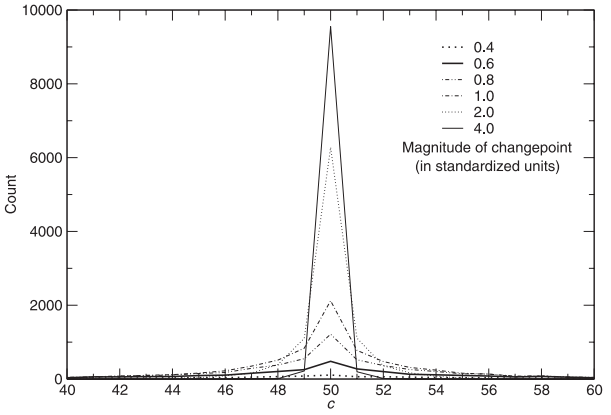


FIG. 2. Histogram of the most likely changepoint date identified by the SNHT for 10 000 series with $n = 100$ and a step change $\Delta$ at position 50. The magnitude of $\Delta$ was varied systematically from 0.2 to 4.0.

This means that testing a group of target–neighbor difference series often leads to a range of undocumented changepoint dates clustered around the time of the actual change. The simulations summarized in Fig. 2 were used to estimate the confidence limits of a changepoint date as a function of the magnitude of step change.

To determine which change dates likely refer to the same discontinuity, an interim estimate of step-change magnitude is therefore necessary. The estimate is calculated using the most appropriate change model (M3, M4, or M5) according to the BIC, which is used to determine the range of uncertainty for a particular undocumented shift date. The cluster of dates falling within overlapping confidence limits is then conflated to a single date at the target in one of two ways: 1) it is assigned to the date of a known event in the target station's history that occurs within the confidence limits for a shift of that magnitude, or 2) it is assigned to the most common changepoint date that falls within these simulated confidence limits, which means that the discontinuity appears to be truly undocumented.

### *f. Calculation of adjustments*

Steps a–e are necessary simply to identify undocumented changepoints in all temperature series. In many applications, however, station histories also may be available, which might provide additional information regarding possible discontinuities. When available, the dates of documented events should be combined with evidence of undocumented changepoints because the impact of documented events may be too subtle for the tests for undocumented shifts to detect. Potential adjustments can then be calculated for all undocumented and documented shifts at the same time.

Adjustments are determined by calculating multiple estimates of $\Delta$ using segments from neighboring series that are homogeneous for at least 24 months before and after the target changepoint. (When two changepoints occur within 24 months in the target series, an adjustment is made for their combined effect.) The range of pairwise estimates for a particular step change is considered to be a measure of the confidence with which the magnitude of the discontinuity can be estimated. As in step e, the step model found to be most appropriate (i.e., M3, M4, or M5) according to the BIC can be used to calculate a final estimate of the shift for each relevant $\{D_t\}$ segment to avoid biased estimates of $\Delta$ when a relative trend is also present. At least three separate pairwise estimates of step-change magnitude are required for each target changepoint because the distribution of estimates is used to determine the significance of the adjustment (when fewer than 3 estimates are available, the shift is considered "unadjustable"). Moreover, because the distribution of step-change estimates is not necessarily symmetric, the median estimate is used to adjust the target series.

The consistency of the pairwise estimates for $\Delta$ is determined by comparing the median estimate to either the 5th percentile (median $> 0$) or to the 95th percentile (median $< 0$) of all estimates, subject to an initial outlier check. Because fewer than 20 estimates may be available for any given changepoint, a multiple of the difference between the median and the first quartile ($Q_1$) or between the median and third quartile ($Q_3$) serves as an estimate of the 5th or 95th percentile, respectively. A factor of 2.5 is used because it approximates a one-tailed test at the 5% ($\alpha = 0.05$) significance level (assuming independent estimates). When the median and the tail of the distribution closest to zero are of the same sign (i.e., median $-$ Q1 $\times$ 2.5 or median $+$ Q3 $\times$ 2.5), the step change is considered to be significant, and an adjustment is made to the target series. This approach is similar to the Tukey (box plot) outlier test (Tukey 1977), but allows for asymmetry in the distribution of estimates. Alternatively, one could simply use the median $\Delta$ estimate when all estimates are of the same sign. Both approaches appear to yield comparable results.

### *g. Example of changepoint detection and adjustment*

Application of the pairwise algorithm to the group of series shown in Fig. 1 revealed two significant changepoints in Chula Vista maximum temperatures, both of which were associated with documented station moves, first on 1 January 1982 and then again on 25 April 1985. Difference series between the pairwise-adjusted mean monthly maximum temperatures for Chula Vista and its

neighbors are shown in Fig. 3, which suggests that the algorithm has removed the major step inhomogeneities from all series in the group.

## 4. Evaluation of the algorithm

To evaluate the performance of the pairwise algorithm more generally, temperature series were simulated under a number of trend and step-change scenarios. The simulations were designed to test the skill of changepoint detection as well as to facilitate comparison of the results to previous investigations regarding the use of a reference series as well as the identification of the type of changepoint.

### *a. Evaluation under monthly temperature simulations*

The performance of the pairwise algorithm was first evaluated using two different sets of simulated monthly temperature anomalies. One set was comprised of series with step changes, while the second set contained series with both trend and step inhomogeneities. Both sets consisted of 1000 groups of 21 correlated ''red noise'' series generated as in MW05. The average correlation between each series within a group was about 0.7. For all series the mean ($\mu$) was zero and the standard deviation ($\sigma$) was one; the number of values in each series ($n$) was equal to 1200, the equivalent of 100 yr of monthly means.

A random number of step changes was imposed on each series at random dates. The number of steps per series varied symmetrically about a peak frequency of 5, with as few as 0 and as many as 10. The magnitude of each step change was also assigned randomly by sampling from the standard normal distribution, which means that about two-thirds of the imposed steps were equal to one $\sigma$ or less. As discussed in MW05, the standard normal distribution is a good proxy for the distribution of known impacts to U.S. temperature series (Karl and Williams 1987). All imposed step changes were treated as undocumented, and 10 neighbors were identified by the pairwise algorithm for all 21 series in the groups.

In the ''monthly steps and trends'' simulations, a trend inhomogeneity was added to roughly 60% of the simulated series. The magnitude of this trend was varied randomly from $0.001\sigma$ month$^{-1}$ up to about $0.18\sigma$ month$^{-1}$, while the trend interval varied randomly from 2 months up to the full period of record. Usually the trend inhomogeneity did not initiate with a step change, although steps frequently occurred randomly within the intervals of a creeping inhomogeneity. In total, about 25% of all series segments were characterized by a trend.

Figure 4 illustrates the impact of random step-only shifts on one group of simulated series. Prior to imposing step changes, the true trend in each series was zero. After imposing shifts, the trends ranged from $-7.62\sigma$ century$^{-1}$ to $+4.34\sigma$ century$^{-1}$. The pairwise algorithm correctly identified 34 of the 43 imposed step changes. Of the nine shifts not identified, six had a magnitude of less than $0.3\sigma$, which is below the sensitivity of most tests for undocumented changepoints (DeGaetano 2006; Ducré-Robitaille et al. 2003). Furthermore, the largest undetected changepoint ($+0.696\sigma$) was preceded 10 time steps earlier by another undetected changepoint of $-0.451\sigma$; that is, the two changepoints essentially masked one another. The overall effectiveness of the pairwise adjustments is evident in Fig. 5, which depicts the 10 series after homogenization by the pairwise algorithm. Note that changepoints have been adjusted relative to the latest mean level in each series, the convention in climate data homogenization. In general, the adjusted series all have trends much closer to the true ''climate'' trend of zero.

Table 2 more generally summarizes the detection skill of the pairwise approach for both the step-only and the step-/trend-change scenarios. The hit rate (the ratio of the number of changepoints correctly identified relative to the total number imposed) is roughly 67% for both scenarios. The false-alarm rate (FAR; the ratio of falsely detected changepoints to the total number detected) is 6.77% for the step-only scenario (only slightly higher than the expected type-I error rate at the $\alpha = 0.05$ significance level) and 19.65% for the step/trend scenario. The increase in false alarms when trend inhomogeneities are present occurs for two main reasons. First, the beginning or end of a trend inhomogeneity is often identified as a step change by the pairwise algorithm. Second, short interval trends of about 24 months or less tend to be virtually indistinguishable from step changes and are therefore adjusted as an abrupt change. Indeed, the largest magnitude false alarms under the steps-and-trend inhomogeneity simulations result from short interval, but large magnitude trend inhomogeneities that are approximated by a step change.

Histograms indicating the magnitude of hits, misses, and false alarms for the step-only and step/trend simulations are shown in Figs. 6 and 7, respectively. In both cases, changes in excess of $0.5\sigma$ are readily detected, and most misses are generally less than $0.5\sigma$. The number of false alarms is also generally small, suggesting that they will have little impact on the homogenized trends for the simulated series.

Regarding the series trends, two measures of error are provided in Table 2. The first is the root-mean-square error (RMSE) for a trend calculated using the unadjusted series and the second is the RMSE for trends calculated using the adjusted series. As shown in the

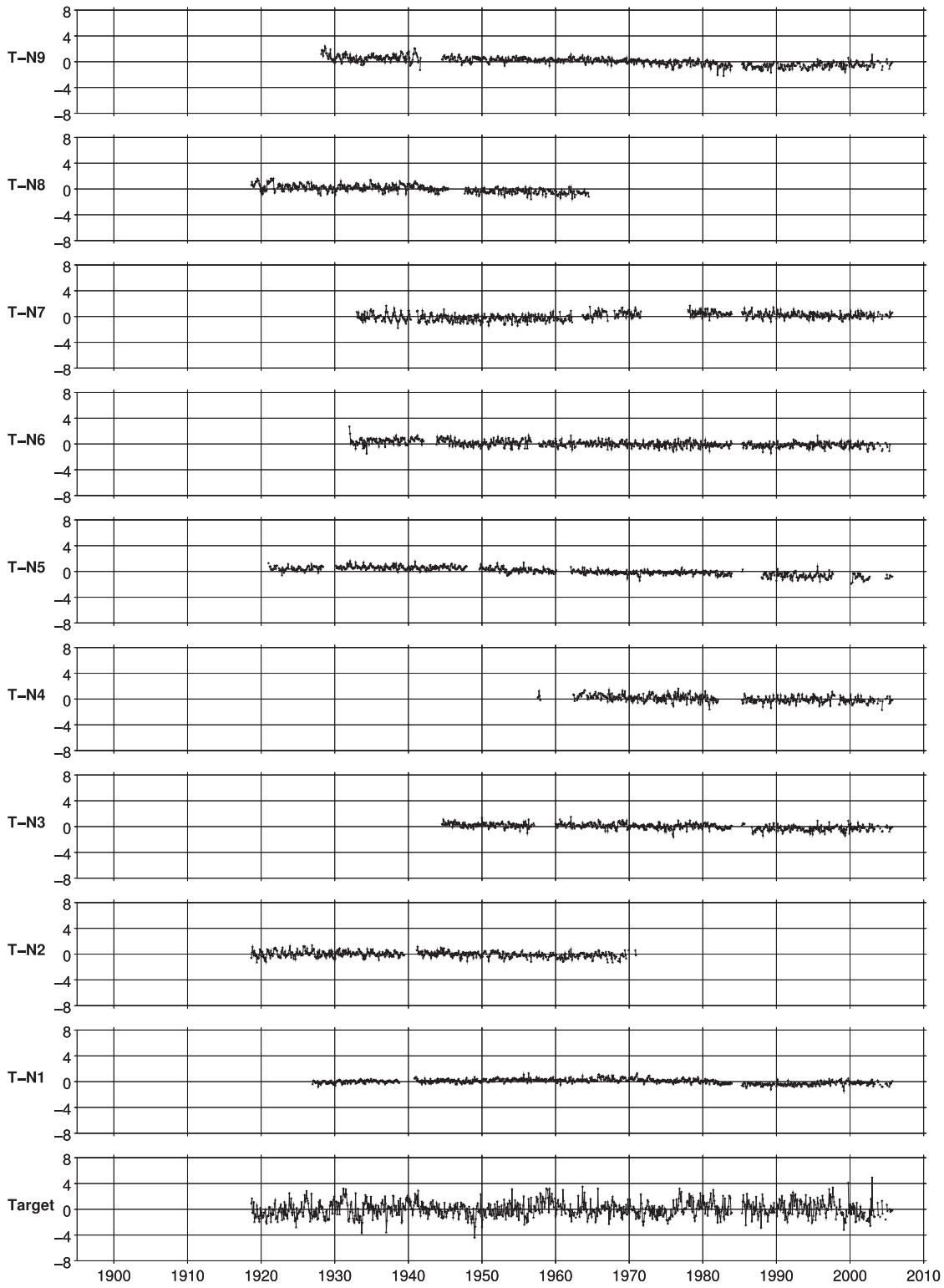


FIG. 3. As in Fig. 1, following adjustments by the pairwise algorithm.

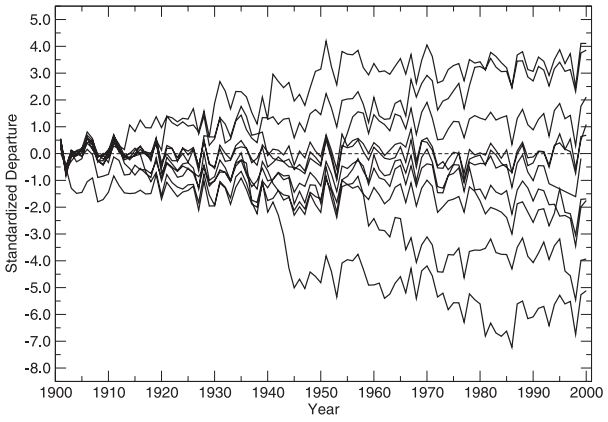


FIG. 4. ''Annual'' averages of simulated monthly series with a random number of changepoints imposed at random times and with random magnitudes. The true trend in all 10 correlated series is zero. Simulations are treated as beginning in January 1901 and ending December 2000.

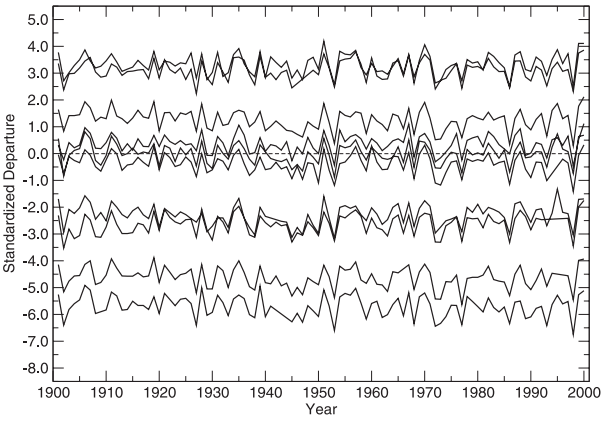


FIG. 5. As in Fig. 4, after homogenization by the pairwise algorithm.

table, the pairwise homogenization process greatly reduces the error associated with the calculation of the true background climate trend. Table 2 also indicates that the RMSE for changepoint estimates in series with trends is about as good as in the series with no trend inhomogeneities, which suggests that the model identification is reasonably successful at identifying step changes that occur within local trends. A more thorough assessment of changepoint-type identification is provided in section 4c.

### *b. Pairwise versus reference series changepoint detection skill*

The use of a reference series is the most widely employed approach to relative changepoint detection, and MW05 evaluated the implications of such an approach for undocumented changepoint detection. The pairwise approach was therefore evaluated using the same simulations and scenarios as in MW05 to directly compare its skill of undocumented changepoint detection against the reference series approach. Table 3 depicts the seven scenarios evaluated in MW05. Each case was comprised of 1000 groups of six correlated series (one target and five neighbors) with $n = 100$ values. Of the three reference series formulations evaluated by MW05, the one based on a correlated weighted average of the five neighbors (Alexandersson and Moberg 1997) is compared here. As in the pairwise algorithm, the SNHT was used to test the target-minus-weighted-average reference $\{D_t\}$ series ($\alpha = 0.05$). All changepoints detected in the $\{D_t\}$ series were attributed to the target series to test the consequences of assuming reference series homogeneity.

Table 4 summarizes the pairwise and reference series detection skill for the MW05 target series. Two statistics are presented for each case: the FAR (previously described) and the correct changepoint (CRC) power statistic (R07), which is the percentage of time that either (a) the changepoint date in the target series was selected within $\pm 2$ time steps of the correct date or (b) the target was correctly identified as homogeneous. Basically, the CRC is synonymous with hit rate except that it also credits the number of times that the target series was successfully identified as homogeneous.

In general, the pairwise algorithm has a much higher success rate in identifying homogeneous target series than the reference series approach as indicated by the higher CRC percentages for cases 1, 3, and 5. This is true when the neighbor series are themselves homogeneous as in cases 1 and 5, but especially when all the neighbors have changepoints as in case 3, which cause numerous inhomogeneities in the reference series. More generally, Table 4 indicates the degree to which

TABLE 2. Changepoint detection and magnitude estimation skill for monthly temperature. RMSE of $\Delta$ and $\beta$ expressed in standardized units ($\sigma$). The RMSE of $\beta$ is calculated with respect to the true trend of zero.

| Case study | HR (%) | FAR (%) | RMSE of $\Delta$ ($\sigma$) | RMSE of $\beta$ for unadjusted values ($\sigma$) | RMSE of $\beta$ for adjusted values ($\sigma$) |
|---|---|---|---|---|---|
| Monthly data with step changes | 67.11 | 6.77 | 0.284 | 2.455 | 0.401 |
| Monthly data with step and trend changes | 67.56 | 19.65 | 0.313 | 2.899 | 0.757 |

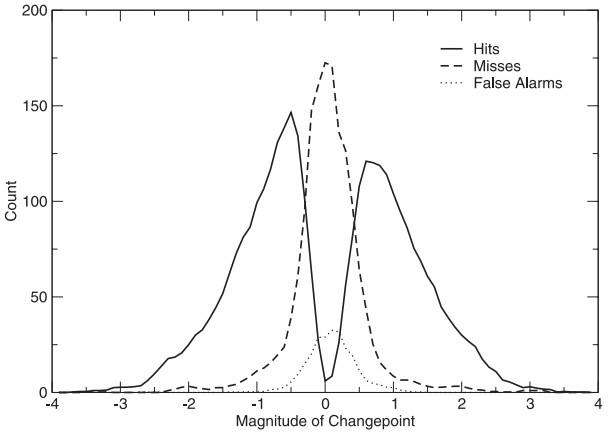


FIG. 6. Changepoint detection results for the monthly "step only" simulations.

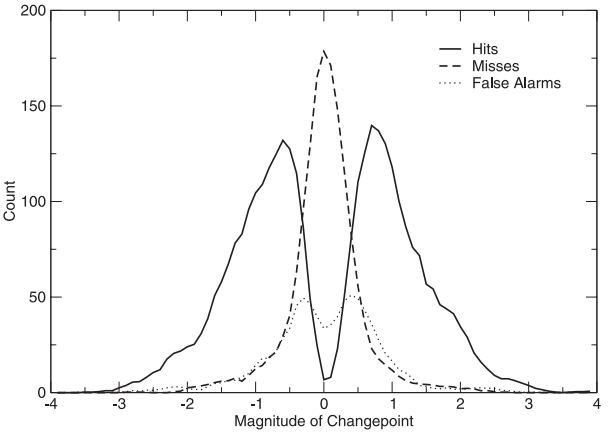


FIG. 7. Changepoint detection results for the monthly steps and trends simulations.

the pairwise approach limits the number of false alarms whenever the neighboring series are impacted by undocumented changepoints as evidenced by the low FAR for cases 4 and 6 relative to the reference series approach.

As shown in Fig. 8, the pairwise hit rate meets or exceeds that of the reference series approach when there are at least seven viable neighbors available at all times during a target station's history. (This is the foundation for the number of neighbors selected for comparison as described in section 3a.) The relatively steep increase in the power of detection as the number of comparisons increases illustrates an advantage of pairwise testing, namely, that there are multiple chances to detect a changepoint in any particular target series. If the SNHT misses a changepoint in one target–neighbor difference series, or if it misidentifies the date, there are a number of additional chances to test for the same undocumented break. The chances are not completely independent, however, because any two $\{D_t\}$ series with a common target will have an expected correlation of 0.5 (Menne and Duchon 2001). Moreover, the power of pairwise detection can be further improved by increasing the sample size between changepoints, which can be achieved by testing serial monthly values rather than annual or seasonal averages. This accounts for the higher hit rate in the "monthly" simulations, that is, 67% (Table 2) compared to the rate of a little less than 50% shown in Fig. 8 when 10 neighbors are available.

### c. Skill in identifying the type of changepoint

The magnitude of a step change will not be accurately estimated if the type of changepoint has been misidentified. Consequently, the skill of the pairwise algorithm in classifying changepoint type was assessed for the range of models in Table 1. As in section 4b, a set of 1000 groups of target and neighbor series with $n = 100$ values were used for each scenario. In this case different magnitudes of trend and step parameters, that is, $c$, $\Delta$, $\beta$, $\beta_1$, and $\beta_2$, were imposed on the target series as shown in Table 5; the five neighbor series, in contrast, were always homogeneous (M1). The magnitudes of the parameters imposed on the target series were the same as those used by R07, although only a portion of the results are summarized here.

A comparison of the CRC's in Table 5 for the $\Delta = 1\sigma$ simulations indicates that the pairwise algorithm correctly identified more than 85% of these step changes regardless of whether the target series followed M3, M4, or M5. Moreover, the algorithm also correctly identified more than 85% of the M2 (constant trend) target series as homogeneous (no steps). On the other hand, there is

TABLE 3. Number of changepoints imposed on each target and/or neighbor series for various case studies. The cases comprise 1000 simulations of six correlated series with $n = 100$ as described in MW05.

| Scenario | Number of imposed changepoints | |
|---|---|---|
| | Target series | Each neighbor series |
| Case 1 (null case) | 0 | 0 |
| Case 2 | 2 | 0 |
| Case 3 | 0 | 2 |
| Case 4 | 2 | 2 |
| Case 5 (null case with missing values) | 0 | 0 |
| Case 6 | 0–6* | 0–6* |
| Case 7 | 6** | 0 |

* The number of changepoints in each series is symmetrically distributed about a peak frequency of 3.

** Changepoint position and magnitude are fixed as in Caussinus and Mestre (2004): +2.0 at $c = 20$, +2.0 at $c = 40$, −2.0 at $c = 50$, −2.0 at $c = 70$, +2.0 at $c = 75$, and +2.0 at $c = 85$.

TABLE 4. Skill scores from the pairwise homogenization algorithm for the case studies described in Table 3. The subscripts "pw" and "ref" refer to the pairwise and reference series approaches, respectively.

| Case study (and scenario description) | $CRC_{pw}$ (%) | $FAR_{pw}$ (%) | $CRC_{ref}$ (%) | $FAR_{ref}$ (%) |
|---|---|---|---|---|
| Case 1 (homogeneous target and neighbor series) | 99.5 | 100.0 | 88.8 | 100.0 |
| Case 2 (two random changepoints in target; homogeneous neighbor series) | 44.0 | 5.6 | 55.4 | 21.0 |
| Case 3 (homogeneous target series; two random changepoints in each neighbor series) | 95.2 | 100.0 | 0.0 | 100.0 |
| Case 4 (two random changepoints in all series) | 37.3 | 8.5 | 50.3 | 46.0 |
| Case 5 (homogeneous target and neighbor series with missing values) | 100.0 | Undefined (zero false alarms) | 87.2 | 100.0 |
| Case 6 (up to six changepoints in all series) | 31.6 | 7.0 | 45.4 | 41.0 |
| Cse 7 (six changepoints in target [$\Delta=2\sigma$]; homogeneous neighbors) | 84.6 | 1.1 | 70.4 | 6.0 |

more variability in the skill of classifying the type of changepoint as indicated by the correct type percentages shown in bold. The percentages indicate that the algorithm had somewhat less success in classifying M4- and M5-type changepoints relative to M3-type changepoints and series that follow M2.

Under the M2 scenarios, the pairwise algorithm correctly classified more than 85% of the $\{D_t\}$ series when $\beta$ was greater than or equal to 0.01 (a slope yielding a change of $1\sigma$ in 100 time steps), but less than 50% when $\beta = 0.005$ (a change of $0.5\sigma$ in 100 time steps). The reason for the difference is that the BIC does not always distinguish a sloped line from a flat line when $\beta$ is small. This kind of misclassification, however, does not impact the CRC because there is no step change assigned to the target. On the other hand, when $\beta$ is larger, the SNHT tends to partition the $\{D_t\}$ trend into one or more step-type changes. The BIC correctly reclassifies most of these breaks as M2, but also cannot always distinguish a trend (M2) from a step change (M3, M4, or M5). Consequently, the pairwise algorithm classifies only 91% of M2 target series as homogeneous (no step) when $\beta = 0.01$ and 86.9% when $\beta = 0.02$. The impact of this type of misclassification is to inadvertently remove some of the unique target series trend as a step adjustment, thereby bringing the target series more in line with the regional background climate trend captured by the neighbors (DeGaetano 2006; Pielke et al. 2007).

For target series under M3 (step change with no trend), the overall power of detection is a function of the magnitude of the step, as shown in previous investigations (e.g., DeGaetano 2006). In the pairwise algorithm, most (> 88%) of the $\{D_t\}$ series with a step change of $1\sigma$ or greater were correctly identified as M3, and the CRC exceeds 90% in such cases. On the other hand, many (about 45%) of the $0.5\sigma$ magnitude step changes are misclassified as a trend change (M2).

When the target series follows M4 (step change within a constant trend), the pairwise CRC varies between 85% and 90% for the $1\sigma$ step changes, close the M3 rate. However, in the M4 simulations, the algorithm frequently (about 80% of the time when $\beta = 0.005$) misclassifies the $\{D_t\}$ series as M3, especially when $\beta$ is small. This type of misclassification also leads to a biased estimate of the magnitude of the jump by aliasing the unique target trend on to the estimate of the step change. Much like a false alarm when the target follows M2, the biased estimate would bring the adjusted target more in agreement with the background trend captured by the neighbors (DeGaetano 2006; Pielke et al. 2007).

Under M5, the target series has a step change within a trend change, but there is also a change in trend coincident with the step. In this scenario, the CRCs are comparable to the M4 simulations, but in this case, the pairwise algorithm tended to misclassify the $\{D_t\}$ series as M3 or M4 in roughly equal proportions. Consequently, some of the target series trends would be

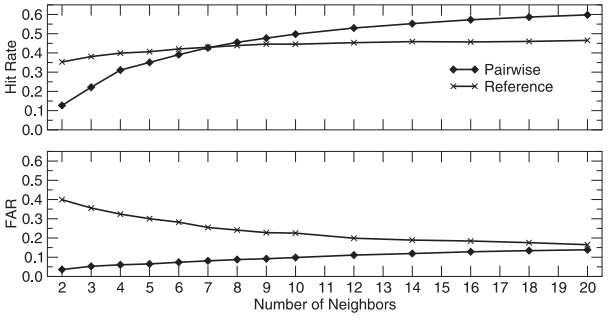


FIG. 8. Relationship between the hit rate (HR) and FAR for changepoints attributed to the target series as a function of the number of neighbors used to compute a composite reference series or in pairwise comparisons. Results are based on 1000 groups of series ($n = 100$) simulated under case 6 (between 0 and 6 random changepoints added to the target and all neighbor series).

TABLE 5. Changepoint detection and model identification results (%) for 1000 sets of five target–neighbor difference ($\{D_t\}$) series ($n = 100$). Parameters were added as indicated to the target series and $c = 50$ for the target simulated under M3, M4, and M5. The neighbor series always followed M1 (constant mean with no breaks). CRC refers to the pairwise algorithm's detection results for the target series. The percentage of $\{D_t\}$ identified correctly is given in bold.

| Target series follows M2 | | | | | | | |
|---|---|---|---|---|---|---|---|
| | $\beta$ | M1 | M2 | M3 | M4 | M5 | CRC |
| | 0.005 | 51.25 | **44.36** | 3.86 | 0.24 | 0.30 | 95.70 |
| | 0.010 | 2.45 | **88.23** | 7.30 | 0.72 | 1.31 | 91.00 |
| | 0.020 | 0.55 | **85.75** | 3.48 | 4.56 | 5.67 | 86.90 |
| Target series follows M3 | | | | | | | |
| $\Delta$ | | M1 | M2 | M3 | M4 | M5 | CRC |
| 0.5 | | 11.71 | 45.16 | **39.03** | 1.66 | 2.44 | 30.30 |
| 1.0 | | 0.06 | 4.83 | **88.59** | 1.82 | 4.70 | 90.90 |
| 2.0 | | 0.00 | 0.10 | **93.21** | 1.85 | 4.85 | 99.90 |
| Target series follows M4 | | | | | | | |
| $\Delta$ | $\beta$ | M1 | M2 | M3 | M4 | M5 | CRC |
| 1.0 | 0.005 | 0.04 | 6.56 | 80.08 | **5.24** | 8.08 | 89.30 |
| 1.0 | 0.010 | 0.06 | 7.32 | 51.07 | **24.56** | 16.99 | 87.90 |
| 1.0 | 0.020 | 0.05 | 7.46 | 19.75 | **52.87** | 19.87 | 85.50 |
| Target series follows M5 | | | | | | | |
| $\Delta$ | $\beta_1$ | $\beta_2$ | M1 | M2 | M3 | M4 | M5 | CRC |
| 1.0 | 0.010 | 0.015 | 0.11 | 8.16 | 34.84 | 33.17 | **23.72** | 87.11 |
| 1.0 | 0.010 | 0.020 | 0.07 | 8.45 | 25.17 | 33.51 | **32.79** | 86.20 |
| 1.0 | 0.010 | 0.030 | 0.07 | 7.47 | 19.34 | 23.22 | **49.91** | 85.70 |

aliased onto the estimate of the M5 step changes, as in the case of the M4 target series simulations.

Overall, the results in Table 5 are consistent with the changepoint-type identification capabilities of the generalized methods investigated by R07, namely, that it is more challenging to classify M4- and M5-type changepoints. As shown in R07, the lower identification skill occurs even when changepoint tests specifically designed for these types of change are used, that is, Wang (2003) for M4 and Lund and Reeves (2002) for M5. Nevertheless, from Table 5 and results (not shown) based on directly testing a target series as in R07, it appears that the pairwise approach (SNHT plus BIC) has comparable skill at model identification compared to the methods evaluated by R07. The advantage of the pairwise approach is that the SNHT's superior power of detection is exploited.

The skill of identifying changepoint type, like the power of detection, can also be improved by increasing the sample size of the test series, that is, by testing serial monthly series. For example, the percentage of correctly identified M4 difference series is about 70% at $\beta = 0.02$ when $n = 240$ and $c = 120$ versus 50% for $n = 100$ and $c = 50$. Similarly, when $\beta_1 = 0.01$ and $\beta_2 = 0.03$ under M5, the percentage of series correctly identified increases to 75% for $n = 240$ versus about 50% for $n = 100$. In addition, the skill of changepoint detection and identification increases with increasing correlation between series, which reduces the variance of the $\{D_t\}$ series. As noted by DeGaetano (2006), the correlation between temperature series in the United States is typically higher than in the simulations used here.

## 5. Application to U.S. temperature series

A number of recent studies have focused on the impact of land use change on the temperature record (e.g., Peterson and Owen 2005; Kalnay et al. 2006; Parker 2006; Pielke et al. 2007), yet no general assessment of the frequency of the various types of changepoints in observed temperature series has been conducted. For this reason, the pairwise algorithm was applied to monthly temperature series from the Coop Network in order to assess relative frequency of the type of inhomogeneity (including local trends) in U.S. temperature records. Monthly mean maximum and minimum values from over 7000 stations covering the period from 1895 to 2006 were used, although the specific period of record varied from station to station. The nature of the shifts for a commonly used subset of the Coop network, that is, the U.S. Historical Climatology Network (HCN; Easterling et al. 1996) was examined in detail.

An analysis of the more than 100 000 $\{D_t\}$ series segments used to calculate the shift magnitudes for HCN temperature series indicates that about 50% of the step changes follow M3 (step change with no trend), while approximately 40% follow M5 (step change accompanied by a trend change) and about 10% follow M4 (step change within a general trend). While these percentages were calculated on a segment-by-segment basis, the models M4 and M5 also minimized the BIC statistic about 50% of the time when calculated across each series' full period of record (shown in Table 7). In other words, the trend models appear to be a better fit about 50% of the time even for observed $\{D_t\}$ that are generally decades long and incorporate shifts identified in both HCN targets and their Coop neighbors (and are thus highly penalized by the BIC).

To further evaluate the pairwise adjustments for these types of shifts, the adjusted series were also manually inspected. In brief, this entailed graphing each HCN series and its Coop neighbors as in Fig. 3, and then subjectively deeming the adjusted series as plausible or implausible. This subjective evaluation revealed that roughly 15%–20% of the adjusted series exhibited

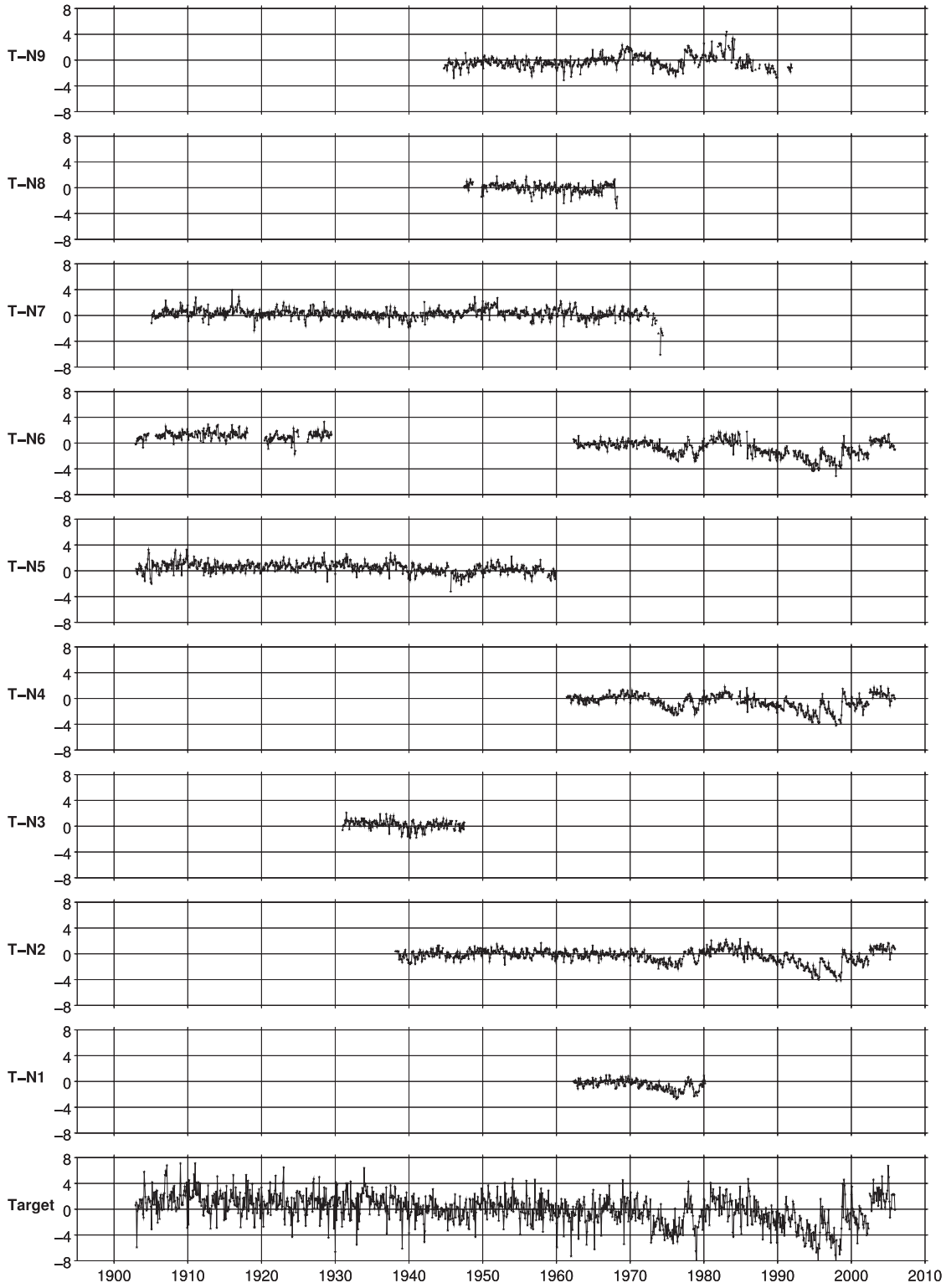


FIG. 9. Mean monthly minimum temperature anomalies (°C) for Cheesman (target) and differences between monthly temperature anomalies at Cheesman and nine neighboring series (T-N1 to T-N9).

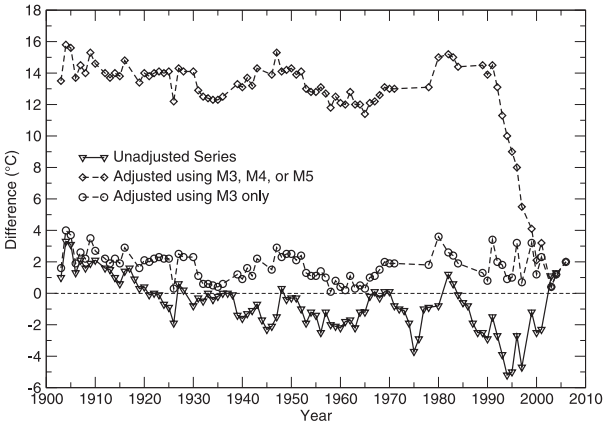


FIG. 10. Differences between annual minimum temperatures at Cheezman and 20 neighboring stations, and following adjustments for step changes using the most appropriate model determined by the pairwise algorithm (M3, M4, or M5) and using M3 only.

physically unrealistic trends that were clearly inconsistent with neighboring stations. The minimum temperature series at Cheesman, Colorado, is an extreme example. As shown in Fig. 9, a sawtooth pattern is evident in the $\{D_t\}$ series formed between the Cheesman series and its neighbors. The increasing difference between Cheesman and the surrounding stations (particularly after 1980) sometimes exceeded 4°C in 5 yr, a relative change that was easily classified as M5 (step change with a trend change) by the pairwise algorithm. The consequence of adjusting the series using M5 (i.e., removal of the step and retention of the trend) is shown in Fig. 10. The result is clearly unrealistic.

Given that preserving local trends (i.e., trend inhomogeneities) can often result in undesirable adjusted series, the pairwise algorithm was modified to employ the more commonly used M3 adjustment for all step changes (DeGaetano 2006). (Note that M3, M4, and M5 were still employed to detect step changes.) The impact of the M3-only approach on the Cheesman series is also shown in Fig. 10. Although the sawtooth signature remains in the adjusted data, the trend at Cheesman using the M3 adjustments is clearly in sync with the average of trends in surrounding series. A similar visual inspection of all HCN temperature series suggests that an M3-only adjustment approach works well for all situations in which there is evidence of a step change because any associated trend inhomogeneity is consistently aliased onto estimates of the step change in a way that favors the background climate signal.

The same result occurs when M3 alone is used to adjust the simulated series in the ''monthly steps and trends'' simulations, as shown in Table 6. From a comparison of the RMSE for the adjusted trends in Tables 2 and 6, it is evident that using M3 for all step-change adjustments removes the impact of most trend inhomogeneities because the error for the adjusted trends is roughly the same for the step-only and steps and trends simulations. Still, while the temperature series that result using the M3-only adjustments arguably approximate the best theoretical climate series for each location, the local trend signal is nevertheless aliased out of the original series, thus limiting the use of the adjusted series in some attribution studies of observed temperature change. Ultimately, a better solution would be to remove trend inhomogeneities via trend adjustments and step inhomogeneities via step adjustments. Unfortunately, unlike step changes that occur at the same time within a group of target/neighbor $\{D_t\}$ series, a trend inhomogeneity at a given target station may begin and end at different times with respect to each of its neighbors. This makes identifying the true interval of trend inhomogeneity more difficult than detecting step changes, and is beyond the scope of this paper.

In any case, another reason to use only M3-type adjustments is that it appears that at least some apparent trends may in fact be artifacts of unidentified step changes. This conclusion comes from an evaluation of the capability of the BIC statistic to determine the true dimensions of the simulated target-minus-neighbor period of record $\{D_t\}$ series when the shifts are treated as wholly undocumented (and identified by the pairwise algorithm) versus when the shift times are known perfectly. Table 7 summarizes the frequency that each model minimized the BIC statistic in the 420 000 unique $\{D_t\}$ series that comprise each set of monthly simulations. Based on these results, it appears that M5 rarely minimizes the BIC when there are no relative trends in the simulated data, but M4 is identified as the ''best'' model in over 16% of cases when the shifts are treated as undocumented. Conversely, when there is perfect

TABLE 6. Changepoint detection and magnitude estimation skill for monthly temperature series using a constant mean model (M3) for all step change adjustments regardless of the identified type.

| Case study | HR(%) | FAR (%) | RMSE of $\Delta$ ($\sigma$) | RMSE of $\beta$ for unadjusted values ($\sigma$) | RMSE of $\beta$ for adjusted values ($\sigma$) |
|---|---|---|---|---|---|
| Monthly data with step changes | 67.22 | 6.77 | 0.291 | 2.455 | 0.401 |
| Monthly data with step and trend changes | 67.58 | 20.14 | 0.349 | 2.899 | 0.488 |

TABLE 7. Frequency (%) that the model minimizes the BIC statistic for the period of record difference series formed between all target and neighbor series.

| Scenario | Model | | | | |
|---|---|---|---|---|---|
| | M1 | M2 | M3 | M4 | M5 |
| Observed HCN monthly temperatures (pairwise identification of steps) | 5.12 | 10.06 | 34.18 | 25.29 | 25.35 |
| Monthly step-only simulations (pairwise identification of steps) | 0.45 | 0.27 | 78.58 | 16.71 | 3.99 |
| Monthly steps-only simulations (perfect knowledge of steps) | 0.45 | 0.34 | 96.24 | 2.96 | 0.00 |
| Monthly steps and trends simulations (pairwise identification of steps) | 0.09 | 0.27 | 20.47 | 17.21 | 61.96 |
| Monthly steps and trends simulations (perfect knowledge of steps) | 0.09 | 0.41 | 41.79 | 21.72 | 35.99 |

knowledge of the timing of all shifts no matter how small, M4 rarely minimizes the BIC when only shifts occur (in this case M5 was selected as the ''best'' model in only 2 of 420 000 cases).

The sloped models more frequently minimize the BIC in the steps and trends scenarios than in the steps-only simulations. Given that in this case approximately 63% of the full $\{D_t\}$ series have a trend segment somewhere in the period of record, Table 7 suggests that the frequency of M3-type models is nevertheless underestimated when shifts are treated as undocumented (because of unidentified step changes). However, when there is perfect knowledge of all shifts in the monthly steps and trends simulations, the models minimize the BIC in way that suggests that the frequency of M3 solutions is approximately correct (although M4 is selected too often at the expense of M5). Based on these results, we conclude that, while perhaps prevalent, the frequency of apparent trend inhomogeneities in the HCN is inflated by the presence of unidentified (i.e., small and perhaps unidentifiable) step changes.

## 6. Conclusions

Our evaluation of the pairwise algorithm suggests that it is a robust, reliable, and accurate approach to detecting step-type inhomogeneities under a wide variety of circumstances. Relative to the more traditional use of a climate reference series, a pairwise approach to undocumented changepoint detection reduces the number of false alarms in general and is particularly successful at identifying homogeneous segments. In addition, unlike the reference approach, there are no requirements for a group of series to have a common base period. As a result, the estimation of step-change magnitude is not confined to the shortest homogeneous interval within a group of neighboring series. In this regard, the pairwise method is similar to the graph theory approach used by Christy et al. (2006) except that the pairwise algorithm makes no attempt to compare climate series that do not overlap in time.

Moreover, because each climate series is paired with a unique set of neighboring series in the algorithm, it is possible to determine whether more than one nearby station series shares a particular shift date because both stations will have been implicated multiple times on or about the same date. This property of the algorithm is important when a widespread and near simultaneous change in observation practice occurs in a network. Such a situation arose in the U.S. Cooperative Network when liquid-in-glass thermometers were replaced with electronic thermistors at roughly two-thirds of sites during the mid- and late 1980s (Quayle et al. 1991; Hubbard and Lin 2006). Of course, if a change is implemented on exactly the same date at all stations, relative homogeneity testing will not be effective.

Results from applying the pairwise algorithm to observed temperature series suggest that while there is evidence of relative trends between series in the U.S. surface temperature record, some apparent trends may be an artifact of unidentified, small shifts. Although there is some interest in preserving such trend inhomogeneities for land use/land change impact studies (e.g., Pielke et al. 2007), the results of this analysis indicate that physically implausible trends can result when apparent trend inhomogeneities are preserved. On the other hand, if the goal is to produce an accurate estimate of the background climate signal, all identified shifts can nevertheless be removed using the step-only model. While this necessarily leads to the aliasing of any associated trend inhomogeneity onto the estimate of the step change, a reliable estimate of the background climate signal is obtained.

Finally, we reiterate that the pairwise algorithm was designed to solve the practical problem of adjusting temperature series to remove the impacts of artificial changes in a holistic way. Because the algorithm is modular, it is possible to enhance its various components. For example, shifts in the target-minus-neighbor difference series might be resolved using optimal methods (e.g., Caussinus and Mestre 2004) and/or by incorporating tests for periodicity in the serial monthly difference series. The

latter may be important because the monthly adjustments calculated by the pairwise algorithm are currently constant for all months. Although the increased sample size afforded by testing serial monthly data likely overwhelms any benefit to testing seasonal values separately (cf. Karl and Williams 1987; Begert et al. 2005; Brunet et al. 2007), there is evidence that bias changes often have impacts that vary seasonally and/or synoptically (Trewin and Trevitt 1996; Guttman and Baker 1996). As shown by Della-Marta and Wanner (2006), it is possible to estimate the differential impacts indirectly by evaluating the magnitude of change as a function of the frequency distribution of *daily* temperatures. Such a method requires knowledge of the timing of shifts as a starting point, which can be provided by the pairwise results.

*Acknowledgments.* Thanks to Russell Vose, Imke Durre, and Tamara Houston for constructive comments on earlier drafts for this manuscript. The reviews by Dr. Robert Lund and an anonymous reviewer also greatly improved the paper. Partial support for this work was provided by the Office of Biological and Environmental Research, U.S. Department of Energy (Grant DE-AI02-96ER62276).

## REFERENCES

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.

Begert, M., T. Schlegel, and W. Kirchhofer, 2005: Homogeneous temperature and precipitation series of Switzerland from 1864 to 2000. *Int. J. Climatol.,* **25,** 65–80.

Brunet, M., and Coauthors, 2007: Temporal and spatial temperature variability and change over Spain during 1850–2003. *J. Geophys. Res,* **112,** D12117, doi:10.1029/2006JD008249.

Carretero, J. C., and Coauthors, 1998: Changing waves and storms in the northeast Atlantic? *Bull. Amer. Meteor. Soc.,* **79,** 741–760.

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

Christy, J. R., W. B. Norris, K. Redmond, and K. P. Gallo, 2006: Methodology and results of calculating central California surface temperature trends: Evidence of human-induced climate change? *J. Climate,* **19,** 548–563.

Conrad, V., and L. W. Pollak, 1962: *Methods in Climatology*. Harvard University Press, 459 pp.

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

Della-Marta, P. M., and H. Wanner, 2006: A method of homogenizing the extremes and mean of daily temperature measurements. *J. Climate,* **19,** 4179–4197.

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., T. R. Karl, E. H. Mason, P. Y. Hughes, and D. P. Bowman, 1996: United States Historical Climatology Network (U.S. HCN) monthly temperature and precipitation data. ORNL/CDIAC-87, NDP-019/R3, Carbon Dioxide Information Analysis Center, Oak Ridge National Laboratory, U.S. Department of Energy, Oak Ridge, TN, 280 pp.

González-Rouco, J. F., J. L. Jiménez, V. Quesada, and F. Valero, 2001: Quality control and homogeneity of precipitation data in the southwest of Europe. *J. Climate,* **14,** 964–978.

Guttman, N. B., and C. B. Baker, 1996: Exploratory analysis of the difference between temperature observations recorded by ASOS and conventional methods. *Bull. Amer. Meteor. Soc.,* **77,** 2865–2873.

Hanssen-Bauer, I., and E. J. Førland, 1994: Homogenizing long Norwegian precipitation series. *J. Climate,* **7,** 1001–1013.

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

Hubbard, K. G., and X. Lin, 2006: Reexamination of instrument change effects in the U.S. Historical Climatology Network. *Geophys. Res. Lett.,* **33,** L15710, doi:10.1029/2006GL027069.

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

Kalnay, E., M. Cai, H. Li, and J. Tobin, 2006: Estimation of the impact of land-surface forcings on temperature trends in eastern United States. *J. Geophys. Res.,* **111,** D06106, doi:10.1029/2005JD006555.

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.

——, H. F. Diaz, and G. Kukla, 1988: Urbanization: Its detection and effect in the United States climate record. *J. Climate,* **1,** 1099–1123.

Livezey, R. E., and W. Y. Chen, 1983: Statistical field significance and its determination by Monte Carlo techniques. *Mon. Wea. Rev.,* **111,** 46–59.

Lu, Q., and R. B. Lund, 2007: Simple linear regression with multiple level shifts. *Can. J. Stat.,* **35,** 447–458.

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

——, C. Gallagher, X. L. Wang, Y. Feng, Q. Lu, and J. Reeves, 2007: Changepoint detection in periodic and autocorrelated time series. *J. Climate,* **20,** 5178–5190.

McCarthy, M. P., H. A. Titchner, P. W. Thorne, S. F. B. Tett, L. Haimberger, and D. E. Parker, 2008: Assessing bias and uncertainty in the HadAT-adjusted radiosonde climate record. *J. Climate,* **21,** 817–832.

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

Parker, D. E., 2006: A demonstration that large-scale warming is not urban. *J. Climate,* **19,** 2882–2895.

Peterson, T. C., and T. W. Owen, 2005: Urban heat island assessment: Metadata are important. *J. Climate,* **18,** 2637–2646.

——, 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,** 25 967–25 974.

Pielke, R. A., Sr.,and Coauthors, 2007: Documentation of uncertainties and bias associated with surface temperature measurement sites for climate change assessment. *Bull. Amer. Meteor. Soc.,* **88,** 913–928.

Quayle, R. G., D. R. Easterling, T. R. Karl, and P. Y. Hughes, 1991: Effects of recent thermometer changes in the Cooperative Station Network. *Bull. Amer. Meteor. Soc.,* **72,** 1718–1723.

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

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

Slonosky, V. C., P. D. Jones, and T. D. Davies, 1999: Homogenization techniques for European monthly mean surface pressure series. *J. Climate,* **12,** 2658–2672.

Thorne, P. W., D. E. Parker, J. R. Christy, and C. A. Mears, 2005: Uncertainties in climate trends: Lessons from upper-air temperature records. *Bull. Amer. Meteor. Soc.,* **86,** 1437–1442.

Trewin, B. C., and A. C. F. Trevitt, 1996: The development of composite temperature records. *Int. J. Climatol.,* **16,** 1227–1242.

Tukey, J. W., 1977: *Exploratory Data Analysis*. Addison-Wesley, 688 pp.

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

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