Good Practices and Common Pitfalls in Climate Time Series Changepoint Techniques: A Review

ROBERT B. LUND,[a] CLAUDIE BEAULIEU,[b] REBECCA KILLICK,[c] QIQI LU,[d] AND XUEHENG SHI[e,f]

[a] *Department of Statistics, University of California, Santa Cruz, California*
[b] *Department of Ocean Sciences, University of California, Santa Cruz, California*
[c] *Department of Mathematics and Statistics, Lancaster University, Lancaster, United Kingdom*
[d] *Department of Statistical Sciences and Operations Research, Virginia Commonwealth University, Richmond, Virginia*
[e] *Department of Statistics, University of California, Davis, California*
[f] *Department of Statistics and Department of Biological Systems Engineering, University of Nebraska–Lincoln, Lincoln, Nebraska*

(Manuscript received 13 January 2023, in final form 7 June 2023, accepted 14 July 2023)

ABSTRACT: Climate changepoint (homogenization) methods abound today, with a myriad of techniques existing in both the climate and statistics literature. Unfortunately, the appropriate changepoint technique to use remains unclear to many. Further complicating issues, changepoint conclusions are not robust to perturbations in assumptions; for example, allowing for a trend or correlation in the series can drastically change changepoint conclusions. This paper is a review of the topic, with an emphasis on illuminating the models and techniques that allow the scientist to make reliable conclusions. Pitfalls to avoid are demonstrated via actual applications. The discourse begins by narrating the salient statistical features of most climate time series. Thereafter, single- and multiple-changepoint problems are considered. Several pitfalls are discussed en route and good practices are recommended. While most of our applications involve temperatures, a sea ice series is also considered.

SIGNIFICANCE STATEMENT: This paper reviews the methods used to identify and analyze the changepoints in climate data, with a focus on helping scientists make reliable conclusions. The paper discusses common mistakes and pitfalls to avoid in changepoint analysis and provides recommendations for best practices. The paper also provides examples of how these methods have been applied to temperature and sea ice data. The main goal of the paper is to provide guidance on how to effectively identify the changepoints in climate time series and homogenize the series.

KEYWORDS: Climate; Changepoint analysis; Time series

1. Introduction

Climate time series often contain sudden structural changes (shifts or changepoints) in their behavior. These shifts may reflect linear or nonlinear dynamics in the climate system and need to be identified for an accurate depiction of long-term changes in any associated climate time series (Beaulieu et al. 2012; Beaulieu and Killick 2018; Cahill et al. 2015; Mudelsee 2019). Some structural changes may be artificial discontinuities induced by changes in measurement practices (e.g., station relocations, gauge changes, observer changes) (Menne and Williams 2009; Ribeiro et al. 2016; Peterson et al. 1998; Venema et al. 2012). Some (but not necessarily all) artificial changes induce shift discontinuities into the series. If these shifts are not detected and removed from the series, conclusions about long-term trends can be biased or erroneous. Regardless of the shift cause, changepoint techniques are used to estimate the true number of structural changes and their timings. If the change is artificial, the number of changepoints and their locations are needed to adjust (homogenize) climate records a priori for realism. If the structural change is caused by natural forcings in the climate system, the number of changepoints and their timings are needed to accurately quantify long-term changes.

Changepoint detection is a rapidly growing field in the data science literature (Chen and Gupta 2012; Truong et al. 2020) and applications to climate time series are numerous. This paper contains a modern statistical review of the changepoint topic in climate settings. The overarching goal is to accurately estimate the number of changepoints and their locations, and to accessibly present the methods for the climate scientists and experts with a minimum of jargon and technicalities (some technical methods, of course, are needed). The paper intends to serve as a technical guide to changepoint detection, informing the researcher of the appropriate methods to use based on the statistical properties of the time series. Unfortunately, changepoints are a thorny modeling issue: seemingly small changes in model assumptions can yield very different conclusions (Lund and Reeves 2002; Beaulieu et al. 2012; Beaulieu and Killick 2018). Because of this, it is important that researchers be aware of common pitfalls with changepoint/

Denotes content that is immediately available upon publication as open access.

Corresponding author: Robert B. Lund, rolund@ucsc.edu

DOI: 10.1175/JCLI-D-22-0954.1

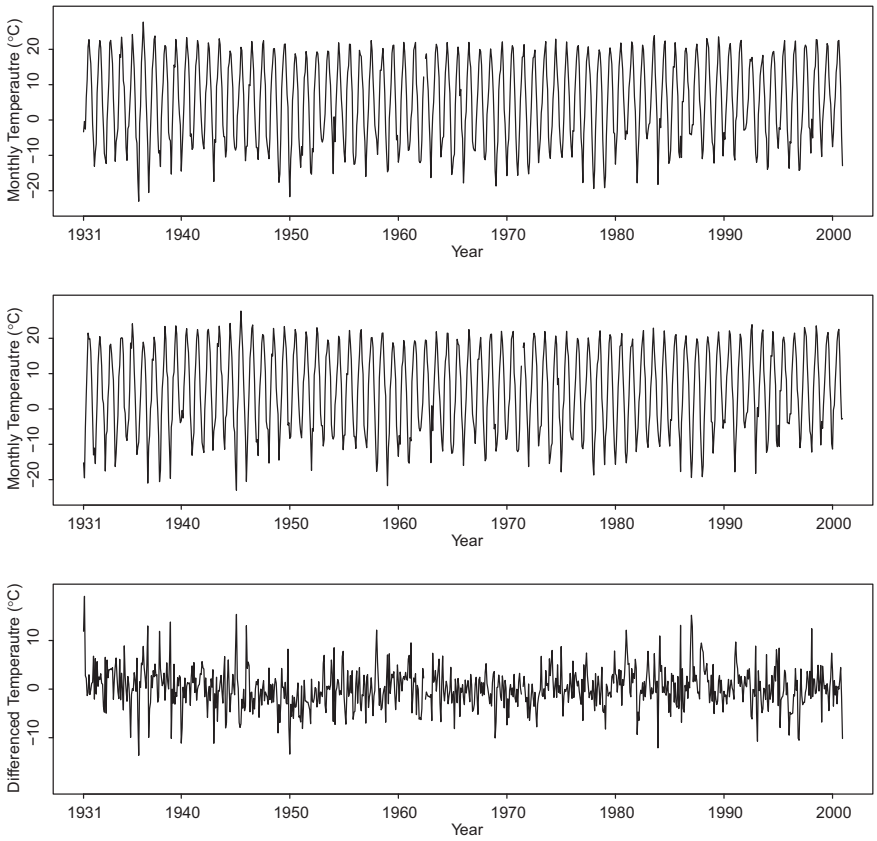


FIG. 1. Monthly averaged temperatures at the (top) Mott and (middle) Richardton-Abbey stations in west-central North Dakota. (bottom) The Mott minus Richardton-Abby series in a target minus reference subtraction.

homogenization analyses. This paper illuminates some common mistakes in the field and makes recommendations on best general practices.

Even in a review paper such as this, concessions must be made for length. In particular, this paper will not compare or classify the many software packages used today to homogenize climate time series; see Ribeiro et al. (2016) and Domonkos et al. (2021) for such lists. Indeed, our focus is on the techniques themselves, the intent being to illuminate the concepts that underlie sound changepoint analyses. Also, the paper will not delve into attribution of any discovered changepoints in our examples—what caused the changepoints is immaterial in this discussion. Toward this, most homogenizations aim to remove artificial changepoint features from the record (e.g., station moves); changepoints reflecting "true fluctuations" (e.g., natural variability) should be retained in the series. This can be done by subtracting a reference series from a nearby location from the target series to be homogenized before analysis. The target minus reference subtraction eliminates naturally occurring fluctuations in the series being analyzed and can reduce the correlation present. These so-called absolute versus relative homogenization procedures, and the "target" and "reference" series involved in them, are discussed in Menne et al. (2009). More is said about these in the next section. Finally, our analysis of some series may employ suboptimal assumptions at times. This is primarily done to show that different assumptions can produce very different changepoint conclusions. We rehash this issue in the discussion, indicating which features seem important for each series that is scrutinized for changepoints in this paper.

The rest of this paper proceeds as follows. The next section discusses the statistical properties of typical climate time series, delving into correlation, trends, seasonality, and changepoints. Here, target and reference series are introduced and absolute versus relative homogenization procedures are distinguished. Section 3 introduces a time series regression model that describes a wide suite of climate series. This model provides the mathematical backdrop for our discourse. Section 4 considers the case of a single changepoint, presenting what is generally viewed as the best (most powerful) single-changepoint detector. Section 5 moves to multiple-changepoint cases, which arise when the number of changepoints is a priori unknown, the typical setting in practice. Section 6 closes with conclusions and comments, including some remarks about future research.

2. Statistical properties of climate time series

Figure 1 presents 71 years of monthly averaged temperatures from two nearby stations in west-central North Dakota: Mott and Richardton-Abby. These stations are in the U.S. Historical Climatology Network (USHCN) database and can

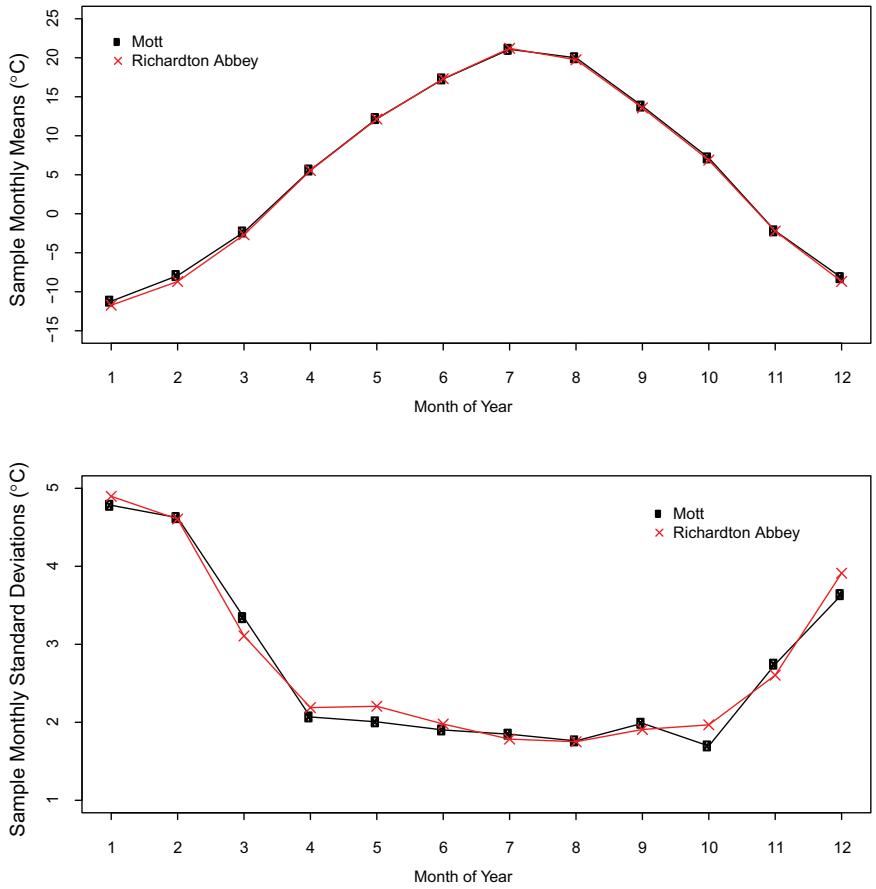


FIG. 2. (top) Monthly sample means and (bottom) standard deviations for the Mott and Richardton-Abbey stations.

be downloaded at https://www.ncei.noaa.gov/cdo-web/search. We consider the January 1931–December 2001 subspan of their records. These series will be used to illustrate our list of salient statistical features in climate series, which is needed to construct an accurate estimated changepoint configuration.

a. Seasonality

A prominent seasonal mean cycle exists in the plotted data in Fig. 1. In fact, the yearly range of the monthly sample means exceeds 30°C: from a January minimum of less than −10°C to a July maximum of more than 20°C. This seasonal cycle can visually mask some small shifts, say on the order of a degree or two (the typical discontinuity magnitude induced by a changepoint), in the record. These small shifts become critical when assessing long-term changes in temperatures. Figure 2 shows the sample means and standard deviations for each month of the data in Fig. 1—they are close to one another.

Seasonality is also present in the variability of many climate series. The sample standard deviations (or equivalently, the square root of the variabilities) in Fig. 2 show that winter temperatures are much more variable than summer temperatures; examples exist of stations in the temperate zone where January standard deviations are roughly 5 times larger than July standard deviations (Lund et al. 1995). This same paper shows that a stationary time series model modified to allow for periodicities in mean and variance adequately describes many periodic climate series $\{X_t\}$:

$$X_{nT+\nu} = \mu_\nu + \sigma_\nu \epsilon_{nT+\nu}. \tag{1}$$

Here, our notation has $X_{nT+\nu}$ as the series observation during the νth phase (season) of the nth data cycle, T is the known period ($T = 12$ for monthly data; $\nu \in \{1, \ldots, 12\}$ refers to a specific month), $\{\epsilon_t\}$ is a zero-mean unit variance stationary time series in time t ($t = nT + \nu$), and σ_ν is the standard deviation of the data at phase ν within the cycle. Trends and changepoint features are neglected (for the moment) in the above model.

Seasonal features complicate changepoint detection when not taken into account. Elaborating, it can be difficult to visually discern the impact of a changepoint in a plotted temperature series, which typically shifts a series only by a degree or two, when the series has a seasonal cycle magnitude of say 30°. Figure 3 demonstrates this by adding a 2°C mean shift to the Mott series at time index 600. It is harder to see this shift in the series containing the seasonal cycle, becoming easier to see after the seasonal cycle has been removed. In a multiple-changepoint analysis of a daily series, methods may flag many

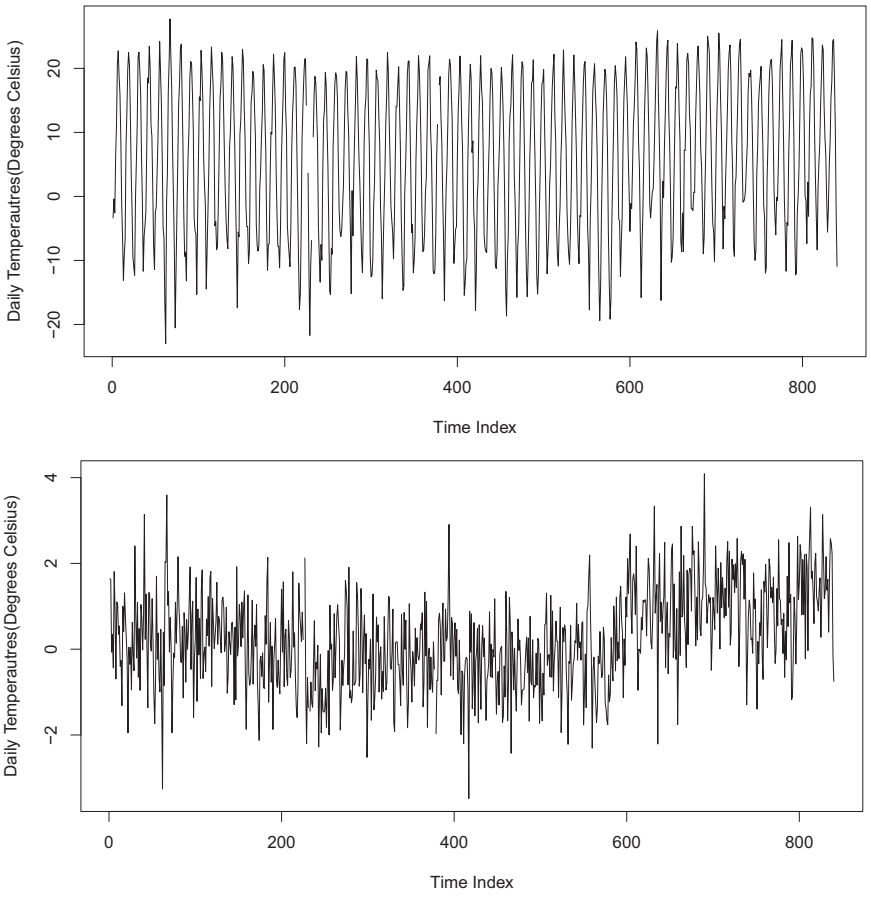


FIG. 3. The Mott series with an artificial mean change of 2°C added after time 600, showing (a) raw data and (b) the series in (a) after the seasonal cycle has been estimated and removed.

spurious changepoints within a year in an attempt to track the seasonal mean cycle should it be ignored in the modeling procedure.

b. *Autocorrelation*

Temporal autocorrelation, which measures the tendency for adjacent observations in time to be similar/dissimilar, is often present in climate data. Autocorrelation is typically positive in temperature and other climate series; for example, hot and cold periods often cluster in runs of days or months. Like seasonality, autocorrelation hinders detection of mean shifts. This is because long runs of above/below normal temperatures, often attributable to correlation, can be mistaken as a mean shift.

The correlation between X_t and X_{t+h} is defined as

$$\mathrm{Corr}(X_t, X_{t+h}) = \frac{\mathrm{Cov}(X_t, X_{t+h})}{\sqrt{\mathrm{Var}(X_t)}\sqrt{\mathrm{Var}(X_{t+h})}},$$

where $\mathrm{Cov}(X_t, X_{t+h}) = E[X_tX_{t+h}] - E[X_t]E[X_{t+h}]$ and $E[Z]$ denotes the statistical mean of Z. Due to the constancy of the other model parameters in (1), $\mathrm{Corr}(X_t, X_{t+h}) = \mathrm{Corr}(\epsilon_t, \epsilon_{t+h})$. A clarification here: data should be deseasonalized (i.e., subtracting the seasonal mean cycle) before correlations are calculated, a practice followed here. This is because seasonal mean cycles are deemed fixed and not a contributor to variability; this said, some authors view the seasonal cycle as a part of a more robust annual variability. For concreteness, our estimates of the seasonal mean and variance at season ν in the cycle are, respectively,

$$\widehat{\mu}_\nu = d^{-1}\sum_{n=0}^{d-1} X_{nT+\nu} = \hat{E}[X_{nT+\nu}],$$

$$\widehat{\sigma}_\nu^2 = \frac{\sum_{n=0}^{d-1}(X_{nT+\nu} - \widehat{\mu}_\nu)^2}{d},$$

and our estimate of the lag $h \geq 0$ autocovariance/autocorrelation in $\{\epsilon_t\}$ is

$$\widehat{\mathrm{Corr}}(\epsilon_t, \epsilon_{t+h}) = \widehat{\mathrm{Cov}}(\epsilon_t, \epsilon_{t+h}) = \frac{1}{dT}\sum_{t=1}^{dT-h}\widehat{\epsilon}_t\widehat{\epsilon}_{t+h},$$

$$\widehat{\epsilon}_{nT+\nu} = \frac{X_{nT+\nu} - \widehat{\mu}_\nu}{\widehat{\sigma}_\nu}. \quad (2)$$

Here, d denotes the number of complete cycles of data (we assume that no partial years of data are observed simply to

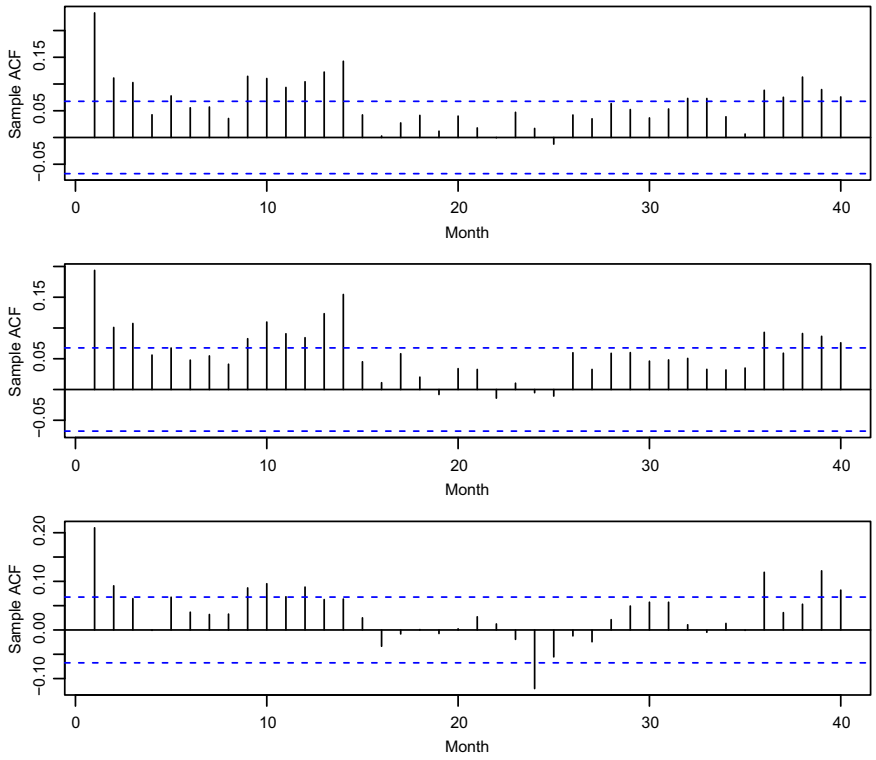


FIG. 4. Sample autocorrelations over the first 40 months at the (top) Mott, (middle) Richardton-Abbey, and (bottom) Mott minus Richardton-Abby series after the seasonal standardization in (2). While the autocorrelations in the two individual series are similar, correlation does not completely vanish in the target minus reference series in the bottom plot.

avoid trite work) and hats signify estimators of quantities. Note that the first cycle of data is indexed with $n = 0$ and the last with $n = d - 1$. Some authors use $d - 1$ in place of d in the denominator of $\widehat{\sigma}_{\nu}$, which yields an unbiased estimator); others use $dT - h$ in place of dT in the denominator of $\widehat{\text{Corr}}(\epsilon_t, \epsilon_{t+h})$ (which yields an unbiased estimator). Regardless of the denominator used, all estimators are asymptotically unbiased.

Figure 4 shows sample correlations from the monthly Mott and Richardton-Abby stations along with 95% pointwise confidence bounds for zero correlation (white noise). Notice that significant nonzero correlation exists at both stations.

Statistical methods for changepoint detection often lose detection power when autocorrelation is present. Examples below are shown where a changepoint declaration is repealed once autocorrelation is taken into account. Similarly, when mean shifts are taken into account, estimates of autocorrelation can be drastically reduced (Norwood and Killick 2018). A key aspect of this paper deals with cases where both autocorrelation and mean shifts are present.

c. *Target minus reference comparisons*

Climate homogenization is a procedure for adjusting time series for artificial features only, such as station relocations and instrumentation changes. Natural/anthropogenic-attributed changepoints occasionally exist in series and are generally viewed as part of the record that should be retained. To facilitate this, climatologists often make target–reference comparisons. A reference series is a record of like data collected geographically near the target series (that hopefully experiences similar weather). Subtracting a reference series from a target series serves to remove natural fluctuations, especially if the target and reference series experience similar weather. This subtraction reduces or altogether eliminates seasonal cycles and long-term trends (more on trends below), helping to highlight changepoints in the record. Of course, any changepoint in either the target or reference series becomes a changepoint in the target minus reference series. The additional changepoints inherited from the reference series create challenges, especially if changepoints shift both series in the same direction or the changes occur close in time (which reduces detection power).

The bottom plot in Fig. 1 shows the Mott series subtracted from the Richardton-Abby series. Observe that the seasonal cycles have lessened, if not altogether disappeared. If the target minus reference comparison is good, any long-term trend experienced by the target series should also be experienced by the reference series and removed (or greatly reduced) in the subtraction. The lower plot in Fig. 4 shows sample autocorrelations from the Mott minus Richardton-Abby stations along with 95% pointwise confidence bounds for zero correlation (white noise). These correlations are computed after the

standardization in (2) mean, which puts all variation measures on the same mean zero unit variance scale. Notice that significant nonzero autocorrelation exists at both stations. Unfortunately, a target minus reference subtraction will not generally eliminate autocorrelations. Mathematically, let $\{T_t\}$ and $\{R_t\}$ be the target and reference series, respectively. Suppose that they are jointly stationary with the same marginal covariance function: $\text{Cov}(T_t, T_{t+h}) = \text{Cov}(R_t, R_{t+h}) = \gamma(h)$. This assumption holds approximately for good reference series. The variance of the target/reference series is then $\gamma(0)$ and the variance of the target minus reference series is $2\gamma(0) - 2\text{Cov}(T_t, R_t)$. This latter quantity will be less than $\gamma(0)$ precisely when $\text{Corr}(T_t, R_t) > 1/2$. In short, if the correlation between the target and reference series is not at least 1/2, using a reference series will introduce additional variability into the series being analyzed. Of course, no one would subtract an uncorrelated reference series! Arguments for the other lags are similar but more cumbersome as one has to contend with the asymmetry of the cross-covariance function [the fact that $\text{Cov}(T_t, R_{t+h})$ is in general not equal to $\text{Cov}(T_{t+h}, R_t)$].

Since the statistical methods to conduct a changepoint analysis on the target series alone or the target minus reference series are the same, this point is essentially moot in the rest of this paper; nonetheless, its practical implications are profound. We refer the reader to Menne and Williams (2005, 2009) for more on target minus reference comparisons. Some modern methods use multiple reference series, sometimes as many as 40 (Menne and Williams 2005, 2009).

d. Trends

Many climate series have long-term trends (Gulev et al. 2021). For example, in the Mott series in Fig. 1, a long-term linear trend of 0.86°C century^{-1} is estimated (computed neglecting any changepoints). Of course, many temperature series exhibit recent warming, and many other climatic series also have trends. Trend features will be important to account for in changepoint analyses: a multiple-changepoint procedure applied to a series with a trend that is ignored in the modeling procedure will typically flag multiple mean shifts that attempt to track the trend.

e. Normality

Climate time series may or may not be Gaussian (normally distributed). A series is Gaussian if its marginal distributions come from the multivariate normal distributional family. Series that are averaged—like monthly or annual series that are obtained by averaging daily data—are often very close to normally distributed by the central limit effect (Kwak and Kim 2017). Normality can be visually checked by plotting a histogram of the series; normal data should have a unimodal symmetric histogram. A Q-Q (quantile–quantile) plot provides a graphical check for normality; points scattered closely about the main diagonal indicate that the data are well described by a Gaussian model. A commonly used and powerful nonparametric statistical test for normality is the Shapiro–Wilks test. The p value for the Shapiro–Wilks test for the Mott series is 0.73 (computed neglecting any changepoints), reinforcing that normality is very reasonable for the Mott series [see Yazici and Yolacan (2007) for more on normality tests]. Most normality tests assume zero-mean series; thus, trends, seasonal cycles, and/or changepoint features should be removed before testing.

Some climate series are decisively non-Gaussian. Examples include discrete categorical series of cloud cover, ordered from zero (say clear sky) to ten (say complete overcast), zero to one series describing an on/off phenomena like snow cover/absence, series whose marginal distributions are skewed (such as annual precipitation), and series of minima or maxima. Averaging tends to induce normality. For example, while the monthly averaging of daily data above rendered the Mott series essentially Gaussian, daily data are often skewed and/or nonnormal. In fact, daily temperatures at temperate zone stations often have a distribution with a heavy left tail (skewed), especially in winter; see Lund et al. (2006) for an example.

3. Time series models

Having introduced the typical elements of climate time series, we now address their representation. The classical decomposition of a time series $\{X_t\}$ has the form

$$X_t = \mu_t + s_t + \epsilon_t, \tag{3}$$

where $\{X_t\}$ is the observed series, $\{\mu_t\}$ is a long-term trend (not necessarily linear), $\{s_t\}$ is a deterministic seasonal cycle having known period T, and $\{\epsilon_t\}$ is zero-mean random error that is possibly correlated in time. Most changepoint scenarios for univariate series can be worked into the form in (3). The seasonal cycle $\{s_t\}$ is periodic in that $s_{t+T} = s_t$ for all times t. When the parameterization for $\{\mu_t\}$ contains a location parameter, one typically assumes that $\sum_{t=1}^{T} s_t = 0$ so that all regression parameters are statistically identifiable. This is the so-called classical decomposition model in Brockwell and Davis (1991).

For a simple example, suppose that one is examining an annual series for multiple mean shifts, permitting a possible background linear time trend. Then $T = 1$, $s_t \equiv 0$, and $\mu_t = \beta_0 + \beta_1 t$ for a location parameter β_0 and trend parameter β_1. The regression model can be written as

$$X_t = \beta_0 + \beta_1 t + \delta_t + \epsilon_t, \tag{4}$$

where the mean shift changepoint component $\{\delta_t\}$ has the form

$$\delta_t = \begin{cases} \Delta_1 = 0, & 0 < t \le \tau_1, \\ \Delta_2, & \tau_1 < t \le \tau_2, \\ \vdots & \\ \Delta_{m+1}, & \tau_m < t \le N. \end{cases}$$

The above setup takes data at the times $1, 2, \ldots, N$ and allows for m mean shift changepoints occurring at the ordered times $\tau_1, \tau_2, \ldots, \tau_m$; the changepoint count m and the changepoint occurrence times $\tau_1, \ldots, \tau_m$ are all unknown. If the location

parameter β_0 is omitted from the long-term trend expression, one need not require $\Delta_1 = 0$.

A prominent seasonal mean cycle $\{s_t\}$ exists in most temperate zone series; in general, variation induced by the seasonal cycle makes changepoints harder to "visually see and detect." The random errors $\{\epsilon_t\}$ in climate data are generally correlated. Positive autocorrelation reduces the effective number of independent observations, also making it harder to detect changepoints.

Our primary focus lies with the detection of mean changes in a series—the so-called mean shift problem. This problem keeps the autocovariance structure of $\{\epsilon_t\}$ constant across the entire series. Changepoint methods exist for autocovariance changes (Davis et al. 2006) or even changes in the marginal distributions of the series (Gallagher et al. 2012), but the major focus within the climate literature to date has been on mean shifts. Our model shifts all subsequent series values by the same amount; shifts have no seasonal character. While the methods here could be modified to allow shifts to have seasonal magnitude, this extension is not considered here.

When $T > 1$, such as for a monthly series, it is convenient to rewrite the regression model in a periodic form:

$$X_{nT+\nu} = \mu_{nT+\nu} + s_\nu + \delta_{nT+\nu} + \epsilon_{nT+\nu}, \tag{5}$$

where $\nu \in \{1, 2, \ldots, T\}$ denotes the season in a cycle and n indicates the cycle number corresponding to time $nT + \nu$. For example, a regression model allowing for a different linear trend between all consecutive changepoint times has the form

$$\mu_t = \begin{cases} \beta_1 + \alpha_1 t, & 0 < t \le \tau_1, \\ \beta_2 + \alpha_2 t, & \tau_1 < t \le \tau_2, \\ \vdots & \vdots \\ \beta_{m+1} + \alpha_{m+1} t, & \tau_m < t \le N. \end{cases}$$

The time series component $\{\epsilon_t\}$ is typically assumed to be stationary when $T = 1$, or periodically stationary when $T > 1$. A flexible and parsimonious model class for stationary series are the autoregressive (AR) series (Brockwell and Davis 1991). A pth-order zero-mean autoregression is uniquely characterized by the pth-order linear difference equation

$$\epsilon_t = \phi_1 \epsilon_{t-1} + \ldots + \phi_p \epsilon_{t-p} + Z_t,$$

where $\phi_1, \ldots, \phi_p$ are the autoregressive coefficients and $\{Z_t\}$ is a zero-mean white noise sequence with variance σ^2. When $T > 1$, AR models are replaced with periodic AR models (PAR):

$$\epsilon_{nT+\nu} = \phi_1(\nu)\epsilon_{nT+\nu-1} + \ldots + \phi_p(\nu)\epsilon_{nT+\nu-p} + Z_{nT+\nu},$$

where $\phi_1(\nu), \ldots, \phi_p(\nu)$ are the autoregressive parameters during season ν and $\{Z_{nT+\nu}\}$ is periodic white noise having the periodic variance $\mathrm{Var}(Z_{nT+\nu}) = \sigma_\nu^2$. PAR models can have a large number of parameters and are generally nonparsimonious. For example, a PAR(3) model for a monthly series has 36 AR parameters and 12 more white noise parameters; Lund et al. (2006) shows how to parsimonize PAR model fits.

4. Single-changepoint detection

a. A single mean shift

The simplest changepoint test discerns whether a series has no mean shifts (the null hypothesis) against the alternative hypothesis that there exists precisely one mean shift occurring at an unknown time. These are the so-called at most one changepoint (AMOC) methods. For the moment, assume that no long-term trends exist in the series. Almost all AMOC mean shift changepoint methods essentially compare sample means of the series before and after all candidate changepoint times. That is, after some scaling, they compare differences between $(1/k)\sum_{t=1}^{k} X_t$ and $[1/(N-k)]\sum_{t=k+1}^{N} X_t$ for each admissible changepoint time k, selecting the k where this difference is statistically maximal as the changepoint time estimate. If the maximal statistic is larger than some preset threshold, then a changepoint is declared; otherwise, the series is deemed changepoint free.

Formalizing this, suppose first that $\{\epsilon_t\}$ is independent and identically distributed (IID) with zero mean and variance σ^2. One scaled version of sample mean differences that takes into account the differing number of observations in the two segments is the cumulative sum (CUSUM) statistic having a changepoint at time k:

$$\mathrm{CUSUM}_X(k) = \frac{1}{\widehat{\sigma}}\sqrt{N}\left[\sum_{t=1}^{k} X_t - \frac{k}{N}\sum_{t=1}^{N} X_t\right],$$

where

$$\widehat{\sigma}^2 = \frac{\sum_{t=1}^{N} (X_t - \overline{X})^2}{N-1}$$

is the no changepoint null hypothesis estimate of the series' variance and $\overline{X} = (1/N)\sum_{t=1}^{N} X_t$ is the overall sample mean. One takes the argument k that maximizes $|\mathrm{CUSUM}_X(k)|$ as the estimated changepoint time.

1) PITFALL 1

Some authors examine a "maximum statistic" akin to $D_{\max} = \max_{2 \le k \le N} |\mathrm{CUSUM}_X(k)|$ to check for a single changepoint. The location where the maximum occurs is estimated as the time of the changepoint. While this is fine, incorrect null hypothesis distribution percentiles for $D_{\max}$ abound in the climate literature, often producing unjustifiable conclusions; see Lund and Reeves (2002) and Robbins et al. (2011) for discussion. When a changepoint is known to occur at time k, $\mathrm{CUSUM}_X(k)$ can be used as the test statistic. One could even scale CUSUM(k) to a z, t, or even F distribution to make valid conclusions. However, when the time of the changepoint is unknown, the maximum statistic $D_{\max}$ must be used. The correct null hypothesis percentiles for $D_{\max}$ must account for the many times k where the maximum could happen—these percentiles are much larger than those for a fixed k. Easterling and Peterson (1995) is one example where the randomness of the changepoint time is not taken into

account. Other examples of incorrect percentiles include Wang et al. (2007) and Rodionov (2004); this list is not exhaustive. The correct asymptotic quantification of AMOC changepoint statistics is often unwieldy as the scenario is not readily scalable to an extreme value distribution, even though the statistic is a maximum. Indeed, $\{\text{CUSUM}_X(k)\}$ is highly autocorrelated in k (they are not IID). The limit distribution of AMOC tests often converges to the supremum of some Gaussian process. The reader is referred to Csörgo and Horváth (1997) and MacNeill (1974) for historical technical development.

2) BEST PRACTICE 1

Several legitimate statistics can be used to test for a single mean shift. One test with superior detection power uses a sum of squared CUSUM statistics to assess whether a changepoint is present:

$$\text{SCUSUM} = \sum_{k=1}^{N} \text{CUSUM}_X^2(k).$$

Note that CUSUM and SCUSUM are distinct acronyms. The time of the changepoint is still estimated as the location $k \geq 2$ that maximizes $|\text{CUSUM}_X(k)|$. This test won the single-changepoint comparison competition in Shi et al. (2022b), is developed further in Kirch (2006), and has good false detection properties and superior detection power.

Under the null hypothesis of no changepoints, the asymptotic distribution of the SCUSUM test converges to that of $\int_0^1 B^2(t)dt$, the integrated square of a standard Brownian bridge stochastic process (Shi et al. 2022b). Null hypothesis percentiles of this distribution are presented in Table 1 for convenience and are simulated. While the SCUSUM test does not appear to be frequently used in today's climate literature, summing CUSUM statistics over all times increases detection power. As such, we recommend this test in single-changepoint analyses. Additional discussion is contained in Shi et al. (2022b).

b. Autocorrelation

We now move to AMOC tests for correlated data (the errors are not IID). A significant body of statistical research modifies the limit theory for IID data to account for autocorrelation (Robbins et al. 2011; Shi et al. 2022b). Much of this literature has the following flavor. With the SCUSUM test above (and other AMOC tests), simply replace $\widehat{\sigma}$ with an estimate of the long-run variance parameter τ^2 defined by

$$\tau^2 = \lim_{N\to\infty} N\text{Var}\left(\frac{1}{N}\sum_{t=1}^{N} X_t\right).$$

Most asymptotic statistical laws still hold with this simple modification. For example, should $\{X_t\}$ be a short-memory covariance stationary series with lag-h covariance $\gamma(h) = \text{Cov}(X_t, X_{t+h})$ (such as an ARMA model), then

$$\tau^2 = \gamma(0) + 2\sum_{h=1}^{\infty} \gamma(h).$$

These tests should not be applied to long-memory series where τ^2 can be infinite. In practice, it is not clear how to best estimate τ^2, which is the notorious spectral density at frequency zero.

TABLE 1. Critical values for the SCUSUM statistic.

Percentile	Critical value
90.0th	0.347 304 6
95.0th	0.461 374 4
97.5th	0.580 616 8
99.0th	0.743 434 8

In some asymptotic tests, convergence to the limit law can be slow, making application to even a century of annual data questionable [how slow depends on many things; Shi et al. (2022b) give further details]. An alternative way to handle correlation involves prewhitening techniques. Statistical references for prewhitening are Robbins et al. (2011) and Gallagher et al. (2022). To account for correlation in an AMOC changepoint analysis, prewhitening first fits a pth-order autoregressive [AR(p)] model to the series (this assumes nonperiodic data). This fit is conducted under the null hypothesis of no changepoints and is easily accomplished with many standard time series analysis packages. This procedure yields estimates of the autoregressive parameters $\phi_1, \ldots, \phi_p$ and the white noise variance σ^2; hats over these symbols demarcate estimators of these parameters. Next, the estimated one-step-ahead predictions

$$\hat{X}_{t+1} = \overline{X} + \widehat{\phi}_1(X_t - \overline{X}) + \cdots + \widehat{\phi}_p(X_{t-p+1} - \overline{X}), \quad t \geq p, \tag{6}$$

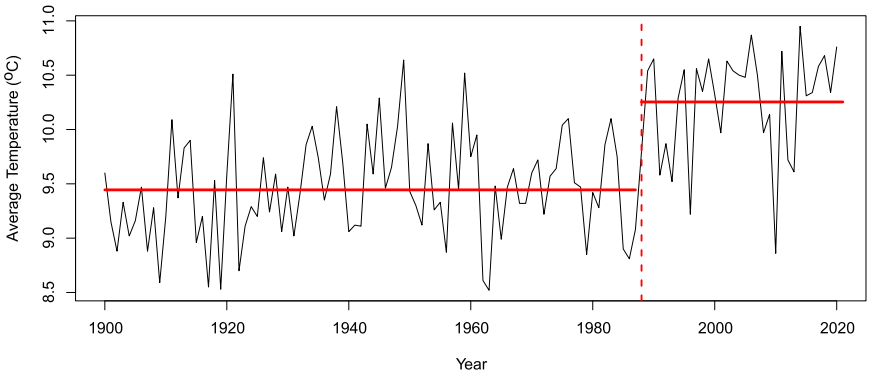
$\overline{X} = (1/N)\sum_{t=1}^{N} X_t$, are calculated with $\widehat{\phi}_j$ replacing ϕ_j and the one-step-ahead prediction errors $Y_t = X_t - \hat{X}_t$ are formed. The $\{Y_t\}$ are often called innovations. When the AR(p) parameters are known, the one-step-ahead prediction errors $\{Y_t\}$ are independent. Using estimated AR parameters leaves the $\{Y_t\}$ slightly dependent, but this dependence is usually negligible. The series $\{Y_t\}$ is also called the prewhitened series. To compute the startup values $\hat{X}_1, \ldots, \hat{X}_p$, one uses the time series prediction equations; see chapter 3 of Brockwell and Davis (1991) for details.

Next, one simply applies the SCUSUM (or some other AMOC) test to the prewhitened $\{Y_t\}$ using the percentiles for IID errors to make conclusions. Robbins et al. (2011) proves that this procedure is statistically valid asymptotically and shows that the limit laws typically "kick in more quickly" than asymptotic laws that replace $\widehat{\sigma}$ with $\widehat{\tau}$.

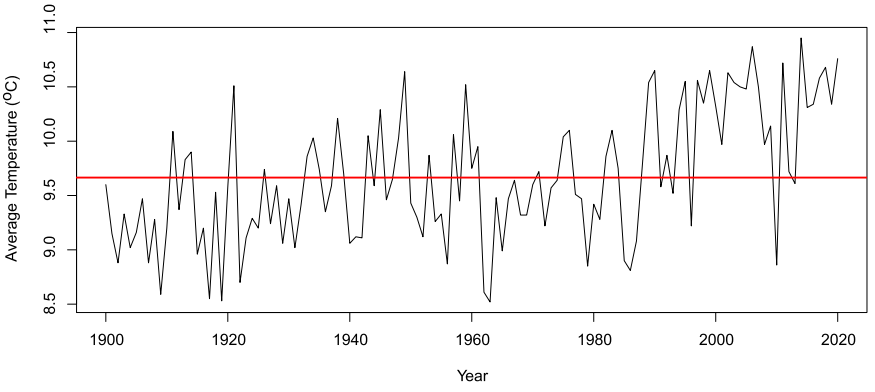
While prewhitening adds to the analysis burden, our next pitfall notes the importance of taking correlation into account.

1) PITFALL 2

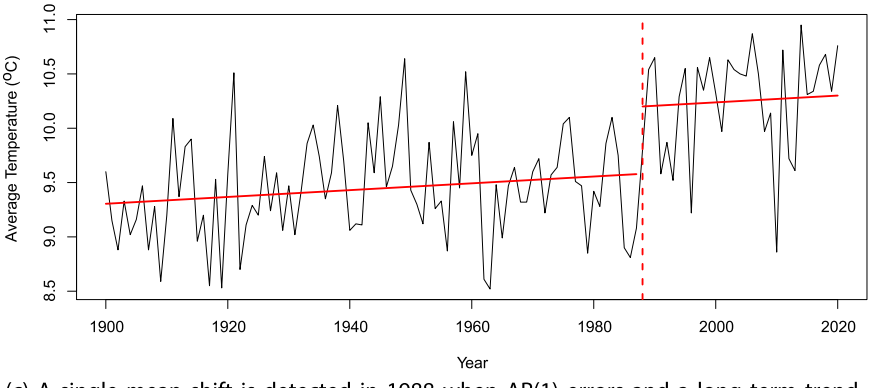
Ignoring positive autocorrelation in a series will often produce spurious changepoint conclusions. In fact, series that are heavily positively correlated tend to make long sojourns above and below the long-term mean of the series, inducing the appearance of a changepoint. Ignoring correlation may induce the spurious conclusion that a changepoint exists when in truth it does not.



(a) A single mean shift is detected in 1988 when IID errors are assumed.



(b) No changepoint is detected when AR(1) errors are assumed.



(c) A single mean shift is detected in 1988 when AR(1) errors and a long term trend are assumed.

FIG. 5. Annual central England temperatures (1900–2020). Single-changepoint tests give different conclusions depending on the mean structure and autocorrelation properties assumed.

2) BEST PRACTICE 2

Prewhiten autocorrelated series before applying any AMOC IID changepoint tests. As shown below, dubious conclusions can arise when autocorrelation is ignored. A general theme for AMOC tests with positively autocorrelated data, which entail the majority of climate cases, is clear: one risks concluding that a changepoint exists when in truth it does not when positive correlation is ignored. The situation reverses itself should negatively autocorrelated data be encountered.

c. *An example*

We now examine the annual central England temperature (CET) series from 1900 to 2020 with a single-changepoint test. The CET record was obtained from the Met Office at https://www.metoffice.gov.uk/hadobs/hadcet/. For a multiple-changepoint analysis of the entire CET series dating back to the 1600s, see Shi et al. (2022a). Figure 5 and Table 2 display this series against several single-changepoint configurations explored below. Conclusions will be heavily dependent on the assumptions made.

TABLE 2. Single-changepoint tests for the central England temperature series.

Model assumptions	Test	p value
Mean shift + IID errors	SCUSUM	$\leq 10^{-6}$
Mean shift + AR(1) errors	SCUSUM_Z	0.31
Fixed trend + AR(1) errors	CUSUM_D	0.039

As a first step, we examine the series for a single mean shift assuming IID errors. The CUSUM(k) statistic is maximized at $k = 1988$ and the SCUSUM statistic is 3.577. Comparing to the 95th percentile of SCUSUM statistic, which is 0.461, one concludes that a mean shift exists with confidence at least 95% (in fact, the p value of erroneously rejecting a no changepoint null hypothesis is zero to about six decimal places). The estimated series mean is plotted against the series in the top panel of Fig. 5.

If one plots the residuals from this fit, autocorrelation is clearly present. Indeed, the estimated correlation between consecutive raw series values is 0.425, which entails moderate autocorrelation [this would be the estimated ϕ_1 coefficient in an AR(1) fit should there be no changepoints]. This correlation estimate drops to $\widehat{\phi}_1 = 0.252$ when the 1988 changepoint is taken into account, which is still significantly positive. Thus, we rerun the single mean shift test allowing for autocorrelated errors, this time using a simple AR(1) structure for the model errors. A SCUSUM test was applied to the prewhitened AR(1) one-step-ahead prediction errors and gives $\text{SCUSUM}_Z = 0.180$, which is well below the 0.461 threshold needed to declare statistical significance with 95% confidence (the p value for this test is 0.31). This essentially repeals the 1988 mean shift. The conflicting conclusions illustrate why one needs to allow for correlation in changepoint tests when correlation is present. Neglecting to account for positive correlation can lead to an overestimation of the number of changepoints.

d. Trends

As previously mentioned, trends can also influence changepoint conclusions. In particular, one should not apply a changepoint test to data with a trend without accounting for the trend. For example, should the linear trend $\mu_t = \beta_1 t$ exist in (3) but not be modeled, then an AMOC test tends to signal a single changepoint in the center of the record with a positive mean shift when $\beta_1 > 0$, and flag a negative mean shift in the center of the record when $\beta_1 < 0$. The methods are simply rejecting that the mean is constant (which is why some authors use changepoint tests as a check for a homogeneous mean). When a seasonal cycle exists in the data, the situation becomes even more nebulous, with multiple-changepoint techniques flagging multiple changes in an attempt to "track the seasonal mean and long-term trend." In short, changepoint techniques are not robust to assumption changes in $\mu_t = E[X_t]$. Unfortunately, in changepoint analyses, each different form of μ_t requires a different set of null hypothesis percentiles. For example, for a simple mean shift where $\mu_t = \beta_0$, the 95th percentile of the CUSUM test is 1.358 [this percentile comes from Robbins et al. (2011)]; when the linear trend $\mu_t = \beta_0 + \beta_1 t$ is considered, the 95th CUSUM percentile becomes 0.902 (Gallagher et al. 2013).

1) PITFALL 3

Applying an AMOC changepoint test to series with trends or seasonality that does not account for the trend or seasonality can result in spurious changepoint declarations. Here, the methods are simply declaring that the series' mean is time-varying.

2) BEST PRACTICE 3

Account for all potential features in the mean of a series. If in doubt, allow for a trend and/or seasonality and use the statistical methods to distinguish which features are present in the series.

We now take a deeper look at the CET series with an analysis that allows for trends. Global warming posits a slow temperature increase; as such, an AMOC analysis with the linear trend $\mu_t = \beta_0 + \beta_1 t$ is explored. Based on our previous analysis, AR(1) errors are again used to account for autocorrelation. An AMOC CUSUM-type mean shift test for IID errors in linear trend models is developed in Gallagher et al. (2012) (we are unaware of anyone studying SCUSUM tests in the linear trend setting). This test statistic will be denoted by CUSUM_D. Estimating the linear trend and AR(1) parameters under the null hypothesis of no changepoints provides $\widehat{\beta}_0 = 9.1°\text{C}$, $\widehat{\beta}_1 = 0.009°\text{C yr}^{-1}$ (we will not address the statistical significance of this estimate), and $\widehat{\phi}_1 = 0.194$.

One needs to be careful to account for the trend when prewhitening this series. Specifically, our estimated one-step-ahead predictions with AR(1) errors become [cf. to (6)].

$$\begin{aligned}\hat{X}_t &= \widehat{\mu}_t + \widehat{\phi}_1(X_{t-1} - \widehat{\mu_{t-1}}) \\ &= \widehat{\beta}_0 + \widehat{\beta}_1 t + \widehat{\phi}_1[X_{t-1} - \widehat{\beta}_0 - \widehat{\beta}_1(t-1)],\end{aligned} \tag{7}$$

for $t \geq 2$, with the start-up condition $\hat{X}_1 = \widehat{\beta}_0 + \widehat{\beta}_1$. The prewhitened series is always $Y_t = X_t - \hat{X}_t$.

The CUSUM_D test applied to $\{Y_t\}$ gives a statistic of 0.929, occurring in 1988, which is slightly above the 95th percentile null hypothesis threshold of 0.903. The p value for this test is 0.038. With 95% confidence, the 1988 changepoint is detected again. The bottom line in Table 2 shows this result. The bottom panel in Fig. 5 displays the fit to this data. This configuration is the best fitting of our three models since it takes into account both trends and autocorrelation. As an aside, we comment that model fitting is not about maximizing or minimizing p values, but rather making sure that all relevant statistical features are accounted for in a parsimonious model. See Shi et al. (2022a) for a detailed analysis of the CET series. The thresholds used in the CUSUM_D test were calculated via simulation under the null hypothesis with 10 000 repetitions. See Shi et al. (2022b) for additional details.

Obviously, the assumptions made in changepoint analyses are extremely important and influence conclusions. While issues become more complex with multiple changepoints, the topic of our next section, much of the AMOC intuition carries over to that setting.

5. Multiple-changepoint detection

Many climate series have more than one changepoint. United States climate series average a station move or gauge change once every 17 years (Mitchell 1953); see also the findings in Menne and Williams (2005, 2009), and O'Neill et al. (2022). As in the AMOC case, multiple-changepoint (MCPT) detection is fraught with challenges and pitfalls, perhaps more than the single-changepoint case. While MCPT analyses are less developed than AMOC tests, the problem is being actively researched in statistical settings.

Initially, AMOC techniques were extended to MCPT problems via binary segmentation methods (Scott and Knott 1974). Binary segmentation examines the entire series first for a single changepoint with some AMOC test. If a changepoint is found, the series is then split into two subsegments about the identified changepoint time and the two subsegments are further scrutinized for a single changepoint with the AMOC test. The procedure continues iteratively until all subsegments are declared changepoint free. We now know that binary segmentation is one of the poorer ways to handle multiple-changepoint problems (Shi et al. 2022b). This point is further reinforced in section 5a.

Other approaches to the MCPT problem can be classified into distinct camps. One camp examines recursive segmentation procedures that improve upon binary segmentation methods; these include wild binary segmentation (Fryzlewicz 2014) and wild contrast maximization (Cho and Fryzlewicz 2020). Wild binary segmentation draws random subintervals of varying lengths of the data, conducts an AMOC test on each subinterval, and reconciles across all subintervals analyzed to produce an estimated changepoint configuration. Wild contrast maximization is built upon wild binary segmentation and uses an AMOC test that accounts for autocorrelation. These methods are computationally quick and often yield reasonable results. Unfortunately, many of these techniques declare an excessive number of changepoints when the true number of changepoints is small (Shi et al. 2022b), essentially rendering them unusable in climate cases where say two changepoints exist in a 100-yr climate series.

Another camp applies dynamic programming techniques to MCPT problems. Here, an objective function associated with the problem is optimized. The segment neighborhood algorithm of Auger and Lawrence (1989) and the pruned exact linear time of Killick et al. (2012) are two examples. Dynamic programming techniques provide optimal (relative to the chosen objective function) changepoint configurations and runs quickly. Unfortunately, these techniques often make unrealistic assumptions (uncorrelated series or all model parameters must shift at every changepoint time) that make them unfeasible in some climate applications. Advances to these methods are currently being pursued. Model selection approaches such as Harchaoui and Lévy-Leduc (2010) and Shen et al. (2014) and scan statistics procedures based on moving sum statistics (Eichinger and Kirch 2018) exist among other techniques (Cho and Kirch 2021)—changepoint research is a huge field and this list is not exhaustive.

Like the AMOC case, assumptions are crucial in MCPT analyses. Many MCPT techniques assume IID $\{\epsilon_t\}$, which is often unrealistic in climate applications. As with the AMOC case, MCPT techniques assuming independent $\{\epsilon_t\}$ can give suboptimal answers for autocorrelated series (Davis et al. 2006; Li and Lund 2012; Chakar et al. 2017). While one can prewhiten the series, estimation of the correlation structure and the multiple mean shift sizes and locations confound each other. No simple null and alternative hypotheses suggest themselves in the MCPT setting, as opposed to the AMOC setting where the null and alternative hypotheses have zero and one changepoint, respectively. In the MCPT case, the null hypothesis could be zero, exactly one, at most one, two, etc., changepoint counts. In the AMOC case, estimates of the series' correlation structure are computed under the null hypothesis of no changepoints and models containing no and one changepoints are statistically compared.

Penalized likelihood methods, another MCPT camp, tackle the problem by minimizing a likelihood objective function that is penalized when the model contains too many changepoints. Elaborating, statisticians often estimate model parameters via likelihood techniques. Let $L(m; \tau_1, \ldots, \tau_m)$ denote the likelihood of the best time series model having m changepoints at the times $1 < \tau_1 < \tau_2 < \ldots < \tau_m \leq N$. Likelihoods for $\{X_t\}_{t=1}^N$ take the classical time series form

$$L(m; \tau_1, \ldots, \tau_m) = (2\pi)^{-N/2}\left(\prod_{t=1}^{N} V_t\right)^{-1/2} \exp\left[-\frac{1}{2}\sum_{t=1}^{N}\frac{(X_t - \hat{X}_t)^2}{V_t}\right],$$

where $\hat{X}_t$ is the best linear prediction of X_t from past observations in (7) and $V_t = E[(X_t - \hat{X}_t)^2]$ is its unconditional mean squared error. While CUSUM tests do not assume a Gaussian distribution for their errors, penalized likelihood methods are parametric and often assume a Gaussian structure.

As the number of changepoints m increases, the model fit improves: $L(m; \tau_1, \ldots, \tau_m)$ increases in m. However, after a while, additional changepoints do not appreciably improve the likelihood. This is where the penalty term comes in. The penalty for having m changepoints at the times $\tau_1, \ldots, \tau_m$ is denoted by $P(m; \tau_1, \ldots, \tau_m)$ and increases as m increases. Penalized likelihood methods look to minimize the penalized objective function

$$O(m; \tau_1, \ldots, \tau_m) = -2\ln[L(m; \tau_1, \ldots, \tau_m)] + P(m; \tau_1, \ldots, \tau_m)$$

over all feasible values of m and $\tau_1, \ldots, \tau_m$. When there are no changepoints ($m = 0$), the penalty term is taken as zero.

Development of penalty functions is a well-studied statistical problem. Commonly used penalties in the literature for the mean shift problem include the AIC, BIC, mBIC, and MDL penalties. Their formulas are

$$\begin{aligned}
\text{AIC}: &\quad P(m; \tau_1, \ldots, \tau_m) = 2(2m + p + 2),\\
\text{BIC}: &\quad P(m; \tau_1, \ldots, \tau_m) = (2m + p + 2)\ln(N),\\
\text{mBIC}: &\quad P(m; \tau_1, \ldots, \tau_m) = (3m + p + 2)\ln(N) + \sum_{i=1}^{m+1} \ln\left(\frac{\tau_i - \tau_{i-1}}{N}\right),\\
\text{MDL}: &\quad P(m; \tau_1, \ldots, \tau_m) = (p + 1)\ln(N) + \sum_{i=1}^{m+1} \ln(\tau_i - \tau_{i-1}) + 2\ln(m) + 2\sum_{i=2}^{m} \ln(\tau_i),
\end{aligned} \tag{8}$$

Here $\tau_0 = 1$ and $\tau_{m+1} = N$ are defined for convenience. The above penalties are for mean shift models and AR(p) errors (if the errors are IID, $p = 0$); should the regression structure change at each changepoint time, the above formulas require modifications. While other penalties exist, these are the most popular penalties used in today's literature. Note that the mBIC and MDL penalties depend on where the changepoints lie, but that the AIC and BIC penalties are simple multiples of the number of changepoints. A detailed discussion of penalty performance is found in Shi et al. (2022b). For specifics, AIC is well known to overestimate the true number of changepoints and should not be used. For a penalty that does not depend on the changepoint location times, BIC performs surprisingly well in a variety of settings (Shi et al. 2022b).

Optimizing $O(m; \tau_1, \ldots, \tau_m)$ requires significant computations. To compute $O(m; \tau_1, \ldots, \tau_m)$, an optimal time series model with m changepoints at the times $\tau_1, \ldots, \tau_m$ needs to be fitted. While this is a straightforward task for most time series packages, there are 2^{N-1} distinct changepoint configurations that need to be evaluated as candidates in an exhaustive search for the best MCPT configuration. This total is immense for even N as large as 100, making exhaustive searches a strenuous task. Authors have used genetic algorithms (Davis et al. 2006; Li and Lund 2012) to overcome these difficulties. Today, despite computational issues, penalized likelihoods are considered the gold standard for MCPT problems.

Estimates of the mean and seasonal parameters in a penalized likelihood procedure are not corrupted/degraded by the presence of changepoints. This is because a genetic algorithm search first fixes the changepoint configuration and then estimates all other model parameters in a manner that takes the changepoint structure into account. In the AMOC case, where hypothesis testing logic applies, all parameters are estimated under the null hypothesis of no changepoints. In AMOC tests, if a changepoint is found, the estimates of the mean and seasonal parameters should be revised to take into account the identified shift.

a. *Binary segmentation*

As the earliest invented and still widely used MCPT technique, binary segmentation's popularity rests on two ingredients: simplicity and rapid computation. Binary segmentation is a "greedy algorithm" that optimizes an objective function stagewise. Such a procedure often does not find the globally optimal solution. An attempted remedy to binary segmentation, wild binary segmentation (Fryzlewicz 2014), injects randomization into the changepoint search to avoid local optimums. However, simulation studies in Lund and Shi (2020) suggest that wild binary segmentation overestimates changepoint counts for IID model errors, and becomes dysfunctional in settings with correlated errors. Wild contrast maximization (Cho and Fryzlewicz 2020), another improvement of wild binary segmentation designed for autocorrelated processes, is capable of handling serial dependence. While we will not discourage this technique, we also comment that it has not been fully vetted as of 2023.

1) PITFALL 4: USING ORDINARY BINARY SEGMENTATION IN MCPT PROBLEMS

Binary segmentation is generally an inferior MCPT problem approach, regardless of assumptions. Unfortunately, binary segmentation is used in many engineering, computer science, and climate applications. To illustrate binary segmentation pitfalls, a simulation is constructed. Here, Gaussian series of length 500 were simulated with white noise errors with a unit variance. Three equally spaced mean shifts were added, shifting the series by a unit length in alternating directions. This partitions the series into four equal length segments of 125 points each; Fig. 6 displays a sample generated series.

We randomly generated 1000 such series and applied several different changepoint methods. The estimated changepoint configurations were compared to the true changepoint configuration with the distance metric in Shi et al. (2022b). This distance incorporates both m and the changepoint locations $\tau_1, \ldots, \tau_m$. Smaller distances indicate better performance; a perfectly estimated configuration has zero distance to the truth. Boxplots of distances between the estimated changepoint configuration and the true configuration over the 1000 simulations are summarized in Fig. 7. The boxplots show that binary segmentation underperforms all penalized likelihood methods. The number of detected changepoints for each method are listed in Table 3.

2) BEST PRACTICE 4: USE PENALIZED LIKELIHOOD MCPT METHODS IN LIEU OF BINARY SEGMENTATION

In fitting a penalized likelihood MCPT model, the autocorrelation structure of the series is estimated in the fit. Binary segmentation does not give such estimates, but they are not difficult to obtain after the piecewise regime means are subtracted from the series. Some MCPT techniques only allow special time series structures. For example, Chakar et al. (2017) requires AR(1) errors. While the AR order is not believed to be as important as

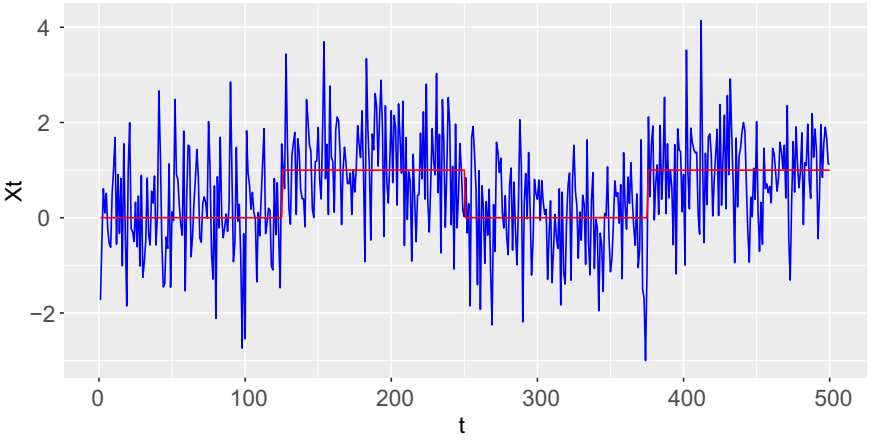


FIG. 6. A series with three equally spaced mean shifts of unit size that shift the series in alternating directions. The true series mean is plotted for reference. Regression errors are uncorrelated white noise with a unit variance.

other issues in most climate applications, it is also infeasible that an AR(1) correlation structure adequately describes all climate series. Hewaarachchi et al. (2017) push genetic algorithms to their limit by homogenizing daily temperatures via penalized likelihoods. While Cho and Fryzlewicz (2020) allow general AR(p) errors, simulations indicate that wild contrast maximization tends to estimate too many changepoints, inheriting this flaw from wild binary segmentation. While it is widely understood in the statistical literature that binary segmentation is inherently flawed [see Shi et al. (2022b) for comparisons], the technique is still widely used. In what follows, we focus on penalized likelihood techniques estimated by a genetic algorithm.

b. *Atlanta airport temperatures*

To see differences between the approaches in practice, annual mean surface temperatures from 1879 to 2013 at Atlanta, Georgia's Hartsfield International Airport station will be analyzed. This dataset was provided by Berkeley Earth at http://berkeleyearth.lbl.gov/station-list/, and contains "raw" temperatures, unadjusted for potential artifacts. Mean shift models with AR(1) errors were fitted via penalized likelihood techniques and binary segmentation approaches. The results are depicted in Fig. 8. Binary segmentation flags a single changepoint in the early 1980s, while a BIC penalized likelihood approach estimates three changepoints, occurring in the 1920s, 1960s, and 1980s. Our binary segmentation algorithm uses the SCUSUM AMOC test with a 95% confidence threshold and accounts for autocorrelation via an AR(1) model. While detailed simulations illustrating the inferiority of binary segmentation are supplied in Shi et al. (2022b), binary segmentation often has trouble identifying multiple mean shifts that move the series in opposite directions, which seems to be the case here: the successive changepoints estimated by penalized likelihood move the series up, down, and then up. This leads us to conclude that two changepoints are missed by binary segmentation here.

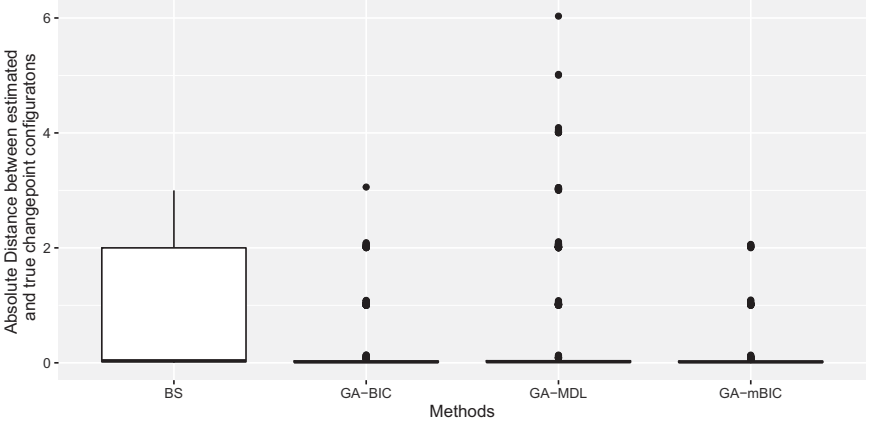


FIG. 7. A comparison of binary segmentation and penalized likelihood methods. The biggest errors occur with binary segmentation. A 95% threshold is used for binary segmentation; the BIC, MDL, and mBIC penalized likelihoods were optimized by a genetic algorithm (GA).

TABLE 3. Distribution of detected number ($\hat{m}$) of changepoints in 1000 simulations (all values in %). Binary segmentation is the worst-performing method. The true configuration (boldface) has three changepoints ($m = 3$)

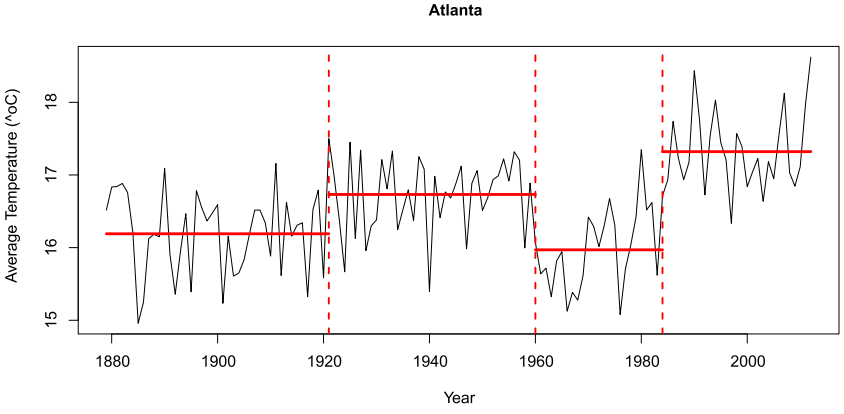
Methods	$\hat{m} < 2$	$\hat{m} = 2$	$\hat{\boldsymbol{m}} = \mathbf{3}$	$\hat{m} = 4$	$\hat{m} < 4$
BIC	0	0	**92.3**	5.7	2
mBIC	0	0	**96.5**	2.5	1
MDL	0	0	**87.6**	3.8	8.6
BS	25.9	0.3	**72.4**	1.5	0

c. *Ignoring trends*

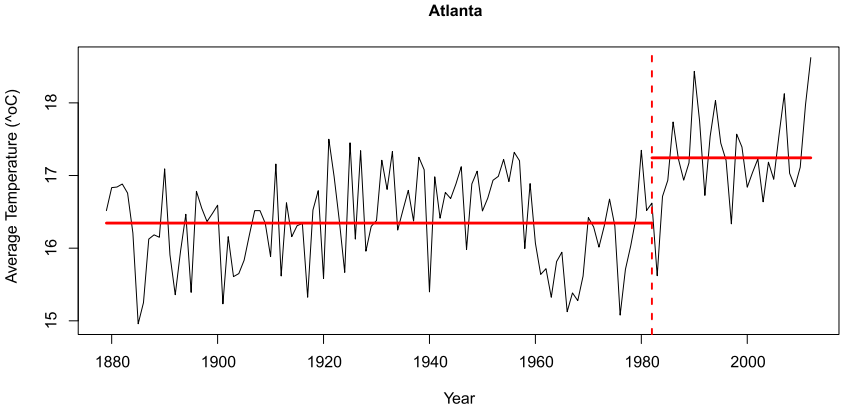
As in the AMOC case, ignoring trends in the MCPT setting may produce spurious changepoint declarations. For if the long-term trend is decisively increasing or decreasing, but ignored in the analysis, then MCPT procedures typically flag one or more changepoints in an attempt to track the series mean. In the AMOC case, each different mean functional form changes the asymptotic percentiles of the statistical test (Tang and MacNeill 1993). In the MCPT case, as long as the same trend parameters apply to all series subsegments, the penalties in (8) can be used without adjustment (adjustments to the penalties will not alter the estimated changepoint configuration). Should one desire models where all parameters shift at the changepoint times—one example would allow the trend slope to depend on the regime—then the penalties in (8) must be modified. The reader is referred to Shi et al. (2022a) for the technicalities.

1) PITFALL 5: APPLYING MEAN SHIFT MCPT TECHNIQUES TO SERIES WITH TRENDS OR SEASONALITY WITHOUT ACCOUNTING FOR THESE FEATURES.

Similar to AMOC techniques in pitfall 2, applying a MCPT technique that neglects trends and seasonality can result in spurious changepoint declarations. For example, an increasing long-term trend will likely be estimated as a series of changepoints acting as an increasing stairway.

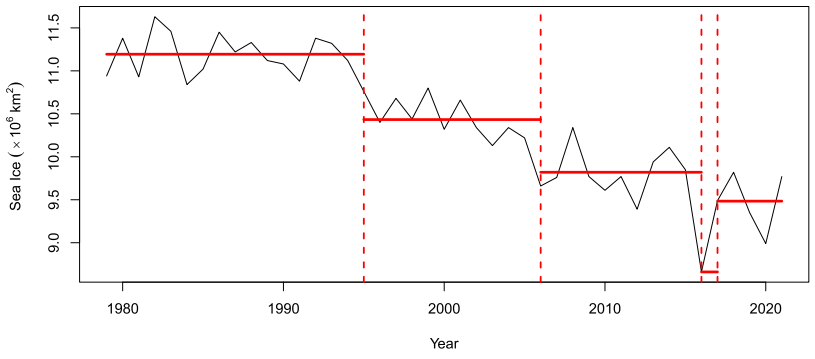


(a) BIC penalized likelihood flags three mean shift changepoints in 1921, 1960, 1984.

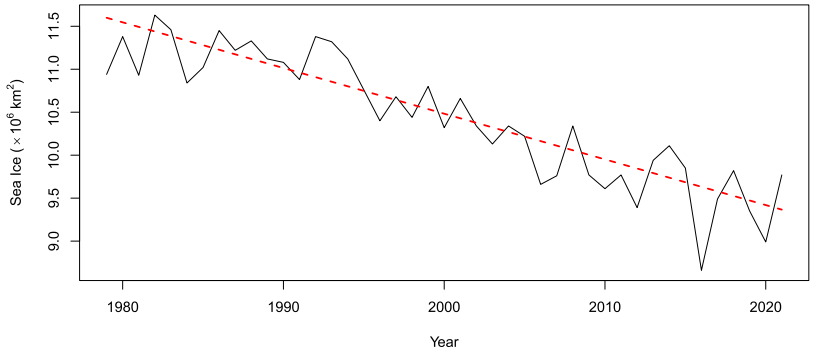


(b) Binary segmentation detects one mean shift changepoint in 1982.

FIG. 8. A changepoint analysis of the Atlanta airport temperature series. When AR(1) errors are assumed, changepoints flagged by (a) a BIC penalized likelihood and (b) binary segmentation. Binary segmentation flags one changepoint, while a BIC penalized likelihood flags three.



(a) Four mean shifts are detected at 1995, 2006, 2016, 2017 when the trend is ignored and a mean shift with AR(1) error model is assumed. The changepoints attempt to "follow the trend".



(b) No changepoints are detected when a trend with AR(1) error model is assumed.

FIG. 9. A changepoint analysis of the Arctic sea ice series.

2) BEST PRACTICE 5: ALLOW FOR TRENDS AND/OR SEASONALITY IN SERIES HAVING THESE FEATURES.

d. *Arctic sea ice*

To illustrate the importance of accounting for trends, we analyze a series of September sea ice extent in the Northern Hemisphere from 1979 to 2021. The data were provided by the National Snow and Ice Data Center and downloaded from: https://nsidc.org/data. The sea ice extent represents the total area of all grid cells with at least 15% sea ice concentration. Since 1979, the Northern Hemisphere sea ice has shown declines (Meredith et al. 2019). In Reid et al. (2016), this series was used to illustrate a rapid, large-scale change in Earth's biophysical systems in the 1980s, and a mean shift was suggested to have occurred in 1989. Here, this analysis is revisited to assess whether one or multiple mean shifts are still detected when a long-term trend is taken into account. Figure 9 shows the series and some MCPT fits. The top plot, a BIC penalized likelihood estimated MCPT configuration with AR(1) errors, identifies four changepoints, when no trend is put in the model. When a linear trend is added to the model (the bottom plot), all four changepoints are repealed. The estimated trend slope of sea ice retreat was -0.05 million km^2 yr^{-1}. The linear trend fit here is preferred as this model is more parsimonious; see Shi et al. (2022a) for more on comparing different model types.

6. Discussion and comments

This paper highlighted some common pitfalls in changepoint analysis/homogenization methods and suggests best practices to avoid them. In general, changepoint methods are not robust to assumptions on the structure of a series, especially its mean, and care is needed in their proper application. Issues considered in the paper include correlation, trends, distributions of maximum statistics, and the type of multiple-changepoint analysis employed. The general mantra is that if a series feature is not obvious (say existence of trends or correlation), it is best to put that feature in a model and let statistical methods discern whether or not it is present. While the paper attempts to put forth a best practice, any user of changepoint methods in the climate sciences should be aware of the litany of mistaken and/or dubious analyses in this field.

Indeed, the number of changepoint declarations that would be repealed due to failure to consider positive autocorrelation would be extensive.

Revisiting the individual series scrutinized for changepoints in this paper for end conclusions, the CET series seems to have nonignorable trends, but little autocorrelation ($\widehat{\phi}_1 = 0.055$). The Atlanta series has slightly more autocorrelation ($\widehat{\phi}_1 = 0.11$), multiple changepoints, and little long-term trend. One may wish to explore models with trends for the more recent years of this series further. In comparing changepoint versus trend models for the Arctic sea ice series, trends seem more physically plausible than changepoints. There is also little autocorrelation ($\widehat{\phi}_1 = 0.05$) in the trend model. Shi et al. (2022a) shows how to compare these two distinctly different models in a statistical fashion (this is more involved and is not done here).

It is worth rehashing target minus reference series analyses versus target series analyses only (absolute versus relative homogenization). While the statistical procedures to analyze both settings are the same, subtraction of a reference series often reduces trends and/or seasonal cycles, making some issues clearer. Nonetheless, as was shown here, formation of a target minus reference series may not eliminate or even reduce autocorrelation, nor need it totally eliminate long-term trends and/or seasonal cycles. Existence of metadata is another issue. While most authors tend to eschew metadata in their changepoint analyses, Beaulieu et al. (2010) and Li and Lund (2015) show how an informative Bayesian prior can be constructed from it and used this information to increase changepoint detection power.

Multiple-changepoint techniques are actively being researched in statistics. Computational advances are expected in the near future, especially in regard to penalized likelihood methods for more complex models. Other aspects of the problem are also being studied. A clear point from the literature lies with the inferiority of ordinary binary segmentation techniques in multiple-changepoint problems. Here, we simply urge researchers to use better methods.

Acknowledgments. Robert Lund thanks National Science Foundation Grant DMS-2113592 for partial support. Claudie Beaulieu thanks National Science Foundation Grant AGS-2143550 for partial support. Rebecca Killick thanks EPSRC (EP/R01860X/1) and NERC (NE/T012307/1) for partial support. Xueheng Shi thanks National Science Foundation Grant CCF-1934568 for partial support.

Data availability statement. Webpages where the series in this paper can be downloaded were listed where the series first appeared. R code is available at https://github.com/rkillick/2023JCLreview.

REFERENCES

Auger, I. E., and C. E. Lawrence, 1989: Algorithms for the optimal identification of segment neighborhoods. *Bull. Math. Biol.*, **51**, 39–54, https://doi.org/10.1007/BF02458835.

Beaulieu, C., and R. Killick, 2018: Distinguishing trends and shifts from memory in climate data. *J. Climate*, **31**, 9519–9543, https://doi.org/10.1175/JCLI-D-17-0863.1.

——, T. B. M. J. Ouarda, and O. Seidou, 2010: A Bayesian normal homogeneity test for the detection of artificial discontinuities in climatic series. *Int. J. Climatol.*, **30**, 2342–2357, https://doi.org/10.1002/joc.2056.

——, J. Chen, and J. L. Sarmiento, 2012: Change-point analysis as a tool to detect abrupt climate variations. *Philos. Trans. Roy. Soc.*, **370**, 1228–1249, https://doi.org/10.1098/rsta.2011.0383.

Brockwell, P. J., and R. A. Davis, 1991: *Time Series: Theory and Methods.* 2nd ed. Springer, 580 pp.

Cahill, N., S. Rahmstorf, and A. C. Parnell, 2015: Change points of global temperature. *Environ. Res. Lett.*, **10**, 084002, https://doi.org/10.1088/1748-9326/10/8/084002.

Chakar, S., E. Lebarbier, C. Lévy-Leduc, and S. Robin, 2017: A robust approach for estimating change-points in the mean of an AR(1) process an. *Bernoulli*, **23**, 1408–1447, https://doi.org/10.3150/15-BEJ782.

Chen, J., and A. K. Gupta, 2012: *Parametric Statistical Change Point Analysis.* Birkhäuser Boston, 273 pp.

Cho, H., and P. Fryzlewicz, 2020: Multiple change point detection under serial dependence: Wild contrast maximisation and gappy Schwarz algorithm. arXiv, 2011.13884v6, https://doi.org/10.48550/arXiv.2011.13884.

——, and C. Kirch, 2021: Data segmentation algorithms: Univariate mean change and beyond. *Econometrics Stat.*, https://doi.org/10.1016/j.ecosta.2021.10.008, in press.

Csörgo, M., and L. Horváth, 1997: *Limit Theorems in Change-Point Analysis.* John Wiley and Sons, 438 pp.

Davis, R. A., T. C. M. Lee, and G. A. Rodrigues-Yam, 2006: Structural break estimation for nonstationary time series models. *J. Amer. Stat. Assoc.*, **101**, 223–239, https://doi.org/10.1198/016214505000000745.

Domonkos, P., J. A. Guijarro, V. Venema, M. Brunet, and J. Sigró, 2021: Efficiency of time series homogenization: Method comparison with 12 monthly temperature test datasets. *J. Climate*, **34**, 2877–2891, https://doi.org/10.1175/JCLI-D-20-0611.1.

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, https://doi.org/10.1002/joc.3370150403.

Eichinger, B., and C. Kirch, 2018: A MOSUM procedure for the estimation of multiple random change points. *Bernoulli*, **24**, 526–564, https://doi.org/10.3150/16-BEJ887.

Fryzlewicz, P., 2014: Wild binary segmentation for multiple change-point detection. *Ann. Stat.*, **42**, 2243–2281, https://doi.org/10.1214/14-AOS1245.

Gallagher, C., R. Lund, and M. Robbins, 2012: Changepoint detection in daily precipitation series. *Environmetrics*, **23**, 407–419, https://doi.org/10.1002/env.2146.

——, ——, and ——, 2013: Changepoint detection in climate series with long-term trends. *J. Climate*, **26**, 4994–5006, https://doi.org/10.1175/JCLI-D-12-00704.1.

——, R. Killick, R. Lund, and X. Shi, 2022: Autocovariance estimation in the presence of changepoints. *J. Korean Stat. Soc.*, **51**, 1021–1040, https://doi.org/10.1007/s42952-022-00173-5.

Gulev, S. K., and Coauthors, 2021: Changing state of the climate system. *Climate Change 2021: The Physical Science Basis*, V. Masson-Delmotte et al., Eds., Cambridge University Press, 287–422, https://doi.org/10.1017/9781009157896.004.

Harchaoui, Z., and C. Lévy-Leduc, 2010: Multiple change-point estimation with a total variation penalty. *J. Amer. Stat. Assoc.*, **105**, 1480–1493, https://doi.org/10.1198/jasa.2010.tm09181.

Hewaarachchi, A. P., Y. Li, R. Lund, and J. Rennie, 2017: Homogenization of daily temperature data. *J. Climate*, **30**, 985–999, https://doi.org/10.1175/JCLI-D-16-0139.1.

Killick, R., P. Fearnhead, and I. A. Eckley, 2012: Optimal detection of changepoints with a linear computational cost. *J. Amer. Stat. Assoc.*, **107**, 1590–1598, https://doi.org/10.1080/01621459.2012.737745.

Kirch, C., 2006: Resampling methods for the change analysis of dependent data. Ph.D. thesis, Universität zu Köln, 221 pp., https://kups.ub.uni-koeln.de/1795/1/kirchdiss.pdf.

Kwak, S. G., and J. H. Kim, 2017: Central limit theorem: The cornerstone of modern statistics. *Korean J. Anesthesiol.*, **70**, 144–156, https://doi.org/10.4097/kjae.2017.70.2.144.

Li, S., and R. Lund, 2012: Multiple changepoint detection via genetic algorithms. *J. Climate*, **25**, 674–686, https://doi.org/10.1175/2011JCLI4055.1.

Li, Y., and R. Lund, 2015: Multiple changepoint detection using metadata. *J. Climate*, **28**, 4199–4216, https://doi.org/10.1175/JCLI-D-14-00442.1.

Lund, R., and J. Reeves, 2002: Detection of undocumented changepoints: A revision of the two-phase regression model. *J. Climate*, **15**, 2547–2554, https://doi.org/10.1175/1520-0442(2002)015<2547:DOUCAR>2.0.CO;2.

——, and X. Shi, 2020: Short communication: Detecting possibly frequent change-points: Wild binary segmentation 2 and steepest-drop model selection. *J. Korean Stat. Soc.*, **49**, 1090–1095, https://doi.org/10.1007/s42952-020-00081-6.

——, H. Hurd, P. Bloomfield, and R. Smith, 1995: Climatological time series with periodic correlation. *J. Climate*, **8**, 2787–2809, https://doi.org/10.1175/1520-0442(1995)008<2787:CTSWPC>2.0.CO;2.

——, Q. Shao, and I. Basawa, 2006: Parsimonious periodic time series modeling. *Aust. N. Z. J. Stat.*, **48**, 33–47, https://doi.org/10.1111/j.1467-842X.2006.00423.x.

MacNeill, I. B., 1974: Tests for change of parameter at unknown times and distributions of some related functionals on Brownian motion. *Ann. Stat.*, **2**, 950–962, https://doi.org/10.1214/aos/1176342816.

Menne, M. J., and C. N. Williams Jr., 2005: Detection of undocumented changepoints using multiple test statistics and composite reference series. *J. Climate*, **18**, 4271–4286, https://doi.org/10.1175/JCLI3524.1.

——, and ——, 2009: Homogenization of temperature series via pairwise comparisons. *J. Climate*, **22**, 1700–1717, https://doi.org/10.1175/2008JCLI2263.1.

——, J. Williams, N. Claude, and R. S. Vose, 2009: The U.S. Historical Climatology Network monthly temperature data, version 2. *Bull. Amer. Meteor. Soc.*, **90**, 993–1008, https://doi.org/10.1175/2008BAMS2613.1.

Meredith, M., and Coauthors, 2019: Polar regions. *IPCC Special Report on the Ocean and Cryosphere in a Changing Climate*, H.-O. Pörtner et al., Eds., Cambridge University Press, 203–320.

Mitchell, J. M., Jr., 1953: On the causes of instrumentally observed secular temperature trends. *J. Atmos. Sci.*, **10**, 244–261, https://doi.org/10.1175/1520-0469(1953)010<0244:OTCOIO>2.0.CO;2.

Mudelsee, M., 2019: Trend analysis of climate time series: A review of methods. *Earth-Sci. Rev.*, **190**, 310–322, https://doi.org/10.1016/j.earscirev.2018.12.005.

Norwood, B., and R. Killick, 2018: Long memory and changepoint models: A spectral classification procedure. *Stat. Comput.*, **28**, 291–302, https://doi.org/10.1007/s11222-017-9731-0.

O'Neill, P., and Coauthors, 2022: Evaluation of the homogenization adjustments applied to European temperature records in the Global Historical Climatology Network dataset. *Atmosphere*, **13**, 285, https://doi.org/10.3390/atmos13020285.

Peterson, T. C., and Coauthors, 1998: Homogeneity adjustments of in situ atmospheric climate data: A review. *Int. J. Climatol.*, **18**, 1493–1517, https://doi.org/10.1002/(SICI)1097-0088(19981115)18:13<1493::AID-JOC329>3.0.CO;2-T.

Reid, P. C., and Coauthors, 2016: Global impacts of the 1980s regime shift. *Global Change Biol.*, **22**, 682–703, https://doi.org/10.1111/gcb.13106.

Ribeiro, S., J. Caineta, and A. C. Costa, 2016: Review and discussion of homogenisation methods for climate data. *Phys. Chem. Earth*, **94**, 167–179, https://doi.org/10.1016/j.pce.2015.08.007.

Robbins, M., C. Gallagher, R. Lund, and A. Aue, 2011: Mean shift testing in correlated data. *J. Time Ser. Anal.*, **32**, 498–511, https://doi.org/10.1111/j.1467-9892.2010.00707.x.

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

Scott, A. J., and M. Knott, 1974: A cluster analysis method for grouping means in the analysis of variance. *Biometrics*, **30**, 507–512, https://doi.org/10.2307/2529204.

Shen, J., C. M. Gallagher, and Q. Lu, 2014: Detection of multiple undocumented change-points using adaptive LASSO. *J. Appl. Stat.*, **41**, 1161–1173, https://doi.org/10.1080/02664763.2013.862220.

Shi, X., C. Beaulieu, R. Killick, and R. Lund, 2022a: Changepoint detection: An analysis of the Central England temperature series. *J. Climate*, **35**, 6329–6342, https://doi.org/10.1175/JCLI-D-21-0489.1.

——, C. Gallagher, R. Lund, and R. Killick, 2022b: A comparison of single and multiple changepoint techniques for time series data. *Comput. Stat. Data Anal.*, **170**, 107433, https://doi.org/10.1016/j.csda.2022.107433.

Tang, S. M., and I. B. MacNeill, 1993: The effect of serial correlation on tests for parameter change at unknown time. *Ann. Stat.*, **21**, 552–575, https://doi.org/10.1214/aos/1176349042.

Truong, C., L. Oudre, and N. Vayatis, 2020: Selective review of offline change point detection methods. *Signal Process.*, **167**, 107299, https://doi.org/10.1016/j.sigpro.2019.107299.

Venema, V. K. C., and Coauthors, 2012: Benchmarking homogenization algorithms for monthly data. *Climate Past*, **8**, 89–115, https://doi.org/10.5194/cp-8-89-2012.

Wang, X. L., Q. H. Wen, and Y. Wu, 2007: Penalized maximal *t* test for detecting undocumented mean change in climate data series. *J. Appl. Meteor. Climatol.*, **46**, 916–931, https://doi.org/10.1175/JAM2504.1.

Yazici, B., and S. Yolacan, 2007: A comparison of various tests of normality. *J. Stat. Comput. Simul.*, **77**, 175–183, https://doi.org/10.1080/10629360600678310.