GOOD PRACTICES AND COMMON PITFALLS IN CLIMATE TIME SERIES CHANGEPOINT TECHNIQUES: A REVIEW

Robert Lund Department of Statistics, University of California, Santa Cruz rolund@ucsc.edu Claudie Beaulieu Department of Ocean Sciences, University of California, Santa Cruz **Rebecca Killick** Department of Statistics, Lancaster University, UK

QiQi Lu Department of Statistical Sciences and Operations Research, Virginia Commonwealth University Xueheng Shi Department of Statistics, University of California, Davis xhshi@ucdavis.edu

December 7, 2022

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 small perturbations in assumptions; for example, allowing for a trend or correlation in the series can drastically change conclusions. This paper is a review of the changepoint 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 the majority of our applications involve temperature series, other settings are mentioned.

1 Introduction

Climate time series often contain nonlinearities or sudden structural changes in their behavior. Such features may reflect linear or nonlinear dynamics in the climate system, and need to be detected and modeled for an accurate depiction of long-term changes in the time series [1, 2, 3, 4]. Alternatively, these features may reflect artificial discontinuities induced by changes in measurement practices (e.g., station relocations, gauge changes, observer changes) [5, 6, 7, 8]. Some (but not necessarily all) of these artificial changes induce shift discontinuities into the series. If these shifts are not detected and removed from the data, results of analysis using these data may be biased or erroneous. Regardless of the underlying cause of the shift, if the true number of changes and their timings are unknown, changepoint detection techniques are typically used to estimate these key quantities. 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 change is caused by variability or forcings in the climate system, the number of changepoints and their timings are needed to accurately represent long-term changes in the data.

Changepoint detection is a rapidly growing field in the statistical literature, and applications to climate time series are numerous. This paper contains a modern statistical review of the changepoint topic in climate settings. Our overarching goal is to develop a good estimate of 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 about changepoint detection, informing the researcher of the appropriate methods to use given statistical properties of the time series. Unfortunately, changepoints are a thorny statistical issue: small changes in assumptions can yield very different fitted changepoint configurations

(e.g., [1, 2, 9]). Because of this, it is important for researchers to be aware of changepoint/homogenization pitfalls. This paper aims to illuminate some common changepoint mistakes, and to make recommendations on the best practices to avoid them.

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. Indeed, our focus is on the techniques themselves, the intent being to illuminate the concepts that underlie sound changepoint analyses. 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 only want to remove changepoint features from the record that can be attributed to man-made (artificial) influences — changepoints caused by natural fluctuations should be retained. This is best done by subtracting references series from nearby locations from the target series to be homogenized before analysis to eliminate naturally occurring fluctuations. These so-called absolute versus relative homogenization procedures, and the "target" and "reference" series involved in them, are discussed in [10]. While a bit more is said about this in the next section, the issue is not a prominent feature of 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 all changepoint analyses. 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 the multiple changepoint case, which arises when one does not *a priori* know how many changepoints are present, the typical setting in practice. Section 6 then transitions to a list of pitfalls to avoid when homogenizing climate series. Section 7 closes with conclusions and comments, including avenues for 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. The records span January, 1931 — December, 2001. These series will serve to illustrate our list of salient statistical features in climate series.

2.1 Seasonality

A prominent seasonal mean cycle exists in the plotted data in Figure 1. In fact, the yearly range of the sample means exceeds 30°C: from a January minimum of less than -10°C to a July maximum of more than 20.0°C. This seasonal cycle makes it difficult to see small shifts, say on the order of a degree of two (the typical discontinuity magnitude induced by a changepoint), in the record. These small changepoint shifts become critical in assessing long term changes in temperatures. Figure 2 shows the monthly sample means and standard deviations of these two series.

Seasonality is also present in the variability of many climate series. The sample standard deviations in Figure 2 show that winter temperatures are much more variable than summer temperatures; temperate zone examples exist where the January standard deviations of winter temperatures can be some five times larger than summer standard deviations [11]. This same reference also shows that a stationary model modulated for periodicities in mean and variance (or equivalently, the standard deviation - the square root of the variance) adequately describes many periodic climate series $\{X_t\}$:

$$X_{nT+\nu} = \mu_{\nu} + \sigma_{\nu} S_{nT+\nu}.$$
(1)

Here, $X_{nT+\nu}$ is the series observation during the ν th season of the *n*th cycle, *T* is a known period (*T* = 12 for monthly data), {*S*_t} is a zero mean unit variance stationary time series in time *t* (*t* = *nT* + ν), and σ_{ν} is the process standard deviation during season ν . Trends and changepoint features are neglected (for the moment) in the above model.

Seasonal features make changepoint analyses difficult if not taken into account. Elaborating, in time series plots, it can be difficult to visually discern the impact of a changepoint in a temperature series, which typically shifts a series only by a degree or two, when the series has a seasonal cycle of 40 degrees. In a multiple changepoint analysis of a daily series, the methods may flag many spurious changepoints within a year in an attempt to follow the seasonal mean cycle should the seasonal cycle be ignored in the modeling procedure.

2.2 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,



Figure 1: Monthly averaged temperatures at the Mott (Top) and Richardton-Abbey (Middle) stations in west-central North Dakota. The bottom graphic shows the Mott minus Richardton-Abby series in a target minus reference comparison.

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, attributable to correlation, can be mistaken as a shift.

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

$$\operatorname{Corr}(X_t, X_{t+h}) = \frac{\operatorname{Cov}(X_t, X_{t+h})}{\operatorname{Var}(X_t)^{1/2} \operatorname{Var}(X_{t+h})^{1/2}},$$

where $Cov(X_t, X_{t+h}) = E[X_t X_{t+h}] - E[X_t]E[X_{t+h}]$. In terms of (1), $Corr(X_t, X_{t+h}) = Corr(S_t, S_{t+h})$. A clarification here: data should be deseasonalized (i.e., subtracting the seasonal mean cycle) before correlation calculations, a practice followed here. This is because seasonal mean cycles are deemed fixed and not a contributor to variability; however, some authors view the seasonal cycle as a part of annual variability. For more concreteness, our estimates of



Figure 2: Monthly sample means (Top) and standard deviations (Bottom) for the Mott and Richardton-Abbey stations.

the seasonal mean and variance during season ν are, respectively,

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

and our estimate of the lag *h* correlation in $\{S_t\}$ is

$$\widehat{\operatorname{Corr}}(S_t, S_{t+h}) = \frac{1}{dT} \sum_{\ell=1}^{dT-h} \widehat{S}_t \widehat{S}_{t+h}.$$

Here, d is the number of complete cycles of data (we assume that no partial years of data are observed to avoid trite work) and hats indicate estimates of quantities. Note that the first cycle of data is indexed by n = 0 and the last by



Figure 3: Sample correlations at the Mott (Top), Richardton-Abbey (Middle), and the Mott minus Richardton-Abby (Bottom) series after a seasonal standardization for each series over the first 40 months. While the auto correlations in the two individual series are similar, correlation does not vanish in the target minus reference series in the bottom plot.

n = d - 1. Some authors use d in place of d - 1 in the definition of $\hat{\sigma}_{v}$; others use dT - h in place of dT in the definition of $\widehat{\text{Corr}}(S_t, S_{t+h})$.

Figure 2 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 non-zero correlation exists at both stations.

Statistical methods for changepoint detection often degrade when correlation is present. Indeed, an example is shown below where a changepoint declaration is repealed once correlation is taken into account. A key aspect of this paper is how to deal with cases where both correlation and mean shift changepoints are plausible.

2.3 Target-Reference Comparisons

Climate homogenization seeks to adjust time series for artificial features only, such as station relocations and instrumentation changes. Natural/anthropogenic changepoints occasionally exist in series and are generally thought to be 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). When one series is subtracted from the other, the target minus reference subtraction serves to remove natural fluctuations, especially if the target and reference series experience similar weather. This subtraction reduces or altogether eliminates seasonal cycles and 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 differenced series, so some negatives are incurred in the comparison.

The bottom plot in Figure 1 shows the Mott series subtracted from the Richardton-Abby series. Observe that the seasonal cycle has lessened, if not altogether disappeared. If the target-reference comparison is good, any 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 Figure 3 shows sample correlations from the Mott minus Richardton-Abby stations along with 95% pointwise confidence bounds for zero correlation (white noise). These are correlations computed after a sample monthly mean has been subtracted from the series and the series has been further divided by a monthly sample standard deviation to put all temperatures on the same mean zero unit variance scale. Notice that significant non-zero correlations; in fact, they often serve to increase them.

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, it's practical implications are profound. The authors are strong proponents of using target minus reference comparisons. We refer the reader to [5, 12] for more on target-reference comparisons. Some modern methods use multiple references series, sometime as many as forty [5, 12].

2.4 Trends

Many climate series have long term trends. For example, in the Mott series in Figure 1, a long term linear trend of $0.858^{\circ}C$ /Century is estimated (computed neglecting any changepoints). Of course, many temperature series exhibit recent warming, and many other climatic series have substantial trends as well. 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 in an attempt to follow the trend.

While here, trends and seasonal means are features in the series mean $E[X_t]$. Since a changepoint analysis typically tries to find abrupt changes in $E[X_t]$, it is imperative that these features should be competently described in the modeling scheme. Unfortunately, even for single changepoint tests, the best test statistic to use will change depending on the true form of $E[X_t]$. While one goal of this article is to deconvolve this issue for the reader, the limit distribution (and test percentiles) for a test with a long term trend are substantially different from those without a trend.

2.5 Normality

Climate time series may or may not be Gaussian (normal). A series is Gaussian if its marginal distributions come from the normal distributional family. Series that are averaged — like monthly or annual series that are obtained by averaging daily data — are often very close to normal by the central limit effect [13]. This 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 gives a graphical check for normality where points lying on the diagonal indicates the data is well described by a Gaussian model. A commonly used and powerful non-parametric statistical test for normality is the Shapiro-Wilks test. The *p*-value for the Shapiro-Wilks test is 0.734, reinforcing that normality is quite plausible for the Mott series (see [14] for more on normality tests). We do however need to stress that the Gaussian assumption is placed on the residuals (from $E[X_t]$) thus if there is trend, seasonality or changepoints within the data these must before removed before testing.

Some climate series are decisively non-Gaussian. Examples include discrete categorical series of cloud cover, ordered from zero (clear sky) to ten (complete overcast), zero-one series describing an on/off phenomenon, and series whose marginal distributions are skewed (such as annual precipitation). Averaging tends to induce normality. For example, while the monthly averaging of daily data above rendered the Mott series essentially Gaussian, daily data is often skewed and non-normal. In fact, daily temperatures at temperate zone stations often have a heavy left tail (are skewed), especially in winter; see [15] for an example.

3 Time Series Models

The classical decomposition of a time series $\{X_t\}$ has the form

$$X_t = \mu_t + s_t + \epsilon_t, \tag{2}$$

where $\{X_t\}$ is the observed series, $\{\mu_t\}$ is a long-term trend, $\{s_t\}$ 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 (2). The seasonal cycle $\{s_t\}$ is periodic in that $s_{t+T} = s_t$ for all times t. When the parametrization for $\{\mu_t\}$ contains a location parameter, one typically assumes that $\sum_{t=1}^{T} s_t = 0$ so that regression parameters are statistically identifiable. This is the so-called classical decomposition model in [16].

For a simple example, suppose that one is examining an annual series for possible 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

$$X_t = \beta_0 + \beta_1 t + \delta_t + \epsilon_t, \tag{3}$$

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

$$\delta_t = \begin{cases} \Delta_1 = 0, & \tau_0 \le t < \tau_1, \\ \Delta_2, & \tau_1 \le t < \tau_2, \\ \vdots \\ \Delta_{m+1}, & \tau_m \le t < N. \end{cases}$$

The above setup takes data observed at the times 1, 2, ..., N and allows for *m* mean shift changepoints occurring at the times $\tau_1 < \tau_2 < \cdots < \tau_m$; the changepoint count *m* and their occurrence times $\tau_1, ..., \tau_m$ are all unknown. If the location parameter β_0 is omitted from the long-term trend expression, one need not require that $\Delta_1 = 0$.

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

In what follows, 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 [17], or even changes in the marginal distributions of the series [18], but the major focus on the climate literature to date has been on mean shifts. Toward this, the mean shift changepoints here shift all subsequent series values by the same amount — shifts are not seasonal. While the methods here could be modified to induce a seasonal change, this extension is not considered here.

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

$$X_{nT+\nu} = \mu_{nT+\nu} + s_{\nu} + \delta_{nT+\nu} + \epsilon_{nT+\nu}, \tag{4}$$

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

$$\mu_t = \begin{cases} \beta_1 + \alpha_1 t, & \tau_0 \le t < \tau_1, \\ \beta_2 + \alpha_2 t, & \tau_1 \le t < \tau_2, \\ \vdots \\ \beta_{m+1} + \alpha_{m+1} t, & \tau_m \le t < N. \end{cases}$$

The time series component $\{\epsilon_t\}$ is typically assumed to be stationary when T = 1, or periodically stationary when T > 1. A good and flexible model class for stationary series are the autoregressive (AR) series [16]. A *p*th order zero mean autoregression is uniquely characterized by the *p*-th order linear difference equation

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

where ϕ_1, \ldots, ϕ_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 $\phi_1(\nu), \dots, \phi_p(\nu)$ are the autoregressive parameters during season ν and $\{Z_{nT+\nu}\}$ is periodic white noise having a periodic variance $\operatorname{Var}(Z_{nT+\nu}) = \sigma_{\nu}^2$. 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. Our later examples will exhibit some techniques to reduce these parameter counts.

4 Single Changepoint Detection

4.1 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 some 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 compare sample means of the series before and after all possible candidate changepoint times. That is, they compare differences between $k^{-1} \sum_{t=1}^{k} X_t$ and $(N-k)^{-1} \sum_{t=k+1}^{N} X_t$ for each admissible changepoint time k, selecting the k where this difference is statistically maximal as the changepoint time estimate.

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

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

where

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

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

Pitfall 1: Many past climate changepoint authors examine a "maximum statistic" akin to $D^* = \max_{2 \le k \le N} |\text{CUSUM}_X(k)|$ to check for a single changepoint. Where the maximum occurs is estimated as the time of the changepoint. While this is fine, incorrect null hypothesis distribution percentiles abound in the climate literature, often producing unjustifiable conclusions [9, 19]. When a changepoint is known to occur at time k, $\text{CUSUM}_X(k)$ should be used as the test statistic. Scaling $\text{CUSUM}_X(k)$ to a z, t, or even F distribution is used to make conclusions. When the time of the changepoint is unknown, the maximum statistic D^* is used. The correct null hypothesis percentiles for D^* must account for the many times k where the maximum could arise — these percentiles are much larger than those for a fixed k. The correct asymptotic quantification of AMOC changepoint statistics is often unwieldy as the scenario is not readily scaleable to an extreme value distribution. Indeed, $\text{CUSUM}_X(k)$ are highly correlated in k (they are not IID). The distribution of AMOC tests often converges to the supremum of some Gaussian process. The reader is referred to [20, 21] for historical technical development.

Best Practice 1: Several legitimate statistics can be used to detect a single mean shift changepoint. One test with great detection power uses a sum of squared CUSUM statistics to assess whether a changepoint is present:

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

The time of the changepoint is still estimated as the location $k \ge 2$ that maximizes $|\text{CUSUM}_X(k)|$. This test won the single changepoint comparison competition in [22], originates from [23], 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 [22]. The null hypothesis percentiles of this distribution are presented in Table 1 for convenience and are extracted from [22]. 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 [22].

| Percentile | Critical Value |
|------------|----------------|
| 90.0% | 0.3473046 |
| 95.0% | 0.4613744 |
| 97.5% | 0.5806168 |
| 99.0% | 0.7434348 |

Table 1: Critical Values for SCUSUM Statistics

4.2 Autocorrelation

We now move to AMOC tests in correlated data. A significant body of statistical research modifies the limit theory for IID data to account for autocorrelation [19, 22]. Much of this literature takes the following flavor. With the SCUSUM test above (and other AMOC tests), simply replace $\hat{\sigma}$ with an estimate of the long-run variance parameter τ defined by

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

Then most asymptotic limit laws apply 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. A recent statistical reference on this topic is [24].

In some asymptotic tests, convergence to the limit law can be slow, making application to even a century of annual data questionable. A preferable way to handle correlation involves pre-whitening techniques. Statistical reference for pre-whitening are [19, 25]. To account for correlation in an AMOC changepoint analysis, pre-whitening first fits a *p*-th order autoregressive (AR(*p*)) model to the series (this assumes annual or non-periodic 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 ϕ_1, \ldots, ϕ_p and the white noise variance σ^2 . Next, the one-step-ahead predictions

$$\hat{X}_{t+1} = \hat{\phi}_1 X_t + \ldots + \hat{\phi}_p X_{t-p+1}, \quad t \ge p,$$

are calculated with $\hat{\phi}_j$ replacing ϕ_j and the one-step-ahead prediction errors $Y_t = X_t - \hat{X}_t$ are formed. 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 called the pre-whitened series. To compute the startup values $\hat{X}_1, \dots, \hat{X}_p$, one uses the time series prediction equations; see Chapter 3 of [16] for details.

Next, one simply applies the SCUSUM (or some other AMOC) test to the pre-whitened $\{Y_t\}$ using the percentiles for IID errors to make conclusions. [19] proves that this procedure is statistically valid asymptotically. More importantly, this paper shows that the limit laws typically "kick in more quickly" than asymptotic laws that replace $\hat{\sigma}$ with $\hat{\tau}$.

While pre-whitening adds another layer to the analysis, our next pitfall notes the importance of taking correlation into account.

Pitfall 2: Ignoring positive correlation in a series will often produce spurious changepoint conclusions. In fact, series that are heavily positively correlated tend to have 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.

Best Practice 2: Pre-whiten any autocorrelated series before applying AMOC IID changepoint tests. As shown below, dubious conclusions can arise when autocorrelation is ignored. A general theme for AMOC tests with positively correlated data, which entail the majority of climate cases, is clear: one risks concluding that a changepoint exits when in truth it does not when positive correlation is ignored. The situation reverses itself should negatively correlated data be encountered.

4.3 An Example

We now examine the annual Central England temperature (CET) series from 1900-2020 with a single changepoint test. The CET record was provided by the UK Met Office at https://www.metoffice.gov.uk/hadobs/hadcet/. For a multiple changepoint analysis of the entire series dating back to the 1600s, see [26]. Figure 4 plots this series against several single changepoint configurations explored below. Conclusions are shown below to be heavily dependent on the assumptions made.

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 is 0.4614, 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 mean configuration is plotted against the series in the top panel of Figure 4.

We now rerun the single mean shift test allowing for autocorrelated errors, this time using a simple AR(1) structure for the model errors. An SCUSUM test was applied to the pre-whitened AR(1) one-step-ahead prediction errors and gives SCUSUM_Z = 0.1799, which is well below the 0.4614 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 changepoint declaration made in the above paragraph. Here, the estimated AR(1) correlation between consecutive series observations is $\hat{\phi} = 0.425$, which is not extreme autocorrelation. The conflicting conclusions illustrate why one needs to be careful to allow for correlation in changepoint tests when correlation is present; neglecting to account for positive correlation often leads to an overestimation of the number of changepoints.

Table 2: Single Changepoint Tests for the Central England Series

| Model Assumptions | Test | <i>p</i> -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 |

4.4 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 the CET series but not 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 constant 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 "follow" the seasonal mean and long term trend. In short, changepoint techniques are not robust to changes in $\mu_t = E[X_t]$. Unfortunately, in changepoint analyses, each different form of m_t requires a different set of null hypothesis percentiles. For example, in the case of a simple mean shift where $\mu_t = \beta_0$, the 95th percentile of the CUSUM test is 1.358; when there is a linear trend $\mu_t = \beta_0 + \beta_1 t$, the 95th CUSUM percentile becomes 0.902.

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 does not have a constant mean.

Best Practice 3: Account for all 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.

4.5 CET Example Rejoinder

As an example, we return to the 1900-2020 CET series. Global warming posits a slow temperature increase; as such, an AMOC analyses with the linear trend $\mu_t = \beta_0 + \beta_1 t$ is explored. AR(1) errors are also allowed in the procedure. An AMOC CUSUM-type mean shift test for IID errors is developed in (**author?**) [18] (we are unaware of anyone studying SCUSUM tests in the linear trend setting). This test is denoted by CUSUM_D. Estimating the linear trend and AR(1) specifications under the null hypothesis of no changepoints provides $\hat{\beta}_0 = 9.1^\circ \text{ C}$, $\hat{\beta}_1 = -0.0092^\circ \text{ C/year}$, and $\hat{\phi} = 0.194$.



(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.

Figure 4: Single changepoint tests flag a mean shift in 1988.

One needs to be careful to account for the trend when pre-whitening this series. Specifically, our estimated one-stepahead prediction with AR(1) errors is

$$\hat{X}_{t} = \hat{\mu}_{t} + \hat{\phi}(X_{t-1} - \hat{\mu}_{t-1}) = \hat{\beta}_{0} + \hat{\beta}_{1}t + \hat{\phi}\left[X_{t-1} - \hat{\beta}_{0} - \hat{\beta}_{1}(t-1)\right],$$
(5)

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

The CUSUM_D test applied to $\{Y_t\}$ gives a statistic of 0.929, which is slightly above the 95th percentile null hypothesis threshold of 0.9028. The *p*-value for this test is 0.038. With 95% confidence, the 1988 changepoint has come back. The bottom panel in Figure 4 displays the fit to this data. This configuration is the best fitting of our three models. Overall, it is the most reliable fit as it takes into account both trends and autocorrelation. See [26] for a detailed analysis of the CET series.

Obviously, the assumptions made in changepoint analyses are extremely important and may influence conclusions. While issues become even more complex in multiple changepoint settings, 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, roughly [27]. As in the AMOC case, multiple changepoint (MCPT) detection is also 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 [28]. 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 with AMOC tests. 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 [22]. This point is further reinforced below.

Other approaches to the MCPT problem can be classified into distinct camps. One camp examines recursive segmentation procedures that improve upon binary segmentation methods include wild binary segmentation [29] and wild contrast maximization [30]. 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 [22], essentially rendering them unusable in climate cases where say two changepoints exist in a hundred year climate series.

Another camp applies dynamic programming to MCPT problems. Here, an objective function associated with the problem is optimized. The segmented neighborhood algorithm of [31] and the pruned exact linear time of [32] are two examples. Dynamic programming techniques provide optimal (relative to the chosen objective function) changepoint configurations and can be computed very 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 in [33]. Model selection approaches such as [34] and [35] and scan statistics procedures based on moving sum statistics [36] also exist. Other methods also exist — 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. MCPT techniques for independent $\{\epsilon_t\}$ can give suboptimal answers for correlated series [17, 37, 38]. While one can still pre-whiten the series, estimation of the correlation structure and the multiple mean shift sizes and locations confound each other. No null hypothesis suggests itself in the MCPT setting. In the AMOC case, estimates of the series' correlation structure were computed under the null hypothesis of no changepoints and models containing no and one changepoints were statistically compared. In the MCPT case, the number of changepoints in the null hypothesis is unclear. Trends and seasonality further impede issues.

Penalized likelihood methods, another MCPT camp, attack 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 $L(m; \tau_1, ..., \tau_m)$ denote the likelihood of the best time series model having *m* changepoints at the times $1 < \tau_1 < \tau_2 < \cdots < \tau_m \leq N$. Likelihoods for a Gaussian series $\{X_t\}_{t=1}^N$ take the classical time series form

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

where \hat{X}_t is the best linear prediction of X_t from past observations in (5) and $V_t = E[(X_t - \hat{X}_t)^2]$ is its unconditional mean squared error.

As the number of changepoints *m* increases, the model fit improves: $L(m; \tau_1, ..., \tau_m)$ increases in *m*. However, after a while, adding additional changepoints does not appreciably improve the likelihood. This is where the penalty term comes in. The penalty for having *m* changepoints at the times $\tau_1, ..., \tau_m$ is denoted by $P(m; \tau_1, ..., \tau_m)$ and grows 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 τ_1, \ldots, τ_m . When there are no changepoints (m = 0), the penalty term is taken as zero.

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

AIC:
$$P(m; \tau_1, ..., \tau_m) = 2(2m+2)$$

BIC: $P(m; \tau_1, ..., \tau_m) = (2m+2)\ln(N)$

mBIC:
$$P(m; \tau_1, ..., \tau_m) = 3m \ln(N) + \sum_{i=1}^{m+1} \ln\left(\frac{\tau_i - \tau_{i-1}}{N}\right)$$

MDL:
$$P(m;\tau_1,\ldots,\tau_m) = \sum_{i=1}^{m+1} \ln(\tau_i - \tau_{i-1}) + 2\ln(m) + 2\sum_{i=2}^m \ln(\tau_i)$$
(6)

The penalties above are for the changepoint configuration portion of the model only; should the time series structure change at each changepoint time, the above formulae require modifications. When the time series structure is constant across all series segments, the above formulae are appropriate. While other penalties exist, these are the major 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 performances is found in [22]; however, AIC typically overestimates the true number of changepoints and is not used. For a penalty that does not depend on the changepoint location times, BIC performs surprisingly well in a variety of settings [22].

In optimizing $O(m; \tau_1, ..., \tau_m)$, significant computational issues arise. To compute $P(m; \tau_1, ..., \tau_m)$, an optimal time series model with *m* changepoints at the times $\tau_1, ..., \tau_m$ 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. This total is immense for even *N* as large as 100, making exhaustive identification of the best changepoint configuration a strenuous task. Authors have used genetic algorithms [17, 39] to overcome these difficulties. Today, despite computational issues, penalized likelihood is considered the gold standard of MCPT problems. Moreover, work is currently being pursued to bring rapid computation to the setting where the time series parameters are constant across all regimes [33].

5.1 Binary Segmentation

As the earliest invented and still widely used MCPT technique, binary segmentation's popularity rests on two ingredients: ease of interpretation 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 [29], injects randomization into the changepoint search to avoid local optimums. However, simulation studies in [40] suggest that wild binary segmentation overestimates changepoint counts for IID model errors, and becomes dysfunctional in settings with correlated errors. Wild contrast maximization [30], another improvement of wild binary segmentation designed for dependent processes, is capable of handling serial dependence. While we will not discourage the user from using this technique, we also comment that it has not be fully vetted as of today.

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 was constructed. Here, Gaussian series of length 500 were simulated with white noise errors. Three equally spaced changepoints were inserted shifting the series by a unit length in alternating directions. This partitions the series into four equal length segments of 125 points each; Figure 5 displays a sample generated series.



Figure 5: A series with three equally spaced mean shifts of unit size shifting the series in alternating directions. The regression errors are uncorrelated white noise with a unit variance.



Figure 6: Binary segmentation and penalized likelihood methods compared. The biggest errors occur with binary segmentation. Note 95% threshold is used for binary segmentation and BIC, MDL and mBIC penalized likelihoods are optimized by the genetic algorithm (GA).

We randomly generated 100 such series and applied several different changepoint methods. The estimated changepoint configurations were compared to the true changepoint configuration with the distance metric in [22]). This distance incorporates both *m* and the changepoint locations τ_1, \ldots, τ_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 100 simulations are summarized in Figure 6. The red dots in the boxplots demarcate the average distance to the correct changepoint configuration. The boxplots show that binary segmentation underperforms all penalized likelihood methods.



Figure 7: Changepoints flagged by a BIC penalized likelihood (red) and binary segmentation (blue). Binary segmentation flags one changepoint, while a BIC penalized likelihood flags three.

Best Practice 4: Use penalized likelihood MCPT methods — or at least one of the binary segmentation improvements like wild binary segmentation or wild contrast maximization.

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, [38] requires AR(1) errors. While the AR order is not believed to be as important as other issues in most climate applications, it is also infeasible that an AR(1) correlation structure works for all series. See [41] for daily temperature homogenization via penalized likelihoods. While [30] allows general AR(p) errors, simulations indicate that wild contrast maximization tends to estimate too many changepoints. In what follows, we concentrate on general penalized likelihood techniques estimated by a genetic algorithm.

5.2 Atlanta Airport Temperatures

To see differences between the approaches in practice, annual mean surface temperatures from 1879-2013 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 reflects "raw" temperatures that were not adjusted for potential artifacts. BIC penalized likelihoods and binary segmentation were fitted and compared, with results depicted in Figure 7. Binary segmentation flags only one 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 [22], binary segmentation often has trouble identifying mean shifts that move the series in opposite directions. In this case, the successive changepoints estimated by penalized likelihood move the series up, down, and then up, two changepoints are apparently missed by binary segmentation.

5.3 Ignoring Trends

As in the AMOC case, ignoring trends in the MCPT setting will produce spurious results. For if the trend $\mu_t = E[X_t]$ is decisively increasing or decreasing, but ignored in the analysis, then MCPT procedure should flag one or more changepoints in an attempt to shift with the series mean. In the AMOC case, each different form of $\{\mu_t\}$ changes the asymptotic percentiles of the statistical test [42]. In the MCPT case, as long as the trend has the same form and parameters in all series subsegments, the penalties in (6) can be used. Should one want models where all parameters shift at the changepoint times — an example would allow the trend slope to depend on the regime — then the penalties in (6) must be modified. The reader is referred to [26] for the appropriate penalties.



Figure 8: A changepoint analysis of the Arctic sea ice series. When trends are ignored and AR(1) errors are assumed, binary segmentation flags numerous mean shift changepoints in 1994, 2001, 2015 that attempt to "follow the trend".

Pitfall 5: Applying MCPT techniques to series with trends or seasonality that neglects these features.

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

Best Practice 5: Allow for trends and/or seasonality in series with these features.

5.4 Arctic Sea Ice

To illustrate the importance of accounting for trends, we analyse a series of September sea ice extent in the Northern hemisphere from 1979 - 2021. The data was provided by the National Snow and Ice Data Center, and downloaded from: https://nsidc.org/data. The series exhibits a decreasing trend that is partly attributable to increasing greenhouse gases emissions [43]. Figure 8, 9 and 10 show the series and some MCPT fits. Figure 8 is a binary segmentation fit with AR(1) errors that estimates three changepoints. A BIC-type penalized likelihood estimated MCPT configuration with AR(1) errors identifies four changepoints (Figure 9). Both of these fits were calculated assuming no trend. When a linear trend is allowed in the BIC penalized likelihood, all changepoints are repealed (Figure 10). The estimated trend slope of sea ice retreat was -0.053 million km²/year.

6 Discussion and Comments

This paper highlights some common pitfalls in changepoint analysis/homogenization methods and suggests best practices to avoid them. In general, changepoint methods are not robust to changes in the mean structure of a series 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 that feature truly exists. 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 or dubious analyses in the field. For example, the number of changepoint declarations that would be repealed due to failure to consider positive autocorrelation would be extensive.

It is worth rehashing target minus reference series analyses versus target series analyses only (absolute versus relative homogenization). While the statistical procedure to analyse both settings are the same, subtraction of a references series often reduces trends and/or seasonal cycles, making some issues clearer. Nonetheless, as shown here, formation of a target minus reference series often does not reduce series autocorrelation, nor need it totally eliminate trends and/or



Figure 9: A changepoint analysis of the Arctic sea ice series. When trends are ignored and AR(1) errors are assumed, BIC penalized likelihood flags numerous mean shift changepoints in 1995, 2006, 2016, 2017 that attempt to "follow the trend".



Figure 10: After a linear decreasing trend is added to the analysis and AR(1) errors are still assumed, all changepoints are repealed in a BIC penalized likelihood fit.

seasonal cycles. Existence of metadata is another issue. While most authors tend to eschew metadata in their analyses, [39] show how an informative Bayesian prior can be constructed from it and used with an MDL penalized likelihood to enhance changepoint detection power.

Multiple changepoint techniques are actively being researched in statistics. Computational advances are expected within the next few years, especially in regard to penalized likelihood methods [33]. Other aspects about the problem are also being studied. One thing that is already clear from the literature is the inferiority of ordinary binary segmentation techniques in multiple changepoint problems. Here, we urge researchers to use better methods.

Acknowledgement

Robert Lund thanks National Science Foundation Grant DMS-2113592 for partial support. Xueheng Shi thanks National Science Foundation Grant CCF-1934568 for partial support. Claudie Beaulieu thanks National Science Foundation Grant AGS-2143550 for support.

Data statement

The Central England data used in this study is available at https://www.metoffice.gov.uk/hadobs/hadcet/. We used the annual records from 1659-2020.

References

- [1] Claudie Beaulieu, Jie Chen, and Jorge L. Sarmiento. Change-point analysis as a tool to detect abrupt climate variations. *Philosophical Transactions of the Royal Society*, 370:1228–1249, 2012.
- [2] Claudie Beaulieu and Rebecca Killick. Distinguishing trends and shifts from memory in climate data. *Journal of Climate*, 31(23):9519 9543, 2018.
- [3] Niamh Cahill, Stefan Rahmstorf, and Andrew C Parnell. Change points of global temperature. *Environmental Research Letters*, 10(8):084002, 2015.
- [4] Manfred Mudelsee. Trend analysis of climate time series: A review of methods. *Earth-Science Reviews*, 190:310–322, 2019.
- [5] Matthew J. Menne and Claude N Williams, Jr. Homogenization of temperature series via pairwise comparisons. *Journal of Climate*, 22:1700–1717, 2009.
- [6] S. Ribeiro, J. Caineta, and A.C. Costa. Review and discussion of homogenisation methods for climate data. *Physics and Chemistry of the Earth, Parts A/B/C*, 94:167–179, 2016. 3rd International Conference on Ecohydrology, Soil and Climate Change, EcoHCC'14.
- [7] Thomas C. Peterson, David R. Easterling, Thomas R. Karl, Pavel Groisman, Neville Nicholls, Neil Plummer, Simon Torok, Ingeborg Auer, Reinhard Boehm, Donald Gullett, Lucie Vincent, Raino Heino, Heikki Tuomenvirta, Olivier Mestre, Tamás Szentimrey, James Salinger, Eirik J. Førland, Inger Hanssen-Bauer, Hans Alexandersson, Philip Jones, and David Parker. Homogeneity adjustments of in situ atmospheric climate data: a review. *International Journal of Climatology*, 18(13):1493–1517, 1998.
- [8] V.K. Venema, O. Mestre, E. Aguilar, I. Aue, J.A. Guijarro, P. Domonkos, G. Vertacnik, T. Szentimrey, P. Stepanek, P. Zahradnicek, and J. Viarre. Benchmarking homogenization algorithms for monthly data. *Climate of the Past*, 8:89–115, 2012.
- [9] Robert Lund and Jaxk Reeves. Detection of undocumented changepoints: A revision of the two-phase regression model. *Journal of Climate*, 15:2547–2554, 2002.
- [10] Matthew J. Menne, Jr. Williams, Claude N., and Russell S. Vose. The U.S. Historical Climatology Network Monthly Temperature Data, Version 2. *Bulletin of the American Meteorological Society*, 90(7):993–1008, 07 2009.
- [11] R. B. Lund, H. Hurd, P. Bloomfield, and R. L. Smith. Climatological time series with periodic correlation. *Journal of Climate*, 11:2787–2809, 1995.
- [12] Matthew J. Menne and Claude N. Williams, Jr. Detection of undocumented changepoints using multiple test statistics and composite reference series. *Journal of Climate*, 18:4271–4286, 2005.

- [13] Sang Gyu Kwak and Jong Hae Kim. Central limit theorem: the cornerstone of modern statistics. *Korean Journal of Anesthesiology*, 70(2):144, 2017.
- [14] Berna Yazici and Senay Yolacan. A comparison of various tests of normality. Journal of Statistical Computation and Simulation, 77(2):175–183, 2007.
- [15] Robert Lund, Qin Shao, and Ishwar Basawa. Parsimonious periodic time series modeling. Australian & New Zealand Journal of Statistics, 48(1):33–47, 2006.
- [16] Peter J. Brockwell and Richard A. Davis. Time Series: Theory and Methods. Springer, 2 edition, 1991.
- [17] R. A. Davis, T. C. M. Lee, and G. A. Rodrigues-Yam. Structural break estimation for nonstationary time series models. *Journal of the American Statistical Association*, 101:223–239, 2006.
- [18] C. Gallagher, R. B. Lund, and M. Robbins. Changepoint detection in daily precipitation series. *Environmetrics*, 23:407–419, 2012.
- [19] Michael Robbins, C. Gallagher, R. B. Lund, and A. Aue. Mean shift testing in correlated data. *Journal of Time Series Analysis*, 32:498–511, 2011.
- [20] Miklos Csörgo and Lajos Horváth. Limit Theorems in Change-point Analysis. John Wiley & Sons, 1997.
- [21] Ian B MacNeill. Tests for change of parameter at unknown times and distributions of some related functionals on Brownian motion. *The Annals of Statistics*, pages 950–962, 1974.
- [22] Xuesheng Shi, Colin Gallagher, Robert Lund, and Rebecca Killick. A comparison of single and multiple changepoint techniques for time series data. *Computational Statistics & Data Analysis*, 170:107433, 2022.
- [23] Claudia Kirch. *Resampling methods for the change analysis of dependent data*. PhD thesis, Universität zu Köln, 2006.
- [24] Gianfranco Pierobon. Codes for zero spectral density at zero frequency (corresp.). *IEEE Transactions on Information Theory*, 30(2):435–439, 1984.
- [25] Colin Gallagher, Rebecca Killick, Robert Lund, and Xueheng Shi. Autocovariance estimation in the presence of changepoints. *Journal of the Korean Statistical Society*, pages 1–20, 2022.
- [26] Xueheng Shi, Claudie Beaulieu, Rebecca Killick, and Robert Lund. Changepoint detection: An analysis of the Central England temperature series. *Journal of Climate*, pages 1–46, 2022.
- [27] J. Murray Mitchell. On the causes of instrumentally observed secular temperature trends. *Journal of Meteorology*, 10:244–261, 1953.
- [28] Andrew Jhon Scott and M Knott. A cluster analysis method for grouping means in the analysis of variance. *Biometrics*, pages 507–512, 1974.
- [29] Piotr Fryzlewicz. Wild binary segmentation for multiple change-point detection. Annals of Statistics, 42:2243– 2281, 2014.
- [30] Haeran Cho and Piotr Fryzlewicz. Multiple change point detection under serial dependence: Wild energy maximisation and gappy schwarz criterion. arXiv preprint arXiv:2011.13884, 2020.
- [31] Ivan E Auger and Charles E Lawrence. Algorithms for the optimal identification of segment neighborhoods. *Bulletin of Mathematical Biology*, 51(1):39–54, 1989.
- [32] Rebecca Killick, Paul Fearnhead, and Idris A Eckley. Optimal detection of changepoints with a linear computational cost. *Journal of the American Statistical Association*, 107(500):1590–1598, 2012.
- [33] Colin Gallagher, Rebecca Killick, Robert Lund, and Xeuheng Shi. Rapid dynamic programming changepoint estimation with gradient step and search. *In Preperation*, 2022+.
- [34] Zaid Harchaoui and Céline Lévy-Leduc. Multiple change-point estimation with a total variation penalty. *Journal* of the American Statistical Association, 105(492):1480–1493, 2010.
- [35] Jie Shen, Colin M Gallagher, and QiQi Lu. Detection of multiple undocumented change-points using adaptive Lasso. *Journal of Applied Statistics*, 41(6):1161–1173, 2014.
- [36] Birte Eichinger and Claudia Kirch. A MOSUM procedure for the estimation of multiple random change points. *Bernoulli*, 24(1):526–564, 2018.
- [37] Shanghong Li and Robert Lund. Multiple changepoint detection via genetic algorithms. *Journal of Climate*, 25:674–686, 2012.
- [38] Souhil Chakar, E Lebarbier, Céline Lévy-Leduc, and Stéphane Robin. A robust approach for estimating changepoints in the mean of an AR(1) process. *Bernoulli*, 23(2):1408–1447, 2017.

- [39] Yingbo Li and Robert Lund. Multiple changepoint detection using metadata. *Journal of Climate*, 28:4199–4216, 2015.
- [40] Robert Lund and Xueheng Shi. Detecting possibly frequent change-points: Wild Binary Segmentation 2 and steepest-drop model selection. *Journal of the Korean Statistical Society*, 49(4):1090–1095, 2020.
- [41] Anuradha P Hewaarachchi, Yingbo Li, Robert Lund, and Jared Rennie. Homogenization of daily temperature data. *Journal of Climate*, 30(3):985–999, 2017.
- [42] S. M. Tang and I. B. MacNeill. The effect of serial correlation on tests for parameter change at unknown time. *The Annals of Statistics*, 21:552–575, 1993.
- [43] M. Meredith, M. Sommerkorn, S. Cassotta, C. Derksen, A. Ekaykin, Hollowed, G. Kofinas, A. Mackintosh, J. Melbourne-Thomas, M.M.C. Muelbert, G. Ottersen, H. Pritchard, and E.A.G. Schuur. *IPCC Special Report* on the Ocean and Cryosphere in a Changing Climate, chapter 3, pages 203–320. Cambridge University press, Cambridge, UK and New York, NY, 2019. https://doi.org/10.1017/9781009157964.005.