Changepoint Detection: An Analysis of the Central England Temperature Series

Xueheng Shi

Department of Statistics, University of California, Santa Cruz

Claudie Beaulieu[*]

Department of Ocean Sciences, University of California, Santa Cruz

Rebecca Killick

Department of Statistics, Lancaster University

Robert Lund

Department of Statistics, University of California, Santa Cruz

[*] *Corresponding author*: Claudie Beaulieu, beaulieu@ucsc.edu

Early Online Release: This preliminary version has been accepted for publication in *Journal of Climate* cited, and has been assigned DOI 10.1175/JCLI-D-21-0489.1. The final typeset copyedited article will replace the EOR at the above DOI when it is published.

ABSTRACT

This paper presents a statistical analysis of structural changes in the Central England temperature series, one of the longest surface temperature records available. A changepoint analysis is performed to detect abrupt changes, which can be regarded as a preliminary step before further analysis is conducted to identify the causes of the changes (e.g., artificial, human-induced or natural variability). Regression models with structural breaks, including mean and trend shifts, are fitted to the series and compared via two commonly used multiple changepoint penalized likelihood criteria that balance model fit quality (as measured by likelihood) against parsimony considerations. Our changepoint model fits, with independent and short-memory errors, are also compared with a different class of models termed long-memory models that have been previously used by other authors to describe persistence features in temperature series. In the end, the optimal model is judged to be one containing a changepoint in the late 1980s, with a transition to an intensified warming regime. This timing and warming conclusion is consistent across changepoint models compared in this analysis. The variability of the series is not found to be significantly changing, and shift features are judged to be more plausible than either short- or long-memory autocorrelations. The final proposed model is one including trend-shifts (both intercept and slope parameters) with independent errors. The analysis serves as a walk-through tutorial of different changepoint techniques, illustrating what can be statistically inferred.

1. Introduction

Climate time series often contain abrupt changes and other nonlinearities in their behavior. Changepoints are times of abrupt shifts in a series' characteristics, including means, trends, variances, and autocorrelations. For examples, a sudden change from a cooling period (i.e., decreasing trend) to a warming period can be characterised by a changepoint in the trend; a sudden increase due to the relocation of a station may be characterised as a changepoint in the mean. Abrupt changes may be caused by changes in climate forcings, related to climate variability in the ocean and atmosphere, or induced by artificial changes in measurement procedures such as station relocations or instrumentation changes.

It is crucial to know changepoint times in climate series, especially when assessing long-term trends, as their presence may grossly alter trend estimates, which impedes our understanding of external forcings and climate variability over the instrumental record (Lund et al. 2007; Beaulieu et al. 2012; Cahill et al. 2015; Beaulieu and Killick 2018). Series with artificial changes merit adjustment via homogenization methods, as trends and extreme quantiles are more accurately estimated from homogenized data (Hewaarachchi et al. 2017; Trewin et al. 2020; Vincent et al. 2020). On average, approximately six station relocations or instrumentation changes occur over a century in a randomly selected US climate station (Mitchell Jr. 1953; Menne and Williams Jr. 2009). As such, a changepoint analysis of a climate series is often a worthy initial exploratory endeavor.

Statistical methods to detect changepoints have rapidly evolved over the last few decades. These include methods to detect a single shift in the series' mean (Chernoff and Zacks 1964), in its variance (Hsu 1977), or in a general linear regression model (Quandt 1958; Robbins et al. 2016). In the climate literature, changepoint detection has most often been used to detect mean shifts.

However, this may result in misinterpreting a long-term climate trend as a sequence of mean shifts that follows (approximates) the trend (Beaulieu and Killick 2018).

Much of the changepoint literature assumes independent and identically distributed model errors (termed white noise here). However, climate time series are often autocorrelated, inducing memory at time scales longer than the measurement frequency (Hasselmann 1976). This memory is often modeled as a first-order autoregressive (AR(1)) process in climate studies (Lund et al. 2007; Robbins et al. 2011; Hartmann et al. 2013). In an AR(1) model, autocorrelation geometrically decays to zero with increasing time, representing one type of short-term memory. In the climate setting, it is important to allow autocorrelation and mean shift model features in tandem as both can inject similar run patterns into a climate series. An alternative is to use pre-whitening techniques that mitigate the effects of autocorrelation (Robbins et al. 2011; Serinaldi and Kilsby 2016). Beaulieu and Killick (2018), Shi et al. (2022), and Gallagher et al. (2021) show that changepoint inferences can be drastically wrong if autocorrelation in a series is ignored. The memory in climate series has also been modeled as a long-memory process, where autocorrelation decays as a power law (Yuan et al. 2015). Long-memory processes and changepoint models can be confused as they both have similar spectrums. Unfortunately, this ambiguity may lead to mislead inferences. Beaulieu et al. (2020) discuss how to distinguish changepoints and long-memory in surface temperatures.

Multiple changepoints may be present in climate series. Methods designed to detect a single changepoint have been applied iteratively to estimate multiple changepoint configurations through a process known as binary segmentation (Scott and Knott 1974; Rodionov 2004). Binary segmentation is now known to perform poorly in multiple changepoint problems (Shi et al. 2022) (see Fryzlewicz (2014) for an interesting attempt to fix binary segmentation). Penalized likelihood methods, the approach taken here, were developed in Davis et al. (2006); Lu et al. (2010); Killick et al. (2012); Li and Lund (2012) and tend to perform better (Shi et al. 2022). Here, a likelihood, which measures the goodness of the statistical model fit, is balanced against a penalty that prevents fitting too many changepoints. Penalized likelihood methods can allow for autocorrelation. Bayesian approaches to the multiple changepoint problem also exist. Most of these place some sort of prior distribution on the changepoint times, for instance a spike and slab prior (see Barry and Hartigan (1993); Chib (1998); Fearnhead (2006); and Cappello et al. (2021) and the references within). Li et al. (2019) construct an informative prior on the changepoint times from the station's metadata record. The references above are by no means exhaustive; indeed, the changepoint literature is vastly expanding.

As most methodological statistics papers are not written with user comprehension in mind, the technical changepoint literature can seem impenetrable to non-statisticians, making it challenging to select an appropriate approach for the climate scientist. Compounding difficulties, Lund and Reeves (2002) and Beaulieu and Killick (2018) show that spurious changepoint inferences easily occur when prominent data features (e.g. autocorrelation, long-term trend) are ignored — the choice of model and method is critical in changepoint analyses. Indeed, changepoint techniques can produce different results when the models and assumptions are only slightly changed.

The aim of this paper is to present, through an example, a comprehensive changepoint analysis of a climate series. To this end, we analyze the Central England temperature (CET) series by fitting different changepoint models capable of detecting shifts in trends. We also compare our changepoint fits with long-memory models. Our focus is on penalized likelihood multiple changepoint techniques, enabling us to compare several models while preventing overestimation of the number of changepoints. We also discuss mean shift models and how they fit data containing a long-term trend such as the CET series. Emphasis is placed on implementation and interpretation over the theoretical foundations of penalized likelihoods. Nonetheless, references to the formal statistical literature are provided.

Accepted for publication in *Journal of Climate.* DOI 10.1175/JCLI-D-21-0489.1.

The rest of this paper proceeds as follows. The CET series used here is introduced in the next section. Section 3 then provides some rudimentary background on changepoint models, describing the penalized likelihood methods used here. The next three sections present fits of various multiple changepoint models. Results for each type of model motivate the subsequent fits. Remarks about the optimal model are made in the final section along with concluding comments.

2. The CET Series

The CET time-series is perhaps the longest instrumental record of surface temperatures in the world, commencing in 1659 and spanning 362 years through 2020. The CET series is a benchmark for European climate studies, as it is sensitive to atmospheric variability in the North Atlantic (Parker et al. 1992). This record has been previously analyzed for long-term changes (Plaut et al. 1995; Harvey and Mills 2003; Hillebrand and Proietti 2017); however, to our knowledge, no detailed changepoint analysis of it has been previously conducted. Changepoints are plausible in the CET record for several reasons. First, artificial shifts near the record's onset may exist when data quality was lower (Parker et al. 1992). Furthermore, an increase in the pace of climate warming arising globally during the 1960s-1970s (Beaulieu and Killick 2018; Cahill et al. 2015) may be present. The length of the CET record affords us the opportunity to explore a variety of temperature features.

The CET series, available at `https://www.metoffice.gov.uk/hadobs/hadcet/`, was provided by the UK Met Office. Measurements commenced in 1659 and were mostly compiled by Manley (1953, 1974) until 1973, then continued and updated to 1991 in Parker et al. (1992). The series is now kept by the Hadley Centre, Met Office. The CET time series is an annual composite of 15 stations in the UK, located over a roughly triangular area bounded by Lancashire, London, and Bristol. The series is thus representative of the climate of the English Midlands.

The station locations used to form the composite series are depicted in the top graphic in Figure 1. The CET temperatures, presented in the bottom graphic of Figure 1, have been previously adjusted for inhomogeneities due to changes in measurement practices through time (Manley 1953, 1974; Parker et al. 1992), and for urban warming since 1960 (Parker and Horton 2005). However, until 1722, available instrumental records used in the CET time series did not overlap. As such, non-instrumental weather diaries and the Utrecht instrumental series were used to adjust the CET series and fill the gaps (Parker et al. 1992). Between 1722 and 1760, there are no gaps in the composite record of all stations, but observations were generally collected in unheated rooms as opposed to outdoors. A few outdoor temperature measurements were collected and used to establish relationships between temperatures in unheated rooms and outdoors. These relationships were then used to adjust the CET time series (Parker et al. 1992). The daily CET time series starts in 1772, and has been used to update the monthly series (Parker et al., 1992). As such, some authors use only the data post-1772 for their analyses (Hillebrand and Proietti 2017). In this paper, we conduct a changepoint analysis on both the full CET time series (1659-2020) and the truncated series (1772-2020) that excludes the poorer quality data at the beginning of the record.

3. Structural Change Models

To explore structural changes in the CET series, a hierarchical changepoint analysis, gradually building on past findings, will be conducted. Let X_t denote the annual temperature observed at time t and suppose that data from the years $1, \ldots, N$ are available. In general, a changepoint analysis partitions the series into $m+1$ distinct regimes, each regime having homogeneous characteristics. The number of changepoints m is unknown and needs to be estimated from the series. Let τ_i denote the ith changepoint time; boundary conditions take $\tau_0 = 0$ and $\tau_{m+1} = N$.

All regression models in this paper have the time series regression form

$$X_t = f(t) + \epsilon_t, \qquad t = 1, 2, \ldots, N, \tag{1}$$

where $f(t) = E[X_t]$ is the mean of the series at time t. The structural form of f will vary, generally containing location and/or trend parameters and their shifts; each model form will be discussed as we proceed. The model errors $\{\epsilon_t\}_{t=1}^{N}$ have zero mean and may be correlated in time. We work with AR(1) errors for simplicity, but more complex time series models are possible. While it is important to allow for autocorrelation in annual data, the form of the correlation structure is typically not as crucial as its presence.

The AR(1) difference equation governing the errors $\{\epsilon_t\}$ is

$$\epsilon_t = \phi\epsilon_{t-1} + Z_t,$$

where $\phi \in (-1, 1)$ and $\{Z_t\}$ is zero mean white noise (WN) with unknown variance σ^2. Solutions to the AR(1) equation have exponentially decaying correlations: $\mathrm{Corr}(\epsilon_t, \epsilon_{t+h}) = \phi^h$ for $h \geq 0$. Because the data are annually averaged, Gaussian distributed errors $\{\epsilon_t\}$ are statistically realistic. An implication of this is that future model likelihood functions will be Gaussian based.

Methods for handling multiple changepoint analyses without penalized likelihoods exist. One popular technique is termed binary segmentation (Scott and Knott 1974). Binary segmentation works with any single changepoint technique, termed an at most one change (AMOC) method. Many AMOC tests have been developed, including cumulative sums (CUSUM) (Page 1954), likelihood ratios (Jandhyala et al. 2013), Chow tests (Chow 1960), and sum of squared CUSUM tests (Shi et al. 2022). Binary segmentation first analyzes the entire series for a changepoint. If a changepoint is found, the series is split into subsegments about the identified changepoint time and the two subsegments are further scrutinized for additional changepoints. The procedure is repeated iteratively until no subsegments are deemed to have changepoints. While simple and computationally convenient, binary segmentation is one of the poorer performing multiple changepoint techniques (Shi et al. 2022), often being fooled by changepoints that occur close to one another or multiple shifts that move the series in opposite directions. There have been attempts to fix binary segmentation — see the wild binary segmentation and related methods in Fryzlewicz (2014) and Eichinger and Kirch (2018). Unfortunately, these techniques typically assume independent model errors or are restricted to single parameter changes per regime (for example, mean shifts only). Perhaps worse, wild binary segmentation tends to overestimate changepoint numbers when they are in truth infrequent (Lund and Shi 2020).

To estimate the changepoint structure and model parameters from the data, penalized likelihood methods will be used. Likelihood methods choose the model parameters that make seeing the observed data most likely; a penalty is imposed on the changepoint configuration to keep the fitted model parsimonious (from having too many changepoints). Our penalized likelihoods have the following form

$$-2\log(L^*(m;\tau_1,\ldots,\tau_m))+P(m;\tau_1,\ldots,\tau_m). \tag{2}$$

The notation here is as follows: $L^*(m;\tau_1,\ldots,\tau_m)$ is the optimal Gaussian likelihood that can be achieved from a model with m changepoints that occur at the times $\tau_1,\ldots,\tau_m$. Here, the data sample $X_1,X_2,\ldots,X_N$ is regarded as fixed. To determine $L^*(m;\tau_1,\ldots,\tau_m)$, one must estimate all parameters in the mean function f and the AR(1) model errors assuming that m changepoints occur at the times $\tau_1,\ldots,\tau_m$. This procedure will be discussed further below. The quantity $P(m;\tau_1,\ldots,\tau_m)$ is the penalty for having a model with m changepoints at the times $\tau_1,\ldots,\tau_m$. As more and more changepoints are added to the model, the overall fit gets better ($-2\log(L^*)$ gets smaller); the penalty, which is positive and increases with the number of changepoints, prevents an overfitted model (one with too many changepoints).

Many penalty structures have been proposed in the statistics and climate literature. These include the Akaike information criterion (AIC), the Bayesian information criterion (BIC), the modified Bayesian information criterion (mBIC), and Minimum description lengths (MDL). We will use BIC and MDL here. These two penalties were judged as "winners" in a recent changepoint detection comparison in Shi et al. (2022). AIC penalties are not considered here because they often erroneously estimate an excessive number of changepoints (Shi et al. 2022). The BIC penalty for having m changepoints at the times $\tau_1, \ldots, \tau_m$ is $m \log(N)$ and is proportional to the number of changepoints; additional parameters are penalized at the rate of $\log(N)$ per model parameter. Our penalized likelihood objective functions for structural changes are summarized in Table 1. The individual models will be explained in subsequent sections. The boxed quantities are the model penalties. When $m = 0$, penalties for any changepoint quantities are taken as zero since changepoint features are absent from the model.

When comparing models via BIC (or any other model selection criterion), one computes the BIC statistic for all fitted models and chooses the one with the smallest BIC score. Differences between BIC values can give a sense of uncertainty between different model fits. The “posterior model probabilities" of Burnham and Anderson (2004) can further highlight differences. Elaborating, we label the compared models as $g_i, (i = 1, \ldots, R)$ and let ΔBIC_i denote the difference between the BIC score of model g_i and the model having the smallest BIC score. The posterior model probabilities of Burnham and Anderson (2004) are

$$p_i = \frac{exp(-\Delta BIC_i/2)}{\sum_{r=1}^{R} exp(-\Delta BIC_r/2)}. \tag{3}$$

Then p_i is the inferred probability that model g_i is the quasi-true model in the model set under a prior where all R models are equally likely (prior probabilities are $1/R$ for each model). These BIC posterior model probabilities highlight uncertainties in our model comparisons.

In contrast to the BIC penalty, the MDL penalty is more complex in form, also accounting for the changepoint location times $\tau_1, \ldots, \tau_m$. The MDL penalty depends on the form of f and is rooted in information theory, quantifying the computer memory needed to store the model (good fitting models use minimal space). MDL penalties have previously proven useful in changepoint detection (Davis et al. 2006; Li and Lund 2012)). Posterior model probabilities are not available for the MDL information criterion. Other penalties used in the climate literature for changepoint problems include those in Caussinus and Mestre (2004).

A drawback of penalized likelihood methods involves computation time. There are $\binom{N-1}{m}$ distinct changepoint configurations having m changepoints. Summing this over all m shows that there are 2^{N-1} distinct changepoint configurations that need to be searched in an exhaustive optimization of a penalized likelihood, a daunting task for long time series. As a solution, genetic algorithms (GA) will be used to optimize our penalized likelihoods. GAs are randomized search algorithms that mimic natural selection processes. In a genetic algorithm, an initial collection (generation) of changepoint configurations is randomly evolved towards ones with improved penalized likelihoods. Better fitting models are allowed priority in passing on their changepoints (genes) to children models of the next generation. Occasionally, mutations (very different changepoint configurations) occur; this keeps the GA from converging to local minimums of the penalized likelihood. Ultimately, the GA converges to a model with a very good penalized likelihood. The natural selection mechanism in GAs make it unlikely to visit suboptimal changepoint configurations. While Li and Lund (2012) illustrate how to devise a GA in climate changepoint applications, generally available GAs have now become savvy enough to capably handle our needs. The GA optimizations performed here use the `R` package `GA` (Scrucca 2013).

In contrast to GAs, binary segmentation is a greedy algorithm that often becomes trapped at a local penalized likelihood minimum. Killick et al. (2012) and Maidstone et al. (2017b), two rapid dynamic programming based multiple changepoint configuration optimizers, currently cannot handle our needs: Maidstone et al. (2017b) assumes independent model errors and Killick et al. (2012) assumes all parameters change at each changepoint time (including the AR(1) correlation parameter ϕ and error variance σ^2). GAs are the only optimization method that reasonably handle all models considered in this paper.

4. Models fitted

a. Trend shift models

We start our analysis with models having trends, as a long-term trend in the CET time series has been documented in previous studies (Kendon et al. 2021; Franzke 2012; Karoly and Stott 2006). This model posits $f(\cdot)$ to have the piece-wise linear form

$$f(t) = \begin{cases} \mu_1 + \beta_1 t, & 1 \le t \le \tau_1, \\ \mu_2 + \beta_2 t, & \tau_1 + 1 \le t \le \tau_2, \\ \vdots & \\ \mu_{m+1} + \beta_{m+1} t, & \tau_m + 1 \le t \le N, \end{cases} . \tag{4}$$

More compactly, one can write $E[X_t] = f(t) = \mu_{r(t)} + \beta_{r(t)} t$, where $r(t) \in \{1, 2, \ldots, m+1\}$ denotes the regime being used at time t; for example, $r(t) = 1$ for $1 \le t \le \tau_1$.

The changepoint literature has focused primarily on detecting mean shifts; fewer studies have been dedicated to detecting trend shifts. However, Maidstone et al. (2017a) present a dynamic programming approach that estimates trend shift configurations using a penalty based on absolute distances that is neither the MDL nor BIC. Their $\{\epsilon_t\}$ must be white noise (uncorrelated) with a zero mean and constant variance. See Bai and Perron (1998), Bai and Perron (2003), and their related R package `strucchange` by Zeileis et al. (2015) for more details.

The least squares estimators for the ith regime's parameters are computed from data in this regime only:

$$\hat{\beta}_i = \frac{\sum_{t=\tau_{i-1}+1}^{\tau_i}(X_t-\bar{X}_i)(t-\bar{t}_i)}{\sum_{t=\tau_{i-1}+1}^{\tau_i}(t-\bar{t}_i)^2}, \quad \hat{\mu}_i = \bar{X}_i - \hat{\beta}_i\bar{t}_i, \quad i=1,2,\ldots,m+1, \tag{5}$$

where $\bar{X}_i = (\sum_{t=\tau_{i-1}+1}^{\tau_i} X_t)/(\tau_i - \tau_{i-1})$ and $\bar{t}_i = (\tau_i + \tau_{i-1} + 1)/2$. While these are not the exact maximum likelihood estimators in correlated settings, they are typically very close to them (Lee and Lund 2012). A detailed discussion of least squares versus maximum likelihood estimator differences for time series is contained in Lee and Lund (2012).

One next computes the detrended series via

$$D_t = X_t - \hat{f}(t) = X_t - (\hat{\mu}_{r(t)} + \hat{\beta}_{r(t)}t). \tag{6}$$

The AR(1) parameter is then estimated via

$$\hat{\phi} = \frac{\sum_{t=1}^{N-1} D_t D_{t+1}}{\sum_{t=1}^{N} D_t^2}. \tag{7}$$

One-step-ahead predictions of the time series are now computed by

$$\hat{D}_t = \hat{\phi}\hat{D}_{t-1}, \quad t \geq 2, \tag{8}$$

with the start-up condition $\hat{D}_1 = 0$. The white noise variance in the AR(1) model is estimated as

$$\hat{\sigma}^2 = \frac{1}{N}\sum_{t=1}^{N} \hat{D}_t^2. \tag{9}$$

Plugging $\hat{\mu}_k$, $\hat{\phi}$, and $\hat{\sigma}^2$ into the Gaussian likelihood (see Li and Lund (2012) for details) gives a negative Gaussian log-likelihood of

$$-2\log(L^*(m;\tau_1,\ldots,\tau_m)) = N\log(\hat{\sigma}^2) + \underbrace{N + N\log(2\pi).}_{\text{Constant}} \tag{10}$$

The underbraced constant term above does not change over distinct changepoint configurations and can be neglected in the changepoint configuration comparisons. The above equations show how to estimate model parameters and evaluate model likelihoods given the changepoint configuration; the optimal changepoint configuration is found by a GA search. The penalized likelihoods obtained with two different penalties, MDL and BIC, are presented in Table 1 for the various models used here. Since regression lines are described by two parameters, all regimes are required to be at least three years long (so that fits in any single regime are not perfect).

On the full CET series, GA optimizations of the BIC and MDL penalized likelihoods estimate identical trend shift configurations, both flagging three breaks at the times $1700, 1739$, and 1988 (Table 2). This methodological agreement is convenient, but is not typical in changepoint analyses. Figure 2 graphically depicts our model fit. Cooling occurs during the first 39 years, followed by an increasing-trend second regime, with subsequent shifts to two warming trend regimes. The last regime, which starts in 1989, is warming with a trend of 1.1°C per century. When fitting trend shift models to CET series on post 1772 data only, we find a single changepoint in 1987 (Table 3), which is consistent with our analysis on the full series.

In both cases, the AR(1) correlation estimate is very small ($\hat{\phi} = 0.058$ for the full CET and $\hat{\phi} = 0.073$ for the truncated), and is not significantly different from zero with standard time series tests (Brockwell and Davis 1991). When $\phi = 0$, an AR(1) model reduces to white noise. This point is worth emphasizing: our model fits prefer the trend shift structure over structures involving autocorrelated errors. This is an important point since positive autocorrelation and shifts can induce similar run patterns in series — likelihood methods can decide which feature (or both) is statistically preferable. Should autocorrelation be neglected, one risks flagging spurious changepoints. And while independent model errors is reasonable here, it may not hold in other applications, especially if monthly or daily data are used.

Other assumptions made on the model errors include normality and a constant variance in X_t . To assess normality, we apply a Shapiro-Wilk test to the model residuals. This test does not reject normality (Tables 2- 3) at any common levels of statistical significance. To investigate the constant variance assumption, we apply Leneve's test to the residuals. This test does not find evidence of a changing variance in the residuals of the trend shifts models fitted to the CET series at any appreciable levels of statistical significance. Normality and constant variance assumptions in all future fitted models (Tables 2 and 3 list these) is investigated — these features are not rejected in any of the models compared here.

b. A fixed slope mean shift model

In some cases, it may be appropriate to constrain trends to be identical over all regimes (Wang 2003). This could be the case if artificial changes are expected. For example, a change of instrument may introduce an artificial shift in a time series, but will not necessarily alter the long-term trend in different regimes. A model with a common trend slope in all regimes (Lu and Lund 2007) is

$$f(t) = \begin{cases} \mu_1 + \beta t, & 1 \le t \le \tau_1, \\ \mu_2 + \beta t, & \tau_1 + 1 \le t \le \tau_2, \\ \vdots & \\ \mu_{m+1} + \beta t, & \tau_m + 1 \le t \le N, \end{cases} . \tag{11}$$

where β is the trend slope, which is the same in all regimes.

In compact form, the model can be expressed as

$$X_t = \mu_{r(t)} + \beta t + \epsilon_t, \tag{12}$$

where $\mu_{r(t)}$ is as in (5), and $\{\epsilon_t\}$ is an AR(1) process.

The ordinary least square estimators of β and $\mu_1, \ldots, \mu_{m+1}$ have the explicit form

$$\hat{\beta} = \frac{\sum_{i=1}^{m+1} \sum_{\tau_{i-1}+1}^{\tau_i} (X_t - \bar{X}_i)(t - \bar{t}_i)}{\sum_{i=1}^{m+1} \sum_{t=\tau_{i-1}+1}^{\tau_i} (t - \bar{t}_i)^2}, \qquad \hat{\mu}_i = \bar{X}_i - \hat{\beta}\bar{t}_i, \qquad i = 1, 2, \ldots, m+1, \tag{13}$$

where $\bar{X}_i$ and $\bar{t}_i$ are as before. These are again very close to the maximum likelihood estimators (Lee and Lund 2012). The BIC and MDL penalties are listed in Table 1.

A GA was used to estimate this configuration, which is plotted against the data in Figure 3. For the full CET series, both BIC and MDL flag a single mean shift in 1988, while the single detected shift moves to 1990 in the truncated series (post 1772). Fewer changepoints are detected in this model than with the trend shift models of the previous section, but the time of the single change detected here is consistent with the last changepoint found in the trend shifts models. Since the BIC and MDL penalized likelihoods in Tables 2 and 3 are larger for the constant slope model than for the regime-varying trend slope model, the inference is that regime-varying slopes are preferable.

c. Joinpin models

There is debate over whether trend models should impose continuity in $E[X_t]$ at the changepoint times in temperature series (Rahmstorf et al. 2017). These so-called joinpin models require $E[X_t] = f(t)$ to be continuous in time t. Here, we compare a joinpin model to the trend shifts and fixed slope mean shift models fitted in the previous sections. Unfortunately, it is not clear what an appropriate MDL penalty is for this case, nor does this seem to be an easy matter to rectify; hence, we proceed with BIC penalties only.

To fit a joinpin model, the package in Maidstone et al. (2017a) was used. We fit the same model as (4), but with additional constraints to force continuity at the changepoint time(s). A simple way to enforce this continuity is to view the slopes as determined from $E[X_t]$ at the start and end of each regime. This enforces continuity within a simple form foregoing additional constraints. This formulation fits the model

$$X_t = \gamma_{\tau_i} + \frac{\gamma_{\tau_{i+1}} - \gamma_{\tau_i}}{\tau_{i+1} - \tau_i}(t - \tau_i) + \epsilon_t, \tag{14}$$

where γ_i is the value of the mean at time i. This formulation is equivalent to (4) with an additional continuity constraint at the changepoint locations. Based on Maidstone et al. (2017a), the BIC for the joinpin model is

$$\mathrm{BIC} = N\log(\hat{\sigma}^2) + N + N\log(2\pi) + (2m+1)\log(N), \tag{15}$$

where

$$\hat{\sigma}^2 = \frac{1}{N}\sum_{i=1}^{m+1}\sum_{t=\tau_i}^{\tau_{i+1}}\left[X_t - \frac{\gamma_{\tau_{i+1}} - \gamma_{\tau_i}}{\tau_{i+1} - \tau_i}(t-\tau_i)\right]^2.$$

In the formulation of Maidstone et al. (2017a), the white noise variance is fixed and needs to be estimated. While median absolute deviations could be used for this purpose, we instead use the estimated error variance of 0.29 (Table 2), taken from the discontinuous model fits and BIC penalties of the last section, This fit assumes IID errors, which seems plausible given the results of the previous sections. The fitted model flags a single changepoint in 1973 in the full CET series and none in the truncated series; see Tables 2-3 and Figure 4. These fits are stable against changes from 0.29 in the white noise variance. Compared to our previous model fits, the joinpin model has a much higher BIC than the trend shift and fixed slope mean shifts models (Tables 2-3). As such, joinpin models do not appear to be competitive.

While a changepoint seems plausible towards the end of the record due to an increased warming rate, the joinpin fit to the earliest data is poor, similar to the fixed slope mean shifts model. This is graphically evident in the Figure 4 fits, but is also reflected by the higher BIC scores in Tables 2-3. A joinpin model should be used when a discontinuous mean function is unlikely or physically implausible. With the CET series, it is not evident whether the estimated mean function should be continuous or discontinuous. Elaborating, for series containing "only a single station", mean discontinuities are physically expected. However, when more and more station records are averaged into a composite record, mean function discontinuities are reduced, becoming less pronounced with an increasing number of stations. Should a discontinuous mean function be deemed possible, a trend shift model provides greater flexibility since it can simultaneously approximate a joinpin continuous structure as well as discontinuous shifts (Beaulieu and Killick 2018).

d. Long-memory models

A body of climate literature argues that climate time series exhibit long-memory, where the series' autocorrelation decays slowly in lag, often via a power law (Yuan et al. 2015; Blender and Fraedrich 2003; Franzke 2012). Long-memory correlation and changepoint features can inject similar run properties into a climate series, which is appreciated in the statistical and econometric literatures (Diebold and Inoue 2001; Granger and Hyung 2004; Mills 2007; Yau and Davis 2012). The daily CET series may exhibit long-memory (Syroka and Toumi 2001; Franzke 2012).

To compare our changepoint models to a long-memory model, we fit an autoregressive fractionally integrated moving-average (ARFIMA) model to the CET series. In particular, ARFIMA models with no moving-average component, an integration parameter d with $0 < d < 0.5$, and an autoregressive component of orders zero and one, are considered. The AR(1) long-memory model is characterized as

$$X_t = (1-B)^d(1-\phi B)^{-1}\epsilon_t, \tag{16}$$

where B is the backshift operator applied to X_t.

To fit ARFIMA models, the `R` package `fracdiff` (Maechler 2020) was used. A BIC penalty was calculated and is listed in Table 1. An MDL penalty is not informative since this model does not have any changepoints. Long-memory model fits to the full and truncated CET series are described in Tables 2-3). The long-memory models have the largest BIC score among all models compared on the full CET time series. On the truncated series, they are also amongst the least plausible, although joinpin models have higher BIC scores. These results suggest that changepoints, rather than long-memory, are more plausible in the CET series. For additional evidence that changepoints are preferred over long-memory features, we applied the time varying wavelet spectrum methods in Norwood and Killick (2018) to the CET series. These methods were used on surface temperatures in Beaulieu et al. (2020) and shown to discriminate changepoint and long-memory models well in long series. The results confirm that a changepoint model is more appropriate than a long-memory model. The fitted model of autoregressive order zero was also preferred to the fitted model of order one, reinforcing that correlation aspects in the CET series are minimal.

e. Model selection uncertainty

Among the six models compared, the trend shift model with white noise is judged the most plausible, as suggested by both BIC and MDL scores. The BIC posterior probabilities for all models fitted above are presented in Table 4. For the full series, the model probability for the trend shift model with white noise is 0.64, followed by the joinpin model with probability 0.12 and the trend shift model with AR(1) errors with probability 0.11. The three other models all have a posterior probability of 0.05 or less. This highlights the uncertainty in the model selected, although the trend shifts models with AR(1) and white noise errors are very similar (the autocorrelation estimated in the AR(1) model is small and both configurations identify the same shifts). As for the joinpin model, the fit at the start of the record seems poor.

Moving to the truncated series, the trend shift model with white noise has a posterior probability of 0.68. The next most plausible models are the fixed slope mean shift model with AR(1) errors and the trend shift model with AR(1) errors, having posterior probabilities of 0.1 and 0.09, respectively (Table 4). These models are similar in that estimated changepoints times are very close, giving further evidence for a shift in the late 1980s. However, this suggests that a fixed slope model should not be entirely discarded. Unlike results for the full CET series, the joinpin model ranks very low (0.02) on the truncated CET series. This is not surprising given that no changepoint is detected under the joinpin model in the truncated series (Figure 4).

5. Trends vs Mean Shifts

The simplest changepoint analysis is arguably that of mean shifts. This is the most common model in the changepoint literature and has been widely used to analyze climate series. While this structure is inappropriate for series having trends (such as the CET analyzed here), we include this model here for comparative purposes. The mean shifts model posits $f(\cdot)$ to have form

$$f(t) = \begin{cases} \mu_1, & 1 \le t \le \tau_1, \\ \mu_2, & \tau_1 + 1 \le t \le \tau_2, \\ \vdots & \\ \mu_{m+1}, & \tau_m + 1 \le t \le N. \end{cases} \tag{17}$$

The model's mean structure is compactly written as $f(t) = E[X_t] = \mu_{r(t)}$, where $r(t) \in \{1, 2, \ldots, m+1\}$ denotes the regime being used at time t; for example, $r(t) = 1$ for $1 \le t \le \tau_1$.

Given m and the changepoint times $\tau_1, \ldots, \tau_m$, mean parameters are first estimated via segment averages:

$$\hat{\mu}_i = \frac{1}{\tau_i - \tau_{i-1}} \sum_{t=\tau_{i-1}+1}^{\tau_i} X_t, \qquad i = 1, 2, \ldots, m+1. \tag{18}$$

While sample means are not the exact maximum likelihood estimators of the mean parameters for correlated series, they are typically very close and are easy to compute (unlike maximum likelihood estimators). Next, the regime-wise mean estimated in (18) is subtracted from the series by computing $D_t = X_t - \hat{f}(t) = X_t - \hat{\mu}_{r(t)}$. The variance $\hat{\sigma}^2$ is then estimated as in (9). We do not fit this model with AR(1) errors based on the results from the previous sections. The BIC and MDL penalized likelihoods for this model are

$$\mathrm{BIC} = N\log(\hat{\sigma}^2) + N + N\log(2\pi) + (3m+3)\log(N); \tag{19}$$

$$\mathrm{MDL} = N\log(\hat{\sigma}^2) + N + N\log(2\pi) + \log(N) + 2\log(m) + 2\sum_{i=1}^{m+1}\log(\tau_i - \tau_{i-1}) + 2\sum_{i=2}^{m+1}\log(\tau_i). \tag{20}$$

We discuss only results on the full series here, but conclusions are consistent (i.e., the same changepoints are detected post 1772) if we repeat the analysis on the truncated series only. Fitting this model, seven changepoints are flagged with both MDL and BIC (Figure 5).

Both penalties pinpoint 1989 as a changepoint time, which is consistent with results of the previous section. Here, MDL and BIC both deem the “cold year" in 1740 an outlier, bracketing this time by two changepoints. Because MDL methods are based on information theory (Rissanen 1978) and not large sample statistical asymptotics, they often flag outliers. Shifts are more frequent at the beginning of the record, perhaps suggesting that the data during these times is less reliable. Evident in the fits is that the last three regimes act to move the series higher in a "staircase", which is expected for a series experiencing a long-term warming trend (Figure 5).

The BIC and MDL scores obtained on the full CET series are 648.17 and 656.09, respectively. Should this model be included in our main comparison, one would still prefer the trend shift model should the MDL penalty be used to make conclusions. However, the BIC mean shift score is smaller than the BIC trend shift score in the previous section, indicating preference for the mean shift model. A model containing only mean shifts will flag a sequence of shifts in an attempt to follow a long-term trend should the data have a trend and it not be included in the model. If the trend is not steep, as is the case here, it is especially challenging to distinguish between trends and mean shifts. To illustrate this, we conducted a simulation study where 500 synthetic series with the same trend magnitude and variability (as estimated in the truncated CET time series over 1772-2020) were generated. The mean shifts plus white noise and trend shifts plus white noise models were fitted to each series. In only 18% of the synthetic series, the correct model with a long-term trend was selected by BIC. Figure 6 presents a histogram of the difference between the two fitted models' BIC scores, further demonstrating the bias BIC has for the erroneous mean shifts model. Should there be any suspicion about a trend or “staircase feature" in the record, we recommend using techniques that incorporate trends, as done here.

6. Comments, Conclusions, and Discussion

This study compared and contrasted several common changepoint model fits for data containing trends, as well as a long-memory autocovariance model, to the CET time series. To our knowledge, this is the first time a detailed changepoint analysis has been conducted on this long record. Starting with a trend shift model, several different changepoint structures were fitted, illustrating the techniques and salient points of changepoint analyses.

Tables 2-3 present the log-likelihood, BIC, and MDL scores of all model fits. Depending on the model configuration, we detect either three changepoints (trend shifts models) or one changepoint (fixed slope mean shifts and joinpin models) in the full series. This changepoint count discrepancy traces to the large variations in the series during roughly the first century of the record.

Most models agree on a change to a rapidly warming regime circa 1988, except for the joinpin model (this is also true for the truncated series). Among all fitted models, the optimal one has trend shifts in 1700, 1739, and 1988 (full series), and one in 1988 (truncated series). Table 5 provides estimates of the best fitting model's intercept and slope parameters by regime. While the best fitting model is the trend shifts model, other models are also plausible (Table 4). Models with higher posterior probabilities tend to be consistent in their flagged changepoint times, but highlight that a fixed slope model (as opposed to the varying slopes in the trend shifts models) may be plausible. Long-memory models yield the highest BIC scores, and are less plausible than all other models compared. The results of the full and truncated CET series are consistent, showing that our post 1772 changepoint inferences are not overly sensitive to inclusion of the first century of the series.

Having both BIC and MDL penalties agree on the model type and changepoint configuration adds robustness to our conclusions, suggesting that the fitted segmentations are stable. According to Lavielle (2005), changepoint segmentations that are stable over a range of penalty values should be preferred. Overall, models with shifts were deemed preferable to models having autocorrelated errors.

While our aim is not necessarily directed to the causes of the detected shifts, we provide some interpretations here. Shifts flagged during the first century of the record are likely due to inferior data quality over this early period (Hillebrand and Proietti 2017). Due to lack of overlapping instrumentation coverage before 1722, non-instrumental weather diaries were used to adjust the series (Parker et al. 1992). Observations were generally collected in unheated rooms until 1760, and adjusted by calibrating indoor and outdoor observations later (Parker et al. 1992). Even with the most careful adjustments, one cannot guarantee that all biases were removed from the data. Some authors omit the first century of data altogether due to this issue (Hillebrand and Proietti 2017).

The trend shifts model on the earlier part of the data detects two changepoints in 1700 and 1739, characterizing a steep cooling trend followed by a warming trend. The mean shifts model fitted on the earlier part of the data flags multiple changes (1691, 1699, 1727, 1740, 1741), calling for a closer examination of the earlier part of the record. In data with inhomogeneities, BIC penalties favor mean shift models over trend shift models, even if the trend shifts model is truth. A mean shift model characterizes a warming trend as a staircase of increasing steps. This issue can be troublesome if the trend in the data is weak, as demonstrated in our simulation study (see Figure 6).

The changepoint flagged in 1988 (from multiple models and in both the full and truncated CET series) is not surprising given the warming seen on the global level in the 1960/70s in a range of surface temperature records, as discussed in studies using both trend shift and joinpin models (Cahill et al. 2015; Beaulieu and Killick 2018; Rahmstorf et al. 2017; Ruggieri 2013). While the more recent part of the CET series is considered more reliable and has been adjusted for inhomogeneities, we cannot entirely discard issues in this era either. Overall, it is possible that a combination of natural and artificial causes contribute to shifts in the CET series.

To further rule out artificial changes, one could subtract all $\binom{15}{2} = 105$ pairs of series from one another and examine these differences for changepoints. Then, one can distinguish artificially caused changepoints from those due to natural climate change and variability. See Menne and Williams Jr. (2009) for more details on this procedure. Artificial changes can then be corrected before long-term trends are analyzed. Changes that are not considered artificial can be further investigated through an attribution study (Hartmann et al. 2013).

Residual analyses were conducted to ensure that the underlying assumptions of the model were met. With the CET series, residuals of the trend shift model fit were judged to be uncorrelated (white noise). However, climate time series often exhibit autocorrelation that should be taken into account. We stress the importance of verifying the underlying assumptions in any changepoint model. Indeed, neglecting positive autocorrelation raises the risk of detecting spurious shifts. Also, the series' autocorrelation may be more complex than an AR(1) process and may itself contain shifts (Beaulieu et al. 2012; Beaulieu and Killick 2018). Some climate series may also contain long-memory autocorrelations (Vyushin et al. 2012). An additional challenge lies with the ambiguity between long-memory and changepoint models: both features can produce series with similar run structures. Because of this, a long-memory model was included as part of our comparison. We found that the CET time series is best represented by a multiple trend shift

Accepted for publication in *Journal of Climate*. DOI 10.1175/JCLI-D-21-0489.1.

changepoint structure and not a long-memory model. Such a comparison is not possible for all climate series since lengthy records are required to analyze long-memory series (Beaulieu et al. 2020). The CET time series, which is the longest publicly available surface temperature series, enables this comparison. Other assumptions that were made include constant variance temperatures and normally distributed observations. Both assumptions cannot be rejected in any models fitted (Tables 2-3).

Model selection based on a criteria does not guarantee that the selected model is "truth". All models are an approximation of reality and multiple models can plausibly represent the data. To quantify this, one can calculate posterior model probabilities with BIC that each fitted model is the "quasi-truth". This assumes that all models included in the comparison have the same prior weight, which may not be reasonable. One must also note that this measure is relative to the models included in the comparison, and does not reflect the uncertainty that the "true" model may not be part of the model set. Similarly, uncertainty in the total number of changepoints and their individual occurrence times is a difficult statistics problem. Bayesian methods, which were not considered here, can in principle place uncertainty margins on the number of changepoints and their locations. When several distinct models have similar penalized likelihood scores, inferences about the number of changepoints are likely to be less reliable. Recent statistics work is now studying this issue (Li et al. 2019; Cappello et al. 2021).

Ultimately the choice of "best model" should be arrived at from a judgment made by the researcher(s) based on objective statistical metrics, such as presented in this work, combined with understanding of the data recording practices and physics of the natural system.

Acknowledgments. Rebecca Killick gratefully acknowledges funding from EP/R01860X/1 and NE/T006102/1. Robert Lund and Xueheng Shi thank funding from NSF DMS-2113592. The comments of three referees and the editor substantially improved this manuscript.

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

References

Bai, J., and P. Perron, 1998: Estimating and testing linear models with multiple structural changes. *Econometrica*, **66**, 47–78.

Bai, J., and P. Perron, 2003: Computation and analysis of multiple structural change models. *Journal of Applied Econometrics*, **18 (1)**, 1–22.

Barry, D., and J. A. Hartigan, 1993: A Bayesian analysis for change point problems. *Journal of the American Statistical Association*, **88 (421)**, 309–319.

Beaulieu, C., J. Chen, and J. L. Sarmiento, 2012: Change-point analysis as a tool to detect abrupt climate variations. *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences*, **370 (1962)**, 1228–1249, doi:10.1098/rsta.2011.0383.

Beaulieu, C., and R. Killick, 2018: Distinguishing trends and shifts from memory in climate data. *Journal of Climate*, **31 (23)**, 9519 – 9543.

Beaulieu, C., R. Killick, D. Ireland, and B. Norwood, 2020: Considering long-memory when testing for changepoints in surface temperature: A classification approach based on the time-varying spectrum. *Environmetrics*, **31 (1)**, e2568.

Blender, R., and K. Fraedrich, 2003: Long time memory in global warming simulations. *Geophysical Research Letters*, **30 (14)**.

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

Burnham, K. P., and D. R. Anderson, 2004: Multimodel inference: Understanding AIC and BIC in model selection. *Sociological Methods & Research*, **33 (2)**, 261–304, doi:10.1177/0049124104268644.

Cahill, N., S. Rahmstorf, and A. C. Parnell, 2015: Change points of global temperature. *Environmental Research Letters*, **10 (8)**, 084 002.

Cappello, L., O. H. M. Padilla, and J. A. Palacios, 2021: Scalable Bayesian change point detection with spike and slab priors. *arXiv preprint arXiv:2106.10383*.

Caussinus, H., and O. Mestre, 2004: Detection and correction of artificial shifts in climate series. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, **53 (3)**, 405–425.

Chernoff, H., and S. Zacks, 1964: Estimating the current mean of a normal distribution which is subjected to changes in time. *The Annals of Mathematical Statistics*, **35 (3)**, 999–1018.

Chib, S., 1998: Estimation and comparison of multiple change-point models. *Journal of Econometrics*, **86 (2)**, 221–241.

Chow, G. C., 1960: Tests of equality between sets of coefficients in two linear regressions. *Econometrica: Journal of the Econometric Society*, **28**, 591–605.

Davis, R. A., T. C. M. Lee, and G. A. Rodriguez-Yam, 2006: Structural break estimation for nonstationary time series models. *Journal of the American Statistical Association*, **101 (473)**, 223–239.

Diebold, F. X., and A. Inoue, 2001: Long memory and regime switching. *Journal of Econometrics*, **105 (1)**, 131–159.

Eichinger, B., and C. Kirch, 2018: A MOSUM procedure for the estimation of multiple random change points. *Bernoulli*, **24 (1)**, 526–564.

Fearnhead, P., 2006: Exact and efficient Bayesian inference for multiple changepoint problems. *Statistics and Computing*, **16 (2)**, 203–213.

Franzke, C., 2012: Nonlinear trends, long-range dependence, and climate noise properties of surface temperature. *Journal of Climate*, **25 (12)**, 4172–4183, URL http://www.jstor.org/stable/26191996.

Fryzlewicz, P., 2014: Wild binary segmentation for multiple change-point detection. *Annals of Statistics*, **42 (6)**, 2243–2281.

Gallagher, C. M., R. Killick, R. Lund, and X. Shi, 2021: Autocovariance estimation in the presence of changepoints. *arXiv preprint arXiv:2102.10669*.

Granger, C. W. J., and N. Hyung, 2004: Occasional structural breaks and long memory with an application to the S&P 500 absolute stock returns. *Journal of Empirical Finance*, **11 (3)**, 399–421.

Hartmann, D., and Coauthors, 2013: *Observations: Atmosphere and Surface*, book section 2, 159–254. Cambridge University Press, Cambridge, United Kingdom.

Harvey, D. I., and T. C. Mills, 2003: Modelling trends in Central England temperatures. *Journal of Forecasting*, **22 (1)**, 35–47.

Hasselmann, K., 1976: Stochastic climate models part I. Theory. *Tellus*, **28 (6)**, 473–485.

Hewaarachchi, A. P., Y. Li, R. Lund, and J. Rennie, 2017: Homogenization of daily temperature data. *Journal of Climate*, **30 (3)**, 985–999.

Hillebrand, E., and T. Proietti, 2017: Phase changes and seasonal warming in early instrumental temperature records. *Journal of Climate*, **30 (17)**, 6795–6821.

Hsu, D.-A., 1977: Tests for variance shift at an unknown time point. *Journal of the Royal Statistical Society: Series C*, **26 (3)**, 279–284.

Jandhyala, V., S. Fotopoulos, I. MacNeill, and P. Liu, 2013: Inference for single and multiple change-points in time series. *Journal of Time Series Analysis*, **34 (4)**, 423–446.

Karoly, D. J., and P. A. Stott, 2006: Anthropogenic warming of Central England temperature. *Atmospheric Science Letters*, **7 (4)**, 81–85.

Kendon, M., M. McCarthy, S. Jevrejeva, A. Matthews, T. Sparks, and J. Garforth, 2021: State of the UK climate 2020. *International Journal of Climatology*, **41 (S2)**, 1–76, doi:https://doi.org/10.1002/joc.7285, URL https://rmets.onlinelibrary.wiley.com/doi/abs/10.1002/joc.7285, https://rmets.onlinelibrary.wiley.com/doi/pdf/10.1002/joc.7285.

Killick, R., P. Fearnhead, and I. A. Eckley, 2012: Optimal detection of changepoints with a linear computational cost. *Journal of the American Statistical Association*, **107 (500)**, 1590–1598.

Lavielle, M., 2005: Using penalized contrasts for the change-point problem. *Signal Processing*, **85 (8)**, 1501–1510.

Lee, J., and R. Lund, 2012: A refined efficiency rate for ordinary least squares and generalized least squares estimators for a linear trend with autoregressive errors. *Journal of Time Series Analysis*, **33 (2)**, 312–324.

Li, S., and R. Lund, 2012: Multiple changepoint detection via genetic algorithms. *Journal of Climate*, **25 (2)**, 674–686.

Li, Y., R. Lund, and A. Hewaarachchi, 2019: Multiple changepoint detection with partial information on changepoint times. *Electronic Journal of Statistics*, **13 (2)**, 2462–2520.

Lu, Q., and R. Lund, 2007: Simple linear regression with multiple level shifts. *Canadian Journal of Statistics*, **35 (3)**, 447–458.

Lu, Q., R. Lund, and T. C. M. Lee, 2010: An MDL approach to the climate segmentation problem. *Annals of Applied Statistics*, **4**, 299–319.

Lund, R., and J. Reeves, 2002: Detection of undocumented changepoints: A revision of the two-phase regression model. *Journal of Climate*, **15 (17)**, 2547–2554.

Lund, R., and X. Shi, 2020: Comments on "Detecting possibly frequent change-points: wild binary segmentation 2 and steepest-drop model selection". *Journal of the Korean Statistical Society*, **49 (4)**, 1090–1095.

Lund, R., X. L. Wang, Q. Q. Lu, J. Reeves, C. M. Gallagher, and Y. Feng, 2007: Changepoint detection in periodic and autocorrelated time series. *Journal of Climate*, **20 (20)**, 5178–5190.

Maechler, M., 2020: *fracdiff: Fractionally Differenced ARIMA aka ARFIMA(p,d,q) Models*. URL https://CRAN.R-project.org/package=fracdiff, R package version 1.5-1.

Maidstone, R., P. Fearnhead, and A. Letchford, 2017a: Detecting changes in slope with an L_0 penalty. *Journal of Computational and Graphical Statistics*, **28**, 265–275.

Maidstone, R., T. Hocking, G. Rigaill, and P. Fearnhead, 2017b: On optimal multiple changepoint algorithms for large data. *Statistics and Computing*, **27 (2)**, 519–533.

Manley, G., 1953: The mean temperature of Central England, 1698–1952. *Quarterly Journal of the Royal Meteorological Society*, **79 (340)**, 242–261.

Manley, G., 1974: Central England temperatures: monthly means 1659 to 1973. *Quarterly Journal of the Royal Meteorological Society*, **100 (425)**, 389–405.

Menne, M. J., and C. N. Williams Jr., 2009: Homogenization of temperature series via pairwise comparisons. *Journal of Climate*, **22 (7)**, 1700–1717.

Mills, T. C., 2007: Time series modelling of two millennia of Northern Hemisphere temperatures: long memory or shifting trends? *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, **170 (1)**, 83–94.

Mitchell Jr., J. M., 1953: On the causes of instrumentally observed secular temperature trends. *Journal of Atmospheric Sciences*, **10 (4)**, 244–261.

Norwood, B., and R. Killick, 2018: Long memory and changepoint models: a spectral classification procedure. *Statistics and Computing*, **28 (2)**, 291–302.

Page, E. S., 1954: Continuous inspection schemes. *Biometrika*, **41 (1)**, 100–115.

Parker, D., and B. Horton, 2005: Uncertainties in central england temperature 1878–2003 and some improvements to the maximum and minimum series. *International Journal of Climatology*, **25 (9)**, 1173–1188.

Parker, D. E., T. P. Legg, and C. K. Folland, 1992: A new daily central england temperature series, 1772–1991. *International Journal of Climatology*, **12 (4)**, 317–342.

Plaut, G., M. Ghil, and R. Vautard, 1995: Interannual and interdecadal variability in 335 years of Central England temperatures. *Science*, **268 (5211)**, 710–713.

Quandt, R. E., 1958: The estimation of the parameters of a linear regression system obeying two separate regimes. *Journal of the American Statistical Association*, **53 (284)**, 873–880.

Rahmstorf, S., G. Foster, and N. Cahill, 2017: Global temperature evolution: recent trends and some pitfalls. *Environmental Research Letters*, **12 (5)**, 054 001.

Rissanen, J., 1978: Modeling by shortest data description. *Automatica*, **14 (5)**, 465–471.

Robbins, M., C. Gallagher, R. Lund, and A. Aue, 2011: Mean shift testing in correlated data. *Journal of Time Series Analysis*, **32 (5)**, 498–511.

Robbins, M. W., C. M. Gallagher, and R. Lund, 2016: A general regression changepoint test for time series data. *Journal of the American Statistical Association*, **111 (514)**, 670–683.

Rodionov, S. N., 2004: A sequential algorithm for testing climate regime shifts. *Geophysical Research Letters*, **31 (9)**, doi:10.1029/2004GL019 448.

Ruggieri, E., 2013: A Bayesian approach to detecting change points in climatic records. *International Journal of Climatology*, **33 (2)**, 520–528.

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

Scrucca, L., 2013: GA: a package for genetic algorithms in R. *Journal of Statistical Software*, **53 (4)**, 1–37.

Serinaldi, F., and C. G. Kilsby, 2016: The importance of prewhitening in change point analysis under persistence. *Stochastic Environmental Research and Risk Assessment*, **30 (2)**, 763–777.

Shi, X., C. Gallagher, R. Lund, and R. Killick, 2022: A comparison of single and multiple changepoint techniques for time series data. *Computational Statistics & Data Analysis*, **170**, 107 433.

Syroka, J., and R. Toumi, 2001: Scaling of Central England temperature fluctuations? *Atmospheric Science Letters*, **2 (1)**, 143–154, doi:https://doi.org/10.1006/asle.2002.0047.

Trewin, B., and Coauthors, 2020: An updated long-term homogenized daily temperature data set for Australia. *Geoscience Data Journal*, **7 (2)**, 149–169.

Vincent, L. A., M. M. Hartwell, and X. L. Wang, 2020: A third generation of homogenized temperature for trend analysis and monitoring changes in Canada's climate. *Atmosphere-Ocean*, **58 (3)**, 173–191.

Vyushin, D. I., P. J. Kushner, and F. Zwiers, 2012: Modeling and understanding persistence of climate variability. *Journal of Geophysical Research: Atmospheres*, **117 (D21)**.

Wang, X. L., 2003: Comments on "Detection of undocumented changepoints: A revision of the two-phase regression model". *Journal of Climate*, **16 (20)**, 3383–3385.

Yau, C. Y., and R. A. Davis, 2012: Likelihood inference for discriminating between long-memory and change-point models. *Journal of Time Series Analysis*, **33 (4)**, 649–664.

Yuan, N., M. Ding, Y. Huang, Z. Fu, E. Xoplaki, and J. Luterbacher, 2015: On the long-term climate memory in the surface air temperature records over Antarctica: A nonnegligible factor for trend evaluation. *Journal of Climate*, **28 (15)**, 5922–5934.

Zeileis, A., F. Leisch, K. Hornik, C. Kleiber, B. Hansen, E. C. Merkle, and M. A. Zeileis, 2015: Package 'strucchange'. *R package version*, 1–5.

LIST OF TABLES

TABLE 1: Penalized likelihoods. The boxed terms are the penalties, with the unboxed terms constituting $-2\log(L^*)$. Here, N denotes the length of series, m the number of changepoints, τ_i is the time of the ith changepoint, and $\hat{\sigma}^2$ is the estimated white noise variance.

Criteria	Objective Function
BIC	$N\log(\hat{\sigma}^2)+N+N\log(2\pi)+\boxed{(3m+4)\log(N)}$
MDL	$N\log(\hat{\sigma}^2)+N+N\log(2\pi)+\boxed{2\log(N)+2\log(m)+2\sum_{i=1}^{m+1}\log(\tau_i-\tau_{i-1})+2\sum_{i=2}^{m+1}\log(\tau_i)}$

(a) Penalized likelihoods for the trend shift model with AR(1) errors

Criteria	Objective Function
BIC	$N\log(\hat{\sigma}^2)+N+N\log(2\pi)+\boxed{(3m+3)\log(N)}$
MDL	$N\log(\hat{\sigma}^2)+N+N\log(2\pi)+\boxed{\log(N)+2\log(m)+2\sum_{i=1}^{m+1}\log(\tau_i-\tau_{i-1})+2\sum_{i=2}^{m+1}\log(\tau_i)}$

(b) Penalized likelihoods for the trend shift model with white noise errors

Criteria	Objective function
BIC	$N\log(\hat{\sigma}^2)+N+N\log(2\pi)+\boxed{(2m+4)\log(N)}$
MDL	$N\log(\hat{\sigma}^2)+N+N\log(2\pi)+\boxed{3\log(N)+2\log(m)+\sum_{i=1}^{m+1}\log(\tau_i-\tau_{i-1})+2\sum_{i=2}^{m+1}\log(\tau_i)}$

(c) Penalized likelihoods for the fixed slope mean shift with AR(1) errors

Criteria	Objective function
BIC	$N\log(\hat{\sigma}^2)+N+N\log(2\pi)+\boxed{(2m+1)\log(N)}$

(d) Penalized likelihoods for the Joinpin model with white noise errors

Criteria	Objective function
BIC	$N\log(\hat{\sigma}^2)+N+N\log(2\pi)+\boxed{4\log(N)}$

(e) Penalized likelihoods for the long memory model with AR(1) errors. Minus $\log(N)$ for white noise errors.

TABLE 2: Model fitting results. Here, $\hat{\sigma}^2$ denotes the estimated variance of the white noise (* is assumed rather than estimated). Bolded values are the smallest penalized score. All model residuals have been checked for normality (Shapiro-Wilk's & Kolmogorov-Smirnov test) and constant variance (Levene's test).

Model	Penalty	Flagged Changepoints	$\hat{\sigma}^2$	Log-likelihood	Penalized Score
Trend shifts+AR(1)	BIC	1700,1739,1988	0.290	-288.80	654.19
	MDL	1700,1739,1988	0.290	-288.80	656.52
Trend shifts+WN	BIC	1700,1739, 1988	0.291	-290.02	**650.74**
	MDL	1700,1739, 1988	0.291	-290.02	**653.07**
Fixed slope mean shift+AR(1)	BIC	1988	0.325	-310.11	655.79
	MDL	1988	0.325	-310.11	658.93
Joinpin	BIC	1973	0.291*	-321.19	654.17
Long-memory+AR(1)	BIC	-	0.579	-316.59	656.75
Long-memory	BIC	-	0.584	-319.31	655.93

TABLE 3: Model fitting results based on truncated CET series. Here, $\hat{\sigma}^2$ denotes the estimated variance of the white noise (* is assumed rather than estimated). Bolded values are the smallest penalized score. All model residuals have been checked for normality (Shapiro-Wilk's & Kolmogorov-Smirnov test) and constant variance (Levene's test).

Model	Penalty	Flagged Changepoints	$\hat{\sigma}^2$	Log-likelihood	Penalized Score
Trend shifts+AR(1)	BIC	1987	0.305	-205.44	449.51
	MDL	1987	0.305	-205.44	450.70
Trend shifts+WN	BIC	1987	0.308	-206.13	**445.36**
	MDL	1987	0.308	-206.13	**446.55**
Fixed slope mean shift+AR(1)	BIC	1990	0.306	-208.06	449.23
	MDL	1990	0.306	-208.06	452.51
Joinpin	BIC	-	0.308*	-220.72	452.47
Long-memory+AR(1)	BIC	-	0.333	-217.01	450.57
Long-memory	BIC	-	0.340	-219.41	449.85

TABLE 4: BIC posterior probabilities for models fitted to the full and truncated CET series

Model	Full	Truncated
Trend shifts + AR(1)	0.11	0.08
Trend shifts + WN	0.64	0.68
Fixed slope +mean shifts+AR(1)	0.05	0.10
Joinpin	0.12	0.02
Long-memory+AR(1)	0.03	0.05
Long-memory	0.05	0.07

TABLE 5: Parameter estimates of the best fitting model: trend shifts with white noise errors

Segment	Slope (°C/yr)
1659-1699	-0.027
1700-1738	0.026
1739-1987	0.002
1988-2020	0.011

(a) Full CET

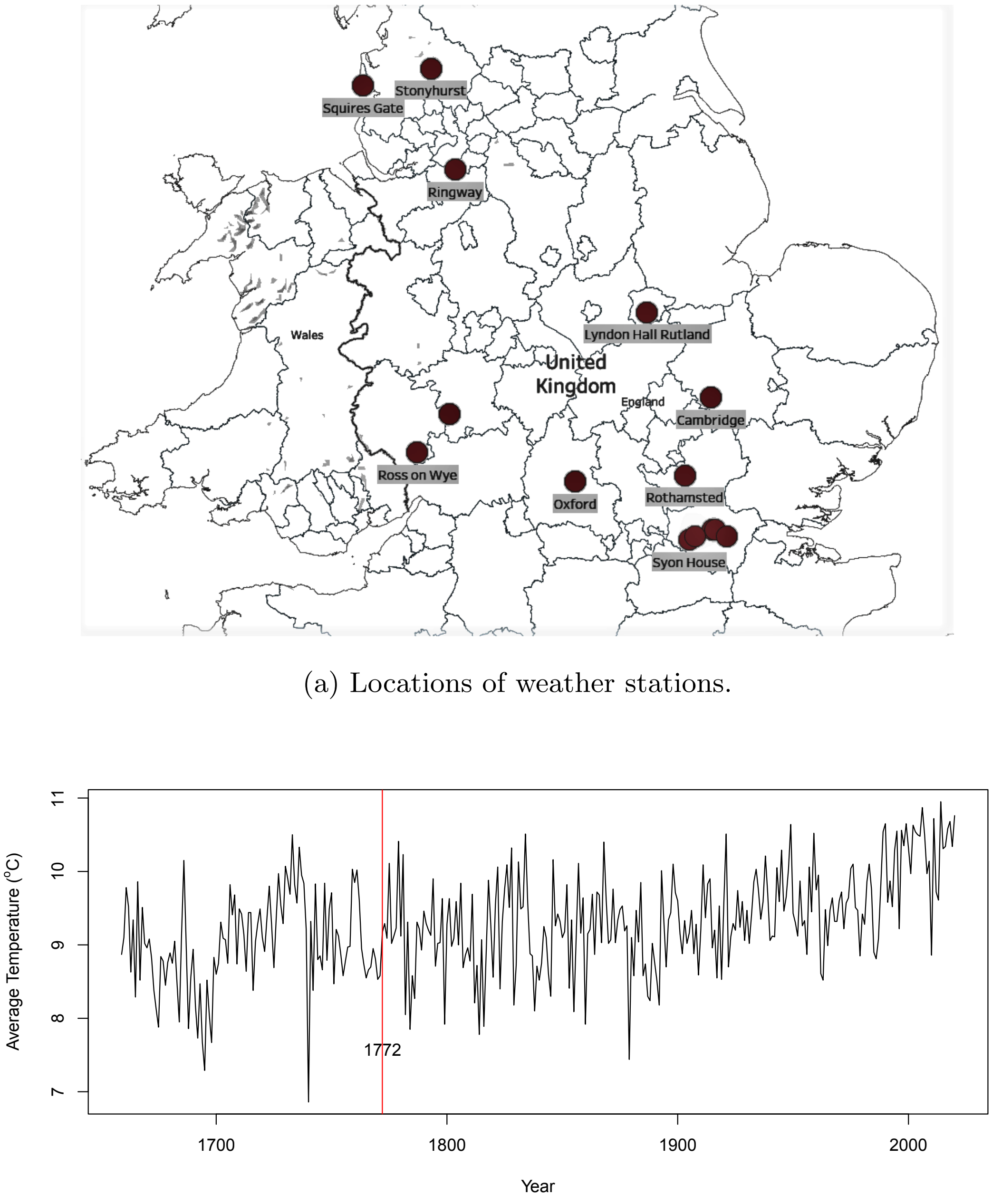
Segment	Slope (°C/yr)
1772-1986	0.002
1987-2020	0.016

(b) Truncated CET

LIST OF FIGURES

Accepted for publication in *Journal of Climate.* DOI 10.1175/JCLI-D-21-0489.1.

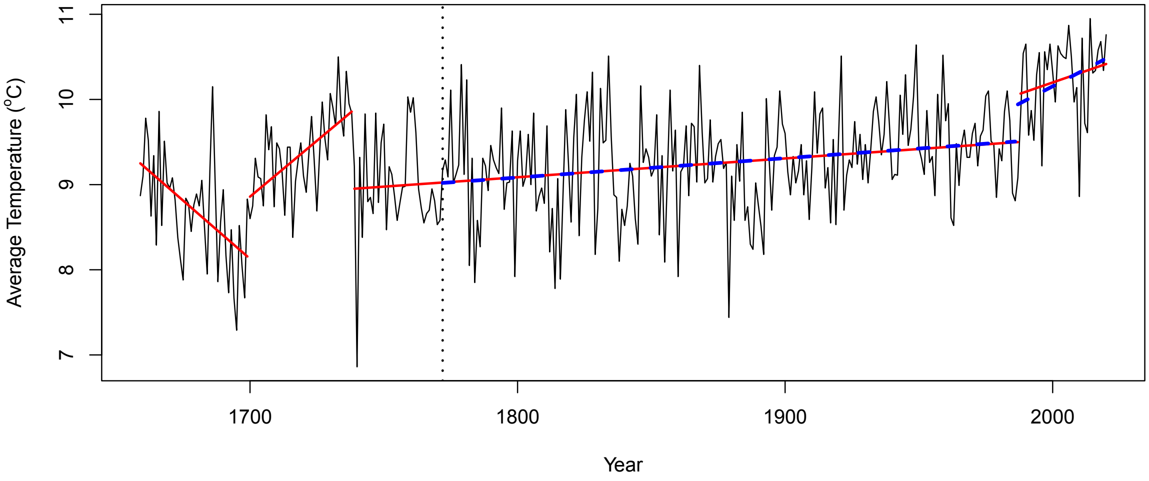
FIG. 1: Station locations and annual average temperatures of Central England.



(a) Locations of weather stations.

(b) Annual average temperature of Central England.

FIG. 2: Estimated CET trend shift structure. BIC and MDL flag the same changepoints in both the CET series (1700, 1739, 1988, red solid line) and truncated CET (1987, blue dashed line) series when assuming either AR(1) or white noise errors.



Accepted for publication in *Journal of Climate*. DOI 10.1175/JCLI-D-21-0489.1.

FIG. 3: The estimated CET trend shift structure for the full (red solid line) and truncated CET (blue dashed line) series when a constant regime trend slope is imposed. Both BIC and MDL flag a single changepoint in 1988 for the full series and 1990 for the truncated series.

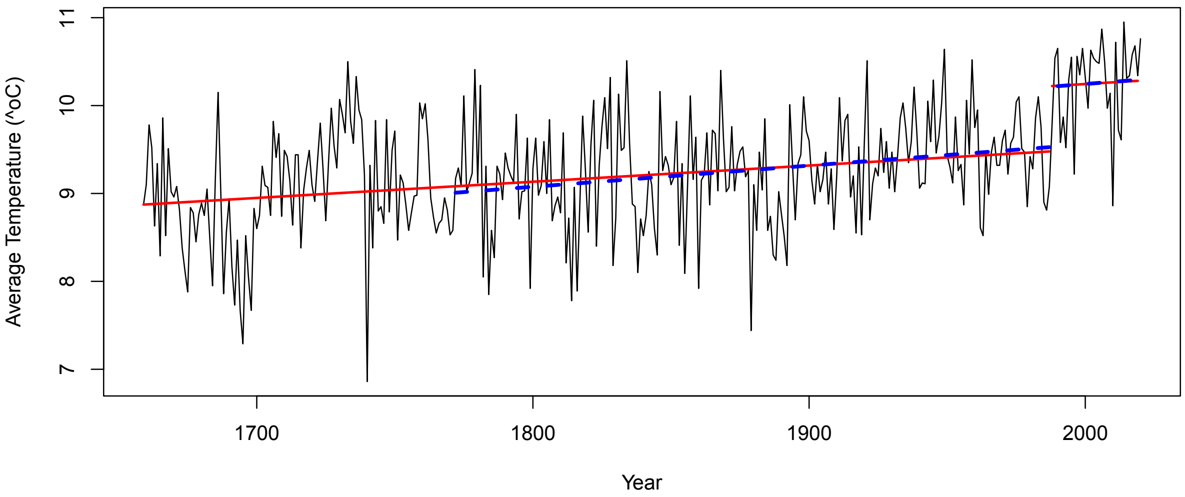
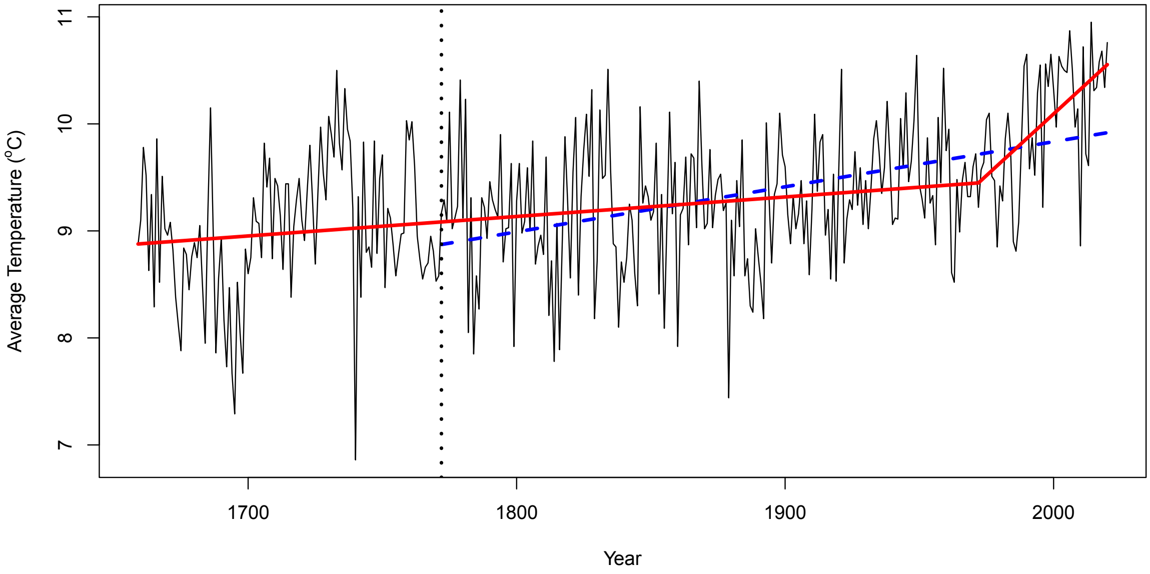
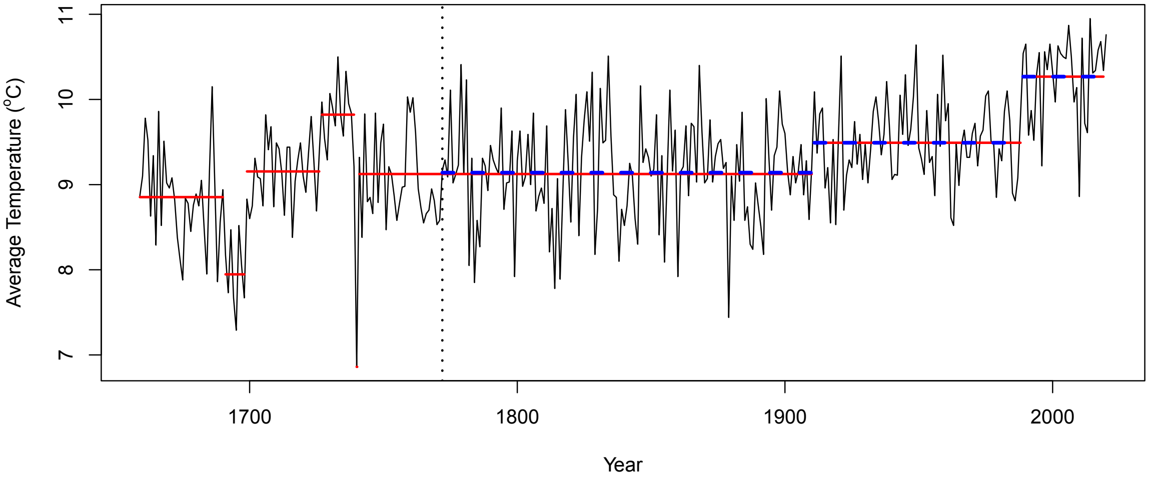


FIG. 4: Estimated CET joinpin shift structure for full (red solid line) and truncated (blue dashed line) series. BIC flags one shift in 1973 in the full series and and none for the truncated series.



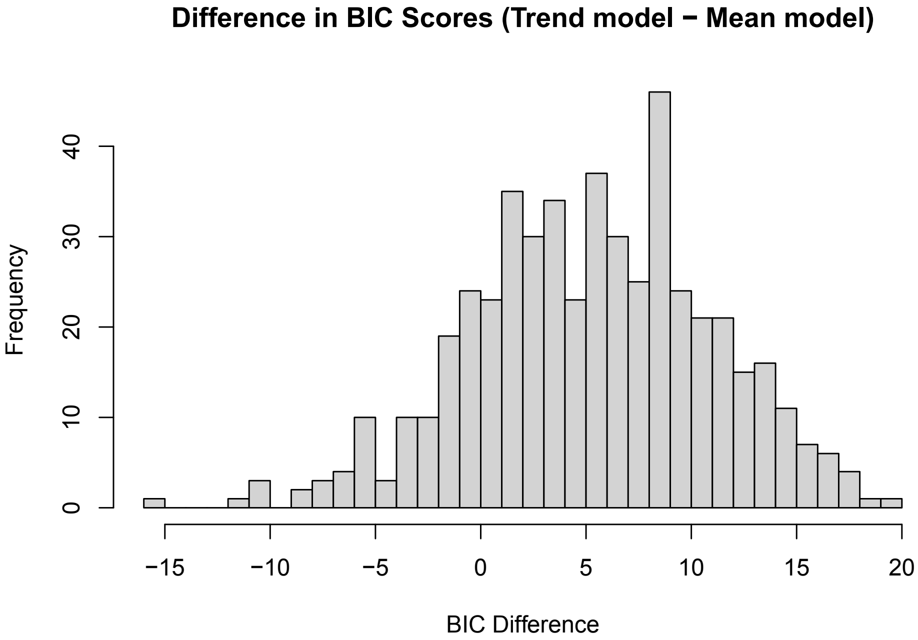
Accepted for publication in *Journal of Climate*. DOI 10.1175/JCLI-D-21-0489.1.

FIG. 5: The estimated CET mean shift structure for full (red solid line) and truncated (blue dashed line) series. BIC and MDL detect the same changepoints for both the CET and truncated CET series assuming white noise errors.



Accepted for publication in *Journal of Climate*. DOI 10.1175/JCLI-D-21-0489.1.

FIG. 6: Histogram of differences in BIC scores between the trend and mean-shift models. The correct model is the trend-shift model; however, BIC selects the mean-shift model the majority of the time.



Accepted for publication in *Journal of Climate.* DOI 10.1175/JCLI-D-21-0489.1.