Distinguishing trends and shifts from memory in climate data

Claudie Beaulieu[1, 2 *] and Rebecca Killick[3]

[1] Ocean Sciences Department, University of California, Santa Cruz, 95064, USA

[2] Ocean and Earth Science, University of Southampton, Waterfront Campus, European Way, Southampton, SO14 3ZH, UK

[3] Department of Mathematics and Statistics, Lancaster University, Lancaster, LA1 4YF, UK

* Corresponding author:
Claudie Beaulieu
Telephone: 831-459-5152
Email: beaulieu@ucsc.edu

Keywords: change-point detection, climate time-series, autocorrelation, memory, long-term trend

Abstract

The detection of climate change and its attribution to the corresponding underlying processes is challenging because signals such as trends and shifts are superposed on variability arising from the memory within the climate system. Statistical methods used to characterize change in time-series must be flexible enough to distinguish these components. Here we propose an approach tailored to distinguish these different modes of change by fitting a series of models and selecting the most suitable one according to an information criterion. The models involve combinations of a constant mean or a trend superposed to a background of white-noise with or without autocorrelation to characterize the memory, and is able to detect multiple change-points in each model configuration. Through a simulation study on synthetic time-series the approach is shown to be effective in distinguishing abrupt changes from trends and memory by identifying the true number and timing of abrupt changes when they are present. Furthermore, the proposed method is better performing than two commonly used approaches for the detection of abrupt changes in climate time-series. Using this approach the so-called "hiatus" in recent global mean surface warming fails to be detected as a shift in the rate of temperature rise but is instead consistent with steady increase since the 1960s/1970s. Our method also supports the hypothesis that the Pacific Decadal Oscillation behaves as a short-memory process, rather than forced mean shifts as previously suggested. These examples demonstrate the usefulness of the proposed approach for change detection and for avoiding the most pervasive types of mistake in detection of climate change.

1. Introduction

The pace of climate change is not smooth; it varies year-to-year and decade-to-decade, naturally. Climate records contain shifts or “abrupt changes” due to internal variability and natural forcings (volcanic and solar) superimposed on the long-term anthropogenic climate change trend (Fyfe et al. 2016; Lean and Rind 2009; Trenberth 2015). For example, the global annual mean surface temperature (GMST) time-series exhibits periods of warming separated by a long pause from approximately mid 1940s to mid 1970s (Kellogg 1993) and potentially a second and shorter one, although highly debated, since the late 1990s/early 2000s (Drijfhout et al. 2014; Karl et al. 2015; Trenberth 2015; Trenberth and Fasullo 2013). Whether this last so-called “hiatus” can be characterized as a slowdown in the rate of climate change is the subject of active debate (Medhaug et al. 2017) and has led to a fast growing number of scientific publications (Lewandowsky et al. 2016; Lewandowsky et al. 2015). Discrepancies between the continued warming in models and apparent slowdown of warming in observations since the late 1990s/early 2000s have been suggested to arise from misrepresentations of forcing or natural variability in models (Huber and Knutti 2014; Meehl et al. 2014; Risbey et al. 2014; Santer et al. 2014; Schmidt et al. 2014) or from data biases in observations (Karl et al. 2015), and such change would unlikely be persistent (Knutson et al. 2016). However, few authors have addressed the problem from a statistical change detection perspective (Cahill et al. 2015; Rahmstorf et al. 2017; Rajaratnam et al. 2015). From this angle, the main question is whether the GMST trend has changed in the late 1990s/early 2000s and whether a significant slowdown of warming can be detected.

The Pacific Decadal Oscillation (PDO) has been suggested as a main driver of variability in the GMST increase (Trenberth 2015), with its cold phases corresponding to periods of paused warming and warm phases corresponding to GMST increase. The PDO has also been suggested to be responsible for widespread ecosystem shifts in the North Pacific with repercussions on the region's fisheries (Mantua et al. 1997) and drought effects of the El Niño Southern oscillation (ENSO) (Wang et al. 2014). Whether PDO shifting patterns arise from internal variability or from a forced bi-stable behavior has also triggered debate in the literature over the last two decades (Mantua et al. 1997; Newman et al. 2016; Rodionov 2006; Rudnick and Davis 2003), and has implications for its predictability.

Statistical approaches to characterize change in time-series behaving as a superposition of several components such as long-term trends, shifts (i.e. either in the rate of change or between two stable states) and internal variability, must be flexible enough to distinguish these components. Internal variability is often characterized by a short-memory process, in which the ocean and other slow components of the climate system (e.g. ice sheets) respond slowly to random atmospheric forcing, producing climate variability at a longer time scale than the white noise atmospheric weather. This mechanism is often referred to as "red noise" in the climate literature (Frankignoul and Hasselmann 1977; Hasselmann 1976; Vallis 2010). Natural fluctuations caused by the internal memory can be large enough to mask the long-term warming trend and create periods of apparent slowdown, possibly akin to a "hiatus", as well as exaggerate the warming trend for short periods, which implies risk for ecosystems (Mustin et al. 2013). Long-term trends and shifts above that level of short-term memory should represent natural or external forcings.

Climate science has typically put greater emphasis on statistical model interpretability rather than flexibility because focus is more on a system-level understanding rather than prediction of single events (Faghmous and Kumar 2014). Therefore, statistical approaches used to quantify long-term change in climate time-series typically assume the change is linear in time (Hartmann et al. 2013), and may not allow for all features described above in the same model, thus leading to five possible misuses of statistics, which are illustrated in Fig. 1.

The first type of misuse can occur when characterizing GMST changes (Seidel and Lanzante 2004), i.e. fitting a linear trend in presence of shifts in the mean or shifts in trend (Fig. 1a), which can potentially bias the estimated rate of change. A series of alternative piecewise linear models has been suggested to represent the GMST time-series including periods of warming separated by a pause from the mid 1940s to 1970s (Seidel and Lanzante 2004). However, the performance of such piecewise models to characterize change in the GMST depends on their ability to identify the timings separating the intervals of different rates of warming. Advances in statistics allow identifying the timing of such changes in time-series using change-point detection (Beaulieu et al. 2012; Reeves et al. 2007), and these approaches have recently been used to analyze the GMST time-series by fitting piecewise linear models to objectively detect the timing of changes in the rate of warming (Cahill et al. 2015; Rahmstorf et al. 2017; Ruggieri 2012). More commonly in climate studies, however, change-point detection has been used to detect only shifts in the mean of a time-series, for example by applying the STARS approach (Rodionov 2004). This often leads to the second type of misuse (Fig. 1b): fitting shifts in the mean in presence of a background trend. Because the null model of the STARS approach is a constant mean and not a secular trend, shifts in the mean will tend to provide a better fit to the trend than a constant mean. As such, the method typically interprets a trend as a "staircase" series of abrupt changes (Beaulieu et al. 2016). However, an approach based on model selection, allowing one to distinguish shifts in the mean from a background trend, can prevent the problem of confusing different types of signals as per the first and second misuses (Beaulieu et al. 2012; Reeves et al. 2007).

In addition to different types of signal that may be confused, internal variability may also be misinterpreted as a forced signal, e.g. as a long-term trend or mean shifts (Fig. 1c-d). Patterns created by the internal memory of the system are challenging signal detection in climate time-series as they pose the risk to be misinterpreted as trends or shifts. The risk is greater in presence of short records (Wunsch 1999). The short-term memory or "red noise" is often represented by a first-order autocorrelation process, AR(1), and challenges signal detection as the risk of false alarms is increased when using statistical techniques designed for independent data (von Storch 1999; von Storch and Zwiers 1999). In trend detection, the internal variability can be distinguished from a secular trend by fitting a regression model containing a trend and AR(1) through generalized least squares (Chatfield 2003) or by adjusting the sample size by the effective number of independent observations, which is reduced in presence of autocorrelation (von Storch and Zwiers 1999), thus avoiding the third misuse. As for detecting abrupt changes, some methods have proposed approaches to distinguish change-points from autocorrelation using information criterion and Monte Carlo methods (Beaulieu et al. 2012; Robbins et al. 2016), or pre-whitening of the time-series (Robbins et al. 2016; Rodionov 2006; Serinaldi and Kilsby 2016; Wang 2008) to prevent from the fourth misuse. Finally, as the natural variability is characterized by an AR(1) process, it carries memory that offers short-term predictability. Forecasting a time-series using a stationary AR(1) model when there is an underlying trend and/or shifts in the mean is the fifth possible misuse (Fig. 1e) and will lead to poor predictions.

Our work is thus motivated by the need for distinguishing signals and internal variability in climate and environmental time-series, which is fundamental to better understanding their behavior and predicting future changes. We investigate the behavior of the GMST and PDO time-series (Fig. 2) by developing an approach, which fits a series of models to a time-series and identifies the most appropriate according to the Akaike information criterion (AIC), which is twice the model likelihood penalized by the number of parameters fitted. The models involve combinations of a constant mean or a trend, with a background of white-noise or an AR(1) process, and include the possibility of change-points in each model configuration so as to yield eight models in total (Fig. 3). When a model with change-points is considered, the number is estimated using an optimal segmentation algorithm (Killick et al. 2012). We refer to our approach as “Environmental time-series change-point detection” (EnvCpt) and have also created software available as an R package on the Comprehensive R Archive Network (CRAN) (Killick et al. 2016). Details on the methodology are provided in the next section. We further demonstrate the appropriateness of the methodology through a simulation experiment in which we apply EnvCpt to synthetic time-series mimicking signals and noise observed in climate time-series such as the GMST and the PDO. We compare our approach to two methodologies that have been used to investigate change-points in the GMST and PDO time-series respectively. More specifically, we compare EnvCpt with the STARS methodology (Rodionov 2004), which has been designed to detect mean change-points and has been used to investigate change-points in the PDO among many other applications in the climate and oceanography literature. We also compare EnvCpt with a Bayesian linear regression multiple change-point detection method (BMCpt), which has been used to investigate change-points in the GMST (Ruggieri 2012).

2. Methods

a. Data

We use five annual GMST datasets:

1) Met Office Hadley Centre and Climatic Research Unit surface temperature dataset (HadCRUT4)

The HadCRUT4 dataset (version HadCRUT.4.5.0.0; available at http://www.metoffice.gov.uk/hadobs/hadcrut4/data/current/download.html) (Morice et al. 2012) comprises sea surface temperatures (SST) from the Hadley Centre SST dataset version

3 (HadSST3; (Kennedy et al. 2011a, 2011b) and land surface temperatures from the Climatic Research Unit version 4 (Jones et al. 2012). The dataset anomalies are relative to 1961-1990.

2) HadCRUT4 infilled by kriging (HadCRUT4krig)

We use a variation of the HadCRUT4 dataset, in which regions with no observations were infilled by kriging, mainly across the Arctic, Antarctic, parts of Africa and other small areas (Cowtan and Way 2014); available at http://www-users.york.ac.uk/~kdc3/papers/coverage2013/series.html). The reference period for the anomalies is the same as for HadCRUT4.

3) Merged Land–Ocean Surface Temperature Analysis (MLOST)

The MLOST dataset from the National Oceanic and Atmospheric Administration National Centers for Environmental Information (Smith et al. 2008; Vose et al. 2012; available at https://www.ncdc.noaa.gov/cag/time-series/global) combines land air temperatures from the Global Historical Climatology Network version 3.3.0 (GHCNv3.3.0) and the Extended Reconstructed Sea Surface Temperature version 4 (ERSST.v4) (Huang et al. 2015; Liu et al. 2015). The anomalies are with respect to the 1971-2000 period.

4) Goddard Institute for Space Studies Surface Temperature Analysis (GISTEMP)

The GISTEMP dataset also combines land and SST temperatures from GHCNv3.3.0 and ERSSTv4, but also includes the Scientific Committee on Antarctic Research (SCAR) stations over Antarctica (Hansen et al. 2010) available at http://data.giss.nasa.gov/gistemp/). The anomalies are relative to 1951-1980.

5) Berkeley Earth Surface Temperatures (BEST)

The BEST dataset (Rohde et al. 2013; available at http://berkeleyearth.org/data/) uses SST derived from HadSST3 combined with air temperatures from CRUTEM4 and stations from the GHCN network. Anomalies are given with respect to 1961-1990.

We use the HadCRUT4, HadCRUT4krig and BEST annual GMST datasets from 1850-2016 and the MLOST and GISTEMP annual GMST datasets from 1880-2016 (Figure 2). These datasets share core common observations, but have been processed, bias-corrected and interpolated independently (Jones and Kennedy 2017; Jones 2016).

The PDO dataset used was derived as the leading principal component of monthly sea surface temperature in the North Pacific (downloaded from: http://jisao.washington.edu/pdo/PDO.latest) (Mantua et al. 1997; Zhang et al. 1997). Annual means from 1900-2016 were calculated from the monthly values as a mean from January to December for each year, and presented in Figure 2.

b. EnvCpt description

EnvCpt fits eight models often used to represent climate and environmental time-series and selects which one provides the best fit to represent the time series. The simplest models for the time-series assume that the series is well represented by either a constant mean or a linear trend in addition to a background white noise. These simple models are also fitted superposed to an AR(1), leading to four types of models without change-points. Then, models including change-points in all model parameters (mean or trend, variance and autocorrelation) are also fitted, leading to a total of eight models that are described below.

1) a constant mean (Mean)

$$y_t = \mu + e_t \quad (1)$$

where y_t represents the time-series, t is the time, μ is the mean and e_t is the white-noise errors, which are independent and identically distributed following a Normal with a mean of zero and variance σ^2.

2) a constant mean with first-order autocorrelation (Mean + AR(1))

$$y_t = \mu + \varphi y_{t-1} + e_t \tag{2}$$

where φ is the first-order autocorrelation coefficient.

3) a linear trend (Trend)

$$y_t = \lambda + \beta t + e_t \tag{3}$$

where λ and β represent the intercept and trend parameters, respectively.

4) a linear trend with first-order autocorrelation (Trend + AR(1))

$$y_t = \lambda + \beta t + \varphi y_{t-1} + e_t \tag{4}$$

5) multiple change-points in the mean

$$y_t = \begin{cases} \mu_1 + e_t & t \leq c_1 \\ \mu_2 + e_t & c_1 < t \leq c_2 \\ \vdots & \\ \mu_m + e_t & c_{m-1} < t \leq n \end{cases} \tag{5}$$

where $\mu_{1,}, \dots, \mu_m$ represent the mean of each of the m-segments with variance $\sigma_1^2, \dots, \sigma_m^2$ respectively, $c_1, \dots, c_{m-1}$the timing of the change-points between segments and n is the length of the time-series.

6) multiple change-points in the mean and first-order autocorrelation

$$y_t = \begin{cases} \mu_1 + \varphi_1 y_{t-1} + e_t & t \leq c_1 \\ \mu_2 + \varphi_2 y_{t-1} + e_t & c_1 < t \leq c_2 \\ \vdots & \\ \mu_m + \varphi_m y_{t-1} + e_t & c_{m-1} < t \leq n \end{cases} \tag{6}$$

where $\varphi_1, \dots, \varphi_m$ represent the autocorrelation in each segment.

7) a trend with multiple change-points in the regression parameters

$$y_t = \begin{cases} \lambda_1 + \beta_1 t + e_t & t \leq c_1 \\ \lambda_2 + \beta_2 t + e_t & c_1 < t \leq c_2 \\ \vdots & \\ \lambda_m + \beta_m t + e_t & c_{m-1} < t \leq n \end{cases} \tag{7}$$

where $\lambda_1, \ldots, \lambda_m$ and $\beta_1, \ldots, \beta_m$ represent the intercept and trend in each segment.

8) a trend with multiple change-points in the regression parameters and first-order autocorrelation (Trend cpt + AR(1))

$$y_t = \begin{cases} \lambda_1 + \beta_1 t + \varphi_1 y_{t-1} + e_t & t \leq c_1 \\ \lambda_2 + \beta_2 t + \varphi_2 y_{t-1} + e_t & c_1 < t \leq c_2 \\ \vdots & \\ \lambda_m + \beta_m t + \varphi_m y_{t-1} + e_t & c_{m-1} < t \leq n \end{cases} \tag{8}$$

The theoretical parameter ranges are real numbers for the means, trends and intercepts, positive real numbers for the variances, [-1,1] for first-order autocorrelation coefficients and [*p, n-p*] for the change-point timings with *p* parameters in the model form. The methodology considers all possible parameters and number of changes across the 8 models.

Each model is fitted according to maximum likelihood estimation. For the change-point models, we find the number and location of change-points using the Pruned Exact Linear Time (PELT) algorithm (Killick et al. 2012), which identifies change-points by performing an exact search considering all options for any possible number of changes (varying from 1 to the maximum number of change-points given the set minimum segment length). The search strategy is exact with a computational cost that is linear in the number of data points. The PELT method is used in combination with the modified Bayesian information criterion (MBIC) as the penalty function (Zhang and Siegmund 2007) to select the optimal number of change-points, as this approach balances the overall fit against the length of each segment. Hence it naturally guards against small segments unless it produces a significantly improved fit. The PELT methodology may choose no change-point as the best model in which it reduces to the same likelihood as the no change equivalent model. The model selection is automated using the Akaike information criterion (AIC), which penalizes the model likelihood by the number of parameters fitted for each model considered (Akaike 1974). The EnvCpt package provides the likelihood and number of parameters fitted for each model. As such, any other criteria or metric based on the likelihood can be used for the model selection. However, we use the MBIC for determining change-points as the AIC has been shown to systematically overestimate the number of changes (Haynes et al. 2017). The pseudo algorithm for EnvCpt and additional details about PELT are presented in Appendix A.

The best model is selected as the one with the smallest AIC. While the choice according to the minimum AIC does not provide a measure of uncertainty, the AIC differences (Δ_i) between the best model and the remaining models can be used to evaluate plausibility of the models fitted:

$$\Delta_i = AIC_i - AIC_{min} \tag{9}$$

where *i* denotes the models fitted (*i=1,…,8*). The larger the difference, the less plausible a model is, given the data and models considered (Burnham and Anderson 2002). As a rule of thumb, a Δ_i of 0-2 provides substantial support for model *i*, while Δ_i of 4-7 has considerably less support, and essentially none if the difference is larger than 10 (Burnham and Anderson 2002). While comparing the differences to a rule of thumb is useful to identify a subset of models at play, we can also quantify the plausibility of the models fitted given the data using Akaike weights:

$$w_i = \frac{\exp\left(-0.5 \cdot \Delta_i\right)}{\sum_{r=1}^{8} \exp\left(-0.5 \cdot \Delta_r\right)} \tag{10}$$

The weights, w_i, represent the evidence in favor of model i being the best model given the data and the set of eight models fitted.

c. Simulation of synthetic series

Synthetic series mimicking typical features observed in GMST and PDO time series issued from the eight general models described in the previous section were generated to assess the performance of EnvCpt. We generated a set of synthetic series inspired by the GMST record with a total of 166 years that corresponds to the four models including a trend component fitted to the GMST (Fig. 3a) with a) a long-term trend, b) a long-term trend with first-order autocorrelation, c) a trend with three change-points in 1906, 1945 and 1976, and d) one change-point in the trend and autocorrelation in 1962. We also generated synthetic time-series inspired by the PDO with a length of 116 years to represent the competing models suggested to characterize the PDO behavior: a) mean change-points in 1948 and 1976 with or without a background of AR(1) (Rodionov 2004, 2006) and b) first-order autocorrelation model (Newman et al. 2016). For completeness, the constant mean model used here represents a "null" model for the two hypotheses. Figure 4 presents the eight cases of synthetic series generated to mimic the GMST and PDO. The specific parameters used to simulate the synthetic series are presented in Appendix A (Table A1). For each category, a total number of 1,000 synthetic series were generated and analyzed.

d. Comparison with STARS

We compare our approach to STARS (Rodionov 2004, 2006) using the code available from http://www.climatelogic.com/download. This approach has been used previously to investigate the presence of mean shifts in the PDO (Rodionov 2004, 2006). STARS uses a binary segmentation algorithm that identifies changes sequentially. As such, this procedure finds the most likely change-point, then splits the data at the change if it is significant, and searches for further changes in each segment. This procedure is repeated iteratively until no more changes are detected or the segments are becoming smaller than the set minimum segment length. The decision rule for the presence of change-points is based on a t-test between segments (Rodionov 2004). A minimum segment length default of 10 observations and a critical level of 5% were used in the present study. Thus we set the same default minimum segment length with EnvCpt to carry out the simulations, although other options can be used. The STARS methodology is developed to detect shifts in the mean, however we present results for all considered models to demonstrate the errors produced when trends are not accounted for within the model. Furthermore, STARS is not originally designed to handle autocorrelation, and pre-whitening of the time-series has been suggested when its presence is suspected (Rodionov 2006). Thus, we also applied STARS with two pre-whitening approaches after some parameter tuning (Appendix C). The results obtained after pre-whitening are presented in Appendix D.

e. Comparison with BMCpt

We also compare our approach to a Bayesian identification of multiple change-points in a regression model (BMCpt), which has been used to investigate the presence of change-points in the GMST (Ruggieri 2012). We use the code made freely available from http://mathcs.holycross.edu/~eruggier/software.html. This approach allows for the detection of changes in the parameters of a regression model and thus can detect changes in the mean, trend and/or variance. The exact solution to the multiple change-point detection is obtained using dynamic programming recursions. Here we use a minimum segment length between two shifts of 10, the same as used for EnvCpt and STARS. This approach necessitates setting several other parameters, which are chosen as per the recommendations in Ruggieri (2012) and are described in Appendix B. The hyper-parameters for the variance prior are optimized, as these have an effect on the number of change-points detected (Fig. A1; Appendix B).

BMCpt is also designed to fit a regression model with independent residuals. Thus, we also apply it to the models with AR(1) after pre-whitening. Again, the choice of pre-whitening parameters is determined by optimizing them to give the best performance and is presented in Appendix C.

3. Results

a. Analysis of the GMST and PDO time-series

The eight EnvCpt models are fitted to the GMST datasets and the PDO in Fig. 3. Table 1 presents the AIC differences for each model and their respective weights. For most datasets, the evidence for the Trend cpt + AR(1) model is strong, with probabilities of 1 for BEST, MLOST and GISTEMP, respectively (Table 1). For these three datasets, none of the seven other models are considered plausible ($\Delta_i > 10$; $w_i = 0; i = 1, \dots, 7$) . The HadCRUT4krig dataset reveals more uncertainty, with substantial evidence for both the Trend cpt + AR(1) and the Trend cpt models ($\Delta_i < 2; i = 7,8$), but a higher probability for the Trend cpt + AR(1) model (0.68 for Trend cpt + AR(1) as opposed to 0.32 for Trend cpt; Table 1). On the opposite, for the HadCRUT4 dataset the best model is Trend cpt with a probability of 0.98, while there is limited evidence for the Trend cpt + AR(1) model (probability of 0.02).

For most GMST datasets, the best model fit has one change-point in both the trend and autocorrelation (Trend cpt + AR(1)) in 1962 or 1972 depending on the source of the GMST data (Fig. 3b-e; Table 1). At that time, the rate of warming increases and is accompanied by a whitening of the GMST, i.e. the AR(1) weakens. The trend and AR(1) parameters associated with this fit are presented in Table 2. The competing model (Trend cpt) exhibits a flat mean until 1906, which was followed by a warming period until 1945, then another period of minimal temperature change that lasted until 1977, followed by a warming trend until now (Fig. 3a-b). It must be noted that all models fitted are valid if their underlying assumptions of normality and independence of the residuals are met. Overall, these assumptions are verified under the Trend cpt + AR(1) fit, but not under the Trend cpt model (Figs A5-A6, Table A2; Appendix E). This further validates a background AR(1) and the occurrence of one change-point in the GMST in 1962 or 1972, as opposed to several changes. The GMST has also been suggested to follow an AR(2) model previously (Karl et al. 2000). We find that while two datasets indicate a potential AR(2) structure in the residuals (Fig. A6a-b; Appendix E), the fits are valid with an AR(1) (Fig. A5, Table A2; Appendix E). Furthermore, an AR(2) does not seem to improve the likelihood of the model enough to be worth including as all models with an AR(2) lead to substantially higher AIC (Table A2; Appendix E).

The only model detecting a change-point in the late 1990s/early 2000s is the "staircase" model (Mean cpt), for which there is essentially no evidence ($w_5 = 0$), given the datasets and other models considered (Fig. 3a-e). As such, this result suggests that the most recent "hiatus" does not emerge as a global signal, but rather indicates that the GMST rate of change has remained approximately constant (linear) since the 1960s/1970s with some fluctuations arising from the memory in the system.

As for the PDO, the best fitting model is a constant mean and autocorrelation (Mean + AR(1)) with a probability of 0.56 (Table 1; Fig. 3f), and has valid underlying assumptions (Fig. A7; Table A2). None of the models including change-points are considered at play, as either no change-points are detected (Mean cpt + AR(1) and Trend cpt + AR(1)) or they are associated with large AIC differences (Table 1). The Trend + AR(1) model is the only competing model ($\Delta_4 = 1.1; w_4 = 0.44$), unveiling some uncertainty about the best way to characterize PDO behavior. However, models including a trend would be counterintuitive to represent PDO behavior (Newman et al. 2016).

b. Simulation study

EnvCpt was also applied to the eight different sets of synthetic series generated. To emphasize the flexibility of the methodology developed, we compare it with two other approaches both detailed in Methods. It must be noted that EnvCpt is developed to distinguish all combinations of trends, change-points and autocorrelation, and thus we expect it to overall outperform BMCpt and STARS, which are both designed for more specific features. Specifically, BMCpt was developed to detect changes in a linear regression model, and it should thus perform similarly to EnvCpt in presence of a constant mean or trend, with or without change-points (cases Mean, Mean cpt, Trend and Trend cpt). Correspondingly, STARS was developed to detect mean shifts only and should be performing in the simulation scenario cases Mean and Mean cpt. Neither STARS nor BMCpt were originally designed to handle a background of autocorrelation. To work around that limitation we also apply the methods on the synthetic series with AR(1) after pre-whitening, which necessitates some parameter tuning (see Appendix D).

Fig. 5 presents the number of shifts detected by EnvCpt, STARS and BMCpt in each simulation case. The results demonstrate that EnvCpt correctly identifies the number of change-points at a higher frequency than STARS and BMCpt in most synthetic series, although BMCpt is equivalent in half of the cases. In presence of a trend only, both EnvCpt and BMCpt succeed at identifying no change (Fig. 5a). However, in presence of three trend change-points (Fig. 5c) EnvCpt detects the three shifts at the highest frequency while BMCpt tends to interpret them as two shifts instead. The rate of false detection with BMCpt increases in presence of autocorrelation (Fig. 5b), illustrating misuse 3. In the simulation case Trend cpt + AR(1), EnvCpt and BMCpt are equivalent (Fig. 5d) even though BMCpt is not designed to handle autocorrelation. We attribute this result to the fact that BMCpt can detect changes in the variance, thus interpreting the changing AR(1) here as a change in variance. Finally, in presence of mean shifts (cases Mean cpt and Mean cpt +AR(1)), BMCpt tends to detect fewer shifts than the true number of change-points (Fig. 5g-h). Indeed, when using a change-point approach fitting a piecewise linear regression model in presence of mean shifts only, consecutive "staircase" mean shifts may be interpreted as a trend as per misuse 1. Pre-whitening reduces the rate of false detection by BMCpt in the Trend + AR(1) scenario, but also diminishes the power of detection for the Trend cpt + AR(1) and Mean cpt + AR(1) cases (Fig. A3; Appendix D).

STARS tends to overestimate the number of change-points and frequently misidentifies an underlying trend as a series of shifts, illustrating misuse 2 (Fig. 5a-d). In the cases of a constant mean or change-points in the mean, STARS should be equivalent to EnvCpt, but tends to detect additional spurious shifts (Fig. 5e,g). This is particularly surprising for the Mean case (Fig. 5e), as the STARS methodology should be able to return a no change model in this case, but rather detects changes in over 34% of the series. However, although a 5% critical level is used when multiple shifts are present this does not correspond to a 5% critical level for the overall segmentation given that the test is applied repetitively. Approaches based on a maximal type t-test or F-test, which accounts for the fact that the test statistic is calculated for each potential change-point timing in the time-series, reduce false alarms to the expected level (Lund and Reeves 2002; Wang et al. 2007). The tendency for spurious detection with STARS is aggravated in presence of autocorrelation (Fig. 5f), where STARS detects changes in 96% of the series when none should be detected, illustrating misuse 4. The rate of false detection is reduced with pre-whitening and the detection power improved for the Mean + AR(1) and Mean cpt +AR(1) cases (Fig. A3; Appendix D).

Whilst the number of positive and false-positive changes detected by a given model provides a picture of the performance, it does not indicate whether the change-points are correctly localized in the time-series. Fig. 6 presents density estimates of the locations of the identified change-points for synthetic series that were generated with change-points. This again demonstrates that EnvCpt outperforms STARS and BMCpt overall. EnvCpt clearly identifies the location of the trend change-points, while both BMCpt and STARS tend to detect spurious changes between the true change-points (Fig. 6a), especially towards the end of the series with STARS (Fig. 6a-b,d). The three methods are equivalent in detecting the location of the mean change-points (Fig. 6c). It must be noted that the height of the density peaks may suggest that BMCpt is better performing in the Mean cpt + AR(1) scenario, but this is due to fewer changes being detected with this approach (Fig. 5h). The density and number of change-points should be considered together.

Discussion

Our results suggest that the GMST rate of change has changed once in 1962 or 1972 and has remained approximately constant since then with fluctuations due to the presence of memory in the system. Furthermore, we find that the GMST is “whitening” around that time, i.e. the AR(1) parameter weakens. This result is consistent across most datasets with high evidence (Table 1). Our GMST characterization is different from previous parametric change-point analysis of the global temperature record (Cahill et al. 2015; Rahmstorf et al. 2017; Ruggieri 2012) that suggested the presence of three change-points in the GMST rate of warming in the 1900s, 1940s and 1970s. The main difference lies in the treatment of autocorrelation: our approach formally takes into account the autocorrelation by the means of an AR(1). Indeed, the optimal fit of the Trend cpt model for the HADCRUT4 dataset (Fig.

3a), which does not take account of AR(1), detects three change-points as in previous studies. However, autocorrelation is present in the residuals such that the underlying assumption of independent residuals is violated under the Trend cpt model. The timings of change-points under this model setting (Trend cpt) are not consistent across all GMST datasets, signaling additional uncertainty. If the BIC is used to select the best model instead of the AIC, the Trend cpt + AR(1) model is selected for all datasets (Table A4). We therefore argue that the Trend cpt model should not be used without AR(1) to characterize the GMST. The GMST has also been suggested to follow an AR(2) model previously (Karl et al. 2000). Here we find that an AR(2) does not improve the likelihood of the model enough to be worth including as the noise term (Table A2; Appendix E). Previous work has also suggested the presence of long-term memory in surface temperature records (e.g. Franzke 2012; Løvsletten and Rypdal 2016), as opposed to the short-term memory detected here. In presence of long-term memory, the autocorrelation function will not decay exponentially as observed here, but rather decays as a power law such that it does not reach zero (Yuan et al. 2015). While we do not find long-term memory in the residuals of the five GMST records analyzed here, we acknowledge that its potential presence presents a risk to misinterpret it as a trend or an abrupt change with EnvCpt, but longer records will be needed to make this distinction (Poppick et al. 2017).

Consequently, our results suggest that the change-points previously detected in the 1900s and 1940s may not be unusual given the background memory. These timings also coincide with the period of highest uncertainty in SST measurements due to corrections applied to account for changes of instrumentation (Jones 2016; Kent et al. 2017; Thompson et al. 2008). Despite different results due to different change-point detection approach, we do agree with previous studies (Cahill et al. 2015; Rahmstorf et al. 2017; Ruggieri 2012) that the most recent “hiatus” in GMST does not emerge as a global signal, regardless of whether or not AR(1) is considered. Hence, the only model fitted that contains a change-point in the late

1990s/early 2000s is a "staircase" in the GMST (Mean cpt) and that model fit is rendered unlikely by its large AIC values (Fig. 3).

It must be noted that the five datasets employed in this study are not independent: they all use in part the same input data for the land and ocean but employ different methodologies for correcting biases and inhomogeneities and for interpolating (Jones 2016). As such, the similar results obtained with the five datasets do not provide independent pieces of evidence that a change-point took place in 1962 or 1972, but rather provides a measure of the uncertainty arising from the different approaches used to create these datasets.

To our knowledge, the whitening of the GMST has not been described in previous studies because methodologies able to detect shifts in the autocorrelation, such as EnvCpt, have not been applied to GMST datasets before. The sudden decrease in memory detected here could be due to changes in SST measurements, as the timing marks the start of a period of SST measurements obtained from a more diverse observing fleet and reduced bias (Kent et al. 2017; Thompson et al. 2008). Future studies should investigate the regions responsible for the change-point in GMST and investigate the underlying causes.

As for the PDO, we show that a model with a flat mean and first-order autocorrelation provides the best fit (Fig. 3f), which is in agreement with previous studies (Newman et al. 2016; Rudnick and Davis 2003). Conversely, a previous study has interpreted the PDO as a series of shifts in the mean in the 1940s and 1970s, superposed to an AR(1) (Rodionov 2006), which was taken as support for the hypothesis of a bi-stable behavior. When focusing on a shorter period of time, the 1970s shift was also suggested to emerge from the background of autocorrelation, although the authors questioned the robustness of this result and emphasized the need of a methodology such as the one presented here (Beaulieu et al. 2016). Our new methodology formally compares the two statistical representations (AR(1) process vs bistability with mean shifts) of the PDO by considering them objectively, and we conclude that it is best modeled as autocorrelation only, without shifts. This result is consistent if the BIC is used to select the best model instead of the AIC (Table A4). Memory in the PDO can offer short-term predictability a few years ahead, depending on the strength of the first-order autocorrelation. Specifically, the first-order autocorrelation of 0.55 in the PDO time-series analyzed here translates into a decorrelation time of 3.5 years (von Storch and Zwiers 1999) after which the current PDO value will be "forgotten". This predictability could be key for management, as PDO patterns have widespread repercussions and have been suggested to be responsible for ecosystem regime shifts in the North Pacific and regional droughts (Mantua et al. 1997; Wang et al. 2014). More recently, it has been suggested that the PDO is "reddening" at the monthly timescale, i.e. the AR(1) is increasing as a sign of critical slowing down (Boulton and Lenton 2015; Lenton et al. 2017). We do not detect this feature here, but this is not surprising since our approach is not designed to detect a trend in autocorrelation and has been applied at the annual timescale.

As the PDO and GMST records become longer, the best fitting model may change. More precisely, EnvCpt is expected to select the true underlying model and detect changes more accurately as the number of observations increase (Killick et al. 2012).

The simulation study demonstrates the advantage of a single comprehensive method to avoid five misuses of statistics in analyzing climate time-series. Our approach reduces the number of pre-assumptions about the presence of trends, shifts and autocorrelation in the time-series. In eight cases of synthetic series mimicking features observed in the GMST and the PDO, our approach shows high skill in selecting the correct number of change-points in mean and slope, and to locate the change-points correctly when present. A drawback is that our conclusions are limited to the synthetic series generated for our simulation study. However, previous simulation studies of change-point detection techniques on synthetic series with shifts having a random timing and magnitude have been carried out before, and revealed expected features that are common to most techniques. First, the signal-to-noise ratio matters the most, i.e. a shift with a large magnitude compared to the background noise has a higher hit rate (Beaulieu et al. 2012; Beaulieu et al. 2008; Reeves et al. 2007; Wang et al. 2010). Second, false alarms occur more often at the beginning or end of the time-series (Beaulieu et al. 2012). Third, successive shifts that are near in time tend to be more difficult to detect, especially if the magnitudes have the same sign (e.g. an increase followed by an other increase is more difficult to detect than an increase followed by a decrease) (Beaulieu et al. 2008).

Here we focus on comparing EnvCpt to STARS and BMCpt, which have been used to investigate changes in PDO and GMST, respectively. Overall, our approach clearly outperforms these two methods. This result was to be expected as STARS and BMCpt only consider a subset of the models fitted within EnvCpt. For example, the STARS methodology is developed to detect shifts in the mean only. In terms of the model fit, it is equivalent to considering only the Mean and Mean cpt models fitted with EnvCpt, thereby ignoring the possibility of and misinterpreting underlying trends. BMCpt is more flexible than STARS and designed to detect changes in the parameters of a regression model, so is also equivalent to fitting the models Trend and Trend cpt. Since both of these approaches were developed for independent data, all the models including an AR(1) are excluded from STARS and BMCpt. While this issue can be mitigated with well-tuned pre-whitening (Appendix C), EnvCpt has the additional advantage of natively supporting AR(1) detection without any parameter tuning. In our attempts to tune the pre-whitening for STARS and BMCpt we used a sub-sample size of 20, which is smaller than the length between the shifts inserted in the synthetic series and shown to be optimal (Appendix C). Knowing *a priori* the minimum distance between two shifts is of great benefit for the tuning, but the necessity of tuning is a great disadvantage for STARS and BMCpt. That is, when the “truth” is unknown the choice of parameter values for the pre-whitening is likely to induce errors (Fig. A2; Appendix C).

Several other methods have been proposed in the literature to detect multiple change-points in environmental time-series (e.g. Beaulieu et al. 2012; Gazeaux et al. 2011; Lu et al. 2010; Reeves et al. 2007; Seidou and Ouarda 2007; Tomé and Miranda 2004; Wang 2008) although these models assume independent errors and thus cannot distinguish signals from autocorrelation, similar to STARS and BMCpt. To mitigate this issue one can use pre-whitening techniques, although we show that pre-whitening has the disadvantage to necessitate some parameters tuning. It has also been argued that an approach that forces the lines of the piecewise linear model to meet assuring continuity between the trends is more physically plausible in the case of the GMST (Cahill et al. 2015; Rahmstorf et al. 2017). Here, we do not force the lines of the piecewise linear model to meet, but we find quasi-continuous trends for the GMST (see Fig. 3). Imposing the continuity condition would restrain our approach and make it unsuitable for the detection of climate regime shifts, which are discontinuous and typically represented by abrupt changes in the mean. The main advantage of the approach suggested here is its flexibility and applicability to a wide-range of climate time-series, as illustrated through the GMST and PDO. The flexibility and breath of applicability extends beyond inferring changes in the mean and trend as illustrated with these two examples. Hence, EnvCpt is designed to detect change-points in all parameters of the models fitted, including changes in autocorrelation and variance. There may be cases in which the variability and/or dependence between successive observations are different after the start of a new regime in the climate system or due to changes in measurements procedures. Keeping the methodology as general as possible ensures these cases can also be analyzed with EnvCpt.

Correctly identifying climate change signals is central to their understanding, as mechanisms responsible for secular trends and abrupt changes are likely to be different (e.g. anthropogenic influence vs natural forcings). However, abrupt changes can also be induced in time-series through gradual increase in anthropogenic forcing when a critical threshold is crossed (Lenton 2011). Further investigation of the forcing-response relationship can help identify threshold and nonlinear dynamics, but correctly identifying the timing of an abrupt change is a crucial first step (Andersen et al. 2009). Our EnvCpt approach is timely, as increasing anthropogenic pressure on the climate system is expected to lead to more frequent occurrences of abrupt changes in the physical climate system (Drijfhout et al. 2015).

Our methodology is flexible as it models different types of signals and memory in the system. However, it assumes that temporal changes in climate time-series are piecewise linear on a background of white noise or first-order autocorrelation, and that measurement errors are random. While these assumptions are reasonable in many instances, there may be cases of climate time-series with additional complexities such as long-term memory. Departures from these assumptions may cause problems with the model selected as serious as the five pervasive mistakes we are trying to avoid with EnvCpt. Thus, it is recommended to combine the model selection with an analysis of the residuals as done here (Appendix E), and to consider models that are physically plausible. Given that model selection is used with EnvCpt, it can be easily extended to consider noise terms with additional parameters such as an autoregressive moving-average (ARMA) models with higher-order and alternative model forms (e.g. nonlinear). The models could be extended to take into account co-variables that may explain part of the variability in climate time-series. For example, ENSO could potentially explain part of the variability both in the GMST and PDO analyzed here, and contribute to reducing the unexplained variability. When modifying the models used here, one must keep in mind that the AIC weights are dependent on the subset of models being compared. As such, if additional models were being considered, the probabilities of the eight models compared here may change. Finally, another advantage of an approach based on model selection is that it can be easily modified to use a different information criterion such as the BIC, but the results may vary. We illustrate this in Appendix F and show that using the BIC instead of the AIC in the simulation study can slightly improve the results for most cases of synthetic series, except for the Mean cpt + AR(1) case, for which the results are worst (Figure A8). We refrain from making a universal recommendation here, as there are many factors affecting the performance of AIC and BIC (Burnham and Anderson 2002) with considerations that are going beyond our simulation study. This aspect should be the focus of future work.

Acknowledgements

We thank SECURE and the EPSRC (EP/M008347/1) for funding this research. CB was also supported by a Marie Curie FP7 Reintegration Grants within the Seventh European Community Framework (project 631466 – TROPHYZ). The authors thank three anonymous reviewers and E. Ruggieri for helpful comments that greatly improved the manuscript. We thank S. Rodionov and E. Ruggieri for making their code freely available.

APPENDIX A

Technical detail on the EnvCpt approach and simulations

The EnvCpt approach fits eight different models to the data and returns the fit and number of parameters for each model. The pseudo-code for the algorithm is as follows:

EnvCpt Pseudo Algorithm

Inputs:	Time series y_t
	msl = Minimum number of time points between changes (default 5)
	pen = Penalty for changepoint algorithms (default MBIC)
Initialize:	Let n = length of time series
Fit:	1. Constant mean with independent errors via maximum likelihood
	2. Constant mean with AR(1) errors via maximum likelihood
	3. Linear trend with independent errors via maximum likelihood
	4. Linear trend with AR(1) errors via maximum likelihood
	5. Constant mean changepoint model with independent errors via PELT algorithm with msl and pen options.
	6. Linear trend changepoint model with independent errors via PELT algorithm with msl and pen options.
	7. Constant mean changepoint model with AR(1) errors via PELT algorithm with msl and pen options.
	8. Linear trend changepoint model with AR(1) errors via PELT algorithm with msl and pen options.
Output:	A matrix of likelihood values and number of parameters for each model fit. A list containing the fit for each of the eight models.

Using the output, one can compute an information criterion to determine the model that best fits the data – in this study we use the AIC. See Appendix E for a sensitivity study to the choice of criterion.

The PELT algorithm used in the EnvCpt procedure is described mathematically in (Killick et al. 2012). Contrary to binary searches, where the most likely change is identified and the time-series is split at that point, the PELT algorithm solves the segmentation problem exactly by performing a search considering all options for any possible number of changes (varying from 1 to the maximum number of change-points given the set minimum segment length). This search is completed efficiently using a combination of dynamic programming and pruning. Dynamic programming allows us to consider the data sequentially from the start to the end and monitor the location of the last change-point only, which reduces the computational time significantly. However, as the size of the data grows it remains time consuming to monitor all potential last change-point locations. Thus, pruning is used to solve this issue. For example, if there is an obvious change-point at, say time point 57, then the probability of the last change being before that (e.g. time point 15) is zero. The definition of "obvious" is controlled by the penalty parameter – a larger value means that a change has to be larger to be considered "obvious". If "obvious" changes occur throughout the data then this dramatically reduces the computational time.

To evaluate the approach, we generate synthetic series from each one of the eight models considered with parameters mimicking the GMST and PDO. For reproducibility, the parameters used are presented in Table A1.

APPENDIX B

Choice of parameters for BMCpt

Hyper-parameters for the prior distributions of the regression parameters and variance used with BMCpt are set following previous recommendations (Ruggieri 2012). We set the variance scaling hyper-parameter for the multivariate Normal prior on the regression parameters to 0.01. The hyper-parameters for the variance prior, i.e. the prior variance (σ_0^2), is set to the variance of the data set being used. As for the pseudo data point of variance (ν_0), which is recommended to be <25% of the minimum segment length (Ruggieri 2012), we vary this parameter between 0 and 2.5 to find the value that optimizes the number of change-points detected (Fig. A1). We focus on the number of change-points here, as these parameters can affect the number of change-points detected, but not the distribution of their positions (Ruggieri 2012). Tuning for ν_0 is performed for the four cases without AR(1) for which BMCpt should perform well at identifying the true underlying model. For the cases scenario with no change-points (i.e. Mean and Trend), the value of ν_0 does not have any impact on the number of changes detected as none are detected for all values of ν_0, thus these results are not shown here. As illustrated in Fig. A1a, all values of ν_0 in the simulation scenario of a trend with change-points (Trend cpt) lead to a low detection of the correct number of change-points, but the most substantial improvement is obtained with a value of 0.25. In the case scenario of mean change-points (Mean cpt), the correct number of change-points is obtained at a highest frequency for any values of ν_0 (Fig. A1b). Setting ν_0 to 0 leads to no change-points. Therefore, a value of 0.25 has been used subsequently in all simulations. Finally, the maximum number of change-points is set to 10.

APPENDIX C

Tuning of parameters for pre-whitening

To reduce false alarms due to the presence of autocorrelation, pre-whitening of the time-series was used with STARS and BMCpt (Rodionov 2006). This consists of removing the first-order autocorrelation in the time-series such as:

$$x_t' = x_t - \hat{\rho}^c x_{t-1} \quad t = 2, \dots, n \tag{1}$$

where x_t and x_t' represent the raw and pre-whitened variable at time *t* respectively, *n* is the length of the raw time-series and $\hat{\rho}^c$ represents the bias-corrected first-order autocorrelation estimate. In a practical situation, the first-order autocorrelation used in pre-whitening is unknown (and may also change over time). To obtain an estimate we used two approaches developed by Marriott and Pope (1954) and Orcutt and Winokur Jr (1969), referred to as MP and INV respectively. The MP estimate is given by:

$$\hat{\rho}^c = \frac{(m-1)\hat{\rho}+1}{(m-4)} \tag{2}$$

where $\hat{\rho}$ is the median of the first-order autocorrelation calculated in each subsample of size m. The INV estimate uses four iterative corrections:

$$\hat{\rho}^{c,1} = \hat{\rho} + \frac{1}{m} \tag{3}$$

$$\hat{\rho}^{c,k} = \hat{\rho}^{c,k-1} + \frac{|\hat{\rho}^{c,k-1}|}{m} \quad k = 2,3,4 \tag{4}$$

In order to find an optimal value for the subsample size used in pre-whitening we conduct simulations over a range of subsample sizes using the Mean cpt + AR(1) scenario. This is done with both MP and INV approaches for pre-whitening using subsample sizes of 5, 10, 20, 30, 50 and 75 and illustrated in Figure A2. With both pre-whitening approaches, very large

(75) and very small (5) subsample size lead to a reduced rate of true positives and increased false negatives towards the end of the time-series. A subsample size of approximately 20 is shown optimal here, which is smaller than the distance between the two shifts (28 years). When the number and location of changes is unknown, the choice of this parameter is rather arbitrary and can have substantial effect on the results (Fig. A2).

APPENDIX D

Results obtained after pre-whitening the synthetic data

For comparison, we apply pre-whitening using both MP and INV in all simulations with both STARS and BMCpt, and with a sub-sample size of 20, as chosen after optimization (Fig. A2). Fig. A3 presents the number of shifts detected for the four simulation cases with AR(1). For the two cases with no shifts: Trend + AR(1) and Mean + AR(1), BMCpt with pre-whitening and EnvCpt are equivalent. The number of shifts detected is reduced for STARS, but there is still a substantial rate of false detection. This is surprising, as STARS should be able to return a no change model for the Mean + AR(1) case, but detects changes in over 34% of the series. Nevertheless, the rate of false detection is reduced with pre-whitening, but remains substantial with STARS. In presence of change-points (cases Trend cpt + AR(1) and Mean cpt + AR(1)), the pre-whitening deteriorates BMCpt performance while it significantly improves STARS ability to detect shifts in the mean.

Fig. A4 presents density estimates of the locations of the identified change-points for synthetic series that were generated with change-points and AR(1). For the case Trend cpt + AR(1), whilst the peaks of the true changes have a similar density to the EnvCpt method, STARS and BMCpt tend to detect spurious changes towards the end of the series. In presence of mean change-points, EnvCpt and both STARS and BMCpt applied with pre-whitening succeed at identifying the correct timing of the change-points. While the densities in Fig. A4b give the impression that BMCpt is performing better than STARS and EnvCpt with higher peaks, this is due to fewer changes being detected with this approach (see Fig. A3d).

APPENDIX E

Goodness-of-fit of the GMST and PDO best models

To validate the models selected, we also verify their underlying assumptions of normality and independence of the residuals with additional testing (Table A2). In all cases, the normality assumption of the residuals is respected, but not the independence for all Trend cpt fits on the GMST and the MLOST Trend cpt + AR(1) fits. To further investigate the autocorrelation structure of the residuals for both the Trend cpt and Trend cpt +AR(1) fits, the autocorrelation and partial autocorrelation functions are presented in Figs. A5-A6, respectively. The autocorrelation and partial autocorrelation functions are consistent with the tests of independence presented in Table A2: the residuals of the Trend cpt + AR(1) fits are independent overall (except for the MLOST dataset) (Fig. A5), while the residuals of the Trend cpt fit are not (Fig. A6). The autocorrelation and partial autocorrelation functions for the HadCRUT4 and HadCRUT4krig datasets (Fig. A6a-b) reveals potential presence of a second-order autocorrelation process (AR(2)) in the residuals. Therefore, our models were also fitted with an AR(2) in the background such as : Mean + AR(2), Trend + AR(2), Mean cpt + AR(2) and Trend cpt + AR(2). Table A3 presents the AIC differences of the models fitted with a background AR(2) as opposed to the previously selected models (Trend cpt and Trend cpt + AR(1); Table 1). These results show that despite a potential AR(2) structure in the residuals, there is no benefit from adding an extra parameter to explain the autocorrelation structure. The AIC differences for the models including an AR(2) are substantially larger than those of the best models selected, i.e mostly larger than 10 indicating essentially no evidence for choosing these models instead. There is one exception for the GISTEMP dataset, for which the Trend cpt +AR(2) model has a Δ of 2.5, which suggests some evidence for this model being the best, but not enough to be at play. Overall, for the five GMST datasets, the Trend cpt + AR(1) fit provides the smallest AIC and meet the underlying assumptions of the model. As for the PDO, the model with the smallest AIC (Mean + AR(1)) respects the underlying assumptions of normality and independence (Fig. A7; Table A2).

APPENDIX F

Sensitivity to the model selection criterion

To evaluate the sensitivity to the choice of model selection criterion, we compare the results obtained on all sets of synthetic series with EnvCpt using the Bayesian Information Criterion (BIC) (Figure A8). In most cases, the EnvCpt performance is slightly improved when using the BIC, except for the Mean cpt + AR(1) case for which the BIC detects no change-points in strong majority while there are two.

We also calculate the BIC for the eight models fitted within EnvCpt to the GMST and PDO datasets (Table A4), For all GMST datasets the model with the smallest BIC is Trend cpt + AR(1). This result is slightly different than the results obtained using the AIC for the HADCRUT4 dataset for which the Trend cpt model has the smallest AIC (Table 1). However, we discarded the Trend cpt model for the HADCRUT4 dataset due to the presence of autocorrelation in the residuals (Table A2; Figs A5-A6) and concluded that the second best model, Trend cpt + AR(1), was more appropriate. Thus, the best models identified using the BIC are consistent with the results obtained with the AIC (Figure 3).

References

Akaike, H., 1974: A new look at the statistical model identification. IEEE Transactions on Automatic Control, **19**, 716-723.

Andersen, T., J. Carstensen, E. Hernandez-Garcia, and C. M. Duarte, 2009: Ecological thresholds and regime shifts: approaches to identification. *Trends Ecol Evol*, **24,** 49-57.

Beaulieu, C., J. Chen, and J. L. Sarmiento, 2012: Change-point analysis as a tool to detect abrupt climate variations. *Philos Trans A Math Phys Eng Sci*, **370,** 1228-1249.

Beaulieu, C., O. Seidou, T. B. M. J. Ouarda, X. Zhang, G. Boulet, and A. Yagouti, 2008: Intercomparison of homogenization techniques for precipitation data. *Water Resources Research*, **44,** W02425.

Beaulieu, C., and Coauthors, 2016: Marine regime shifts in ocean biogeochemical models: a case study in the Gulf of Alaska. *Biogeosciences*, **13,** 4533-4553.

Burnham, K. P., and D. R. Anderson, 2002: *Model selection and multimodel inference. Apractical information-theoretic approach*. 2 ed. Springer, New York.Boulton, C., and T. M. Lenton, 2015: Slowing down of North Pacific variability and its implications for abrupt ecosystem change. *Proc Natl Acad Sci U S A*, **112,** 11496-11501.

Cahill, N., S. Rahmstorf, and A. C. Parnell, 2015: Change points of global temperature. *Environmental Research Letters*, **10,** 084002.

Chatfield, C., 2003: *The Analysis of Time Series: An Introduction.* 7 ed. Chapman and Hall/CRC.

Cowtan, K., and R. G. Way, 2014: Coverage bias in the HadCRUT4 temperature series and its impact on recent temperature trends. *Quarterly Journal of the Royal Meteorological Society*, **140,** 1935-1944.

Drijfhout, S., and Coauthors, 2015: Catalogue of abrupt shifts in Intergovernmental Panel on Climate Change climate models. *Proc Natl Acad Sci U S A*, **112,** E5777-5786.

Drijfhout, S. S., A. T. Blaker, S. A. Josey, A. J. G. Nurser, B. Sinha, and M. A. Balmaseda, 2014: Surface warming hiatus caused by increased heat uptake across multiple ocean basins. *Geophysical Research Letters*, **41,** 7868-7874.

Faghmous, J. H., and V. Kumar, 2014: A Big Data Guide to Understanding Climate Change: The Case for Theory-Guided Data Science. *Big Data*, **2,** 155-163.

Frankignoul, C., and K. Hasselmann, 1977: Stochastic climate models, Part II: Application to sea-surface temperature anomalies and thermocline variability. *Tellus*, **29,** 289-305.

Franzke, C., 2012: Nonlinear trends, long-range dependence, and climate noise properties of surface temperature. *Journal of Climate*, **25**, 4172-4183.

Fyfe, J. C., and Coauthors, 2016: Making sense of the early-2000s warming slowdown. *Nature Climate Change*, **6,** 224-228.

Gazeaux, J., E. Flaounas, P. Naveau, and A. Hannart, 2011: Inferring change points and nonlinear trends in multivariate time series: Application to West African monsoon onset timings estimation. *Journal of Geophysical Research*, **116,** D05101.

Hansen, J., R. Ruedy, M. Sato, and K. Lo, 2010: *Global surface temperature change.* Vol. 48, RG4004 pp.

Hartmann, D. L., and Coauthors, 2013: Observations: Atmosphere and Surface. *Climate Change 2013: The Physical Science Basis. Contribution of Working Group I to the Fifth Assessment Report of the Intergovernmental Panel on Climate Change*, T. F. Stocker, and Coauthors, Eds., Cambridge University Press.

Hasselmann, K., 1976: Stochastic climate models Part I. Theory. *Tellus*, **28,** 473-485.

Haynes, K., I. A. Eckley, and P. Fearnhead, 2017: Computationally Efficient Changepoint Detection for a Range of Penalties. *Journal of Computational and Graphical Statistics*, **26,** 134-143.

Huang, B., and Coauthors, 2015: Extended Reconstructed Sea Surface Temperature Version 4 (ERSST.v4). Part I: Upgrades and Intercomparisons. *Journal of Climate*, **28,** 911-930.

Huber, M., and R. Knutti, 2014: Natural variability, radiative forcing and climate response in the recent hiatus reconciled. *Nature Geoscience*, **7,** 651-656.

Jones, G. S., and J. J. Kennedy, 2017: Sensitivity of Attribution of Anthropogenic Near-Surface Warming to Observational Uncertainty. *Journal of Climate*, **30,** 4677-4691.

Jones, P., 2016: The reliability of global and hemispheric surface temperature records. *Advances in Atmospheric Sciences*, **33,** 269-282.

Jones, P. D., D. H. Lister, T. J. Osborn, C. Harpham, M. Salmon, and C. P. Morice, 2012: Hemispheric and large-scale land-surface air temperature variations: An extensive revision and an update to 2010. *Journal of Geophysical Research: Atmospheres*, **117,** n/a-n/a.

Karl, T. R., K. R. W., and B. Baker, 2000: *The record breaking global temperatures of 1997 and 1998: Evidence for an increase in the rate of global warming?* Vol. 27, 719-722 pp.

Karl, T. R., and Coauthors, 2015: Possible artifacts of data biases in the recent global surface warming hiatus. *Science*, **348,** 1469-1472.

Kellogg, W. W., 1993: An apparent moratorium on the greenhouse warmimg due to the deep ocean. *Climatic Change*, **25,** 85-88.

Kennedy, J. J., N. A. Rayner, R. O. Smith, D. E. Parker, and M. Saunby, 2011a: Reassessing biases and other uncertainties in sea surface temperature observations measured in situ since 1850: 2. Biases and homogenization. *Journal of Geophysical Research*, **116**.

——, 2011b: Reassessing biases and other uncertainties in sea surface temperature observations measured in situ since 1850: 1. Measurement and sampling uncertainties. *Journal of Geophysical Research*, **116**.

Kent, E. C., and Coauthors, 2017: A Call for New Approaches to Quantifying Biases in Observations of Sea Surface Temperature. *Bulletin of the American Meteorological Society*, **98,** 1601-1616.

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,** 1590-1598.

Killick, R., C. Beaulieu, and S. Taylor, 2016: version 0.1.

Knutson, T. R., R. Zhang, and L. Horrowitz, 2016: Prospects for a prolonged slowdown in global warming in the early 21st century. *Nature Communications*, **7,** 13676.

Lean, J. L., and D. H. Rind, 2009: How will Earth's surface temperature change in future decades? *Geophysical Research Letters*, **36**.

Lenton, T. M., 2011: Early warning of climate tipping points. *Nature Climate Change*, **1,** 201-209.

Lenton, T. M., V. Dakos, S. Bathiany, and M. Scheffer, 2017: Observed trends in the magnitude and persistence of monthly temperature variability. *Scientific Reports*, **7,** 5940.

Lewandowsky, S., J. S. Risbey, and N. Oreskes, 2016: The "pause" in global warming: turning a routine fluctuation into a problem for science. *Bulletin of the American Meteorological Society*, **97,** 723-733.

Lewandowsky, S., N. Oreskes, J. S. Risbey, B. R. Newell, and M. Smithson, 2015: Seepage: Climate change denial and its effect on the scientific community. *Global Environmental Change*, **33,** 1-13.

Liu, W., and Coauthors, 2015: Extended Reconstructed Sea Surface Temperature Version 4 (ERSST.v4): Part II. Parametric and Structural Uncertainty Estimations. *Journal of Climate*, **28,** 931-951.

Løvsletten, O., and Rypdal, M., 2016: Statistics of regional surface temperatures post year 1900. Long-range versus short-range dependence, and significance of warming trends. *Journal of Climate*, **29**, 4057–4068, 2016.

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

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

Mantua, N. R., S. R. Hare, Y. Zhang, J. M. Wallace, and R. C. Francis, 1997: A Pacific interdecal oscillation with impacts on salmon production. *Bulletin of the American Meteorological Society*, **78,** 1069-1079.

Marriott, F. H. C., and J. A. Pope, 1954: Bias in the estimation of autocorrelations. *Biometrika*, **41,** 390-402.

Medhaug, I., M. B. Stolpe, E. M. Fischer, and R. Knutti, 2017: Reconciling controversies about the global warming hiatus'. *Nature*, **545,** 41-47.

Meehl, G. A., H. Teng, and J. M. Arblaster, 2014: Climate model simulations of the observed early-2000s hiatus of global warming. *Nature Climate Change*, **4,** 898-902.

Morice, C. P., J. J. Kennedy, N. A. Rayner, and P. D. Jones, 2012: Quantifying uncertainties in global and regional temperature change using an ensemble of observational estimates: The HadCRUT4 data set. *Journal of Geophysical Research: Atmospheres*, **117,** D08101.

Mustin, K., C. Dytham, T. G. Benton, J. M. J. Travis, and J. Watson, 2013: Red noise increases extinction risk during rapid climate change. *Diversity and Distributions*, **19,** 815-824.

Newman, M., and Coauthors, 2016: The Pacific Decadal Oscillation, revisited. *Journal of Climate*, **29,** 4399-4427.

NRC, 2013: *Abrupt impacts of climate change: anticipating surprises.* The National Academies Press.

Orcutt, G. H., and H. S. Winokur Jr, 1969: First order autoregression: inference, estimation and prediction. *Econometrica*, **37,** 1-14.

Poppick, A., E. J. Moyer, and M. L. Stein, 2017: Estimating trends in the global mean temperature record. *Advances in Statistical Climatology, Meteorology and Oceanography*, **3,** 33-53.

Rahmstorf, S., G. Foster, and N. Cahill, 2017: Global temperature evolution: recent trends and some pitfalls. *Environmental Research Letters*, **12,** 054001.

Rajaratnam, B., J. Romano, M. Tsiang, and N. S. Diffenbaugh, 2015: Debunking the climate hiatus. *Climatic Change*, **133,** 129-140.

Reeves, J., J. Chen, X. L. Wang, R. Lund, and Q. Lu, 2007: A review and comparison of changepoint detection techniques for climate data. *Journal of Applied Meteorology and Climatology*, **46,** 900-915.

Risbey, J. S., S. Lewandosky, C. Langlais, D. P. Monselesan, T. J. O'Kane, and N. Oreskes, 2014: Well-estimated global surface warming in climate projections selected for ENSO phase. *Nature Climate Change*, **4,** 835-840.

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

Rodionov, S. N., 2004: A sequential algorithm for testing climate regime shifts. *Geophysical Research Letters*, **31,** L09204.

——, 2006: Use of prewhitening in climate regime shift detection. *Geophysical Research Letters*, **33**.

Rohde, R., and Coauthors, 2013: A New Estimate of the Average Earth Surface Land Temperature Spanning 1753 to 2011. *Geoinformatics & Geostatistics: An Overview*, **01**.

Rudnick, D. L., and R. E. Davis, 2003: Red noise and regime shifts. *Deep-Sea Research Part I*, **50,** 691-699.

Ruggieri, E., 2012: A Bayesian approach to detecting change points in climatic records. *International Journal of Climatology*, **33,** 520-528.

Santer, B. D., and Coauthors, 2014: Volcanic contribution to decadal changes in tropospheric temperature. *Nature Geoscience*, **7,** 185-189.

Schmidt, G. A., D. T. Shindell, and K. Tsigaridis, 2014: Reconciling warming trends. *Nature Geoscience*, **7,** 158-160.

Schwarz, G., 1978: Estimating the dimension of a model. The Annals of Statistics, **6**, 461-464.

Seidel, D. J., and J. R. Lanzante, 2004: An assessment of three alternatives to linear trends for characterizing global atmospheric temperature changes. *Journal of Geophysical Research*, **109,** D14108.

Seidou, O., and T. B. M. J. Ouarda, 2007: Recursion-based multiple changepoint detection in multiple linear regression and application to river streamflows. *Water Resources Research*, **43,** W07404.

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

Smith, T. M., R. W. Reynolds, T. R. Peterson, and J. H. Lawrimore, 2008: Improvements to NOAA's historical merged and–ocean surface temperature analysis (1880–2006). *Journal of Climate*, **21,** 2283-2296.

Thompson, D. W., J. J. Kennedy, J. M. Wallace, and P. D. Jones, 2008: A large discontinuity in the mid-twentieth century in observed global-mean surface temperature. *Nature*, **453,** 646-649.
Tomé, A. R., and P. M. A. Miranda, 2004: Piecewise linear fitting and trend changing points of climate parameters. *Geophysical Research Letters*, **31,** L02207.
Trenberth, K. E., 2015: Has there been a hiatus? *Science*, **349,** 691-692.
Trenberth, K. E., and J. T. Fasullo, 2013: An apparent hiatus in global warming? *Earth's Future*, **1,** 19-32.
Vallis, G. K., 2010: Mechanisms of climate variability from years to decade. *Stochastic Physics and Climate Modelling*, T. Palmer, and P. Williams, Eds., Cambridge University Press, 496.
von Storch, H., 1999: Misuses of statistical analysis in climate research. *Analysis of Climate Variability*, H. von Storch, and A. Navarra, Eds., Springer, 11-26.
von Storch, H., and F. W. Zwiers, 1999: *Statistical analysis in climate research.* Cambridge University Press, 455 pp.
Vose, R. S., and Coauthors, 2012: NOAA's merged land-ocean surface temperature analysis. *Bulletin of the American Meteorological Society*, 1677-1685.
Wang, S., J. Huang, Y. He, and Y. Guan, 2014: Combined effects of the Pacific Decadal Oscillation and El Nino-Southern Oscillation on global land dry-wet changes. *Sci Rep*, **4,** 6651.
Wang, X. L., 2008: Accounting for Autocorrelation in Detecting Mean Shifts in Climate Data Series Using the Penalized MaximaltorFTest. *Journal of Applied Meteorology and Climatology*, **47,** 2423-2444.
Wang, X. L., Q. H. Wen, and Y. Wu, 2007: Penalized maximal t test for detecting undocumented mean change in climate data series. *Journal of Applied Meteorology and Climatology*, **46,** 916-931.
Wang, X. L., H. Chen, Y. Wu, Y. Feng, and Q. Pu, 2010: New Techniques for the Detection and Adjustment of Shifts in Daily Precipitation Data Series. *Journal of Applied Meteorology and Climatology*, **49,** 2416-2436.
Wunsch, C., 1999: The interpretation of short climate records, with comments on the North Atlantic and Southern Oscillations. *Bulletin of the American Meteorological Society*, **80,** 245-255.

Yuan, N., Ding, M., Huang, Y., Fu, Z., Xoplaki, E., 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**, 5922–5934.

Zhang, N. R., and D. O. Siegmund, 2007: A Modified Bayes Information Criterion with Applications to the Analysis of Comparative Genomic Hybridization Data. *Biometrics*, **63,** 22-32.

Zhang, Y., J. M. Wallace, and D. S. Battisti, 1997: ENSO-like Interdecadal Variability: 1900–93. *Journal of Climate*, **10,** 1004-1020.

Tables

Table 1: Comparison of the eight EnvCpt models on the GMST and PDO datasets. AIC differences (Δ) between the model with the smallest AIC and the seven other models, as well as their Akaike weights (*w*) representing the probabilities of each model being the best model given the data and the set of models considered. The model with the smallest AIC has a Δ of 0 and is indicated in bold along with its associated probability. Blanks are left for change-point models that did not detect change-points, as the model fit is the same as the equivalent model without change-points.

Model	Data											
	HadCRUT4		HadCRUT4krig		BEST		MLOST		GISTEMP		PDO	
	Δ	*w*	Δ	*w*	Δ	*w*	Δ	*w*	Δ	*w*	Δ	*w*
1.Mean	355.5	0.00	372.7	0.00	386.5	0.00	340.6	0.00	326.7	0.00	42.5	0.00
2.Mean + AR(1)	46.0	0.00	40.7	0.00	40.0	0.00	35.8	0.00	38.5	0.00	**0.0**	**0.56**
3.Trend	165.2	0.00	162.2	0.00	150.3	0.00	152.1	0.00	136.9	0.00	44.5	0.00
4.Trend +AR(1)	31.3	0.00	25.9	0.00	23.3	0.00	23.2	0.00	24.6	0.00	1.1	0.44
5.Mean cpt	40.7	0.00	45.7	0.00	25.3	0.00	61.3	0.00	43.2	0.00	25.8	0.00
6.Mean cpt +AR(1)												
7.Trend cpt	**0.0**	**0.98**	1.5	0.32	16.8	0.00	26.0	0.00	13.4	0.00	23.4	0.00
8.Trend cpt +AR(1)	7.8	0.02	**0.0**	**0.68**	**0.0**	**1.00**	**0.0**	**1.00**	**0.0**	**1.00**		

Table 2: Trend and first-order autocorrelation (AR(1)) parameter estimates for the model with trend change-points and AR(1) (Trend cpt + AR(1)) in the five GMST datasets.

Dataset	Cpt timing	Trend		AR(1)	
		Before cpt	After cpt	Before cpt	After cpt
HadCRUT4	1962	0.001	0.013	0.653	0.195
HadCRUT4krig	1972	0.001	0.018	0.635	0.083
BEST	1962	0.001	0.015	0.656	0.148
MLOST	1962	0.001	0.015	0.706	0.144
GISTEMP	1962	0.002	0.016	0.644	0.112

Table A1: List of parameters used to simulate the sets of synthetic series.

Variable	Model	Parameters
PDO (n=116 years)	Mean	$\mu = 0.028, \sigma = 0.8$
	Mean + AR(1)	$\mu = 0.049, \varphi = 0.522, \sigma = 0.8$
	Mean cpt	$\mu_1 = 0.222, \mu_2 = -0.652, \mu_3 = 0.271$ $c_1 = 49, c_2 = 77, m = 3, \sigma = 0.3$
	Mean cpt + AR(1)	$\mu_1 = 0.222, \mu_2 = -0.652, \mu_3 = 0.271$ $\varphi_1 = \varphi_2 = \varphi_3 = 0.402$ $c_1 = 49, c_2 = 77, m = 3, \sigma = 0.3$
GMST (n=166 years)	Trend	$\lambda = -0.513, \beta = 0.005, \sigma = 0.1$
	Trend + AR(1)	$\lambda = -0.128, \beta = 0.001, \varphi = 0.756, \sigma = 0.3$
	Trend cpt	$\lambda_1 = -0.299, \lambda_2 = -1.327, \lambda_3 = 0.171, \lambda_4 = -2.124, \beta_1 = -0.001, \beta_2 = 0.014, \beta_3 = -0.002, \beta_4 = 0.016, c_1 = 57, c_2 = 96, c_3 = 127, m = 4, \sigma = 0.4$
	Trend cpt + AR(1)	$\lambda_1 = -0.112, \lambda_2 = -1.707, \beta_1 = -0.001, \beta_2 = 0.013, \varphi_1 = 0.659, \varphi_2 = 0.153, c_1 = 113, m = 2, \sigma = 0.1$

Table A2: Results (p-value) of the Lilliefors (L) and Durbin-Watson (DW) tests applied to the residuals of the best models fitted to the GMST (Trend cpt and Trend cpt + AR(1) and PDO datasets (Mean + AR(1).

Model	Test	Data					
		HadCRUT4	HadCRUT4krig	BEST	MLOST	GISTEMP	PDO
Trend cpt	L	0.50	0.50	0.29	0.39	0.12	
	DW	<0.001*	<0.001*	<0.001*	<0.001*	<0.001*	
Trend cpt + AR(1)	L	0.39	0.50	0.33	0.50	0.08	
	DW	0.53	0.25	0.19	<0.001*	0.66	
Mean + AR(1)	L						0.50
	DW						0.68

*Significant at the 1% critical level.

Table A3: Comparison of the best EnvCpt models (Trend cpt and Trend cpt + AR(1)) with models including a second-order autocorrelation process (AR(2)) on the GMST and PDO datasets. AIC differences (∆) between the model with the smallest AIC and the other models are presented. The model with the smallest AIC has a ∆ of 0 and is indicated in bold.

Model	Data				
	HadCRUT4	HadCRUT4krig	BEST	MLOST	GISTEMP
Trend cpt	**0.0**	1.5	16.8	26.0	13.5
Trend cpt + AR(1)	7.8	**0.0**	**0.0**	**0.0**	**0.0**
Mean + AR(2)	41.6	37.1	37.5	34.4	35.5
Trend + AR(2)	30.5	25.0	24.8	25.4	25.2
Mean cpt + AR(2)	48.0	47.7	42.1	37.8	40.5
Trend cpt + AR(2)	42.5	37.0	36.8	37.4	2.5

Table A4: Bayesian Information Criterion (BIC) differences for the eight models within EnvCpt fitted to the GMST and PDO datasets. The model with the smallest BIC has a Δ of 0 and is indicated in bold. Blanks are left for change-point models that did not detect change-points, as the model fit is the same as the equivalent model without change-points.

Model	Data					
	HadCRUT4	HadCRUT4krig	BEST	MLOST	GISTEMP	PDO
1.Mean	325.8	350.9	364.7	320.2	307.4	39.1
2.Mean + AR(1)	19.5	22.0	21.3	18.3	24.1	**0.0**
3.Trend	138.6	143.6	131.6	134.6	-122.6	43.9
4.Trend +AR(1)	7.8	10.3	7.7	8.6	13.0	3.3
5.Mean cpt	39.1	51.9	40.9	67.1	44.8	30.7
6.Mean cpt +AR(1)						
7.Trend cpt	10.8	20.2	23.0	51.5	23.3	33.8
8.Trend cpt +AR(1)	**0.0**	**0.0**	**0.0**	**0.0**	**0.0**	

Figure Captions

Figure 1: Five possible misuses of statistics when inferring changes in climate time-series exhibiting a long-term linear trend, shifts or memory: a) fitting a linear trend in presence of shifts in the mean or shifts in trend; b) fitting shifts in the mean in presence of a trend; c) fitting a linear trend assuming independent errors (i.e. white noise) in presence of autocorrelation; d) fitting shifts in the mean assuming white noise in presence of autocorrelation; e) fitting a first-order autocorrelation model in presence of mean shifts.

Figure 2: Datasets used in this study a) global mean surface temperature (GMST) from the Met Office Hadley Centre surface temperature (HadCRUT4), HadCRUT4 infilled by kriging (HadCRUT4krig), Berkeley Earth Surface Temperature (BEST), Merged Land–Ocean Surface Temperature Analysis (MLOST), and Goddard Institute of Space Studies Surface Temperature Analysis (GISTEMP) and b) the Pacific Decadal Oscillation (PDO).

Figure 3: Fit of the eight models in EnvCpt to five global mean surface temperature (GMST) datasets: a) Met Office Hadley Centre surface temperature (HadCRUT4), b) HadCRUT4 infilled by kriging (HadCRUT4krig), c) Berkeley Earth Surface Temperature (BEST), d) Merged Land–Ocean Surface Temperature Analysis (MLOST), e) Goddard Institute of Space Studies Surface Temperature Analysis (GISTEMP) and f) the Pacific Decadal Oscillation (PDO). The tick marks indicate where change-points were detected. For each dataset, the Akaike Information Criterion differences (Δ) between each model and the best model (smallest AIC) are also shown on a logarithmic scale adjusted so that the best model has a log difference of zero, and is indicated by a star. The dotted vertical lines indicate cutoffs of models evidence: there is substantial support for models with a difference below the red line and essentially no support for models with differences above the black line.

Figure 4: Synthetic time-series example from each simulation scenario case a) a linear trend, b) a linear trend with first-order autocorrelation, c) a trend with three change-points in the regression parameters, d) a trend with a change-point in the regression parameters and first-order autocorrelation, e) a constant mean, f) a constant mean with first-order autocorrelation, g) two change-points in the mean and h) two change-points in the mean with first-order autocorrelation. For each case, a total number of 1,000 random replications are simulated.

Figure 5: Number of change-points detected with EnvCpt, STARS and BMCpt for each simulated scenario across 1,000 replications a) a linear trend, b) a linear trend with first-order autocorrelation, c) a trend with three change-points in the regression parameters, d) a trend with a change-point in the regression parameters and first-order autocorrelation, e) a constant mean, f) a constant mean with first-order autocorrelation, g) two change-points in the mean and h) two change-points in the mean with first-order autocorrelation. Overall, EnvCpt is closer to the true number of change-points than STARS and BMCpt.

Figure 6: Density of change-point timings detected using EnvCpt, STARS and BMCpt for the four simulated scenarios with change-points across 1,000 replications a) a trend with three change-points in the regression parameters, b) a trend with a change-point in the regression parameters and first-order autocorrelation, c) two change-points in the mean and d) two change-points in the mean with first-order autocorrelation. Overall, EnvCpt identifies correctly the true change-point locations while STARS and BMCpt may detect change-points at timings when none were introduced in the synthetic series in presence of trend change-points.

Figure A1: Number of change-points detected with BMCpt for the a) Trend cpt and b) Mean cpt scenario across 1,000 replications. Change-points were detected using a range of values for the pseudo data point of variance parameter (v_0). A value of 0.25 is shown optimal here.

Figure A2: Density of change-point locations for the change-points in the mean and a background AR(1) (Mean cpt + AR(1)) scenario across 1,000 replications. Change-points were detected with a) STARS and b) BMCpt methodologies using a range of subsample sizes for pre-whitening using the MP and INV approaches. A subsample size of 20 is shown optimal here for both methods. For STARS, very large or very small subsample sizes lead to false detections at the end of the time-series. For BMCpt, very large or very small sample sizes lead to improved detection of one shift to the detriment of the other.

Figure A3: Number of change-points detected with EnvCpt, and STARS and BMCpt with pre-whitening for each simulated scenario across 1,000 replications a) a linear trend, b) a linear trend with first-order autocorrelation, c) a trend with three change-points in the regression parameters, d) a trend with a change-point in the regression parameters and first-order autocorrelation, e) a constant mean, f) a constant mean with first-order autocorrelation, g) two change-points in the mean and h) two change-points in the mean with first-order autocorrelation. The pre-whitening is performed using the using the MP and INV approaches with a subsample size of 20.

Figure A4: Density of change-point timings detected using EnvCpt, STARS and BMCpt with pre-whitening for the two simulated scenarios with change-points and AR(1) across 1,000 replications a) a trend with a change-point in the regression parameters and first-order autocorrelation and b) two change-points in the mean with first-order autocorrelation. The pre-whitening is performed using the using the MP and INV approaches with a subsample size of 20.

Figure A5: Autocorrelation and partial autocorrelation function of the residuals from the Trend cpt + AR(1) model fitted to the global mean surface temperature datasets a) HadCRUT4, b) HadCRUT4krig, c) BEST, d) MLOST and e) GISTEMP. Dashed lines represent the 95% confidence intervals on the partial autocorrelation.

Figure A6: Autocorrelation and partial autocorrelation function of the residuals from the Trend cpt model fitted to the global mean surface temperature datasets a) HadCRUT4, b) HadCRUT4krig, c) BEST, d) MLOST and e) GISTEMP. Dashed lines represent the 95% confidence intervals on the partial autocorrelation.

Figure A7: Autocorrelation and partial autocorrelation function of the residuals from the Mean + AR(1) model fitted to the PDO. Dashed lines represent the 95% confidence intervals on the partial autocorrelation.

Figure A8: Number of change-points detected with EnvCpt with either the Akaike Information Criterion (AIC) vs the Bayesian Information Criterion (BIC) for each simulated scenario across 1,000 replications a) a linear trend, b) a linear trend with first-order autocorrelation, c) a trend with three change-points in the regression parameters, d) a trend with a change-point in the regression parameters and first-order autocorrelation, e) a constant mean, f) a constant mean with first-order autocorrelation, g) two change-points in the mean and h) two change-points in the mean with first-order autocorrelation.

Figures

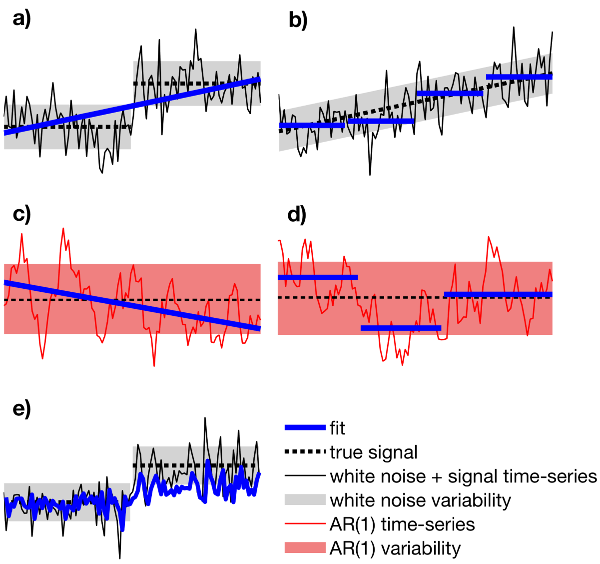

Figure 1: Five possible misuses of statistics when inferring changes in climate time-series exhibiting a long-term linear trend, shifts or memory: a) fitting a linear trend in presence of shifts in the mean or shifts in trend; b) fitting shifts in the mean in presence of a trend; c) fitting a linear trend assuming independent errors (i.e. white noise) in presence of autocorrelation; d) fitting shifts in the mean assuming white noise in presence of autocorrelation; e) fitting a first-order autocorrelation model in presence of mean shifts.

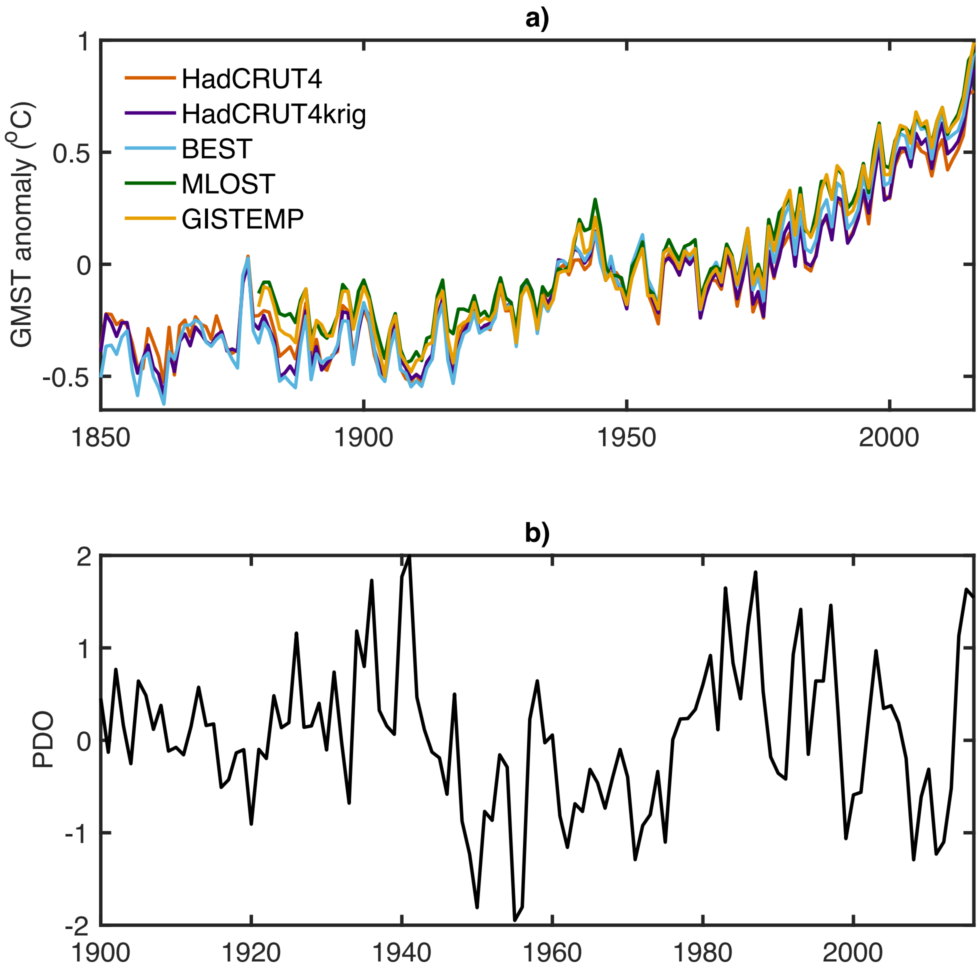

Figure 2: Datasets used in this study a) global mean surface temperature (GMST) from the Met Office Hadley Centre surface temperature (HadCRUT4), HadCRUT4 infilled by kriging (HadCRUT4krig), Berkeley Earth Surface Temperature (BEST), Merged Land–Ocean Surface Temperature Analysis (MLOST), and Goddard Institute of Space Studies Surface Temperature Analysis (GISTEMP) and b) the Pacific Decadal Oscillation (PDO).

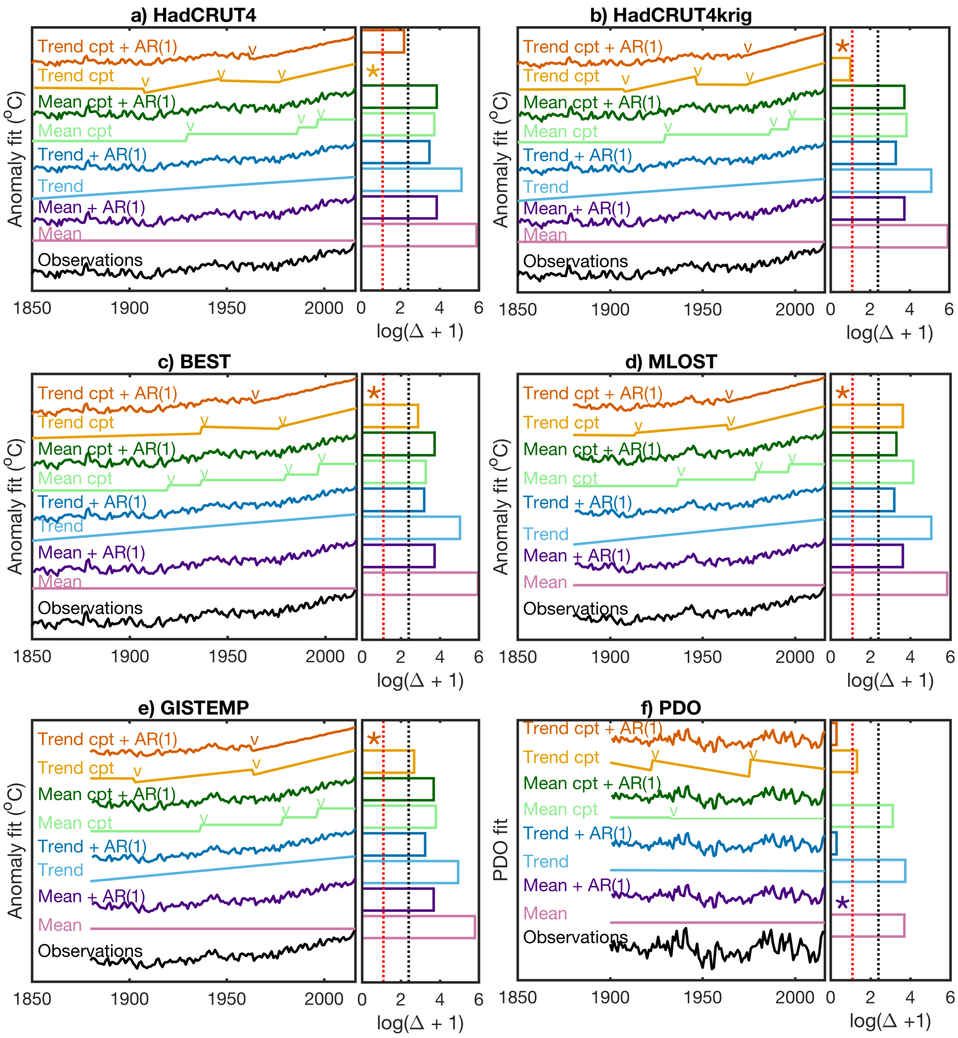

Figure 3: Fit of the eight models in EnvCpt to five global mean surface temperature (GMST) datasets: a) Met Office Hadley Centre surface temperature (HadCRUT4), b) HadCRUT4 infilled by kriging (HadCRUT4krig), c) Berkeley Earth Surface Temperature (BEST), d) Merged Land–Ocean Surface Temperature Analysis (MLOST), e) Goddard Institute of Space Studies Surface Temperature Analysis (GISTEMP) and f) the Pacific Decadal Oscillation (PDO). The tick marks indicate where change-points were detected. For each dataset, the Akaike Information Criterion differences (Δ) between each model and the best model (smallest AIC) are also shown on a logarithmic scale adjusted so that the best model has a log difference of zero, and is indicated by a star. The dotted vertical lines indicate cutoffs of

models evidence: there is substantial support for models with a difference below the red line and essentially no support for models with differences above the black line.

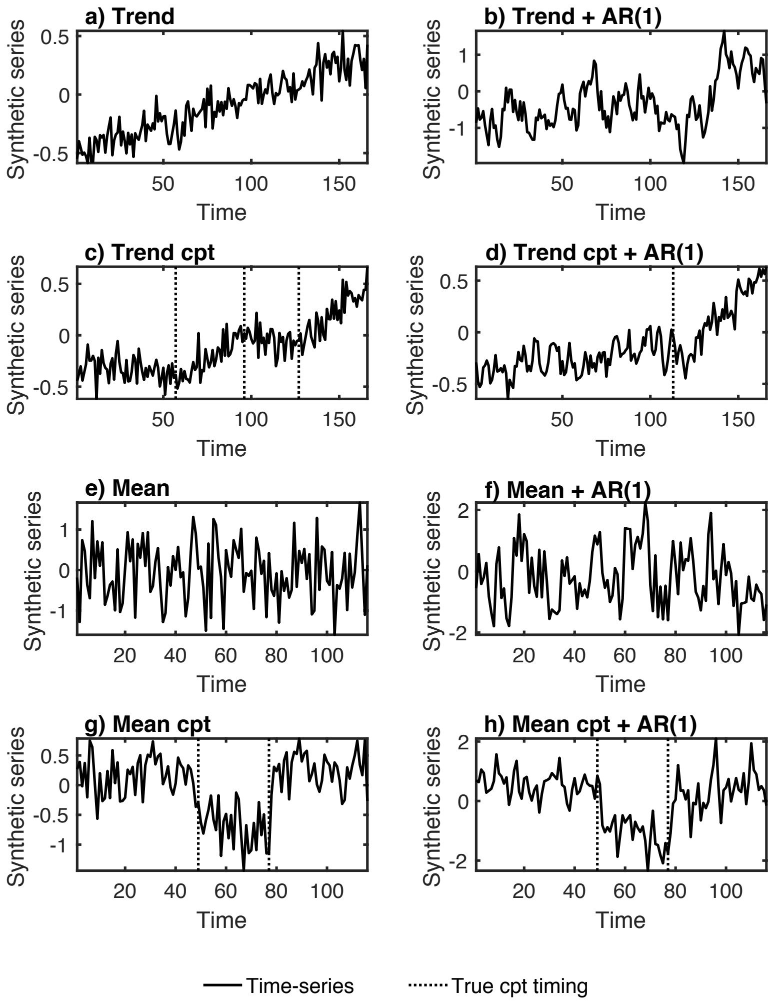

Figure 4: Synthetic time-series example from each simulation scenario case a) a linear trend, b) a linear trend with first-order autocorrelation, c) a trend with three change-points in the regression parameters, d) a trend with a change-point in the regression parameters and first-order autocorrelation, e) a constant mean, f) a constant mean with first-order autocorrelation, g) two change-points in the mean and h) two change-points in the mean with first-order autocorrelation. For each case, a total number of 1,000 random replications are simulated.

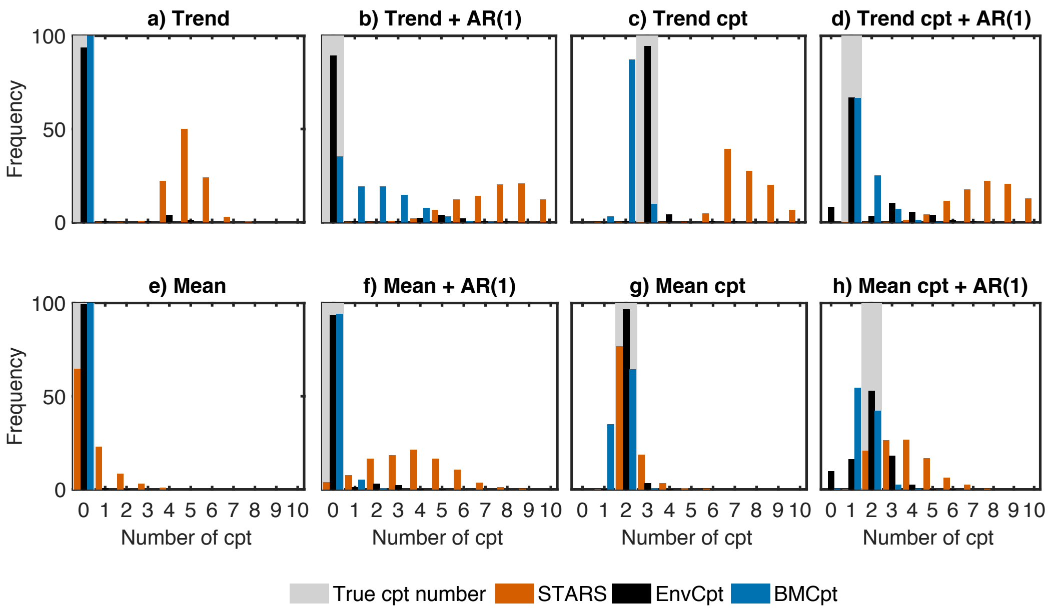

Figure 5: Number of change-points detected with EnvCpt, STARS and BMCpt for each simulated scenario across 1,000 replications a) a linear trend, b) a linear trend with first-order autocorrelation, c) a trend with three change-points in the regression parameters, d) a trend with a change-point in the regression parameters and first-order autocorrelation, e) a constant mean, f) a constant mean with first-order autocorrelation, g) two change-points in the mean and h) two change-points in the mean with first-order autocorrelation. Overall, EnvCpt is closer to the true number of change-points than STARS and BMCpt.

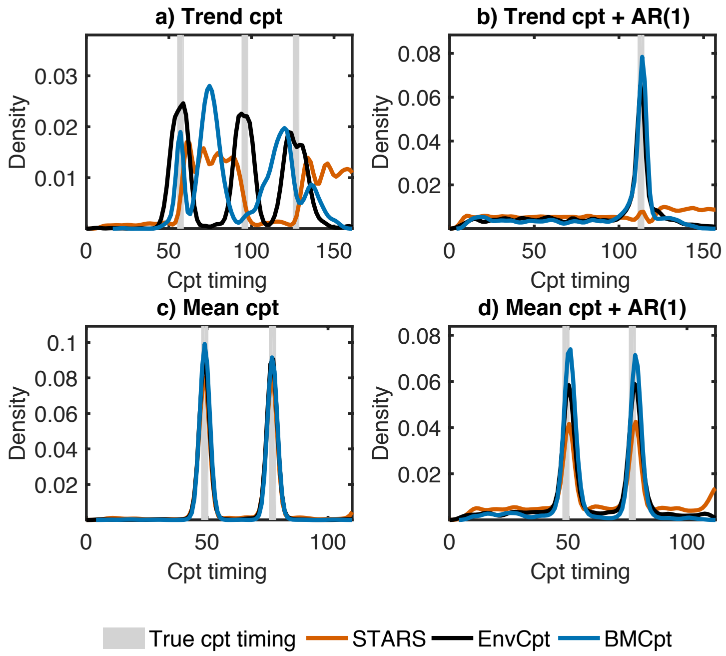

Figure 6: Density of change-point timings detected using EnvCpt, STARS and BMCpt for the four simulated scenarios with change-points across 1,000 replications a) a trend with three change-points in the regression parameters, b) a trend with a change-point in the regression parameters and first-order autocorrelation, c) two change-points in the mean and d) two change-points in the mean with first-order autocorrelation. Overall, EnvCpt identifies correctly the true change-point locations while STARS and BMCpt may detect change-points at timings when none were introduced in the synthetic series in presence of trend change-points.

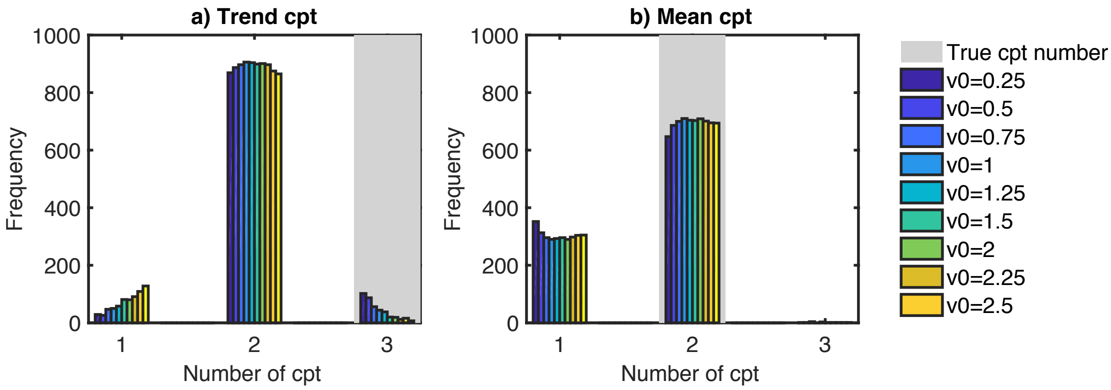

Figure A1: Number of change-points detected with BMCpt for the a) Trend cpt and b) Mean cpt scenario across 1,000 replications. Change-points were detected using a range of values for the pseudo data point of variance parameter (v_0). A value of 0.25 is shown optimal here.

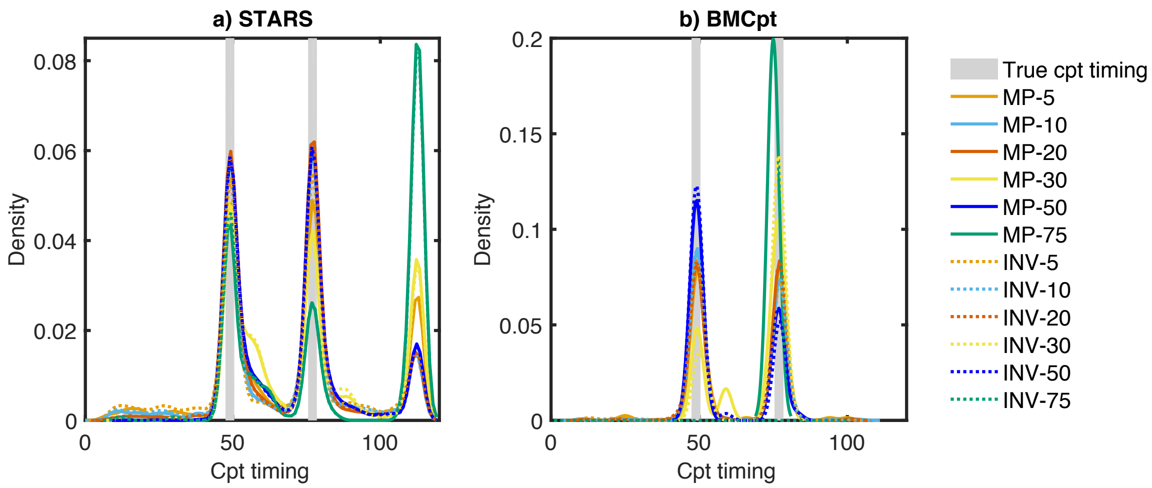

Figure A2: Density of change-point locations for the change-points in the mean and a background AR(1) (Mean cpt + AR(1)) scenario across 1,000 replications. Change-points were detected with a) STARS and b) BMCpt methodologies using a range of subsample sizes for pre-whitening using the MP and INV approaches. A subsample size of 20 is shown optimal here for both methods. For STARS, very large or very small subsample sizes lead to false detections at the end of the time-series. For BMCpt, very large or very small sample sizes lead to improved detection of one shift to the detriment of the other.

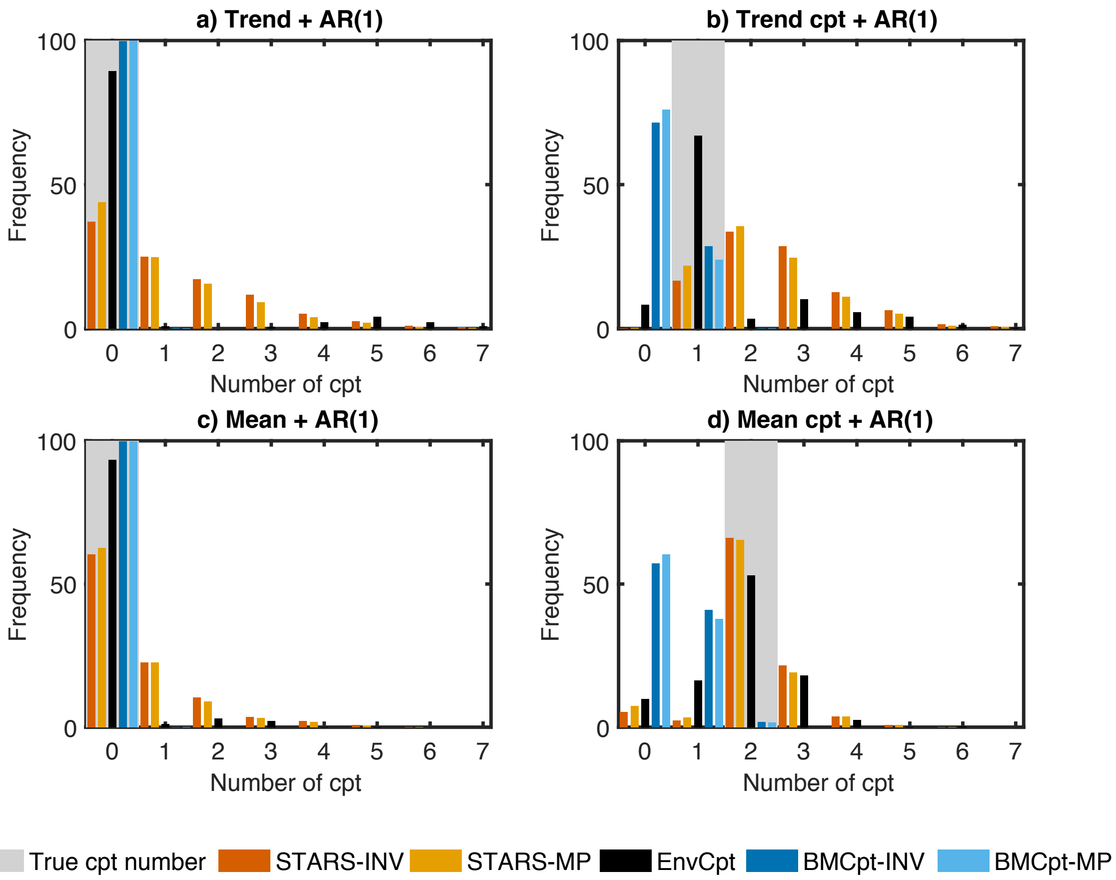

Figure A3: Number of change-points detected with EnvCpt, and STARS and BMCpt with pre-whitening for each simulated scenario across 1,000 replications a) a linear trend, b) a linear trend with first-order autocorrelation, c) a trend with three change-points in the regression parameters, d) a trend with a change-point in the regression parameters and first-order autocorrelation, e) a constant mean, f) a constant mean with first-order autocorrelation, g) two change-points in the mean and h) two change-points in the mean with first-order autocorrelation. The pre-whitening is performed using the using the MP and INV approaches with a subsample size of 20.

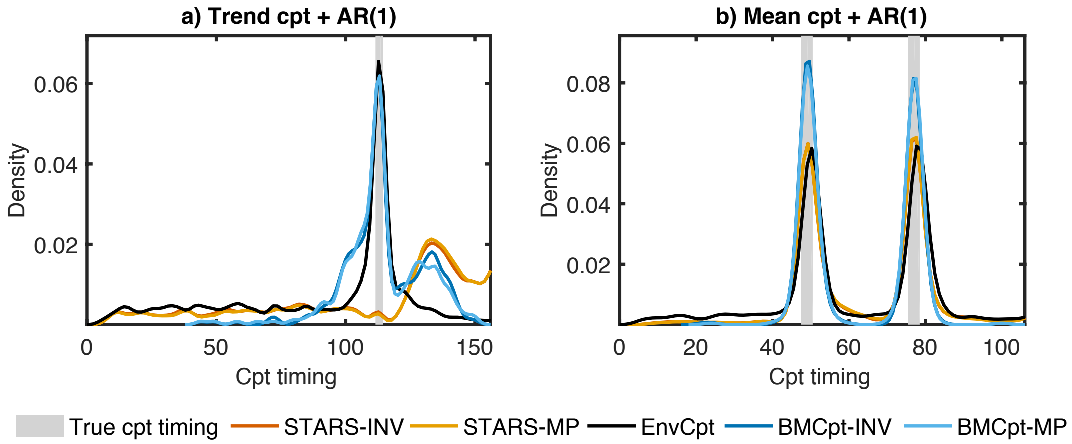

Figure A4: Density of change-point timings detected using EnvCpt, STARS and BMCpt with pre-whitening for the two simulated scenarios with change-points and AR(1) across 1,000 replications a) a trend with a change-point in the regression parameters and first-order autocorrelation and b) two change-points in the mean with first-order autocorrelation. The pre-whitening is performed using the using the MP and INV approaches with a subsample size of 20.

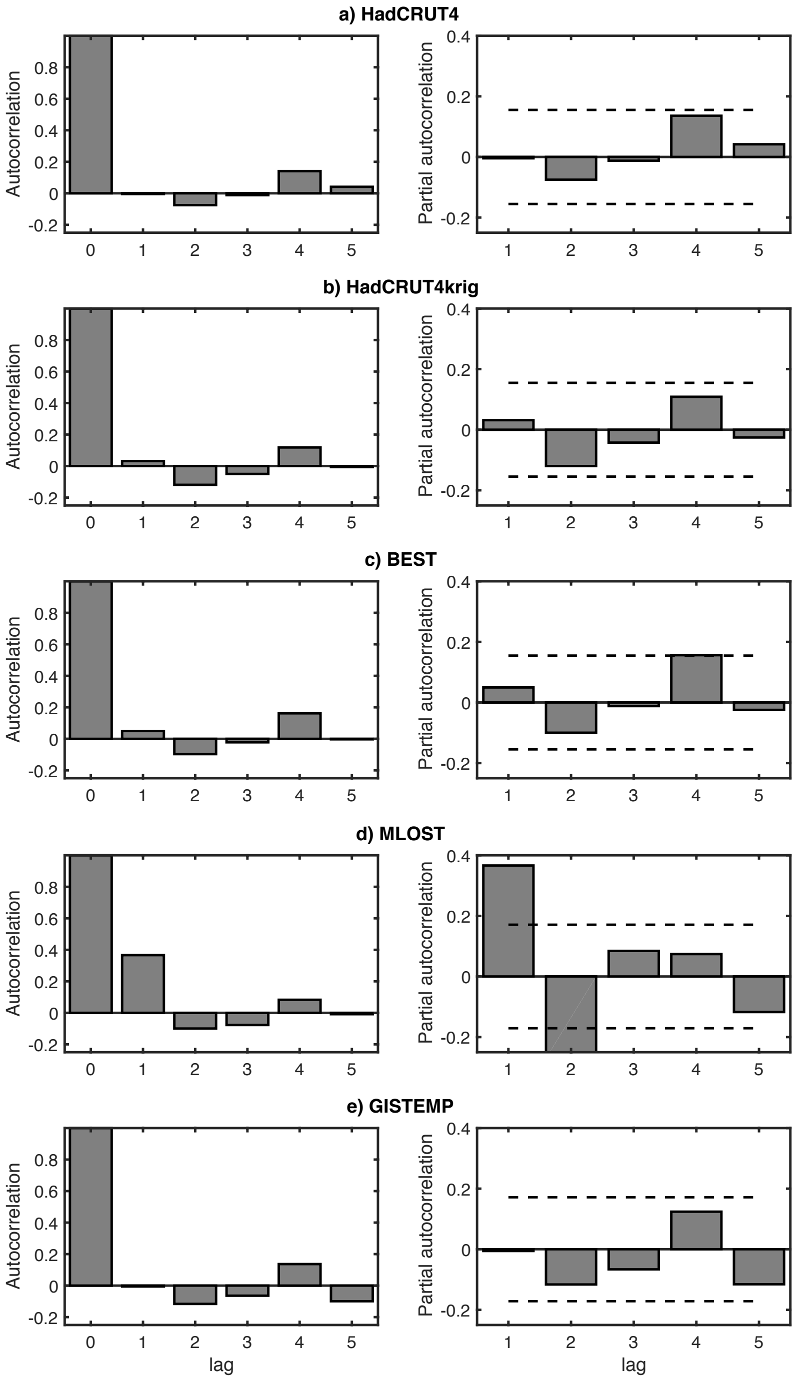

Figure A5: Autocorrelation and partial autocorrelation function of the residuals from the Trend cpt + AR(1) model fitted to the global mean surface temperature datasets a) HadCRUT4, b) HadCRUT4krig, c) BEST, d) MLOST and e) GISTEMP. Dashed lines represent the 95% confidence intervals on the partial autocorrelation.

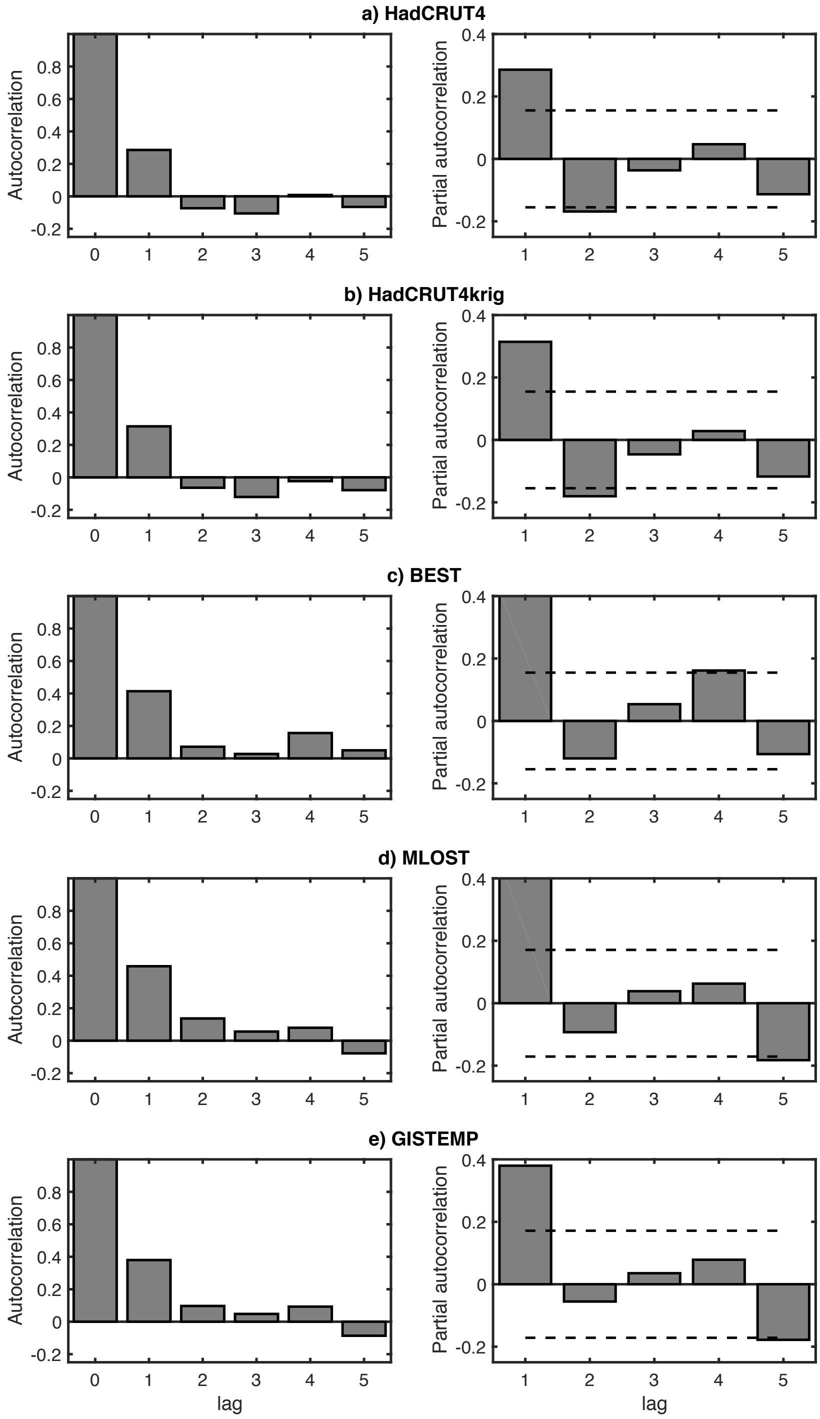

Figure A6: Autocorrelation and partial autocorrelation function of the residuals from the Trend cpt model fitted to the global mean surface temperature datasets a) HadCRUT4, b) HadCRUT4krig, c) BEST, d) MLOST and e) GISTEMP. Dashed lines represent the 95% confidence intervals on the partial autocorrelation.

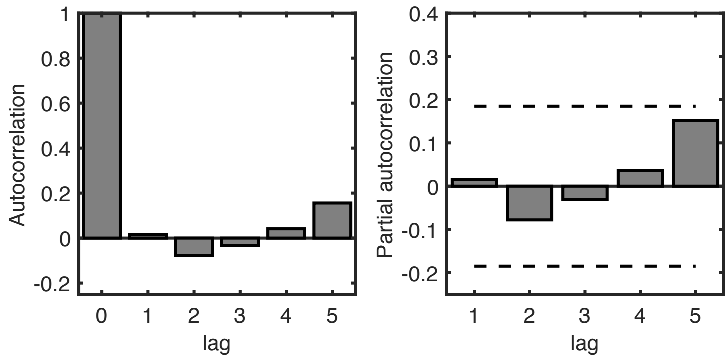

Figure A7: Autocorrelation and partial autocorrelation function of the residuals from the Mean + AR(1) model fitted to the PDO. Dashed lines represent the 95% confidence intervals on the partial autocorrelation.

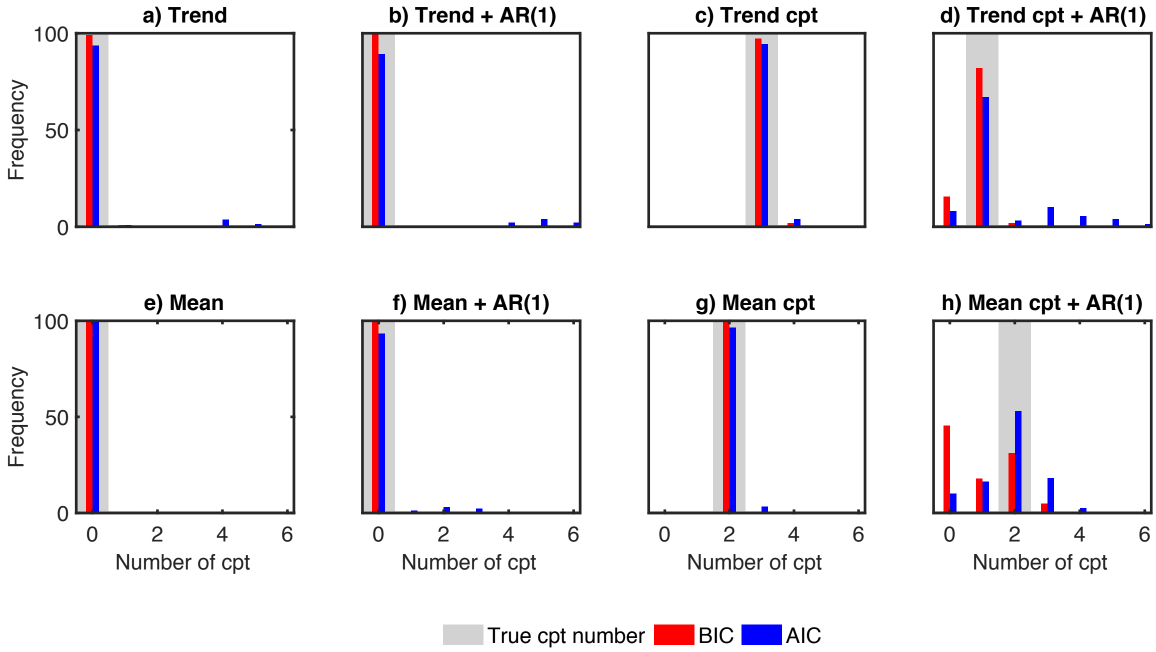

Figure A8: Number of change-points detected with EnvCpt with either the Akaike Information Criterion (AIC) vs the Bayesian Information Criterion (BIC) for each simulated scenario across 1,000 replications a) a linear trend, b) a linear trend with first-order autocorrelation, c) a trend with three change-points in the regression parameters, d) a trend with a change-point in the regression parameters and first-order autocorrelation, e) a constant mean, f) a constant mean with first-order autocorrelation, g) two change-points in the mean and h) two change-points in the mean with first-order autocorrelation.