Considering long-memory when testing for changepoints in surface temperature: a classification approach based on the time-varying spectrum

Claudie Beaulieu[1], Rebecca Killick[2,*], David Ireland[2] and Ben Norwood[2]

[1] Ocean Sciences Department, University of California, Santa Cruz, 1156 High Street, Santa Cruz, CA, 95064, USA

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

* Corresponding author: r.killick@lancs.ac.uk

Abstract

Changepoint models are increasingly used to represent changes in the rate of warming in surface temperatures records. On the opposite hand, a large body of literature has suggested long-memory processes to characterize long-term behavior in surface temperatures. While these two model representations provide different insights into the underlying mechanisms, they share similar spectrum properties that create 'ambiguity', and challenge distinguishing between the two classes of models. This study aims to compare the two representations to explain temporal changes and variability in surface temperatures. To address this question, we extend a recently developed time-varying spectral procedure and assess its accuracy through synthetic series mimicking observed global monthly surface temperatures. We vary the length of the synthetic series to determine the number of observations needed to be able to accurately distinguish between changepoints and long-memory models. We apply the approach to two gridded surface temperature datasets. Our findings unveil regions in the oceans where long-memory is prevalent. These results imply that the presence of long-memory in monthly sea surface temperatures may impact significance of trends, and special attention should be given to the choice of model representing memory (short vs long) when assessing long-term changes.

Keywords: surface temperature, changepoints, long-memory, short-memory, wavelet

Introduction

Quantifying changes in surface temperature records is challenging due to the presence of mixed signals coming from radiative forcings superposed to internal variability. Statistical analyses to characterize changes in such time-series require assumptions for both the signal component and the internal variability. The signal has been commonly characterized as a linear trend (Hartmann et al., 2013; Trenberth et al., 2007), although an increasing number of studies are using piecewise linear trend models with changepoints to describe and quantify the rate of warming (Beaulieu & Killick, 2018; Cahill, Rahmstorf, & Parnell, 2015; Gallagher, Lund, & Robbins, 2013; Karl, Knight, & Baker, 2000; Rahmstorf, Foster, & Cahill, 2017; Ruggieri, 2012; Seidel & Lanzante, 2004) or models with mean changepoints (Jandhyala, Liu, Fotopoulos, & MacNeill, 2014; Khapalova, Jandhyala, Fotopoulos, & Overland, 2018). The model chosen to represent the temporal change is likely to influence estimates of the rate of change, their uncertainty, as well as interpretation of the detected changes.

Internal variability is often characterized as “memory” or “red noise”, in which the ocean and other slow components of the climate system (e.g. ice sheets) respond slowly to random atmospheric forcing, producing variability at a longer time scale than the white noise atmospheric weather (Hasselmann, 1976). The fluctuations caused by the internal memory can be large enough to create periods of apparent slowdowns and surges, and clustering of extreme events (Bunde, Eichner, Kantelhardt, & Havlin, 2005), thus masking or exacerbating the long-term trend with potential risks for ecosystems (Mustin, Dytham, Benton, Travis, & Watson, 2013).

In statistical terms, the memory is often represented by a first-order autocorrelation process (AR(1)) (Mann & Lees, 1996), in which the persistence decays exponentially as a function of the AR(1) parameter, hence representing short-term memory. This assumption has been commonly used in studies quantifying changes in surface temperature (Santer et al., 2008), and adopted to quantify trends in the last Intergovernmental Panel on Climate Change (Hartmann et al., 2013). Some studies even make the simpler assumption of independence (i.e. no memory) in trend detection, but this is well-known to increase the risk of spurious detection if some memory is present (von Storch, 1999; von Storch & Zwiers, 1999). Similarly, the presence of memory increases the risk of spurious detection when applying changepoint models (Tang & MacNeill, 1989, 1993). Another assumption for the internal memory in surface temperatures is that it persists over longer-term such that the autocorrelation function decays as a power law and does not reach zero (Yuan et al., 2015). Long-term memory has been suggested mainly for long climate reconstructions, but also in surface temperature global and gridded observational data sets and model simulations (Blender & Fraedrich, 2003; Efstathiou, Tzanis, Cracknell, & Varotsos, 2011; Fraedrich & Blender, 2003; Huybers & Curry, 2006; Koscielny-Bunde et al., 1998; Lennartz & Bunde, 2009; Rybski, Bunde, Havlin, & von Storch, 2006; Rypdal, Østvand, & Rypdal, 2013; Varotsos & Kirk-Davidoff, 2006; Yuan, Fu, & Liu, 2013).

Research in the statistical and econometric literature has suggested that long-memory processes and changepoint models may be easily confused with one another because both models share some similar properties within the spectrum (Diebold & Inoue, 2001; Granger & Hyung, 2004; Mills, 2007; Smith, 2005; Yau & Davis, 2012). Both representations have been suggested for surface temperatures, and distinguishing between the two has important implications (Ruggieri, 2012) for mechanistic understanding and predictability (Mills, 2007; Smith, 2005). Yau and Davis (2012) proposed a likelihood ratio test for discriminating between the two representations, with a changepoint model as the null hypothesis and long-memory as the alternative hypothesis. Here we instead use a classifying approach (Norwood & Killick, 2018), which does not necessitate to set any models as the null and alternative hypothesis. More specifically, we compare two representations of signals and memory in surface temperatures that have been suggested in the literature: a) piecewise trend with no or short-memory as opposed to b) long-memory with or without a superposed long-term linear trend. We first demonstrate the skill of the method on synthetic series mimicking global surface temperatures with different lengths and determine how many months of observations are necessary to distinguish the true underlying mechanisms described by the two categories of models. We also apply the method to two gridded surface temperature datasets to unveil spatial signatures of the two representations.

Data

We use two monthly gridded surface temperature datasets. The Met Office Hadley Centre and Climatic Research Unit surface temperature (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), combines sea surface temperatures (SST) from the Hadley Centre SST dataset version 3 (HadSST3; (Kennedy, Rayner, Smith, Parker, & Saunby, 2011a, 2011b) and land surface temperatures from the Climatic Research Unit version 4 (Jones et al., 2012). We also use the Merged Land–Ocean Surface Temperature Analysis (MLOST) from the National Oceanic and Atmospheric Administration National Centers for Environmental Information (Smith, Reynolds, Peterson, & Lawrimore, 2008; Vose et al., 2012) available at https://www.ncdc.noaa.gov/cag/time-series/global), which 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).

In both datasets, for each grid cell we retain the longest stretch of data that does not contain missing values. If the length of this stretch of data is below 600 observations (50 years) then we remove that grid point from consideration. This cut-off was chosen, as this is where we saw a tail-off in the accuracy of the classification method for the long-memory model after some preliminary analyses (see Simulation results section). Figure 1 presents the number of observations used in the analysis for each grid cell. The monthly means are deseasonalized to remove a fixed seasonal cycle, i.e. we remove the January average from all January values and so on. The method described below is applied independently to each grid cell to unveil spatial signatures.

Method

Models

We aim to compare two categories of models that have been used to characterize signal and memory in surface temperatures: a) trend changepoints with short-memory and b) trend with long-memory. Since these characteristics may vary in different regions, we use a series of models to generalize how the signal and memory can behave. For the first category we select the best from the following models: mean changepoints and trend changepoints with no or short-term memory as in Beaulieu & Killick (2018). Here the short-memory is represented by an AR(1) process $X_t = \phi X_{t-1} + \epsilon_t$, where $\phi \in (-1,1)$ is the first lag autocorrelation parameter and ϵ_t are the white-noise (WN) errors with variance σ^2. This process is considered short-memory given that its autocovariance decays exponentially with the time-lag τ, such that $\gamma(\tau) = \phi^\tau$ (Brockwell & Davis, 2002). In the absence of memory ($\phi = 0$), the process simplifies to white-noise. The models considered to characterize the surface temperature time-series (Y_t) can be expressed as:

1. multiple changepoints in the mean with WN;

$$Y_t = \begin{cases} \mu_1 + \epsilon_t, & t \leq c_1 \\ \mu_2 + \epsilon_t, & c_1 < t \leq c_2 \\ \vdots & \vdots \\ \mu_m + \epsilon_t, & c_{m-1} < t \leq n \end{cases} \qquad (1)$$

where $\mu_{1,} \dots, \mu_m$ represent the mean of each of the *m*-segments, $c_1, \dots, c_{m-1}$ the timing of the changepoints between segments, ϵ_t are the WN errors with variances $\sigma_1^2, \dots, \sigma_m^2$ depending on the segment and *n* is the length of the time-series.

2. multiple changepoints in the mean with AR(1);

$$Y_t = \begin{cases} \mu_1 + \phi_1 y_{t-1} + \epsilon_t, & t \leq c_1 \\ \mu_2 + \phi_2 y_{t-1} + \epsilon_t, & c_1 < t \leq c_2 \\ \vdots & \vdots \\ \mu_m + \phi_m y_{t-1} + \epsilon_t, & c_{m-1} < t \leq n \end{cases} \qquad (2)$$

where $\phi_1, \dots, \phi_m$ represent the first order autocorrelation in each segment.

3. multiple changepoints in the trend with WN;

$$Y_t = \begin{cases} \lambda_1 + \beta_1 t + \epsilon_t, & t \le c_1 \\ \lambda_2 + \beta_2 t + \epsilon_t, & c_1 < t \le c_2 \\ \vdots & \vdots \\ \lambda_m + \beta_m t + \epsilon_t, & c_{m-1} < t \le n \end{cases} \tag{3}$$

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

4. multiple changepoints in the trend with AR(1);

$$Y_t = \begin{cases} \lambda_1 + \beta_1 t + \phi_1 y_{t-1} + \epsilon_t, & t \le c_1 \\ \lambda_2 + \beta_2 t + \phi_2 y_{t-1} + \epsilon_t, & c_1 < t \le c_2 \\ \vdots & \vdots \\ \lambda_m + \beta_m t + \phi_m y_{t-1} + \epsilon_t, & c_{m-1} < t \le n \end{cases} \tag{4}$$

For all the models listed above, there may be no changepoints detected such that there is only one segment in the time series (m=1).

We use the *EnvCpt* R package (Killick, Beaulieu, Taylor, & Hullait, 2018) to automatically fit the best model among the four models listed above. The methodology considers all possible parameters and number of changes across the 4 models. The number and location of change-points are determined using the Pruned Exact Linear Time (PELT) algorithm (Killick, Fearnhead, & Eckley, 2012), and is used in combination with the modified Bayesian information criterion (MBIC) (Zhang & Siegmund, 2007) as the penalty function to select the optimal number of changepoints. The best model among the four is then selected as the one with the smallest Bayesian Information Criterion (BIC), as shown to be performing well in Beaulieu and Killick (2018). The reader can refer to Beaulieu and Killick (2018) for full details of the methodology.

For the second category of models with long-memory, we either superpose a constant mean or a linear trend to the long-memory process, which we fit using autoregressive fractionally integrated moving average (ARFIMA) models. In its general form, an ARFIMA model can be expressed as:

$$\left(1-\sum_{i=1}^{p}\phi_i B^i\right)(1-B)^d Y_t = \left(1+\sum_{i=1}^{q}\theta_i B^i\right)\epsilon_t \quad (5)$$

where ϵ_t are the WN errors with variance σ^2 and B is the backward operator such that $BY_t = Y_{t-1}$ and $B\epsilon_t = \epsilon_{t-1}$. The ARFIMA model is characterized by the autoregressive (AR) parameters $\boldsymbol{\phi} \in \mathbb{R}^p$, moving average (MA) parameter $\boldsymbol{\theta} \in \mathbb{R}^q$ and the integration (I) parameter is allowed to assume any real value ($d \in \mathbb{R}$). The restriction of d to take only integer values would simplify to an autoregressive integrated moving average (ARIMA) model. For a stationary process, d varies between -0.5 and 0.5 with d=0 indicating no memory, -0.5<d<0 intermediate-memory (anti-persistent) and 0<d<0.5 long-memory. In particular d=0.5 is a discrete-time *1/f* process from (Mandelbrot, 1967). The ARFIMA process with 0<d<0.5 has long-memory because past behavior continues to influence the process for a long time such that the autocovariance decays algebraically as the time lag increases, in contrast to the faster exponentially decaying autocorrelation of a stationary short-memory process (e.g. AR) (Granger & Ding, 1996; Granger & Joyeux, 1980; Hosking, 1981). More specifically, the autocovariance of an ARFIMA (0,d,0) is given by $\gamma(\tau) = |\tau|^{2d-1}$ with a decreasing frequency according to a power law. This is often expressed in terms of the Hurst exponent H (Hurst, 1951), which relates to d as $H = d + 0.5$, and $H \in (0.5,1)$.

Here we restrict the order of the AR process to a maximum of 1 and the order of the MA process to 0 to match the changepoint models (Eqs. 1-4). We fit two long-memory models: one where the long-memory model fluctuates around a constant mean and the other one where long-memory is superposed to a long-term linear trend:

$$Y_t = \mu + ARFIMA(\phi, d, 0), \;\; t \leq n \quad (6)$$

$$Y_t = \lambda + \beta t + ARFIMA(\phi, d, 0), \;\; t \leq n \quad (7)$$

where μ represents a constant mean, λ and β the intercept and linear trend, respectively. For the long-memory models we use the *arfima* R package (Veenstra, 2013). We fit the flat mean (6) and linear trend (7) models separately and choose the model with the smallest BIC value as the best long-memory model.

Classification

Once the best a) trend changepoints with short-memory and b) trend with long-memory models have been identified we use a classification method to select which one is the most appropriate based on examining their time-series spectrum. As changepoint and long-memory models exhibit similar spectral behavior in a standard stationary spectrum, we use the time varying wavelet spectrum to distinguish them (Norwood & Killick, 2018). Heuristically a time varying spectrum is simply the calculation of the traditional spectrum at each individual time point, localized to a small area of information around it. That is, if we take a specific time point we can plot the spectrum across frequency and attain a traditional spectrum but for data localized around that specific time point. To avoid the subjective choice of window size for the localization, as well as other reasons, we use a time varying spectrum based on the locally-stationary wavelet process defined as:

$$Y_{t,N} = \sum_{j=1}^{\infty} \sum_{k} W_j\left(\frac{k}{n}\right) \psi_{j,k-t} \xi_{j,k} \qquad (8)$$

where $j \in 1,2,\ldots$ and $k \in \mathbb{Z}$ are scale and location parameters $\psi_j = \left(\psi_{j,0}, \ldots, \psi_{j,L_j-1}\right)$ are discrete, compactly supported, real-valued non-decimated Daubechies wavelet vectors of support length $L_j = (2^j - 1)(N_h - 1)$ with a Daubechies wavelet filter of size N_h and $\xi_{j,k}$ are orthonormal, zero-mean, identically distributed random variables (Daubechies, 1992). The amplitudes $W_j\left(\frac{k}{n}\right)$ are time-varying, real-valued, piecewise constant functions that have an unknown amount of jumps. The time-varying spectrum is the square of the amplitudes:

$$S_j\left(\frac{k}{N}\right) = \left|W_j\left(\frac{k}{N}\right)\right|^2 \quad (9)$$

and changes over both scale (frequency band) j and location (time) k. The two dimensions of the spectrum (scale and location) allow distinguishing between a changepoint model and a long-memory model. As the long memory model we fit is stationary, the time-varying spectrum is constant over time. In contrast the time-varying spectrum of a model containing changepoints will be piecewise constant. Figure 2 presents examples of time-series simulated from a changepoint model and a long-memory model along with their respective standard stationary spectrum and time-varying spectrum. The ambiguity between their standard stationary spectra is obvious, and notable differences between the time-varying spectra of the two class of models are also highlighted (Figure 2).

To distinguish the two class of models (long-memory vs changepoints), we use a classifier based on these differences, as proposed in Norwood and Killick (2018). This approach involves comparing a dataset to "known" groups through a distance metric. Since the truth is unknown, we simulate 1000 Monte Carlo replications of each of the best models in each category to serve as training data to build a classifier.

For each group, changepoint and long-memory, the time varying spectrum of each of the M=1000 simulated replications is calculated:

$$S_m^g = \left\{S_{k,m}^g\right\}_{k=1,2,\ldots,n*J} \quad (10)$$

Here S is the vector containing the time varying spectrum, g is the group, m is the simulation index from 1 to 1000, k is the index of the time varying spectrum over n time points and J frequency bands.

To get a representation of the time varying spectral behavior of each group, we take the average at each time-frequency point for each of the M=1000 replications:

$$\bar{S}^g = \left\{\frac{1}{M}\sum_{m=1}^{M} S_{k,m}^g\right\}_{k=1,2,\ldots,n*J} \tag{11}$$

Denoting the spectrum of the original data by S^0, based on these average spectra for each group we calculate the variance corrected distance metric across all time-frequency points from Norwood and Killick (2018):

$$D^g = \frac{M}{M+1}\sum_{k=1}^{n*J}\frac{\left(S_k^0-\bar{S}_k^g\right)^2}{\sum_{m=1}^{M}\left(S_{k,m}^g-\bar{S}_k^g\right)^2} \tag{12}$$

This distance metric allows for different variances in each group. Further details on the locally-stationary wavelet process and the time-varying spectrum classifier can be found in Norwood and Killick (2018).

Simulation of synthetic series

Synthetic series were generated to mimic the behavior seen in the HadCRUT4 global monthly surface temperature (GMST) time series for the two categories of models, and evaluate whether the proposed approach would be able to distinguish them. In particular, we fit the best long-memory and changepoint models to the HadCRUT4 GMST, without assuming that one is better than the other, and simulate random series from the fitted models. To evaluate the effect of the record length on the performance, we simulate varying record lengths, from a minimum of 50 years (N=600 months) to the length of the whole record of 168 years (N=2016 months).

The models used for simulation are given as follows where the specific parameters used to simulate the synthetic series are presented in Table 1:

a) Trend changepoint model with AR(1) errors (Trend cpt + AR(1));

b) Trend with long-memory model (Trend + LM).

To investigate how the length of the series affects the classification we take the two models and create 1000 monthly synthetic series for each of N=600, 700, 800, 1000, 1200, 1400, 1600, 1800, 2016 (corresponding to samples varying between 50 to 168 years). For the changepoint series we fix the location of the changepoints relative to the length of the series, as detailed in Table 1. For the Trend + LM scenario, we carry an additional simulation in which we simulate the series with the same parameters (Table 1), except that we vary the long-memory strength (from d=0.1 to d=0.499).

Results

Simulation results

We apply the classification approach detailed above to the two sets of synthetic series generated with varying lengths N. Figure 3 presents the classification hit rates for the two simulation cases. The results demonstrate that overall it is easier to identify models with changepoints than models with long-memory. We show that with 50 years of observations, we can successfully classify the changepoint model (Trend cpt + AR(1)) with hit rates >99%, while the hit rate for the long-memory model (Trend + LM) is ~70%

(Figure 3a). As the series length increases, the classification hit rate improves for the long-memory model. With about 100 years of observations, the classifier's skill improves, reaching ~95% hit rate. With 150 years of observations, the approach correctly classifies the Trend + LM model with a hit rate >99%. Note that the level of long-memory in the Trend + LM case described above is high (d=0.485). To evaluate the effect of long-memory on the classifier's ability, we also run simulations with the Trend + LM model with a varying degree of long-memory (from d=0.1 to d=0.499) (Figure 3b). For a very strong long-memory (d=0.499), the classifier reaches 60% hit rate at best with 168 years of data. For a weaker long-memory (d≤0.4), the classifier produces hit rates >80% with 50 years of data and reaches >97% with 168 years of data.

To demonstrate the importance of distinguishing between the two models for mechanistic understanding, we present how 'wrong' the results get when fitting the changepoint models to the synthetic series with long-memory (Trend + LM). Table 2 presents the percentage of series that detected at least one changepoint when the true model is Trend + LM. We can see that as the sample size increases, the percentage of simulations identifying erroneous changes increases. This is due to the fact that data from long-memory processes are prone to periods of increasing or decreasing trends and thus the longer the simulated long-memory process, the more likely these behaviors will manifest.

GMST gridded datasets

The classification approach detailed above was applied to the HadCRUT4 and MLOST gridded datasets. The results are presented in Figure 4 as a heat map. Results reveal consistent patterns between the two datasets, although more MLOST grid cells were used in the analysis (Figure 1). Overall, the surface temperatures over land are better characterized as changepoint models with short-memory, while long-memory arises in regions of the ocean. For the cases where a changepoint model with short-memory is preferred, the number of changepoints is presented in Figure 5. In most cases, one changepoint is present in the time series, but some regions over land exhibit more than one changepoint. It must be noted that for the cases where no changepoints are detected, our classification approach is considered inconclusive as both series are stationary. These inconclusive areas are mostly located over the ocean around long-memory hot spots, suggesting that the transition zones are especially difficult to classify. Figure 6 presents the memory strength (fractionally-differenced parameter d from the ARFIMA model) in those long-memory hot spots for both datasets. It averages to 0.29 and 0.28 for the HadCRUT4 and MLOST datasets, respectively. This is lower than the long-memory estimated from the global HadCRUT4 time-series used to simulate synthetic series (d=0.485, Table 1). However, since our approach suggests that a changepoint model provides a better fit than a long-memory model at the global level (i.e. the variance corrected distance metric is -1), we hypothesize that the long-memory estimate may be spuriously inflated in the global record.

Discussion

We propose an approach to distinguish between two categories of models commonly used to characterize signal and memory in surface temperatures: a) short-memory superposed by a piecewise trend (Beaulieu & Killick, 2018; Cahill et al., 2015; Karl et al., 2000; Rahmstorf et al., 2017; Ruggieri, 2012; Seidel & Lanzante, 2004) or long-memory that may be superposed by a long-term trend (Franzke, 2012; Ludescher, Bunde, & Schellnhuber, 2017). The ambiguity between changepoints models and long-memory has been widely discussed in the statistical and econometric literature (Diebold & Inoue, 2001; Granger & Hyung, 2004; Mills, 2007; Smith, 2005; Yau & Davis, 2012). In the climate literature, a systematic comparison between the two classes of models on temperature reconstructions datasets showed preference for changepoint models (Rea, Reale, & Brown, 2011), but to our knowledge, there has not been a formal comparison on surface temperature observations. The novelty of the present analysis is to formally and automatically compare both representations on observational records across hundreds of gridded locations. Our results show that the best combination of signal and noise has a strong spatial signature, where changepoints and short-term memory models are mostly appropriate over the land, while long-term memory is more prevalent in the oceans. Rypdal et al. (2013) suggests that the long-memory in the oceans is associated with the thermal inertia of the oceans. The small effective thermal inertia of the land surface compared to the oceans leads to shorter-memory over the continents (Manabe & Stouffer, 1996; Pelletier, 1997). Our results further highlight hot spots where long-memory arises in sea surface temperatures in the extratropical North Pacific and North Atlantic, as well as in the tropical Pacific. These regions were previously shown to exhibit higher persistence (Vyushin, Kushner, & Zwiers, 2012). In oceanic regions away from intense currents and thermal fronts, the persistence is typically explained by a simple model where the ocean slowly responds to atmospheric weather and create short-memory (Frankignoul & Hasselmann, 1977; Hasselmann, 1976). The regions highlighted here are characterized by important currents, such as the Gulf Stream in the North Atlantic for example, and likely need additional complexity to explain the memory structure observed here. This question should be investigated using climate models providing a better spatial coverage. We leave this aspect to future investigation.

The classification used here is inconclusive in some areas (i.e. no changepoints detected, see Figure 5) because the approach is designed to distinguish the shapes of time varying spectra, where changepoints will show a piecewise constant time varying spectrum as opposed to a constant spectrum over time for long-memory. Without changepoints the problem reduces to a comparison between short-memory vs long-memory models, and a time varying spectrum is not appropriate to answer this question. In that case, it is instead recommended to use a test for distinguishing between short-memory and long-memory (Giraitis, Kokoska, Leipus, & Teyssière, 2003). For surface temperature data, a comparison between short-term and long-term memory on reanalysis data sets and model simulations suggest that climate persistence could lie in-between and that the data does not suggest that one representation is superior (Vyushin et al., 2012). However, it must be noted that a significant portion of the inconclusive areas also coincide with grid cells with limited data availability (~50 years/600 months without missing values) (Figure 1), which suggests that the areas of long-memory in the oceans could potentially be underestimated. Hence, classifying the two categories of models is more difficult with shorter time-series as opposed to the full record period (168 years) (Figure 3), and we find that this is emphasized when the "true" underlying model has long-memory. When the "true" model has short-memory and changepoints, fewer observations are required to perform a successful classification. This result is consistent with the simulation study in Norwood and Killick (2018), which demonstrates that this approach provides perfect classification in the case of a true changepoint model and increasingly correct classifications, as n grows, in the case of a true long-memory model. The simulations in Norwood and Killick (2018) were conducted in a constant mean scenario and so we assess the performance of the method for linear trends here. At a lower time resolution such as annual, long-memory may not be detectable due to the reduction in the number of observations and less likely to impact the significance of trends and changepoints. However, this is purely speculative and the time resolution aspect will be left for a future investigation.

The results presented here may be affected by the use of discontinuous piecewise trend models to characterize the behavior of surface temperatures. Some studies have argued that global temperature piecewise trends should be continuous, where the lines of the different segments are forced to meet at the changepoints (Rahmstorf et al., 2017). Here we do not impose the continuity constraint to keep more flexibility, as some regions may exhibit discontinuities (Beaulieu & Killick, 2018). Furthermore, we have previously shown that changes detected under discontinuous models may give quasi-continuous segments, such that even though the continuity constraint is not imposed, the discontinuity is small and may only slightly impact the number and timing of the changepoints. Similarly, the autocorrelation and variances are allowed to vary between segments under our changepoint models, as opposed to simpler models that impose a global autocorrelation and variance and allow changepoints in the trend only. This choice is based on previous findings, where five GMST datasets were shown to be better represented by a trend changepoint model with AR(1), with an intensification in warming in the 1960s/70s accompanied by a reduction of autocorrelation (Beaulieu & Killick, 2018). Forcing a global autocorrelation when it actually varies with time could lead to spurious changepoints, thus we allow the autocorrelation parameters to change in each segments. If in some regions the autocorrelation parameters are constant through the time series, then their estimates will be very similar between segments. Studying the sensitivity of our results to a continuity constraint and constant autocorrelation is out of scope for the present study and is the focus of ongoing work.

Based on our results, it is recommended to verify the presence of long-memory when testing for long-term trends and changepoints in sea surface temperatures, especially over the regions identified here (Figure 4). Hence, assuming a short-memory model such as routinely done in the IPCC (Hartmann et al. 2013) when testing for trends in presence of long-memory may impact their significance (Bloomfield & Nychka, 1992; Franzke, 2012; Lennartz & Bunde, 2009; Ludescher et al., 2017). Similarly, piecewise trends may not hold in the presence of long-memory as demonstrated here. Separating signal and memory in surface temperatures is especially important as there may be implications for the attribution of the signal detected (Imbers, Lopez, Huntingford, & Allen, 2014; Rypdal, 2015).

Throughout this exposition we have concentrated on classifying changepoint models with long-memory models. An interesting statistical avenue to explore would be to include a comparison with long-memory models that also include changepoints (Beran & Terrin, 1996; Horvath, 2001). The challenge here would be in distinguishing between the changepoint model with short-memory and the changepoint model with long-memory as both would present as non-stationary spectra so we may expect the two groups to be close. In the context of modeling surface temperatures we feel that there is currently not enough data to accurately fit changepoint models with long-memory errors. This is due to the fact that a typical segment is unlikely to be longer than 50 years making estimation of the difference between changepoint and long-memory with changes infeasible.

A limiting factor in the modeling presented here is that the estimation and classification require complete data. An interesting avenue for further research would be to develop approaches for identifying changepoints and long-memory in data that contains large periods of missing values. Also, the classification is performed in each grid cell separately, while it is likely that the signal and memory in a given grid cell will be similar to its neighbors. As such, integrating spatial correlation in the analysis has potential to improve the classification for spatial fields such as surface temperatures.

References

Beaulieu, C., & Killick, R. (2018). Distinguishing trends and shifts from memory in climate data. *Journal of Climate, 31*, 9519-9543. doi:10.1175/JCLI-D-17-0863.1

Beran, J., & Terrin, N. (1996). Testing for a change of the long-memory parameter. *Biometrika Trust, 3*, 627-638.

Blender, R., & Fraedrich, K. (2003). Long time memory in global warming simulations. *Geophysical Research Letters, 30*(14). doi:10.1029/2003gl017666

Bloomfield, P., & Nychka, D. (1992). Climate spectra and detecting climate change. *Climatic Change, 21*, 275-287.

Brockwell, P. J., & Davis, R. A. (2002). *Introduction to time series and forecasting* (2nd ed.): Springer.

Bunde, A., Eichner, J. F., Kantelhardt, J. W., & Havlin, S. (2005). Long-term memory: A natural mechanism for the clustering of extreme events and anomalous residual times in climate records. *Phys Rev Lett, 94*(4), 048701. doi:10.1103/PhysRevLett.94.048701

Cahill, N., Rahmstorf, S., & Parnell, A. C. (2015). Change points of global temperature. *Environmental Research Letters, 10*(8), 084002. doi:10.1088/1748-9326/10/8/084002

Daubechies, I. (1992). *Ten lectures on wavelets*. Philadelphia, Pennsylvania: Society for Industrial and Applied mathematics.

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

Efstathiou, M. N., Tzanis, C., Cracknell, A. P., & Varotsos, C. A. (2011). New features of land and sea surface temperature anomalies. *International Journal of Remote Sensing, 32*(11), 3231-3238. doi:10.1080/01431161.2010.541504

Fraedrich, K., & Blender, R. (2003). Scaling of atmosphere and ocean temperature correlations in observations and climate models. *Phys Rev Lett, 90*(10), 108501. doi:10.1103/PhysRevLett.90.108501

Frankignoul, C., & Hasselmann, K. (1977). Stochastic climate models, part 2: Application to sea-surface temperature anomalies and thermocline variability. *Tellus, 29*(4), 289-305.

Franzke, C. (2012). Nonlinear trends, long-range dependence, and climate noise properties of surface temperature. *Journal of Climate, 25*(12), 4172-4183. doi:10.1175/jcli-d-11-00293.1

Gallagher, C., Lund, R., & Robbins, M. (2013). Changepoint detection in climate time series with long-term trends. *Journal of Climate, 26*(14), 4994-5006. doi:10.1175/jcli-d-12-00704.1

Giraitis, L., Kokoska, P., Leipus, R., & Teyssière, G. (2003). Rescaled variance and related tests for long memory in volatility and levels. *Journal of Econometrics, 112*, 265-294.

Granger, C. W. J., & Ding, Z. (1996). Varieties of long memory models. *Journal of Econometrics, 73*, 61-77.

Granger, C. W. J., & Hyung, N. (2004). Occasional structural breaks and long memory with an application to the s&p 500 absolute stock returns. *Journal of Empirical Finance, 11*(3), 399-421. doi:10.1016/j.jempfin.2003.03.001

Granger, C. W. J., & Joyeux, R. (1980). An introduction to long-memory time series models and fractional differencing. *Journal of Time Series Analysis, 1*(1), 15-29.

Hartmann, D. L., Tank, A. M. G. K., Rusticucci, M., Alexander, L. V., Brönnimann, S., Charabi, Y. A.-R., . . . Zhai, P. (2013). Observations: Atmosphere and surface. In T. F. Stocker, D. Qin, G.-K. Plattner, M. M.B.Tignor, S. K. Allen, J. Boschung, A. Nauels, Y. Xia, V. Bex, & P. M. Midgley (Eds.), *Climate change 2013: The physical science basis. Contribution of working group 1 to the fifth assessment report of the intergovernmental panel on climate change*. Cambridge: Cambridge University Press.

Hasselmann, K. (1976). Stochastic climate models part 1. Theory. *Tellus, 28*(6), 473-485. doi:10.1111/j.2153-3490.1976.tb00696.x

Horvath, L. (2001). Change-point detection in long-memory processes. *Journal of Multivariate Analysis, 78*, 218-234. doi:10.1006/jmva.2000.1947

Hosking, J. R. M. (1981). Fractional differencing. *Biometrika, 68*, 165-176.

Huang, B., Banzon, V. F., Freeman, E., Lawrimore, J., Liu, W., Peterson, T. C., . . . Zhang, H.-M. (2015). Extended reconstructed sea surface temperature version 4 (ersst.V4). Part 1: Upgrades and intercomparisons. *Journal of Climate, 28*(3), 911-930. doi:10.1175/jcli-d-14-00006.1

Hurst, H. E. (1951). Long-term storage capacity of reservoirs. *Transactions of the American Society of Civil Engineers, 116*, 770-799.

Huybers, P., & Curry, W. (2006). Links between annual, milankovitch and continuum temperature variability. *Nature, 441*(7091), 329-332. doi:10.1038/nature04745

Imbers, J., Lopez, A., Huntingford, C., & Allen, M. (2014). Sensitivity of climate change detection and attribution to the characterization of internal climate variability. *Journal of Climate, 27*(10), 3477-3491. doi:10.1175/jcli-d-12-00622.1

Jandhyala, V. K., Liu, P., Fotopoulos, S. B., & MacNeill, I. B. (2014). Change-point analysis of polar zone radiosonde temperature data. *Journal of Applied Meteorology and Climatology, 53*(3), 694-714. doi:10.1175/jamc-d-13-084.1

Jones, P. D., Lister, D. H., Osborn, T. J., Harpham, C., Salmon, M., & Morice, C. P. (2012). Hemispheric and large-scale land-surface air temperature variations: An extensive revision and an update to 2010. *Journal of Geophysical Research: Atmospheres, 117*(D5), D05127. doi:10.1029/2011jd017139

Karl, T. R., Knight, R. W., & Baker, B. (2000). The record breaking global temperatures of 1997 and 1998: Evidence for an increase in the rate of global warming? *Geophysical Research Letters, 27*(5), 719-722.

Kennedy, J. J., Rayner, N. A., Smith, R. O., Parker, D. E., & Saunby, M. (2011a). 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*(D14). doi:10.1029/2010jd015218

Kennedy, J. J., Rayner, N. A., Smith, R. O., Parker, D. E., & Saunby, M. (2011b). Reassessing biases and other uncertainties in sea surface temperature observations measured in situ since 1850: 2. Biases and homogenization. *Journal of Geophysical Research, 116*(D14). doi:10.1029/2010jd015220

Khapalova, E. A., Jandhyala, V. K., Fotopoulos, S. B., & Overland, J. E. (2018). Assessing change-points in surface air temperature over alaska. *Frontiers in Environmental Science, 6*. doi:10.3389/fenvs.2018.00121

Killick, R., Beaulieu, C., Taylor, S., & Hullait, H. (2018). Envcpt: Detection of structural changes in climate and environment time series (Version version 1.1.1). CRAN. Retrieved from https://CRAN.R-project.org/package=EnvCpt

Killick, R., Fearnhead, P., & Eckley, I. A. (2012). Optimal detection of changepoints with a linear computational cost. *Journal of the American Statistical Association, 107*(500), 1590-1598. doi:10.1080/01621459.2012.737745

Koscielny-Bunde, E., Bunde, A., Havlin, S., Roman, H. E., Goldreich, Y., & Schellnhuber, H.-J. (1998). Indication of a universal persistence law governing atmospheric variability. *Physical Review Letters, 81*(3), 729-732.

Lennartz, S., & Bunde, A. (2009). Trend evaluation in records with long-term memory: Application to global warming. *Geophysical Research Letters, 36*(16). doi:10.1029/2009gl039516

Liu, W., Huang, B., Thorne, P. W., Banzon, V. F., Zhang, H.-M., Freeman, E., . . . Woodruff, S. D. (2015). Extended reconstructed sea surface temperature version 4 (ersst.V4): Part 2. Parametric and structural uncertainty estimations. *Journal of Climate, 28*(3), 931-951. doi:10.1175/jcli-d-14-00007.1

Ludescher, J., Bunde, A., & Schellnhuber, H. J. (2017). Statistical significance of seasonal warming/cooling trends. *Proc Natl Acad Sci U S A, 114*(15), E2998-E3003. doi:10.1073/pnas.1700838114

Manabe, S., & Stouffer, R. J. (1996). Low-frequency variability of surface air temperature in a 1000-year integration of a coupled atmosphere-ocean-land surface model. *Journal of Climate, 9*, 376-393.

Mandelbrot, B. (1967). Some noises with 1/f spectrum, a bridge between direct current and white noise. *IEEE Transactions on Information Theory, 13*(2), 289-298.

Mann, M. E., & Lees, J. M. (1996). Robust estimation of background noise and signal detection in climatic time-series. *Climatic Change, 33*, 409-445.

Mills, T. C. (2007). Time series modeling of two millennia of northern hemisphere temperatures: Long memory or shifting trends? *Journal of the Royal Statistical Society A, 170*, 83-94.

Mustin, K., Dytham, C., Benton, T. G., Travis, J. M. J., & Watson, J. (2013). Red noise increases extinction risk during rapid climate change. *Diversity and Distributions, 19*(7), 815-824. doi:10.1111/ddi.12038

Norwood, B., & Killick, R. (2018). Long memory and changepoint models: A spectral classification procedure. *Statistics and Computing, 28*, 291-302. doi:10.1007/s11222-017-9731-0

Pelletier, J. D. (1997). Analysis and modeling of the natural variability of climate. *Journal of Climate, 10*, 1331-1342.

Rahmstorf, S., Foster, G., & Cahill, N. (2017). Global temperature evolution: Recent trends and some pitfalls. *Environmental Research Letters, 12*(5), 054001. doi:10.1088/1748-9326/aa6825

Rea, W., Reale, M., & Brown, J. (2011). Long memory in temperature reconstructions. *Climatic Change, 107*(3-4), 247-265. doi:10.1007/s10584-011-0068-y

Ruggieri, E. (2012). A bayesian approach to detecting change points in climatic records. *International Journal of Climatology, 33*(2), 520-528. doi:10.1002/joc.3447

Rybski, D., Bunde, A., Havlin, S., & von Storch, H. (2006). Long-term persistence in climate and the detection problem. *Geophysical Research Letters, 33*(6). doi:10.1029/2005gl025591

Rypdal, K. (2015). Attribution in the presence of a long-memory climate response. *Earth System Dynamics, 6*(2), 719-730. doi:10.5194/esd-6-719-2015

Rypdal, K., Østvand, L., & Rypdal, M. (2013). Long-range memory in earth's surface temperature on time scales from months to centuries. *Journal of Geophysical Research: Atmospheres, 118*(13), 7046-7062. doi:10.1002/jgrd.50399

Santer, B. D., Thorne, P. W., Haimberger, L., Taylor, K. E., Wigley, T. M. L., Lanzante, J. R., . . . Wentz, F. J. (2008). Consistency of modelled and observed temperature trends in the tropical troposphere. *International Journal of Climatology, 28*(13), 1703-1722. doi:10.1002/joc.1756

Seidel, D. J., & Lanzante, J. R. (2004). An assessment of three alternatives to linear trends for characterizing global atmospheric temperature changes. *Journal of Geophysical Research, 109*, D14108. doi:10.1029/2003jd004414

Smith. (2005). Level shifts and the illusion of long memory in economic time series. *Journal of Business & Economic Statistics, 23*(3), 321-335. doi:10.1198/073500104000000280

Smith, Reynolds, R. W., Peterson, T. R., & Lawrimore, J. H. (2008). Improvements to noaa's historical merged and–ocean surface temperature analysis (1880–2006). *Journal of Climate, 21*, 2283-2296. doi:10.1175/2007JCLI2100.1

Tang, S. M., & MacNeill, I. B. (1989). The effect of autocorrelated errors on change-detection statistics. *Environmental Monitoring and Assessment, 12*, 203-226.

Tang, S. M., & MacNeill, I. B. (1993). The effect of serial correlation on tests for parameter change at unknown time. *The Annals of Statistics, 21*(1), 552-575.

Trenberth, K. E., Jones, P. D., Ambenje, P., Bojariu, R., Easterling, D. R., Klein Tank, A., . . . Zhai, P. (2007). Observations: Surface and atmospheric climate change. In S. Solomon, D. Qin, M. Manning, Z. Chen, M. Marquis, K. B. Averyt, M. Tignor, & H. L. Miller (Eds.), *Climate change 2007: The physical science basis. Contribution of working group 1 to the fourth assessment report of the intergovernmental panel on climate change*. Cambridge: Cambridge University Press.

Varotsos, C., & Kirk-Davidoff, D. (2006). Long-memory processes in ozone and temperature variations at the region 60s-60n. *Atmospheric Chemistry and Physics, 6*, 4093-4100.

Veenstra, J. Q. (2013). *Persistence and anti-persistence: Theory and software.* (PhD), The University of Western Ontario.

von Storch, H. (1999). Misuses of statistical analysis in climate research. In H. von Storch & A. Navarra (Eds.), *Analysis of climate variability* (pp. 11-26). Heidelberg, Germany: Springer.

von Storch, H., & Zwiers, F. W. (1999). *Statistical analysis in climate research*. Cambridge: Cambridge University Press.

Vose, R. S., Arndt, D., Banzon, V. F., Easterling, D. R., Gleason, B., Huang, B., . . . Wuertz, D. B. (2012). Noaa's merged land-ocean surface temperature analysis. *Bulletin of the American Meteorological Society*, 1677-1685. doi:10.1175/BAMS-D-11-00241.1

Vyushin, D. I., Kushner, P. J., & Zwiers, F. (2012). Modeling and understanding persistence of climate variability. *Journal of Geophysical Research: Atmospheres, 117*, D21106. doi:10.1029/2012jd018240

Yau, C. Y., & Davis, R. A. (2012). Likelihood inference for discriminating between long-memory and change-point models. *Journal of Time Series Analysis, 33*, 649-664. doi:10.1111/j.1467-9892.2012.00797.x

Yuan, N., Ding, M., Huang, Y., Fu, Z., Xoplaki, E., & Luterbacher, J. (2015). On the long-term climate memory in the surface air temperature records over antarctica: A nonnegligible factor for trend evaluation. *Journal of Climate, 28*(15), 5922-5934. doi:10.1175/jcli-d-14-00733.1

Yuan, N., Fu, Z., & Liu, S. (2013). Long-term memory in climate variability: A new look based on fractional integral techniques. *Journal of Geophysical Research: Atmospheres, 118*(23), 12,962-912,969. doi:10.1002/2013jd020776

Zhang, N. R., & Siegmund, D. O. (2007). A modified bayes information criterion with applications to the analysis of comparative genomic hybridization data. *Biometrics, 63*, 22-32. doi:0.1111/j.1541-0420.2006.00662.x

Tables

Table 1: List of parameters used to simulate the sets of synthetic series.

Variable	Scenario	Model	Parameters
HadCRUT4 GMST (N=2016)	A	Trend cpt + AR(1)	$\lambda_1 = 0.333$, $\lambda_2 = 2.669$, $\lambda_3 = -8.256$, $\lambda_4 = -5.962$, $\beta_1 = -0.000265$, $\beta_2 = -0.00144$, $\beta_3 = 0.00426$, $\beta_4 = 0.00302$, $\varphi_1 = 0.306$, $\varphi_2 = 0.753$,$\varphi_3 = 0.521$, $\varphi_4 = 0.776$, $c_1 = 329(0.163N)$, $c_2 = 806(0.4N)$, $c_3 = 1260(0.625N)$, $m = 4$, $\sigma_1^2 = 0.0302$, $\sigma_2^2 = 0.0123$, $\sigma_3^2 = 0.0130$, $\sigma_4^2 = 0.00907$
	B	Trend + LM	$\lambda = -10.405$, $\beta = 0.00541$, $\sigma = 0.0144$, $d = 0.485$

Table 2: Percentage of synthetic series that detect at least one changepoint over 1000 replications for different sample sizes (N) when the truth is a long memory model.

Scenario	Number of observations expressed in months (years)								
	600 (50y)	700 (58y)	800 (67y)	1000 (83y)	1200 (100y)	1400 (117y)	1600 (133y)	1800 (150y)	2016 (168y)
Trend + LM	70.9	74.4	84.2	91.2	94.7	95.2	97.2	97.9	99.6

Figures

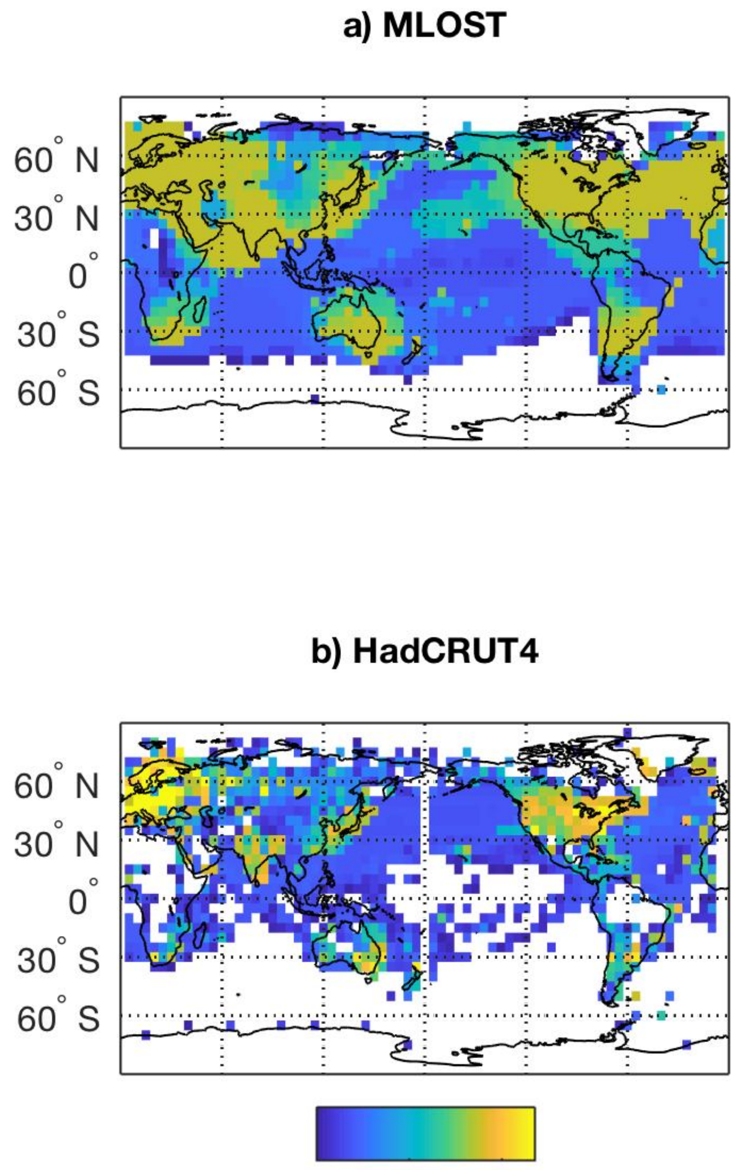

Figure 1: Number of contiguous observations used in each grid cell for two surface temperature datasets a) MLOST and b) HadCRUT4. Grids with an insufficient number of observations (<600) to perform the classification are left blank.

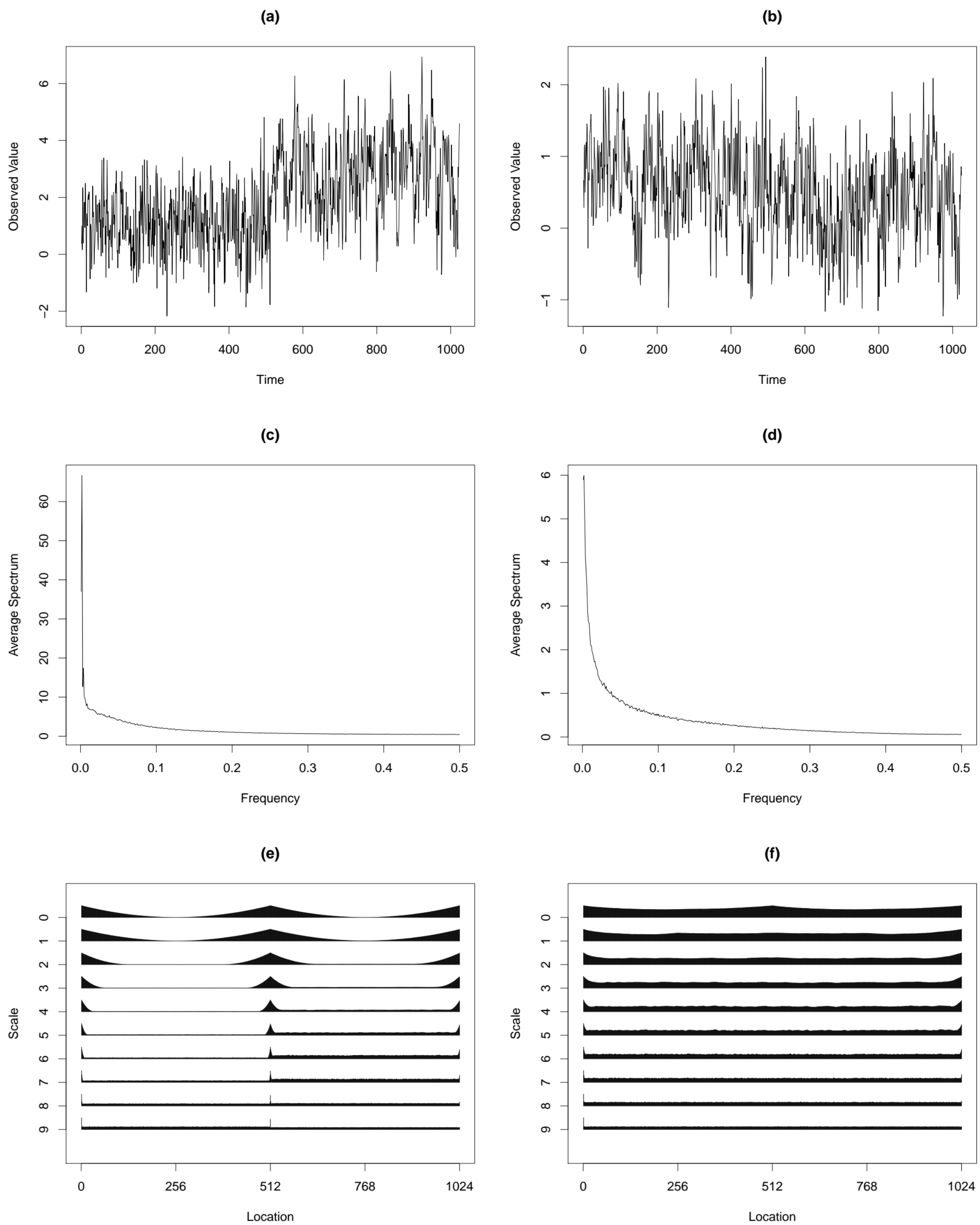

Figure 2: Examples of time series generated a) from a trend changepoint model with AR(1) errors and b) long-memory, their respective average spectrum in c) and d), and the corresponding time-varying spectrum in e) and f).

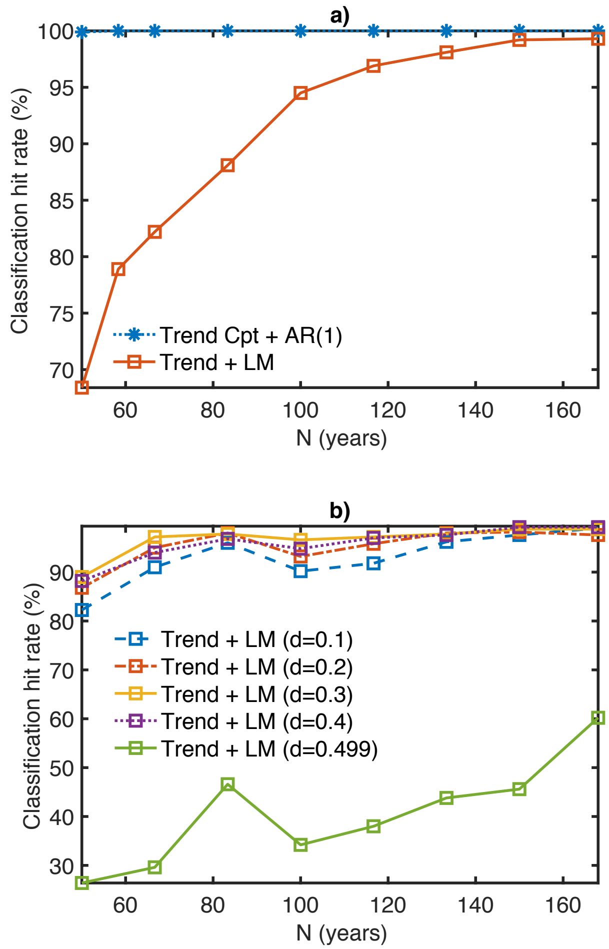

Figure 3: Results of the simulation study for the two scenarios and for time-series with a varying number of years (*N*). (a) For each scenario, the percentage of series classified correctly as either trend changepoint and short-memory (Cpt) or trend and long-memory (LM) is presented taken over 1000 replications. (b) For the scenario with long-memory, the experiment is repeated with varying strengths for the long-memory parameter from low (d= 0.1) to high (d=0.499).

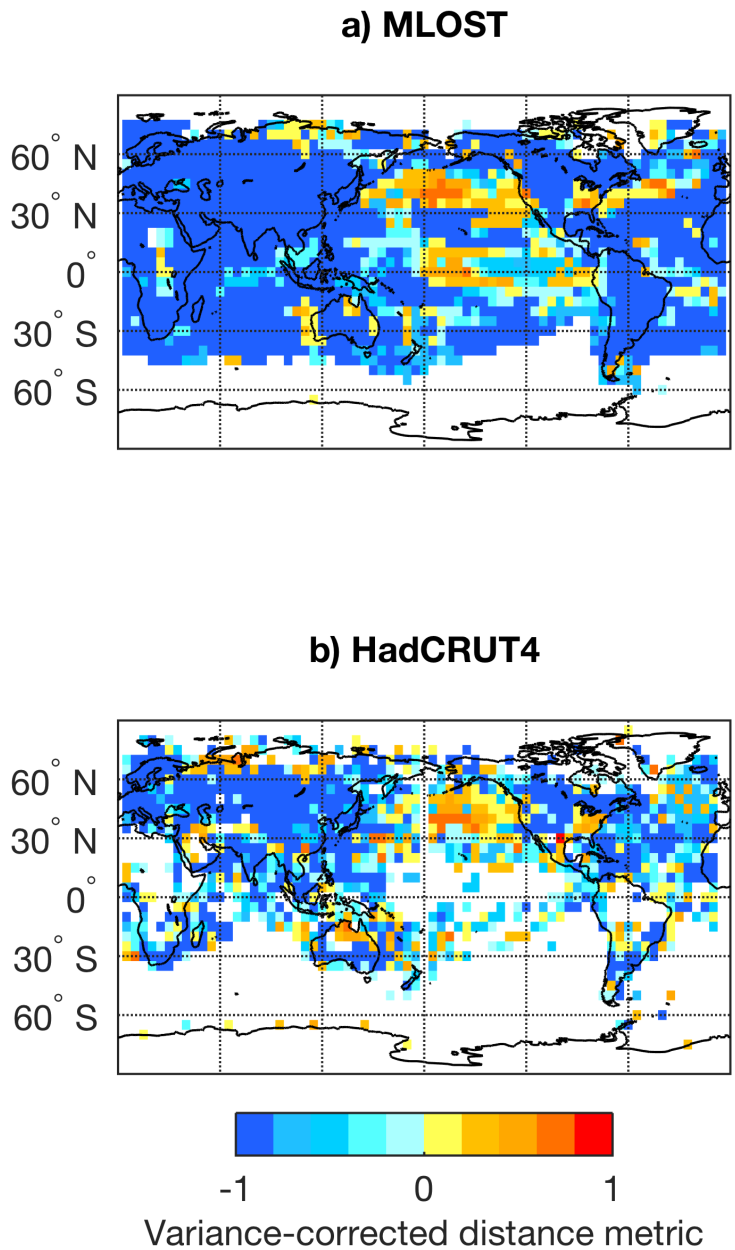

Figure 4: Comparison between changepoint models with short-term memory vs trend with long-term memory for two surface temperature datasets a) MLOST and b) HadCRUT4. The colorbar represents the variance-corrected distance metric presented in Eq. 12, which represents the strength of evidence for the chosen model: negative values indicate evidence for a change-point model (Cpt) while positive values indicate long-memory (LM). Grids with insufficient data to perform the classification are left blank.

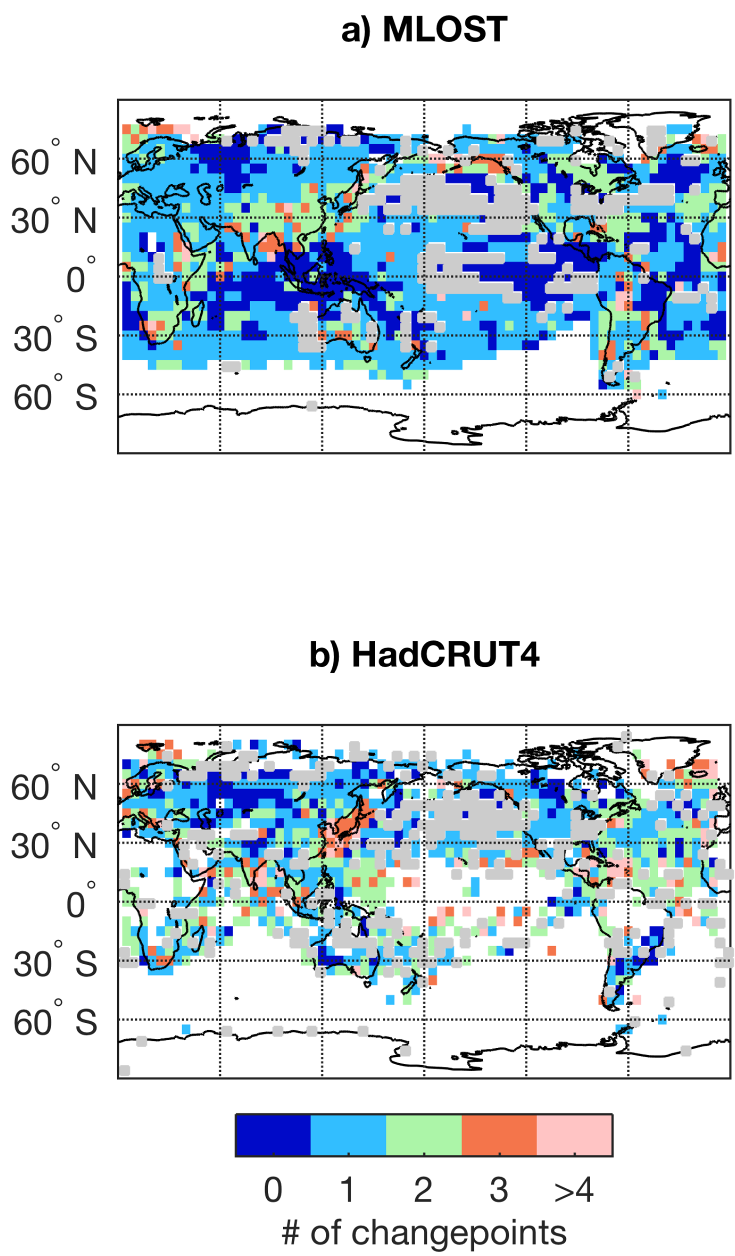

Figure 5: Number of changepoints detected for grid cells where a changepoint model with short-term memory model is more likely than a trend with long-term memory for two surface temperature datasets a) MLOST and b) HadCRUT4. The grey areas indicate grid cells where a long-memory model was preferred. Grids with insufficient data to perform the classification are left blank.

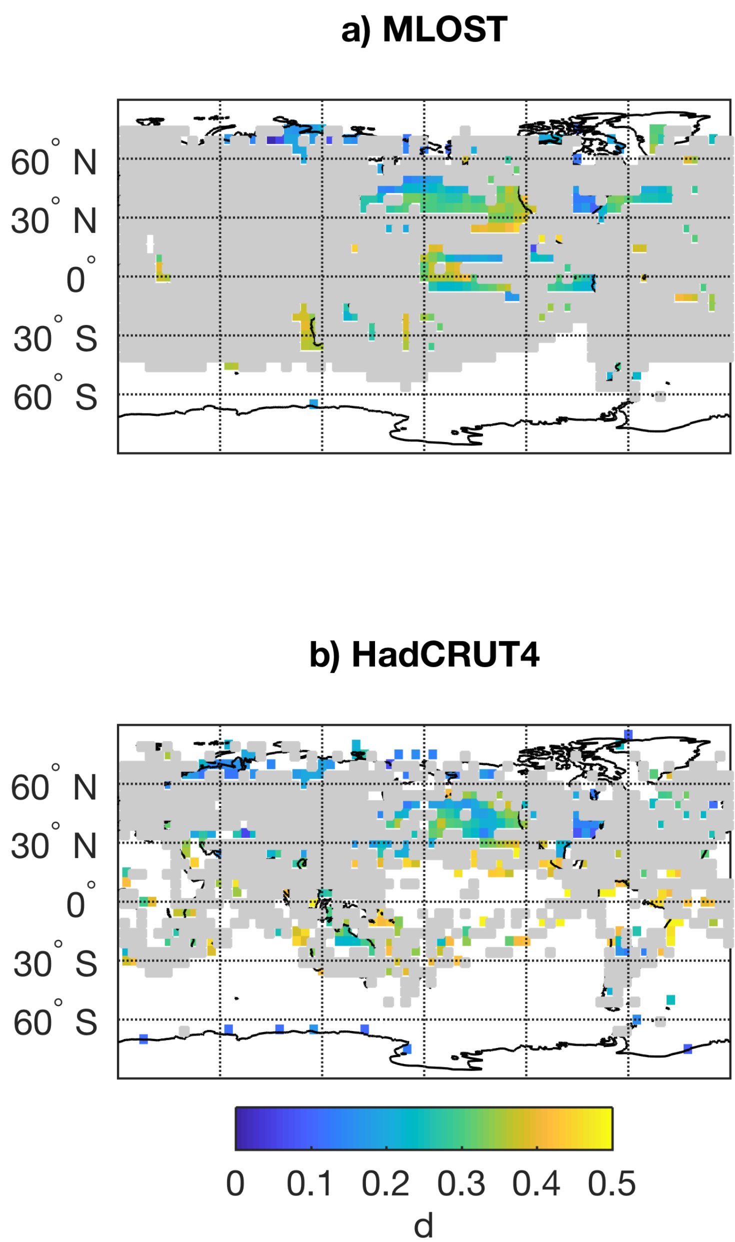

Figure 6: Strength of the memory (given by parameter d) for grid cells where a trend with long-memory is more likely than a changepoint model with short-term memory for two surface temperature datasets a) MLOST and b) HadCRUT4. The grey areas indicate grid cells where a change-point model is more likely. Grids with insufficient data to perform the classification are left blank.