1 Simulations

1.1 Size and power simulations

To assess the finite sample performance of our method, we conduct a number of simulations. In particular, we investigate the test procedure from Section **??**. The simulation design is set up to mimic the situation in the application example of Section 2.1: We generate data from the model $Y_t = m(\frac{t}{T}) + \varepsilon_t$ for different time series lengths T. The errors ε_t are drawn from the AR(1) process $\varepsilon_t = a\varepsilon_{t-1} + \eta_t$, where η_t are independent and normally distributed with mean 0 and variance σ_η^2. We set $a = 0.267$ and $\sigma_\eta^2 = 0.35$, thus matching the estimated values obtained in the application of Section 2.1. To simulate data under the null $H_0 : m' = 0$, we let m be a constant function. In particular, we set $m = 0$ without loss of generality. To generate data under the alternative, we consider the trend functions $m(u) = \beta\,(u-0.6)\,1(0.6 \le u \le 1)$ with $\beta = 1.25, 1.875, 2.5$. These functions are broken lines with a kink at $u = 0.6$ and different slopes β. The slope parameter β corresponds to a trend with the value $m(1) = 0.4\,\beta$ at the right endpoint $u = 1$. We thus consider broken lines with the values $m(1) = 0.5, 0.75, 1.0$. Inspecting the middle panel of Figure 2, the broken line with the slope $\beta = 2.5$ can be seen to resemble the local linear trend estimates in the real-data example of Section 2.1 the most (where we neglect the nonlinearities of the local linear fits at the beginning of the observation period). The broken lines with the smaller slopes $\beta = 1.25$ and $\beta = 1.875$ are closer to the null making it harder for our test to detect these alternatives.
To implement our test, we choose K to be an Epanechnikov kernel and define the set $\mathcal{G}_T$ of location-scale points (u, h) as

$$\begin{aligned}\mathcal{G}_T = \big\{(u,h) : u = 5k/T \text{ for some } 1 \le k \le T/5 \text{ and} \\ h = (3+5\ell)/T \text{ for some } 0 \le \ell \le T/20\big\}.\end{aligned} \tag{1}$$

We thus take into account all rescaled time points $u \in [0, 1]$ on an equidistant grid with step length $5/T$. For the bandwidth $h = (3+5\ell)/T$ and any $u \in [h, 1-h]$, the local linear weights $w'_{t,T}(u, h)$ are non-zero for exactly $5 + 10\ell$ observations. Hence, the bandwidths h in $\mathcal{G}_T$ correspond to effective sample sizes of $5, 15, 25, \ldots$ up to approximately $T/2$ data points. We estimate the long-run error variance σ^2 by the procedure from Section **??**, setting the tuning parameters L_1 and L_2 to $\lfloor\sqrt{T}\rfloor$ and $\lfloor 2\sqrt{T}\rfloor$, respectively. To compute the critical values of the test, we simulate 1000 values of the statistic Φ'_T defined in Section **??** and compute their empirical $(1-\alpha)$ quantile $q'_T(\alpha)$.
Tables 1 and 2 report the simulation results for the sample sizes $T = 250, 350, 500, 1000$ and the significance levels $\alpha = 0.01, 0.05, 0.10$. The sample size $T = 350$ is approximately equal to the time series length 359 in the real-data example of Section 2.1. To produce our simulation results, we generate $S = 1000$ samples for each time series length T and carry out the multiscale test for each simulated sample. The entries of Tables 1 and 2 are computed as the number of simulations in which the test rejects divided by the total number of simulations. As can be seen from Table 1, the actual size of the test is fairly close to the nominal target α even for small values of T. Hence, the test has approximately the correct size. Inspecting Table 2, one can further see that the test has reasonable power properties. For the smallest value $\beta = 1.25$, the deviation from the null is quite small, making it hard for the test to detect the alternative. As a consequence, the power is only moderate for $T = 250$ and $T = 350$. When we move

Table 1: Size of the multiscale test from Section **??** for different sample sizes T and nominal sizes α.

	nominal size α		
T	0.01	0.05	0.1
250	0.004	0.039	0.092
350	0.012	0.051	0.069
500	0.006	0.047	0.094
1000	0.014	0.058	0.105

Table 2: Power of the multiscale test from Section **??** for different sample sizes T and nominal sizes α. Each panel corresponds to a different slope parameter β.

(a) $\beta = 1.25$

	nominal size α		
T	0.01	0.05	0.1
250	0.085	0.252	0.341
350	0.236	0.396	0.470
500	0.315	0.577	0.669
1000	0.763	0.900	0.936

(b) $\beta = 1.875$

	nominal size α		
T	0.01	0.05	0.1
250	0.318	0.621	0.714
350	0.648	0.796	0.865
500	0.793	0.943	0.967
1000	0.997	1.000	1.000

(c) $\beta = 2.5$

	nominal size α		
T	0.01	0.05	0.1
250	0.693	0.898	0.937
350	0.929	0.981	0.990
500	0.986	1.000	1.000
1000	1.000	1.000	1.000

further away from the null by increasing the slope parameter β, the power of the test quickly increases. It can also be seen to rapidly get larger as the sample size grows. For the slope $\beta = 2.5$ and the sample size $T = 350$, which are the values that resemble the real-life data in Section 2.1 the most, the power of the test is above 92.9% for all significance levels α considered and thus comes quite close to 1.

1.2 Comparison to SiZer

In this section, we compare the perfomance of the multiscale method and the SiZer for times series as described in Rondonotti et al. (2007).
First, we assess the finite sample properties of both methods by computing the size and the power from 1000 simulated time series.
In order to make comparison between methods more illustrative, we consider the simplest possible scenario: the observations $Y_t, t \in \{1, \dots, T\}$ are drawn from a model $Y_t = m(\frac{t}{T}) + \varepsilon_t$, where $\{\varepsilon_t\}$ is a stationary and causal AR(1) process

$$\varepsilon_t = a_1 \varepsilon_{t-1} + \eta_t$$

with $|a_1| < 1$. Here η_t are i.i.d standard normal innovations with zero mean and finite variance $\sigma_\eta^2 = 1$. To simulate the data under the null $H_0 : m' = 0$, without loss of generality we set $m = 0$. To generate data under the alternative, we consder the linear trend functions $m(u) = \beta(u - 0.5)$ with $\beta = 3.5, 4.0, 4.5, 5.0$. The smaller slopes $\beta = 3.5$ and $\beta = 4.0$ correspond to the data close to the null, whereas the larger slopes $\beta = 4.5$ and $\beta = 5.0$ result in the data points to be sufficiently different from the null, thus making it easier for both tests to detect the alternative. It is also worth noting that the slope $\beta = 4.0$ leads to the end point $m(1) = 2.0$ being only two times as large as the variance of the innovation η_t.

Both the SiZer testing procedure of Rondonotti et al. (2007) and our multiscale method are applied to find intervals where $m(\cdot)$ is either increasing or decreasing. The result of SiZer is the SiZer map, the graphical representation of the multiple testing procedure. Each pixel of this map corresponds to a particular location $u = t/T$ and a particular bandwidth h, and the color of this pixel reflects statistical significance of the slope at (u, h) obtained by the following procedure: at each point (u, h) the trend is significantly increasing (or decreasing) if the confidence intervals for the derivative of the underlying function is above (or below) 0. If the respective confidence interval contains 0, no statistical inference about this point (u, h) can be made. The exact algorithm for SiZer is provided in the Supplement.

As a preliminary grid we use the same set $\mathcal{G}_T$ of location-scale points (u, h) as in Section 1.2:

$$\mathcal{G}_T = \big\{(u,h) : u = 5k/T \text{ for some } 1 \le k \le T/5 \text{ and } h = (3+5\ell)/T \text{ for some } 0 \le \ell \le T/20\big\}. \tag{2}$$

However, we further restrict our attention only to those locations and bandwidths where the number of "independent blocks", as cited in Rondonotti et al. (2007), is not too sparse for reasonable statistical inference. In particular, we calculate ESS^*, the Effective Sample Size for correlated data, and consider only the following set:

$$\mathcal{G}_T^* = \big\{(u,h) \in \mathcal{G}_T | ESS^*(u,h) \ge 5\big\} \subset \mathcal{G}_T. \tag{3}$$

Detailed discussion of "independent blocks" and ESS^* can be found in Chaudhuri and Marron (1999) and Rondonotti et al. (2007).

As the kernel function K for both methods we use the Epanechnikov kernel. Instead of estimating the autocovariances $\gamma(k)$ for SiZer and the long-run error variance σ^2 for the multiscale test from the data, we use the true theoretical values: $\widehat{\gamma}(k) = \gamma(k) = \frac{\sigma_\eta^2}{1-a_1^2} a_1^{|k|}$ and $\widehat{\sigma}^2 = \sigma^2 = \frac{\sigma_\eta^2}{(1-a_1)^2}$. Furthermore, for each simulation the Gaussian quantile for the multiscale method $q(1-\alpha)$ is obtained with 1000 simulated Gaussian versions of the test statistic Φ_T.

To produce our simulation results, we generate $S = 1000$ samples for each time series length T and carry out the multiscale test and SiZer test for each simulated sample. As before, for the multiscale test the entries of Tables **??** and **??** are computed as the number of simulations in which the test rejects divided by the total number of simulations. For SiZer, the entries are calculated as the number of simulations in which it detects at least one pair location-bandwidth (u, h) with statistically significant increase or decrease divided by the total number of simulations.

As in Section 1.1, we can see from Table **??**, the multiscale test has approximately correct size while SiZer finds statistical significance under the null in roughly 20% of the cases. This observation suggests that SiZer is more liberal, identifying all potential points of increase and decrease for further analysis, whereas our test is more persistent and stable.

As can be expected, the significant advantage of the multiscale method in terms of size come at a cost in terms of power. However, the loss in power is not very pronounced for sufficiently big sample sizes T. Examining Table **??**, we can see that...

As noted in Hannig and Marron (2006), the classic definition of the size of the test that was presented above may not be applicable for the purpose of assessing finite sample

properties of SiZer. In particular, some of the SiZer maps may have a small number of pixels where the method suggests the slope is statistically significant, when, in fact, the underlying trend function is flat. That is the reason for some other comparison between SiZer and our multiscale method.
The most reasonable way to compare these two procedures is to see what exactly are the regions where they detect some significant increase or decrease. However, the linear trend function is not appropriate here since it increases on the whole support $[0, 1]$. One of the alternatives is to choose a sufficiently smooth function that is zero on the bigger part of the interval $[0, 1]$ and non-zero on the smaller part of it. One of the possibilities would be true underlying trend function defined as $m(u) = 2\big(1 - 100(u - 0.5)^2\big)^2$ for $u \in [0.4, 0.6]$ and $m(u) = 0$ for $u \notin [0.4, 0.6]$. Here the trend function reaches its maximum at $u = 0.5$ with $m(0.5) = 2$ which is again only two times larger than the variance of η_t.
The significance level is chosen as $\alpha = 0.05$, and We let $T = 500$ and the .
From Figure 1 we can infer that see that our multiscale procedure works much better in the case of negative coefficient a_1 and has relatively same efficiency for positive coefficients a_1.

2 Applications

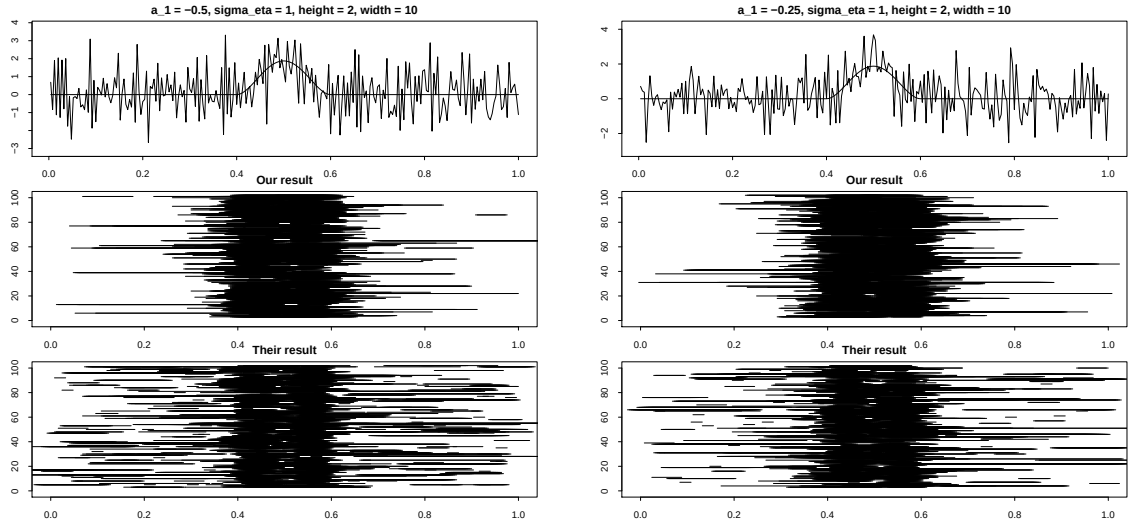
In what follows, we illustrate the multiscale methods from Sections **??** and **??** by two real-data examples. In the first example, we apply the test method from Section **??** to a long time series of temperature data from Central England. In the second, we analyse a sample of temperature time series from 34 different weather stations in Great Britain with the help of the methods from Section **??**.

2.1 Analysis of Central England temperature data

The analysis of time trends in long temperature records is an important task in climatology. Information on the shape of the trend is needed in order to better understand long-term climate variability. The Central England temperature record is the longest instrumental temperature time series in the world. It is a valuable asset for analysing climate variability over the last few hundred years. The data is publicly available on the webpage of the UK Met Office. A detailed description of the data can be found in Parker et al. (1992). For our analysis, we use the dataset of yearly mean temperatures which consists of $T = 359$ observations covering the years from 1659 to 2017. We assume that the data follow the nonparametric trend model

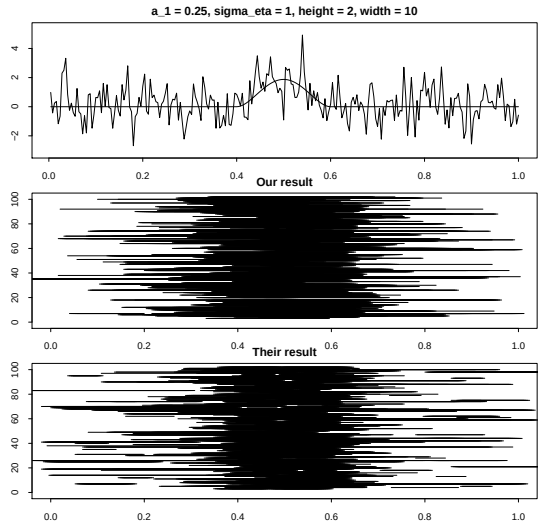
$$Y_t = m\Big(\frac{t}{T}\Big) + \varepsilon_t,$$

where m is the unknown time trend of interest. The error process $\{\varepsilon_t\}$ is supposed to have the AR(1) structure $\varepsilon_t = a\varepsilon_{t-1} + \eta_t$, where η_t are i.i.d. innovations with mean 0 and variance σ_η^2. As pointed out in Mudelsee (2010) among others, this is the most widely used error model for discrete climate time series. We estimate the parameters a and σ_η^2 as described in Section **??** which yields the estimates $\widehat{a} \approx 0.267$ and $\widehat{\sigma}_\eta^2 \approx 0.35$. With the help of our multiscale method from Section **??**, we test the null hypothesis $H_0 : m' = 0$, that is, the hypothesis that m is constant. To do so, we set the significance

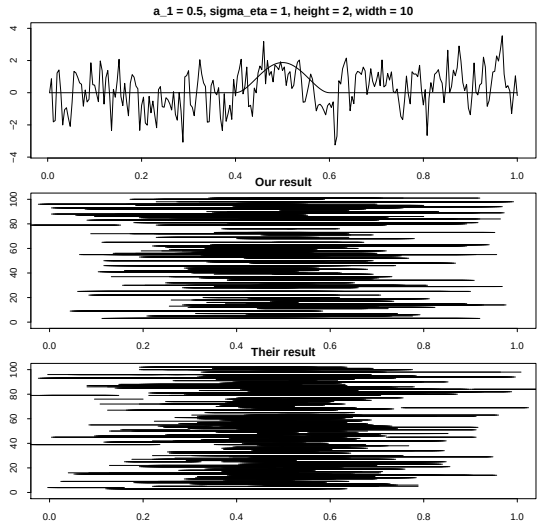


(a) $a_1 = -0.5$

(b) $a_1 = -0.25$



(c) $a_1 = 0.25$



(d) $a_1 = 0.5$

Figure 1: Summary of the simulation results for Section 1.2 for different model settings. In these settings the underlying trend function is $m(x) = 2\big(1 - 100(x - 0.5)^2\big)^2$ for $x \in [0.4, 0.6]$ and $m(x) = 0$ for $x \notin [0.4, 0.6]$. The errors are generated from an AR(1) process with coefficients $a_1 \in \{-0.5, -0.25, 0.25, 0.5\}$. Each of the figures correspond to one of the coefficients a_1 which are provided below the figures. The upper panel on each figure shows an example of the time series generated by the model. The middle panel depicts the union of the minimal intervals in the set Π_T produced by the multiscale test for each of the 100 runs. The lower panel presents the regions where SiZer detected increase or decrease of the underlying trend function for each of the 100 runs.

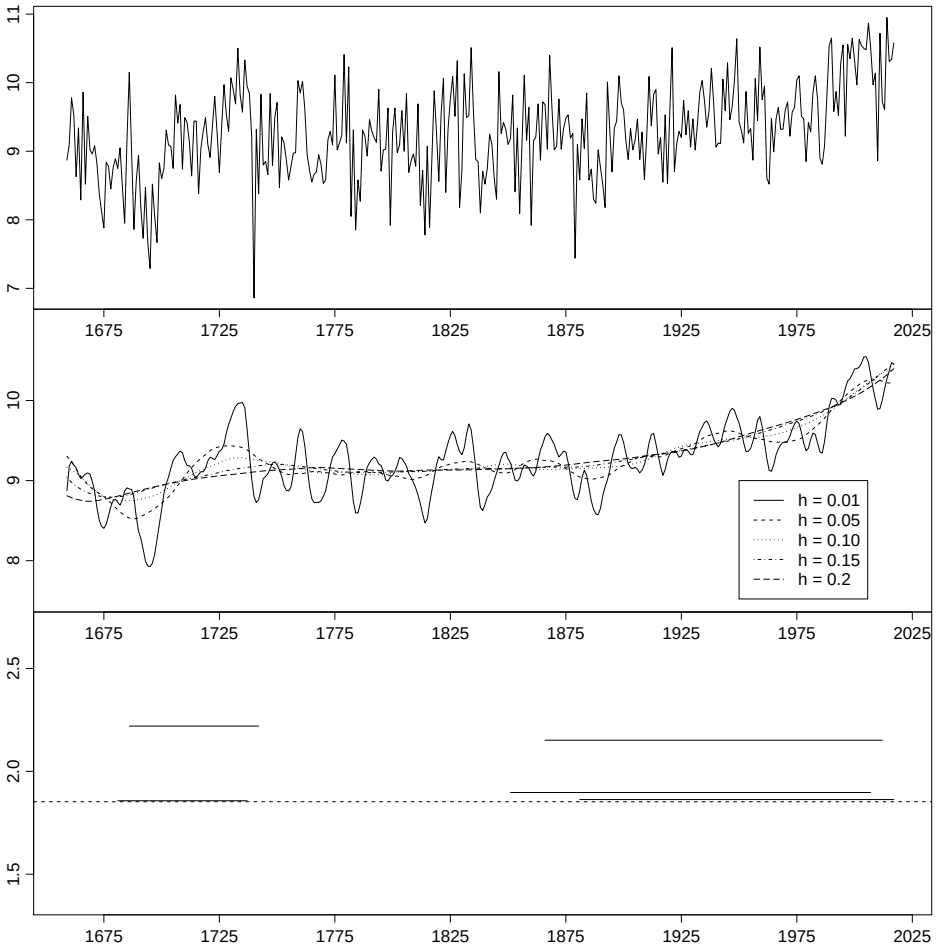


Figure 2: Summary of the application results for Section 2.1. The upper panel shows the Central England mean temperature time series. The middle panel depicts local linear kernel estimates of the time trend for a number of different bandwidths h. The lower panel presents the minimal intervals in the set Π_T^+ produced by the multiscale test. These are $[1681, 1737]$, $[1686, 1742]$, $[1851, 2007]$, $[1866, 2012]$ and $[1881, 2017]$.

level to $\alpha = 0.05$ and implement the test in exactly the same way as in the simulations of Section 1. The results are presented in Figure 2. The upper panel shows the raw temperature time series, whereas the middle panel depicts local linear kernel estimates of the trend m for different bandwidths h. As one can see, the shape of the estimated time trend strongly differs with the chosen bandwidth. When the bandwidth is small, there are many local increases and decreases in the estimated trend. When the bandwidth is large, most of these local variations get smoothed out. Hence, by themselves, the nonparametric fits do not give much information on whether the trend m is increasing or decreasing in certain time regions.

Our multiscale test provides this kind of information, which is summarized in the lower panel of Figure 2. The plot depicts the minimal intervals contained in the set Π_T^+ which is defined in Section **??**. The set of intervals Π_T^- is empty in the present case. The height at which a minimal interval $I_{u,h} = [u-h, u+h] \in \Pi_t^+$ is plotted indicates the value of the corresponding (additively corrected) test statistic $\widehat{\psi}_T'(u,h)/\widehat{\sigma} - \lambda(h)$. The dashed line specifies the critical value $q_T'(\alpha)$, where $\alpha = 0.05$ as already mentioned above. According to Proposition **??**, we can make the following simultaneous confidence statement about

the collection of minimal intervals in Π_T^+. We can claim, with confidence of about 95%, that the trend function m has some increase on each minimal interval. More specifically, we can claim with this confidence that there has been some upward movement in the trend both in the period from around 1680 to 1740 and in the period from about 1880 onwards. Hence, our test in particular provides evidence that there has been some warming trend in the period over approximately the last 140 years. On the other hand, as the set Π_T^- is empty, there is no evidence of any downward movement of the trend.

References

Chaudhuri, P. and Marron, J. S. (1999). SiZer for the exploration of structures in curves. *Journal of the American Statistical Association*, **94** 807–823.

Hannig, J. and Marron, J. S. (2006). Advanced distribution theory for SiZer. *Journal of the American Statistical Association*, **101** 484–499.

Mudelsee, M. (2010). *Climate time series analysis: classical statistical and bootstrap methods*. New York, Springer.

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

Rondonotti, V., Marron, J. S. and Park, C. (2007). SiZer for time series: a new approach to the analysis of trends. *Electronic Journal of Statistics*, **1** 268–289.