Multiscale Inference and Long-Run Variance Estimation in Nonparametric Regression with Time Series Errors

Marina Khismatullina[1] University of Bonn

Michael Vogt[2] University of Bonn

July 2, 2019

In this paper, we develop new multiscale methods to test qualitative hypotheses about the function m in the nonparametric regression model $Y_{t,T} = m(t/T) + \varepsilon_t$ with time series errors ε_t. In time series applications, m represents a nonparametric time trend. Practitioners are often interested in whether the trend m has certain shape properties. For example, they would like to know whether m is constant or whether it is increasing/decreasing in certain time regions. Our multiscale methods allow to test for such shape properties of the trend m. In order to perform the methods, we require an estimator of the long-run error variance $\sigma^2 = \sum_{\ell=-\infty}^{\infty} \text{Cov}(\varepsilon_0, \varepsilon_\ell)$. We propose a new difference-based estimator of σ^2 for the case that $\{\varepsilon_t\}$ belongs to the class of AR(∞) processes. In the technical part of the paper, we derive asymptotic theory for the proposed multiscale test and the estimator of the long-run error variance. The theory is complemented by a simulation study and an empirical application to climate data.

Key words: Multiscale statistics; long-run variance; nonparametric regression; time series errors; shape constraints; strong approximations; anti-concentration bounds.
AMS 2010 subject classifications: 62E20; 62G10; 62G20; 62M10.

1 Introduction

The analysis of time trends is an important aspect of many time series applications. In a wide range of situations, practitioners are particularly interested in certain shape properties of the trend. They raise questions such as the following: Does the observed time series have a trend at all? If so, is the trend increasing/decreasing in certain time regions? Can one identify the regions of increase/decrease? As an example, consider the time series plotted in Figure 1 which shows the yearly mean temperature in Central England from 1659 to 2017. Climatologists are very much interested in learning about

[1]Address: Bonn Graduate School of Economics, University of Bonn, 53113 Bonn, Germany. Email: marina.k@uni-bonn.de.
[2]Corresponding author. Address: Department of Economics and Hausdorff Center for Mathematics, University of Bonn, 53113 Bonn, Germany. Email: michael.vogt@uni-bonn.de.

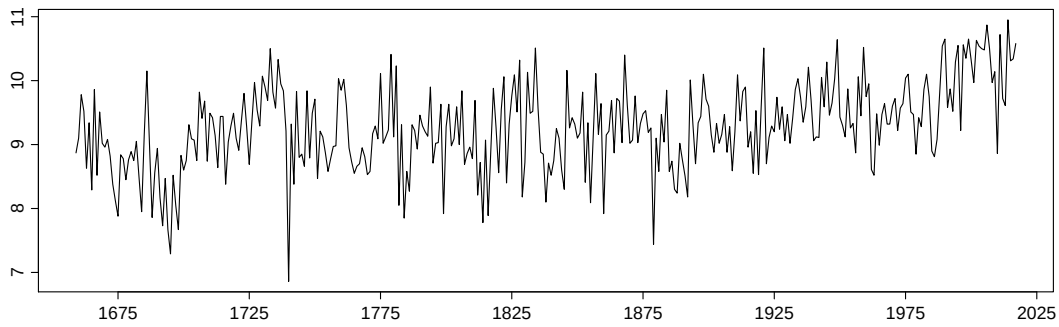


Figure 1: Yearly mean temperature in Central England from 1659 to 2017 measured in °C.

the trending behaviour of temperature time series like this; see e.g. Benner (1999) and Rahmstorf et al. (2017). Among other things, they would like to know whether there is an upward trend in the Central England mean temperature towards the end of the sample as visual inspection might suggest.

In this paper, we develop new methods to test for certain shape properties of a nonparametric time trend. We in particular construct a multiscale test which allows to identify local increases/decreases of the trend function. We develop our test in the context of the following model setting: We observe a time series $\{Y_{t,T} : 1 \leq t \leq T\}$ of the form

$$Y_{t,T} = m\Big(\frac{t}{T}\Big) + \varepsilon_t \tag{1.1}$$

for $1 \leq t \leq T$, where $m : [0,1] \to \mathbb{R}$ is an unknown nonparametric regression function and the error terms ε_t form a stationary time series process with $\mathbb{E}[\varepsilon_t] = 0$. In a time series context, the design points t/T represent the time points of observation and m is a nonparametric time trend. As usual in nonparametric regression, we let the function m depend on rescaled time t/T rather than on real time t. A detailed description of model (1.1) is provided in Section 2.

Our multiscale test is developed step by step in Section 3. Roughly speaking, the procedure can be outlined as follows: Let $H_0(u,h)$ be the hypothesis that m is constant in the time window $[u-h, u+h] \subseteq [0,1]$, where u is the midpoint and $2h$ the size of the window. In a first step, we set up a test statistic $\widehat{s}_T(u,h)$ for the hypothesis $H_0(u,h)$. In a second step, we aggregate the statistics $\widehat{s}_T(u,h)$ for a large number of different time windows $[u-h, u+h]$. We thereby construct a multiscale statistic which allows to test the hypothesis $H_0(u,h)$ simultaneously for many time windows $[u-h, u+h]$. In the technical part of the paper, we derive the theoretical properties of the resulting multiscale test. To do so, we come up with a proof strategy which combines strong approximation results for dependent processes with anti-concentration bounds for Gaussian random vectors. This strategy is of interest in itself and may be applied to other multiscale test problems for dependent data. As shown by our theoretical analysis, our multiscale test is a rigorous level-α-test of the overall null hypothesis

H_0 that $H_0(u,h)$ is simultaneously fulfilled for all time windows $[u-h, u+h]$ under consideration. Moreover, for a given significance level $\alpha \in (0,1)$, the test allows to make simultaneous confidence statements of the following form: We can claim, with statistical confidence $1-\alpha$, that there is an increase/decrease in the trend m on all time windows $[u-h, u+h]$ for which the hypothesis $H_0(u,h)$ is rejected. Hence, the test allows to identify, with a pre-specified statistical confidence, time regions where the trend m is increasing/decreasing.

For independent data, multiscale tests have been developed in a variety of different contexts in recent years. In the regression context, Chaudhuri and Marron (1999, 2000) introduced the so-called SiZer method which has been extended in various directions; see e.g. Hannig and Marron (2006) where a refined distribution theory for SiZer is derived. Hall and Heckman (2000) constructed a multiscale test on monotonicity of a regression function. Dümbgen and Spokoiny (2001) developed a multiscale approach which works with additively corrected supremum statistics and derived theoretical results in the context of a continuous Gaussian white noise model. Rank-based multiscale tests for nonparametric regression were proposed in Dümbgen (2002) and Rohde (2008). More recently, Proksch et al. (2018) have constructed multiscale tests for inverse regression models. In the context of density estimation, multiscale tests have been investigated in Dümbgen and Walther (2008), Rufibach and Walther (2010), Schmidt-Hieber et al. (2013) and Eckle et al. (2017) among others.

Whereas a large number of multiscale tests for independent data have been developed in recent years, multiscale tests for dependent data are much rarer. Most notably, there are some extensions of the SiZer approach to a time series context. Park et al. (2004) and Rondonotti et al. (2007) have introduced SiZer methods for dependent data which can be used to find local increases/decreases of a trend and which may thus be regarded as an alternative to our multiscale test. However, these SiZer methods are mainly designed for data exploration rather than for rigorous statistical inference. Our multiscale method, in contrast, is a rigorous level-α-test of the hypothesis H_0 which allows to make simultaneous confidence statements about the time regions where the trend m is increasing/decreasing. Some theoretical results for dependent SiZer methods have been derived in Park et al. (2009), but only under a quite severe restriction: Only time windows $[u-h, u+h]$ with window sizes or scales h are taken into account that remain bounded away from zero as the sample size T grows. Scales h that converge to zero as T increases are excluded. This effectively means that only large time windows $[u-h, u+h]$ are taken into consideration. Our theory, in contrast, allows to simultaneously consider scales h of fixed size and scales h that converge to zero at various different rates. We are thus able to take into account time windows of many different sizes. In Section 3.4, we compare our approach to SiZer methods in more detail.

Our multiscale approach is also related to Wavelet-based methods: Similar to the latter, it takes into account different locations u and resolution levels or scales h simultaneously.

However, while our multiscale approach is designed to test for local increases/decreases of a nonparametric trend, Wavelet methods are commonly used for other purposes. Among other things, they are employed for estimating/reconstructing nonparametric regression curves [see e.g. Donoho et al. (1995) or Von Sachs and MacGibbon (2000)] and for change point detection [see e.g. Cho and Fryzlewicz (2012)].

The test statistic of our multiscale method depends on the long-run error variance $\sigma^2 = \sum_{\ell=-\infty}^{\infty} \mathrm{Cov}(\varepsilon_0, \varepsilon_\ell)$, which is usually unknown in practice. To carry out our multiscale test, we thus require an estimator of σ^2. Indeed, such an estimator is required for virtually all inferential procedures in the context of model (1.1). Hence, the problem of estimating σ^2 in model (1.1) is of broader interest and has received a lot of attention in the literature; see Müller and Stadtmüller (1988), Herrmann et al. (1992) and Hall and Van Keilegom (2003) among many others. In Section 4, we discuss several estimators of σ^2 which are valid under different conditions on the error process $\{\varepsilon_t\}$. Most notably, we introduce a new difference-based estimator of σ^2 for the case that $\{\varepsilon_t\}$ belongs to the class of AR(∞) processes. This estimator improves on existing methods in several respects.

The methodological and theoretical analysis of the paper is complemented by a simulation study in Section 5 and an empirical application in Section 6. In the simulation study, we examine the finite sample properties of our multiscale test and compare it to the dependent SiZer methods introduced in Park et al. (2004) and Rondonotti et al. (2007). Moreover, we investigate the small sample performance of our estimator of σ^2 in the AR(p) case and compare it to the estimator of Hall and Van Keilegom (2003). In Section 6, we use our methods to analyse the temperature data from Figure 1.

2 The model

We now describe the model setting in detail which was briefly outlined in the Introduction. We observe a time series $\{Y_{t,T} : 1 \leq t \leq T\}$ of length T which satisfies the nonparametric regression equation

$$Y_{t,T} = m\Big(\frac{t}{T}\Big) + \varepsilon_t \tag{2.1}$$

for $1 \leq t \leq T$. Here, m is an unknown nonparametric function defined on $[0, 1]$ and $\{\varepsilon_t : 1 \leq t \leq T\}$ is a zero-mean stationary error process. For simplicity, we restrict attention to equidistant design points $x_t = t/T$. However, our methods and theory can also be carried over to non-equidistant designs. The stationary error process $\{\varepsilon_t\}$ is assumed to have the following properties:

(C1) The variables ε_t allow for the representation $\varepsilon_t = G(\ldots, \eta_{t-1}, \eta_t, \eta_{t+1}, \ldots)$, where η_t are i.i.d. random variables and $G : \mathbb{R}^{\mathbb{Z}} \to \mathbb{R}$ is a measurable function.

(C2) It holds that $\|\varepsilon_t\|_q < \infty$ for some $q > 4$, where $\|\varepsilon_t\|_q = (\mathbb{E}|\varepsilon_t|^q)^{1/q}$.

Following Wu (2005), we impose conditions on the dependence structure of the error process $\{\varepsilon_t\}$ in terms of the physical dependence measure $d_{t,q} = \|\varepsilon_t - \varepsilon_t'\|_q$, where $\varepsilon_t' = G(\ldots, \eta_{-1}, \eta_0', \eta_1, \ldots, \eta_{t-1}, \eta_t, \eta_{t+1}, \ldots)$ with $\{\eta_t'\}$ being an i.i.d. copy of $\{\eta_t\}$. In particular, we assume the following:

(C3) Define $\Theta_{t,q} = \sum_{|s|\geq t} d_{s,q}$ for $t \geq 0$. It holds that $\Theta_{t,q} = O(t^{-\tau_q}(\log t)^{-A})$, where $A > \frac{2}{3}(1/q + 1 + \tau_q)$ and $\tau_q = \{q^2 - 4 + (q-2)\sqrt{q^2 + 20q + 4}\}/8q$.

The conditions (C1)–(C3) are fulfilled by a wide range of stationary processes $\{\varepsilon_t\}$. As a first example, consider linear processes of the form $\varepsilon_t = \sum_{i=0}^{\infty} c_i \eta_{t-i}$ with $\|\varepsilon_t\|_q < \infty$, where c_i are absolutely summable coefficients and η_t are i.i.d. innovations with $\mathbb{E}[\eta_t] = 0$ and $\|\eta_t\|_q < \infty$. Trivially, (C1) and (C2) are fulfilled in this case. Moreover, if $|c_i| = O(\rho^i)$ for some $\rho \in (0,1)$, then (C3) is easily seen to be satisfied as well. As a special case, consider an ARMA process $\{\varepsilon_t\}$ of the form $\varepsilon_t - \sum_{i=1}^{p} a_i \varepsilon_{t-i} = \eta_t + \sum_{j=1}^{r} b_j \eta_{t-j}$ with $\|\varepsilon_t\|_q < \infty$, where $a_1, \ldots, a_p$ and $b_1, \ldots, b_r$ are real-valued parameters. As before, we let η_t be i.i.d. innovations with $\mathbb{E}[\eta_t] = 0$ and $\|\eta_t\|_q < \infty$. Moreover, as usual, we suppose that the complex polynomials $A(z) = 1 - \sum_{j=1}^{p} a_j z^j$ and $B(z) = 1 + \sum_{j=1}^{r} b_j z^j$ do not have any roots in common. If $A(z)$ does not have any roots inside the unit disc, then the ARMA process $\{\varepsilon_t\}$ is stationary and causal. Specifically, it has the representation $\varepsilon_t = \sum_{i=0}^{\infty} c_i \eta_{t-i}$ with $|c_i| = O(\rho^i)$ for some $\rho \in (0,1)$, implying that (C1)–(C3) are fulfilled. The results in Wu and Shao (2004) show that condition (C3) (as well as the other two conditions) is not only fulfilled for linear time series processes but also for a variety of non-linear processes.

3 The multiscale test

In this section, we introduce our multiscale method to test for local increases/decreases of the trend function m and analyse its theoretical properties. We assume throughout that m is continuously differentiable on $[0,1]$. The test problem under consideration can be formulated as follows: Let $H_0(u,h)$ be the hypothesis that m is constant on the interval $[u-h, u+h]$. Since m is continuously differentiable, $H_0(u,h)$ can be reformulated as

$$H_0(u,h) : m'(w) = 0 \text{ for all } w \in [u-h, u+h],$$

where m' is the first derivative of m. We want to test the hypothesis $H_0(u,h)$ not only for a single interval $[u-h, u+h]$ but simultaneously for many different intervals. The overall null hypothesis is thus given by

$$H_0 : \text{ The hypothesis } H_0(u,h) \text{ holds true for all } (u,h) \in \mathcal{G}_T,$$

where $\mathcal{G}_T$ is some large set of points (u, h). The details on the set $\mathcal{G}_T$ are discussed at the end of Section 3.1 below. Note that $\mathcal{G}_T$ in general depends on the sample size T, implying that the null hypothesis $H_0 = H_{0,T}$ depends on T as well. We thus consider a sequence of null hypotheses $\{H_{0,T} : T = 1, 2, \ldots\}$ as T increases. For simplicity of notation, we however suppress the dependence of H_0 on T. In Sections 3.1 and 3.2, we step by step construct the multiscale test of the hypothesis H_0. The theoretical properties of the test are analysed in Section 3.3.

3.1 Construction of the multiscale statistic

We first construct a test statistic for the hypothesis $H_0(u, h)$, where $[u - h, u + h]$ is a given interval. To do so, we consider the kernel average

$$\widehat{\psi}_T(u, h) = \sum_{t=1}^{T} w_{t,T}(u, h) Y_{t,T},$$

where $w_{t,T}(u, h)$ is a kernel weight and h is the bandwidth. In order to avoid boundary issues, we work with a local linear weighting scheme. We in particular set

$$w_{t,T}(u, h) = \frac{\Lambda_{t,T}(u, h)}{\{\sum_{t=1}^{T} \Lambda_{t,T}(u, h)^2\}^{1/2}}, \tag{3.1}$$

where

$$\Lambda_{t,T}(u, h) = K\Big(\frac{\frac{t}{T} - u}{h}\Big)\Big[S_{T,0}(u, h)\Big(\frac{\frac{t}{T} - u}{h}\Big) - S_{T,1}(u, h)\Big],$$

$S_{T,\ell}(u, h) = (Th)^{-1} \sum_{t=1}^{T} K(\frac{\frac{t}{T}-u}{h})(\frac{\frac{t}{T}-u}{h})^\ell$ for $\ell = 0, 1, 2$ and K is a kernel function with the following properties:

(C4) The kernel K is non-negative, symmetric about zero and integrates to one. Moreover, it has compact support $[-1, 1]$ and is Lipschitz continuous, that is, $|K(v) - K(w)| \leq C|v - w|$ for any $v, w \in \mathbb{R}$ and some constant $C > 0$.

The kernel average $\widehat{\psi}_T(u, h)$ is nothing else than a rescaled local linear estimator of the derivative $m'(u)$ with bandwidth h.[3]

A test statistic for the hypothesis $H_0(u, h)$ is given by the normalized kernel average $\widehat{\psi}_T(u, h)/\widehat{\sigma}$, where $\widehat{\sigma}^2$ is an estimator of the long-run variance $\sigma^2 = \sum_{\ell=-\infty}^{\infty} \mathrm{Cov}(\varepsilon_0, \varepsilon_\ell)$ of the error process $\{\varepsilon_t\}$. The problem of estimating σ^2 is discussed in detail in Section 4. For the time being, we suppose that $\widehat{\sigma}^2$ is an estimator with reasonable theoretical properties. Specifically, we assume that $\widehat{\sigma}^2 = \sigma^2 + o_p(\rho_T)$ with $\rho_T = o(1/\log T)$. This is

[3] Alternatively to the local linear weights defined in (3.1), we could also work with the weights $w_{t,T}(u, h) = K'(h^{-1}[u - t/T])/\{\sum_{t=1}^{T} K'(h^{-1}[u - t/T])^2\}^{1/2}$, where the kernel function K is assumed to be differentiable and K' is its derivative. We however prefer to use local linear weights as these have superior theoretical properties at the boundary.

a fairly weak condition which is in particular satisfied by the estimators of σ^2 analysed in Section 4. The kernel weights $w_{t,T}(u,h)$ are chosen such that in the case of independent errors ε_t, $\mathrm{Var}(\widehat{\psi}_T(u,h)) = \sigma^2$ for any location u and bandwidth h, where the long-run error variance σ^2 simplifies to $\sigma^2 = \mathrm{Var}(\varepsilon_t)$. In the more general case that the error terms satisfy the weak dependence conditions from Section 2, $\mathrm{Var}(\widehat{\psi}_T(u,h)) = \sigma^2 + o(1)$ for any u and h under consideration. Hence, for sufficiently large sample sizes T, the test statistic $\widehat{\psi}_T(u,h)/\widehat{\sigma}$ has approximately unit variance.

We now combine the test statistics $\widehat{\psi}_T(u,h)/\widehat{\sigma}$ for a wide range of different locations u and bandwidths or scales h. There are different ways to do so, leading to different types of multiscale statistics. Our multiscale statistic is defined as

$$\widehat{\Psi}_T = \max_{(u,h)\in\mathcal{G}_T} \Big\{ \Big| \frac{\widehat{\psi}_T(u,h)}{\widehat{\sigma}} \Big| - \lambda(h) \Big\}, \tag{3.2}$$

where $\lambda(h) = \sqrt{2\log\{1/(2h)\}}$ and $\mathcal{G}_T$ is the set of points (u,h) that are taken into consideration. The details on the set $\mathcal{G}_T$ are given below. As can be seen, the statistic $\widehat{\Psi}_T$ does not simply aggregate the individual statistics $\widehat{\psi}_T(u,h)/\widehat{\sigma}$ by taking the supremum over all points $(u,h) \in \mathcal{G}_T$ as in more traditional multiscale approaches. We rather calibrate the statistics $\widehat{\psi}_T(u,h)/\widehat{\sigma}$ that correspond to the bandwidth h by subtracting the additive correction term $\lambda(h)$. This approach was pioneered by Dümbgen and Spokoiny (2001) and has been used in numerous other studies since then; see e.g. Dümbgen (2002), Rohde (2008), Dümbgen and Walther (2008), Rufibach and Walther (2010), Schmidt-Hieber et al. (2013) and Eckle et al. (2017).

To see the heuristic idea behind the additive correction $\lambda(h)$, consider for a moment the uncorrected statistic

$$\widehat{\Psi}_{T,\mathrm{uncorrected}} = \max_{(u,h)\in\mathcal{G}_T} \Big| \frac{\widehat{\psi}_T(u,h)}{\widehat{\sigma}} \Big| \tag{3.3}$$

and suppose that the hypothesis $H_0(u,h)$ is true for all $(u,h) \in \mathcal{G}_T$. For simplicity, assume that the errors ε_t are i.i.d. normally distributed and neglect the estimation error in $\widehat{\sigma}$, that is, set $\widehat{\sigma} = \sigma$. Moreover, suppose that the set $\mathcal{G}_T$ only consists of the points $(u_k, h_\ell) = ((2k-1)h_\ell, h_\ell)$ with $k = 1, \ldots, \lfloor 1/2h_\ell \rfloor$ and $\ell = 1, \ldots, L$. In this case, we can write

$$\widehat{\Psi}_{T,\mathrm{uncorrected}} = \max_{1\le\ell\le L} \max_{1\le k\le\lfloor 1/2h_\ell\rfloor} \Big| \frac{\widehat{\psi}_T(u_k,h_\ell)}{\sigma} \Big|.$$

Under our simplifying assumptions, the statistics $\widehat{\psi}_T(u_k,h_\ell)/\sigma$ with $k = 1, \ldots, \lfloor 1/2h_\ell \rfloor$ are independent and standard normal for any given bandwidth h_ℓ. Since the maximum over $\lfloor 1/2h \rfloor$ independent standard normal random variables is $\lambda(h) + o_p(1)$ as $h \to 0$, we obtain that $\max_k \widehat{\psi}_T(u_k,h_\ell)/\sigma$ is approximately of size $\lambda(h_\ell)$ for small bandwidths h_ℓ. As $\lambda(h) \to \infty$ for $h \to 0$, this implies that $\max_k \widehat{\psi}_T(u_k,h_\ell)/\sigma$ tends to be much larger

in size for small than for large bandwidths h_ℓ. As a result, the stochastic behaviour of the uncorrected statistic $\widehat{\Psi}_{T,\text{uncorrected}}$ tends to be dominated by the statistics $\widehat{\psi}_T(u_k, h_\ell)$ corresponding to small bandwidths h_ℓ. The additively corrected statistic $\widehat{\Psi}_T$, in contrast, puts the statistics $\widehat{\psi}_T(u_k, h_\ell)$ corresponding to different bandwidths h_ℓ on a more equal footing, thus counteracting the dominance of small bandwidth values.
The multiscale statistic $\widehat{\Psi}_T$ simultaneously takes into account all locations u and bandwidths h with $(u, h) \in \mathcal{G}_T$. Throughout the paper, we suppose that $\mathcal{G}_T$ is some subset of $\mathcal{G}_T^{\text{full}} = \{(u, h) : u = t/T \text{ for some } 1 \leq t \leq T \text{ and } h \in [h_{\min}, h_{\max}]\}$, where $h_{\min}$ and $h_{\max}$ denote some minimal and maximal bandwidth value, respectively. For our theory to work, we require the following conditions to hold:

(C5) $|\mathcal{G}_T| = O(T^\theta)$ for some arbitrarily large but fixed constant $\theta > 0$, where $|\mathcal{G}_T|$ denotes the cardinality of $\mathcal{G}_T$.

(C6) $h_{\min} \gg T^{-(1-\frac{2}{q})} \log T$, that is, $h_{\min}/\{T^{-(1-\frac{2}{q})} \log T\} \to \infty$ with $q > 4$ defined in (C2) and $h_{\max} < 1/2$.

According to (C5), the number of points (u, h) in $\mathcal{G}_T$ should not grow faster than T^θ for some arbitrarily large but fixed $\theta > 0$. This is a fairly weak restriction as it allows the set $\mathcal{G}_T$ to be extremely large compared to the sample size T. For example, we may work with the set

$$\mathcal{G}_T = \big\{(u, h) : u = t/T \text{ for some } 1 \leq t \leq T \text{ and } h \in [h_{\min}, h_{\max}] \\ \text{with } h = t/T \text{ for some } 1 \leq t \leq T\big\},$$

which contains more than enough points (u, h) for most practical applications. Condition (C6) imposes some restrictions on the minimal and maximal bandwidths $h_{\min}$ and $h_{\max}$. These conditions are fairly weak, allowing us to choose the bandwidth window $[h_{\min}, h_{\max}]$ extremely large. The lower bound on $h_{\min}$ depends on the parameter q defined in (C2) which specifies the number of existing moments for the error terms ε_t. As one can see, we can choose $h_{\min}$ to be of the order $T^{-1/2}$ for any $q > 4$. Hence, we can let $h_{\min}$ converge to 0 very quickly even if only the first few moments of the error terms ε_t exist. If all moments exist (i.e. $q = \infty$), $h_{\min}$ may converge to 0 almost as quickly as $T^{-1} \log T$. Furthermore, the maximal bandwidth $h_{\max}$ is not even required to converge to 0, which implies that we can pick it very large.

Remark 3.1. *The above construction of the multiscale statistic can be easily adapted to hypotheses other than H_0. To do so, one simply needs to replace the kernel weights $w_{t,T}(u, h)$ defined in (3.1) by appropriate versions which are suited to test the hypothesis of interest. For example, if one wants to test for local convexity/concavity of m, one may define the kernel weights $w_{t,T}(u, h)$ such that the kernel average $\widehat{\psi}_T(u, h)$ is a (rescaled) estimator of the second derivative of m at the location u with bandwidth h.*

3.2 The test procedure

In order to formulate a test for the null hypothesis H_0, we still need to specify a critical value. To do so, we define the statistic

$$\Phi_T = \max_{(u,h)\in\mathcal{G}_T} \Big\{ \Big| \frac{\phi_T(u,h)}{\sigma} \Big| - \lambda(h) \Big\}, \tag{3.4}$$

where $\phi_T(u,h) = \sum_{t=1}^T w_{t,T}(u,h)\,\sigma Z_t$ and Z_t are independent standard normal random variables. The statistic Φ_T can be regarded as a Gaussian version of the test statistic $\widehat{\Psi}_T$ under the null hypothesis H_0. Let $q_T(\alpha)$ be the $(1-\alpha)$-quantile of Φ_T. Importantly, the quantile $q_T(\alpha)$ can be computed by Monte Carlo simulations and can thus be regarded as known. Our multiscale test of the hypothesis H_0 is now defined as follows: For a given significance level $\alpha \in (0,1)$, we reject H_0 if $\widehat{\Psi}_T > q_T(\alpha)$.

3.3 Theoretical properties of the test

In order to examine the theoretical properties of our multiscale test, we introduce the auxiliary multiscale statistic

$$\widehat{\Phi}_T = \max_{(u,h)\in\mathcal{G}_T} \Big\{ \Big| \frac{\widehat{\phi}_T(u,h)}{\widehat{\sigma}} \Big| - \lambda(h) \Big\} \tag{3.5}$$

with $\widehat{\phi}_T(u,h) = \widehat{\psi}_T(u,h) - \mathbb{E}[\widehat{\psi}_T(u,h)] = \sum_{t=1}^T w_{t,T}(u,h)\varepsilon_t$. The following result is central to the theoretical analysis of our multiscale test. According to it, the (known) quantile $q_T(\alpha)$ of the Gaussian statistic Φ_T defined in Section 3.2 can be used as a proxy for the $(1-\alpha)$-quantile of the multiscale statistic $\widehat{\Phi}_T$.

Theorem 3.1. *Let (C1)–(C6) be fulfilled and assume that* $\widehat{\sigma}^2 = \sigma^2 + o_p(\rho_T)$ *with* $\rho_T = o(1/\log T)$. *Then*

$$\mathbb{P}\big(\widehat{\Phi}_T \le q_T(\alpha)\big) = (1-\alpha) + o(1).$$

A full proof of Theorem 3.1 is given in the Supplementary Material. We here shortly outline the proof strategy, which splits up into two main steps. In the first, we replace the statistic $\widehat{\Phi}_T$ for each $T \ge 1$ by a statistic $\widetilde{\Phi}_T$ with the same distribution as $\widehat{\Phi}_T$ and the property that

$$\big|\widetilde{\Phi}_T - \Phi_T\big| = o_p(\delta_T), \tag{3.6}$$

where $\delta_T = o(1)$ and the Gaussian statistic Φ_T is defined in Section 3.2. We thus replace the statistic $\widehat{\Phi}_T$ by an identically distributed version which is close to a Gaussian statistic whose distribution is known. To do so, we make use of strong approximation theory for dependent processes as derived in Berkes et al. (2014). In the second step, we show

that

$$\sup_{x\in\mathbb{R}} \left|\mathbb{P}(\widetilde{\Phi}_T \leq x) - \mathbb{P}(\Phi_T \leq x)\right| = o(1), \tag{3.7}$$

which immediately implies the statement of Theorem 3.1. Importantly, the convergence result (3.6) is not sufficient for establishing (3.7). Put differently, the fact that $\widetilde{\Phi}_T$ can be approximated by Φ_T in the sense that $\widetilde{\Phi}_T - \Phi_T = o_p(\delta_T)$ does not imply that the distribution of $\widetilde{\Phi}_T$ is close to that of Φ_T in the sense of (3.7). For (3.7) to hold, we additionally require the distribution of Φ_T to have some sort of continuity property. Specifically, we prove that

$$\sup_{x\in\mathbb{R}} \mathbb{P}\big(|\Phi_T - x| \leq \delta_T\big) = o(1), \tag{3.8}$$

which says that Φ_T does not concentrate too strongly in small regions of the form $[x - \delta_T, x + \delta_T]$. The main tool for verifying (3.8) are anti-concentration results for Gaussian random vectors as derived in Chernozhukov et al. (2015). The claim (3.7) can be proven by using (3.6) together with (3.8), which in turn yields Theorem 3.1.

The main idea of our proof strategy is to combine strong approximation theory with anti-concentration bounds for Gaussian random vectors to show that the quantiles of the multiscale statistic $\widehat{\Phi}_T$ can be proxied by those of a Gaussian analogue. This strategy is quite general in nature and may be applied to other multiscale problems for dependent data. Strong approximation theory has also been used to investigate multiscale tests for independent data; see e.g. Schmidt-Hieber et al. (2013). However, it has not been combined with anti-concentration results to approximate the quantiles of the multiscale statistic. As an alternative to strong approximation theory, Eckle et al. (2017) and Proksch et al. (2018) have recently used Gaussian approximation results derived in Chernozhukov et al. (2014, 2017) to analyse multiscale tests for independent data. Even though it might be possible to adapt these techniques to the case of dependent data, this is not trivial at all as part of the technical arguments and the Gaussian approximation tools strongly rely on the assumption of independence.

We now investigate the theoretical properties of our multiscale test with the help of Theorem 3.1. The first result is an immediate consequence of Theorem 3.1. It says that the test has the correct (asymptotic) size.

Proposition 3.1. *Let the conditions of Theorem 3.1 be satisfied. Under the null hypothesis H_0, it holds that*

$$\mathbb{P}\big(\widehat{\Psi}_T \leq q_T(\alpha)\big) = (1-\alpha) + o(1).$$

The second result characterizes the power of the multiscale test against local alternatives. To formulate it, we consider any sequence of functions $m = m_T$ with the following

property: There exists $(u,h) \in \mathcal{G}_T$ with $[u-h, u+h] \subseteq [0,1]$ such that

$$m_T'(w) \geq c_T \sqrt{\frac{\log T}{Th^3}} \quad \text{for all } w \in [u-h, u+h], \tag{3.9}$$

where $\{c_T\}$ is any sequence of positive numbers with $c_T \to \infty$. Alternatively to (3.9), we may also assume that $-m_T'(w) \geq c_T \sqrt{\log T/(Th^3)}$ for all $w \in [u-h, u+h]$.

Proposition 3.2. *Let the conditions of Theorem 3.1 be satisfied and consider any sequence of functions m_T with the property* (3.9)*. Then*

$$\mathbb{P}\big(\widehat{\Psi}_T \leq q_T(\alpha)\big) = o(1).$$

According to Proposition 3.2, our test has asymptotic power 1 against local alternatives of the form (3.9). The proof can be found in the Supplementary Material.
The next result formally shows that we can make simultaneous confidence statements about the time intervals where the trend m is increasing/decreasing. To formulate it, we define

$$\begin{aligned}
\Pi_T^{\pm} &= \big\{I_{u,h} = [u-h, u+h] : (u,h) \in \mathcal{A}_T^{\pm}\big\} \\
\Pi_T^{+} &= \big\{I_{u,h} = [u-h, u+h] : (u,h) \in \mathcal{A}_T^{+} \text{ and } I_{u,h} \subseteq [0,1]\big\} \\
\Pi_T^{-} &= \big\{I_{u,h} = [u-h, u+h] : (u,h) \in \mathcal{A}_T^{-} \text{ and } I_{u,h} \subseteq [0,1]\big\},
\end{aligned}$$

where

$$\begin{aligned}
\mathcal{A}_T^{\pm} &= \Big\{(u,h) \in \mathcal{G}_T : \Big|\frac{\widehat{\psi}_T(u,h)}{\widehat{\sigma}}\Big| > q_T(\alpha) + \lambda(h)\Big\} \\
\mathcal{A}_T^{+} &= \Big\{(u,h) \in \mathcal{G}_T : \frac{\widehat{\psi}_T(u,h)}{\widehat{\sigma}} > q_T(\alpha) + \lambda(h)\Big\} \\
\mathcal{A}_T^{-} &= \Big\{(u,h) \in \mathcal{G}_T : -\frac{\widehat{\psi}_T(u,h)}{\widehat{\sigma}} > q_T(\alpha) + \lambda(h)\Big\}.
\end{aligned}$$

The object $\Pi_T^{\pm}$ can be interpreted as follows: Our multiscale test rejects the null hypothesis $H_0(u,h)$ if the (corrected) test statistic $|\widehat{\psi}_T(u,h)/\widehat{\sigma}| - \lambda(h)$ lies above the critical value $q_T(\alpha)$. Put differently, it rejects $H_0(u,h)$ for all $(u,h) \in \mathcal{A}_T^{\pm}$. Hence, $\Pi_T^{\pm}$ is the collection of time intervals $I_{u,h} = [u-h, u+h]$ for which our test rejects $H_0(u,h)$. Note that $\Pi_T^{\pm}$ is a random collection of intervals: Whether our test rejects $H_0(u,h)$ for some (u,h) depends on the realization of the random vector $(Y_{1,T}, \ldots, Y_{T,T})$. Hence, whether an interval $I_{u,h}$ belongs to $\Pi_T^{\pm}$ depends on this realization as well. The objects Π_T^{+} and Π_T^{-} can be interpreted analogous to $\Pi_T^{\pm}$ but take into account the sign of the statistic $\widehat{\psi}_T(u,h)/\widehat{\sigma}$. With this notation at hand, we consider the events

$$E_T^{\pm} = \Big\{\forall I_{u,h} \in \Pi_T^{\pm} : m'(v) \neq 0 \text{ for some } v \in I_{u,h} = [u-h, u+h]\Big\}$$

$$E_T^+ = \Big\{\forall I_{u,h} \in \Pi_T^+ : m'(v) > 0 \text{ for some } v \in I_{u,h} = [u-h, u+h]\Big\}$$
$$E_T^- = \Big\{\forall I_{u,h} \in \Pi_T^- : m'(v) < 0 \text{ for some } v \in I_{u,h} = [u-h, u+h]\Big\}.$$

$E_T^\pm$ (E_T^+, E_T^-) is the event that the function m is non-constant (increasing, decreasing) on all intervals $I_{u,h} \in \Pi_T^\pm$ (Π_T^+, Π_T^-). More precisely, $E_T^\pm$ (E_T^+, E_T^-) is the event that for each interval $I_{u,h} \in \Pi_T^\pm$ (Π_T^+, Π_T^-), there is a subset $J_{u,h} \subseteq I_{u,h}$ with m being a non-constant (increasing, decreasing) function on $J_{u,h}$. We can make the following formal statement about the events $E_T^\pm$, E_T^+ and E_T^-, whose proof is given in the Supplement.

Proposition 3.3. *Let the conditions of Theorem 3.1 be fulfilled. Then for $\ell \in \{\pm, +, -\}$, it holds that*

$$\mathbb{P}\big(E_T^\ell\big) \geq (1-\alpha) + o(1).$$

According to Proposition 3.3, we can make simultaneous confidence statements of the following form: With (asymptotic) probability $\geq (1-\alpha)$, the trend function m is non-constant (increasing, decreasing) on some part of the interval $I_{u,h}$ for all $I_{u,h} \in \Pi_T^\pm$ (Π_T^+, Π_T^-). Hence, our multiscale procedure allows to identify, with a pre-specified confidence, time regions where there is an increase/decrease in the time trend m.

Remark 3.2. *Unlike $\Pi_T^\pm$, the sets Π_T^+ and Π_T^- only contain intervals $I_{u,h} = [u-h, u+h]$ which are subsets of $[0,1]$. We thus exclude points $(u,h) \in \mathcal{A}_T^+$ and $(u,h) \in \mathcal{A}_T^-$ which lie at the boundary, that is, for which $I_{u,h} \not\subseteq [0,1]$. The reason is as follows: Let $(u,h) \in \mathcal{A}_T^+$ with $I_{u,h} \not\subseteq [0,1]$. Our technical arguments allow us to say, with asymptotic confidence $\geq 1-\alpha$, that $m'(v) \neq 0$ for some $v \in I_{u,h}$. However, we cannot say whether $m'(v) > 0$ or $m'(v) < 0$, that is, we cannot make confidence statements about the sign. Crudely speaking, the problem is that the local linear weights $w_{t,T}(u,h)$ behave quite differently at boundary points (u,h) with $I_{u,h} \not\subseteq [0,1]$. As a consequence, we can include boundary points (u,h) in $\Pi_T^\pm$ but not in Π_T^+ and Π_T^-.*

Remark 3.3. *The statement of Proposition 3.3 suggests to graphically present the results of our multiscale test by plotting the intervals $I_{u,h} \in \Pi_T^\ell$ for $\ell \in \{\pm, +, -\}$, that is, by plotting the intervals where (with asymptotic confidence $\geq 1-\alpha$) our test detects a violation of the null hypothesis. The drawback of this graphical presentation is that the number of intervals in Π_T^ℓ is often quite large. To obtain a better graphical summary of the results, we replace Π_T^ℓ by a subset $\Pi_T^{\ell,\min}$ which is constructed as follows: As in Dümbgen (2002), we call an interval $I_{u,h} \in \Pi_T^\ell$ minimal if there is no other interval $I_{u',h'} \in \Pi_T^\ell$ with $I_{u',h'} \subset I_{u,h}$. Let $\Pi_T^{\ell,\min}$ be the set of all minimal intervals in Π_T^ℓ for $\ell \in \{\pm, +, -\}$ and define the events*

$$E_T^{\pm,\min} = \Big\{\forall I_{u,h} \in \Pi_T^{\pm,\min} : m'(v) \neq 0 \text{ for some } v \in I_{u,h} = [u-h, u+h]\Big\}$$
$$E_T^{+,\min} = \Big\{\forall I_{u,h} \in \Pi_T^{+,\min} : m'(v) > 0 \text{ for some } v \in I_{u,h} = [u-h, u+h]\Big\}$$

$$E_T^{-,\min} = \Big\{ \forall I_{u,h} \in \Pi_T^{-,\min} : m'(v) < 0 \text{ for some } v \in I_{u,h} = [u-h, u+h] \Big\}.$$

It is easily seen that $E_T^{\ell} = E_T^{\ell,\min}$ *for* $\ell \in \{\pm, +, -\}$. *Hence, by Proposition 3.3, it holds that*

$$\mathbb{P}\big(E_T^{\ell,\min}\big) \geq (1-\alpha) + o(1)$$

for $\ell \in \{\pm, +, -\}$. *This suggests to plot the minimal intervals in* $\Pi_T^{\ell,\min}$ *rather than the whole collection of intervals* Π_T^{ℓ} *as a graphical summary of the test results. We in particular use this way of presenting the test results in our application in Section 6.*

Proposition 3.3 allows to make confidence statements for a fixed significance level $\alpha \in (0,1)$. In some situations, one may be interested in letting $\alpha = \alpha_T \in (0,1)$ tend to zero as $T \to \infty$. We can prove the following corollary to Proposition 3.3 for this case, whose proof can be found in the Supplementary Material.

Corollary 3.1. *Let the conditions of Theorem 3.1 be fulfilled and let* $\alpha = \alpha_T \in (0,1)$ *go to zero as* $T \to \infty$. *Then* $\mathbb{P}(E_T^{\ell}) \to 1$ *for* $\ell \in \{\pm, +, -\}$.

Corollary 3.1 can be interpreted as a consistency result: If we let the significance level $\alpha = \alpha_T$ go to zero, then the event $E_T^{\pm}$ (E_T^{+}, E_T^{-}) occurs with probability tending to 1, that is, the trend m is non-constant (increasing, decreasing) on some part of the interval $I_{u,h}$ for all $I_{u,h} \in \Pi_T^{\pm}$ (Π_T^{+}, Π_T^{-}) with probability tending to 1.

3.4 Comparison to SiZer methods

As already mentioned in the introduction, some SiZer methods for dependent data have been introduced in Park et al. (2004) and Rondonotti et al. (2007), which we refer to as dependent SiZer for short. Informally speaking, both our approach and dependent SiZer are methods to test for local increases/decreases of a nonparametric trend function m. The formal problem is to test the hypothesis $H_0(u,h)$ simultaneously for all $(u,h) \in \mathcal{G}_T$, where we assume that $\mathcal{G}_T = U_T \times H_T$ with $U = U_T$ being the set of locations and $H = H_T$ the set of bandwidths or scales. In what follows, we compare our approach to dependent SiZer and point out the most important differences.
Dependent SiZer is based on the statistics $s_T(u,h) = \widehat{m}'(u,h)/\widehat{\mathrm{sd}}(\widehat{m}'(u,h))$, where $\widehat{m}'(u,h)$ is a local linear kernel estimator of $m'(u)$ with bandwidth h and $\widehat{\mathrm{sd}}(\widehat{m}'(u,h))$ an estimator of its standard deviation. The statistic $s_T(u,h)$ parallels the statistic $\widehat{\psi}_T(u,h)/\widehat{\sigma}$ in our approach. In particular, both can be regarded as test statistics of the hypothesis $H_0(u,h)$. There are two versions of dependent SiZer:

(a) The global version aggegrates the individual statistics $s_T(u,h)$ into the overall statistic $S_T = \max_{h \in H} S_T(h)$, where $S_T(h) = \max_{u \in U} |s_T(u,h)|$. The statistic S_T is the counterpart to the multiscale statistic $\widehat{\Psi}_T$ in our approach.

(b) The row-wise SiZer version considers each scale $h \in H$ separately. In particular, for each bandwidth $h \in H$, a test is carried out based on the statistic $S_T(h)$. A row-wise analogue of our approach would be obtained by carrying out a test for each scale $h \in H$ separately based on the statistic $\widehat{\Psi}_T(h) = \max_{u \in U} |\widehat{\psi}_T(u,h)/\widehat{\sigma}|$.[4]

In practice, SiZer is commonly implemented in its row-wise form. The main reason is that it has more power than the global version by construction. However, this gain of power comes at a cost: Row-wise SiZer carries out a test *separately* for each scale $h \in H$, thus ignoring the simultaneous test problem across scales h. Hence, it is not a rigorous level-α-test of the null H_0. For this reason, we focus on global SiZer in the rest of this section.

Even though related, our methods and theory are markedly different from those of the SiZer approach:

(i) Theory for SiZer is derived under the assumption that the set of bandwidths H is a compact subset of $(0,1)$. As already pointed out in Chaudhuri and Marron (2000) on p.420, this is a quite severe restriction: Only bandwidths h are taken into account that remain bounded away from zero as the sample size T grows. Bandwidths h that converge to zero as T increases are excluded. Our theory, in contrast, allows to simultaneously consider bandwidths h of fixed size and bandwidths h that converge to zero at various different rates. To achieve this, we come up with a proof strategy which is very different from that in the SiZer literature: As proven in Chaudhuri and Marron (2000) for the i.i.d. case and in Park et al. (2009) for the dependent data case, S_T weakly converges to some limit process S under the overall null hypothesis H_0. This is the central technical result on which the theoretical properties of SiZer are based. In contrast to this, our proof strategy (which combines strong approximation theory with anti-concentration bounds as outlined in Section 3.3) does not even require the statistic $\widehat{\Psi}_T$ to have a weak limit and is thus not restricted by the limitations of classic weak convergence theory.

(ii) There are different ways to combine the test statistics $S_T(h) = \max_{u \in I} |s_T(u,h)|$ for different scales $h \in H$. One way is to take their maximum, which leads to the SiZer statistic $S_T = \max_{h \in H} S_T(h)$. We could proceed analogously and consider the multiscale statistic $\widehat{\Psi}_{T,\text{uncorrected}} = \max_{h \in H} \widehat{\Psi}_T(h) = \max_{(u,h) \in U \times H} |\widehat{\psi}_T(u,h)/\widehat{\sigma}|$. However, as argued in Dümbgen and Spokoiny (2001) and as discussed in Section 3.1, this aggregation scheme is not optimal when the set $H = H_T$ contains scales h of many different rates. Following the lead of Dümbgen and Spokoiny (2001), we consider the test statistic $\widehat{\Psi}_T = \max_{(u,h) \in U \times H} \{|\widehat{\psi}_T(u,h)/\widehat{\sigma}| - \lambda(h)\}$ with the additive correction terms $\lambda(h)$. Hence, even though related, our multiscale test statistic $\widehat{\Psi}_T$ differs from the SiZer statistic S_T in important ways.

[4]Note that we can drop the correction term $\lambda(h)$ in this case as it is a fixed constant if only a single bandwidth h is taken into account.

(iii) The main complication in carrying out both our multiscale test and SiZer is to determine the critical values, that is, the quantiles of the test statistics $\widehat{\Psi}_T$ and S_T under H_0. In order to approximate the quantiles, we proceed quite differently than in the SiZer literature. The quantiles of the SiZer statistic S_T can be approximated by those of the weak limit process S. Usually, however, the quantiles of S cannot be determined analytically but have to be approximated themselves (e.g. by the bootstrap procedures of Chaudhuri and Marron (1999, 2000)). Alternatively, the quantiles of S_T can be approximated by procedures based on extreme value theory (as proposed in Hannig and Marron (2006) and Park et al. (2009)). In our approach, the quantiles of $\widehat{\Psi}_T$ under H_0 are approximated by those of a suitably constructed Gaussian analogue of $\widehat{\Psi}_T$. It is far from obvious that this Gaussian approximation is valid when the data are dependent. To see this, deep strong approximation theory for dependent data (as derived in Berkes et al. (2014)) is needed. It is important to note that our Gaussian approximation procedure is not the same as the bootstrap procedures proposed in Chaudhuri and Marron (1999, 2000). Both procedures can of course be regarded as resampling methods. However, the resampling is done in a quite different way in our case.

4 Estimation of the long-run error variance

In this section, we discuss how to estimate the long-run variance $\sigma^2 = \sum_{\ell=-\infty}^{\infty} \mathrm{Cov}(\varepsilon_0, \varepsilon_\ell)$ of the error terms in model (2.1). There are two broad classes of estimators: residual- and difference-based estimators. In residual-based approaches, σ^2 is estimated from the residuals $\widehat{\varepsilon}_t = Y_{t,T} - \widehat{m}_h(t/T)$, where $\widehat{m}_h$ is a nonparametric estimator of m with the bandwidth or smoothing parameter h. Difference-based methods proceed by estimating σ^2 from the ℓ-th differences $Y_{t,T} - Y_{t-\ell,T}$ of the observed time series $\{Y_{t,T}\}$ for certain orders ℓ. In what follows, we focus attention on difference-based methods as these do not involve a nonparametric estimator of the function m and thus do not require to specify a bandwidth h for the estimation of m.

So far, we have assumed that $\{\varepsilon_t\}$ is a general stationary error process which fulfills the weak dependence conditions (C3). Estimating the long-run error variance σ^2 in model (2.1) under general weak dependence conditions is a notoriously difficult problem. Estimators of σ^2 often tend to be quite imprecise. To circumvent this issue in practice, it may be beneficial to impose a time series model on the error process $\{\varepsilon_t\}$. Estimating σ^2 under the restrictions of such a model may of course create some misspecification bias. However, as long as the model gives a reasonable approximation to the true error process, the produced estimates of σ^2 can be expected to be fairly reliable even though they are a bit biased.

Estimators of the long-run error variance σ^2 in model (2.1) have been developed for

different kinds of error models. A number of authors have analysed the case of MA(m) or, more generally, m-dependent error terms. Difference-based estimators of σ^2 for this case were proposed in Müller and Stadtmüller (1988), Herrmann et al. (1992) and Tecuapetla-Gómez and Munk (2017) among others. Presumably the most widely used error model in practice is an AR(p) process. Residual-based methods to estimate σ^2 in model (2.1) with AR(p) errors can be found for example in Truong (1991), Shao and Yang (2011) and Qiu et al. (2013). A difference-based method was proposed in Hall and Van Keilegom (2003).

We consider the class of AR(∞) processes as an error model, which is a quite large and important subclass of linear time series processes. Formally speaking, we let $\{\varepsilon_t\}$ be a process of the form

$$\varepsilon_t = \sum_{j=1}^{\infty} a_j \varepsilon_{t-j} + \eta_t, \tag{4.1}$$

where $a_1, a_2, a_3, \ldots$ are unknown coefficients and η_t are i.i.d. innovations with $\mathbb{E}[\eta_t] = 0$ and $\mathbb{E}[\eta_t^2] = \nu^2$. We assume that $A(z) := 1 - \sum_{j=1}^{\infty} a_j z^j \neq 0$ for all complex numbers $|z| \leq 1 + \delta$ with some small $\delta > 0$, which has the following implications: (i) $\{\varepsilon_t\}$ is stationary and causal. (ii) The coefficients a_j decay to zero exponentially fast, that is, $|a_j| \leq C\xi^j$ with some $C > 0$ and $\xi \in (0, 1)$. (iii) $\{\varepsilon_t\}$ has an MA(∞) representation of the form $\varepsilon_t = \sum_{k=0}^{\infty} c_k \eta_{t-k}$. The coefficients c_k can be computed iteratively from the equations

$$c_k - \sum_{j=1}^{k} a_j c_{k-j} = b_k \tag{4.2}$$

for $k = 0, 1, 2, \ldots$, where $b_0 = 1$, $b_k = 0$ for $k > 0$ and $c_k = 0$ for $k < 0$. Moreover, they decay to zero exponentially fast, that is, $|c_k| \leq C\xi^k$ with some constants $C > 0$ and $\xi \in (0, 1)$.

Notably, the error model (4.1) nests AR(p^*) processes of any finite order p^* as a special case: If $a_{p^*} \neq 0$ and $a_j = 0$ for all $j > p^*$, then $\{\varepsilon_t\}$ is an AR process of order p^*. In what follows, we let $p^* \in \mathbb{N} \cup \{\infty\}$ denote the true AR order of $\{\varepsilon_t\}$ which may be finite or infinite. We can thus rewrite (4.1) as

$$\varepsilon_t = \sum_{j=1}^{p^*} a_j \varepsilon_{t-j} + \eta_t, \tag{4.3}$$

where we treat the AR order p^* as unknown. In particular, it is not known whether p^* is finite or infinite. In order to deal with this, we will fit AR(p) processes to the data whose order $p = p_T$ grows with the sample size T. In the special case that p^* is finite and known, this is of course not needed and we can simply set $p = p^*$.

We now construct a difference-based estimator of σ^2 for the case that $\{\varepsilon_t\}$ is an AR(p^*) process of the form (4.3). To simplify notation, we let $\Delta_\ell Z_t = Z_t - Z_{t-\ell}$ denote the ℓ-th

differences of a general time series $\{Z_t\}$. Our estimation method relies on the following simple observation: If $\{\varepsilon_t\}$ is an AR(p^*) process of the form (4.3), then the time series $\{\Delta_q\varepsilon_t\}$ of the differences $\Delta_q\varepsilon_t = \varepsilon_t - \varepsilon_{t-q}$ is an ARMA(p^*, q) process of the form

$$\Delta_q\varepsilon_t - \sum_{j=1}^{p^*} a_j \Delta_q\varepsilon_{t-j} = \eta_t - \eta_{t-q}. \tag{4.4}$$

As m is Lipschitz, the differences $\Delta_q\varepsilon_t$ of the unobserved error process are close to the differences $\Delta_q Y_{t,T}$ of the observed time series in the sense that

$$\Delta_q Y_{t,T} = \big[\varepsilon_t - \varepsilon_{t-q}\big] + \Big[m\Big(\frac{t}{T}\Big) - m\Big(\frac{t-q}{T}\Big)\Big] = \Delta_q\varepsilon_t + O\Big(\frac{q}{T}\Big). \tag{4.5}$$

Taken together, (4.4) and (4.5) imply that the differenced time series $\{\Delta_q Y_{t,T}\}$ is approximately an ARMA(p^*, q) process of the form (4.4). It is precisely this point which is exploited by our estimation methods.

We first describe our procedure to estimate the AR parameters a_j. For any $q \geq 1$, the ARMA(p^*, q) process $\{\Delta_q\varepsilon_t\}$ satisfies the Yule-Walker equations

$$\gamma_q(\ell) - \sum_{j=1}^{p^*} a_j \gamma_q(\ell - j) = -\nu^2 c_{q-\ell} \qquad \text{for } 1 \leq \ell < q+1 \tag{4.6}$$

$$\gamma_q(\ell) - \sum_{j=1}^{p^*} a_j \gamma_q(\ell - j) = 0 \qquad \text{for } \ell \geq q+1, \tag{4.7}$$

where $\gamma_q(\ell) = \mathrm{Cov}(\Delta_q\varepsilon_t, \Delta_q\varepsilon_{t-\ell})$ and c_k are the coefficients from the MA(∞) expansion of $\{\varepsilon_t\}$. Let $p = p_T \in \mathbb{N}$ grow with the sample size T. The precise conditions on the growth of $p = p_T$ are given below. Combining (4.6)–(4.7) for $\ell = 1, \ldots, p$, we get that

$$\boldsymbol{\Gamma}_q \boldsymbol{a} = \boldsymbol{\gamma}_q + \nu^2 \boldsymbol{c}_q - \boldsymbol{\rho}_q, \tag{4.8}$$

where $\boldsymbol{a} = (a_1, \ldots, a_p)^\top$, $\boldsymbol{\gamma}_q = (\gamma_q(1), \ldots, \gamma_q(p))^\top$ and $\boldsymbol{\Gamma}_q$ denotes the $p \times p$ covariance matrix $\boldsymbol{\Gamma}_q = (\gamma_q(i-j) : 1 \leq i, j \leq p)$. Moreover, $\boldsymbol{c}_q = (c_{q-1}, \ldots, c_{q-p})^\top$ and $\boldsymbol{\rho}_q = (\rho_q(1), \ldots, \rho_q(p))^\top$ with $\rho_q(\ell) = \sum_{j=p+1}^{p^*} a_j \gamma_q(\ell - j)$. Since the AR coefficients a_j as well as the MA coefficients c_k decay exponentially fast to zero, $\boldsymbol{\rho}_q \approx \boldsymbol{0}$ and $\boldsymbol{c}_q \approx \boldsymbol{0}$ for large values of q, implying that $\boldsymbol{\Gamma}_q \boldsymbol{a} \approx \boldsymbol{\gamma}_q$. This suggests to estimate $\boldsymbol{a}$ by

$$\widetilde{\boldsymbol{a}}_q = \widehat{\boldsymbol{\Gamma}}_q^{-1} \widehat{\boldsymbol{\gamma}}_q, \tag{4.9}$$

where $\widehat{\boldsymbol{\Gamma}}_q$ and $\widehat{\boldsymbol{\gamma}}_q$ are defined analogously as $\boldsymbol{\Gamma}_q$ and $\boldsymbol{\gamma}_q$ with $\gamma_q(\ell)$ replaced by the sample autocovariances $\widehat{\gamma}_q(\ell) = (T-q)^{-1} \sum_{t=q+\ell+1}^{T} \Delta_q Y_{t,T} \Delta_q Y_{t-\ell,T}$ and $q = q_T$ goes to infinity as $T \to \infty$. We impose the following formal conditions on the growth of $q = q_T$ and

$p = p_T$:

$$C \log T \leq p \ll \min\{q, T^{1/4}\} \quad \text{and} \quad \log T \ll q \ll \sqrt{T} \tag{4.10}$$

for some sufficiently large constant C, where the symbol $v_T \ll w_T$ means that $v_T/w_T \to 0$ as $T \to \infty$. As already mentioned, if the true AR order p^* is finite and known, we of course do not have to let $p = p_T$ go to infinity but can simple set $p = p^*$.

The estimator $\widetilde{\boldsymbol{a}}_q$ depends on the tuning parameter q, that is, on the order of the differences $\Delta_q Y_{t,T}$. An appropriate choice of q needs to take care of the following two points: (i) q should be chosen large enough to ensure that the vector $\boldsymbol{c}_q = (c_{q-1}, \ldots, c_{q-p})^\top$ is close to zero. As we have already seen, the constants c_k decay exponentially fast to zero and can be computed from the recursive equations (4.2) for given AR parameters $a_1, a_2, a_3, \ldots$ In the special case of an AR(1) process, for example, one can readily calculate that $c_k \leq 0.0035$ for any $k \geq 20$ and any $|a_1| \leq 0.75$. Hence, if we have an AR(1) model for the errors ε_t and the error process is not too persistent, choosing q such that $q \geq 20$ should make sure that $\boldsymbol{c}_q$ is close to zero. Generally speaking, the recursive equations (4.2) can be used to get some idea for which values of q the vector $\boldsymbol{c}_q$ can be expected to be approximately zero. (ii) q should not be chosen too large in order to ensure that the trend m is appropriately eliminated by taking q-th differences. As long as the trend m is not very strong, the two requirements (i) and (ii) can be fulfilled without much difficulty. For example, by choosing $q = 20$ in the AR(1) case just discussed, we do not only take care of (i) but also make sure that moderate trends m are differenced out appropriately.

When the trend m is very pronounced, in contrast, even moderate values of q may be too large to eliminate the trend appropriately. As a result, the estimator $\widetilde{\boldsymbol{a}}_q$ will have a strong bias. In order to reduce this bias, we refine our estimation procedure as follows: By solving the recursive equations (4.2) with $\boldsymbol{a}$ replaced by $\widetilde{\boldsymbol{a}}_q$, we can compute estimators $\widetilde{c}_k$ of the coefficients c_k and thus estimators $\widetilde{\boldsymbol{c}}_r$ of the vectors $\boldsymbol{c}_r$ for any $r \geq 1$. Moreover, the innovation variance ν^2 can be estimated by $\widetilde{\nu}^2 = (2T)^{-1} \sum_{t=p+2}^{T} \widetilde{r}_{t,T}^2$, where $\widetilde{r}_{t,T} = \Delta_1 Y_{t,T} - \sum_{j=1}^{p} \widetilde{a}_j \Delta_1 Y_{t-j,T}$ and $\widetilde{a}_j$ is the j-th entry of the vector $\widetilde{\boldsymbol{a}}_q$. Plugging the expressions $\widehat{\boldsymbol{\Gamma}}_r$, $\widehat{\boldsymbol{\gamma}}_r$, $\widetilde{\boldsymbol{c}}_r$ and $\widetilde{\nu}^2$ into (4.8), we can estimate $\boldsymbol{a}$ by

$$\widehat{\boldsymbol{a}}_r = \widehat{\boldsymbol{\Gamma}}_r^{-1} (\widehat{\boldsymbol{\gamma}}_r + \widetilde{\nu}^2 \widetilde{\boldsymbol{c}}_r), \tag{4.11}$$

where r is any fixed number with $r \geq 1$. In particular, unlike q, the parameter r does not diverge to infinity but remains fixed as the sample size T increases. As one can see, the estimator $\widehat{\boldsymbol{a}}_r$ is based on differences of some small order r; only the pilot estimator $\widetilde{\boldsymbol{a}}_q$ relies on differences of a larger order q. As a consequence, $\widehat{\boldsymbol{a}}_r$ should eliminate the trend m more appropriately and should thus be less biased than the pilot estimator $\widetilde{\boldsymbol{a}}_q$. In order to make the method more robust against estimation errors in $\widetilde{\boldsymbol{c}}_r$, we finally

average the estimators $\widehat{\boldsymbol{a}}_r$ for a few small values of r. In particular, we define

$$\widehat{\boldsymbol{a}} = \frac{1}{\overline{r}} \sum_{r=1}^{\overline{r}} \widehat{\boldsymbol{a}}_r, \tag{4.12}$$

where $\overline{r}$ is a small natural number. For ease of notation, we suppress the dependence of $\widehat{\boldsymbol{a}}$ on the parameter $\overline{r}$. Once $\widehat{\boldsymbol{a}} = (\widehat{a}_1, \ldots, \widehat{a}_p)^\top$ is computed, the long-run variance σ^2 can be estimated by

$$\widehat{\sigma}^2 = \frac{\widehat{\nu}^2}{(1 - \sum_{j=1}^{p} \widehat{a}_j)^2}, \tag{4.13}$$

where $\widehat{\nu}^2 = (2T)^{-1} \sum_{t=p+2}^{T} \widehat{r}_{t,T}^2$ with $\widehat{r}_{t,T} = \Delta_1 Y_{t,T} - \sum_{j=1}^{p} \widehat{a}_j \Delta_1 Y_{t-j,T}$ is an estimator of the innovation variance ν^2 and we make use of the fact that $\sigma^2 = \nu^2/(1 - \sum_{j=1}^{p^*} a_j)^2$ for the AR(p^*) process $\{\varepsilon_t\}$.

We briefly compare the estimator $\widehat{\boldsymbol{a}}$ to competing methods. Presumably closest to our approach is the procedure of Hall and Van Keilegom (2003) which is designed for AR(p^*) error processes of known finite order p^*. For comparing the two methods, we thus assume $\{\varepsilon_t\}$ to be an AR(p^*) process of known finite order p^*. The two main advantages of our method are as follows:

(a) Our estimator produces accurate estimation results even when the AR process $\{\varepsilon_t\}$ is quite persistent, that is, even when the AR polynomial $A(z) = 1 - \sum_{j=1}^{p^*} a_j z^j$ has a root close to the unit circle. The estimator of Hall and Van Keilegom (2003), in contrast, may have very high variance and may thus produce unreliable results when the AR polynomial $A(z)$ is close to having a unit root. This difference in behaviour can be explained as follows: Our pilot estimator $\widetilde{\boldsymbol{a}}_q = (\widetilde{a}_1, \ldots, \widetilde{a}_{p^*})^\top$ has the property that the estimated AR polynomial $\widetilde{A}(z) = 1 - \sum_{j=1}^{p^*} \widetilde{a}_j z^j$ has no root inside the unit disc, that is, $\widetilde{A}(z) \neq 0$ for all complex numbers z with $|z| \leq 1$.[5] Hence, the fitted AR model with the coefficients $\widetilde{\boldsymbol{a}}_q$ is ensured to be stationary and causal. Even though this may seem to be a minor technical detail, it has a huge effect on the performance of the estimator: It keeps the estimator stable even when the AR process is very persistent and the AR polynomial $A(z)$ has almost a unit root. This in turn results in a reliable behaviour of the estimator $\widehat{\boldsymbol{a}}$ in the case of high persistence. The estimator of Hall and Van Keilegom (2003), in contrast, may produce non-causal results when the AR polynomial $A(z)$ is close to having a unit root. As a consequence, it may have unnecessarily high variance in the case of high persistence. We illustrate this difference between the estimators by the simulation exercises in Section **??**. A striking example is Figure 9, which presents the simulation results for the case of an AR(1) process $\varepsilon_t = a_1 \varepsilon_{t-1} + \eta_t$

[5] More precisely, $\widetilde{A}(z) \neq 0$ for all z with $|z| \leq 1$, whenever the covariance matrix $(\widehat{\gamma}_q(i-j) : 1 \leq i, j \leq p^* + 1)$ is non-singular. Moreover, $(\widehat{\gamma}_q(i-j) : 1 \leq i, j \leq p^* + 1)$ is non-singular whenever $\widehat{\gamma}_q(0) > 0$, which is the generic case.

with $a_1 = -0.95$ and clearly shows the much better performance of our method.

(b) Both our pilot estimator $\widetilde{\boldsymbol{a}}_q$ and the estimator of Hall and Van Keilegom (2003) tend to have a substantial bias when the trend m is pronounced. Our estimator $\widehat{\boldsymbol{a}}$ reduces this bias considerably as demonstrated in the simulations of Section **??**. Unlike the estimator of Hall and Van Keilegom (2003), it thus produces accurate results even in the presence of a very strong trend.

We close this section by deriving some basic asymptotic properties of the estimators $\widetilde{\boldsymbol{a}}_q$, $\widehat{\boldsymbol{a}}$ and $\widehat{\sigma}^2$. The following proposition specifies their convergence rates.

Proposition 4.1. *Let $\{\varepsilon_t\}$ be an AR(p^*) process of the form (4.3) with the following properties: $A(z) \neq 0$ for all $|z| \leq 1 + \delta$ with some small $\delta > 0$ and the innovations η_t have a finite fourth moment. Moreover, let m be Lipschitz continuous. If $q = q_T$ and $p = p_T$ satisfy (4.10), then $\widetilde{\boldsymbol{a}}_q - \boldsymbol{a} = O_p(\sqrt{p^2/T})$ as well as $\widehat{\boldsymbol{a}} - \boldsymbol{a} = O_p(\sqrt{p^3/T})$ and $\widehat{\sigma}^2 - \sigma^2 = O_p(\sqrt{p^4/T})$.*

The proof is provided in the Supplementary Material. As one can see, the convergence rate of the second-step estimator $\widehat{\boldsymbol{a}}$ is somewhat slower than that of the pilot estimator $\widetilde{\boldsymbol{a}}_q$. Hence, from an asymptotic perspective, there is no gain from using the second-step estimator. Nevertheless, in finite samples, the estimator $\widehat{\boldsymbol{a}}$ vastly outperforms $\widetilde{\boldsymbol{a}}_q$ as illustrated by our simulations in Section 5.4.

5 Simulations

5.1 Small sample properties of the multiscale test

In this section, we investigate the performance of our multiscale test and compare it to the dependent SiZer methods from Park et al. (2004), Rondonotti et al. (2007) and Park et al. (2009). We consider the following versions of our multiscale test and SiZer:

$\mathcal{T}_{\mathrm{MS}}$: our multiscale test with the statistic $\widehat{\Psi}_T = \max_{h \in H_T}\{\widehat{\Psi}_T(h) - \lambda(h)\}$, where $\widehat{\Psi}_T(h) = \max_{u \in U_T} |\widehat{\psi}_T(u,h)/\widehat{\sigma}|$. Here and in what follows, we write $\mathcal{G}_T = U_T \times H_T$, where U_T is the set of locations and H_T the set of bandwidths.

$\mathcal{T}_{\mathrm{UC}}$: the uncorrected version of our multiscale test with the test statistic $\widehat{\Psi}_{T,\mathrm{uncorrected}} = \max_{h \in H_T} \widehat{\Psi}_T(h)$, which was already introduced in (3.3). The uncorrected test is carried out in exactly the same way as $\mathcal{T}_{\mathrm{MS}}$. The only difference is that the correction terms $\lambda(h)$ are removed.

$\mathcal{T}_{\mathrm{RW}}$: the row-wise (that is, scale-wise or bandwidth-wise) version of our multiscale test as briefly mentioned in Section 3.4. This version carries out a test for each scale $h \in H_T$ separately based on the statistic $\widehat{\Psi}_T(h)$. Note: (i) For each $h \in H_T$,

the test based on $\widehat{\Psi}_T(h)$ can be performed in the same way as the multiscale test $\mathcal{T}_{\mathrm{MS}}$, since it is a degenerate version of the latter with the set of scales H_T replaced by the singleton $\{h\}$. (ii) It does not matter whether we correct the statistic $\widehat{\Psi}_T(h)$ by subtracting $\lambda(h)$ or not, since $\widehat{\Psi}_T(h)$ acts as a fixed constant if only one bandwidth h is taken into account.

$\mathcal{T}_{\mathrm{SiZer}}$: the row-wise version of dependent SiZer as developed in Park et al. (2004), Rondonotti et al. (2007) and Park et al. (2009). We do not consider a global SiZer version since such a version was not introduced in the aforementioned papers.

The simulation setup is as follows: We generate data from the model $Y_{t,T} = m(t/T)+\varepsilon_t$ for different trends m, error processes $\{\varepsilon_t\}$ and sample sizes T. The error terms are supposed to have the AR(1) structure $\varepsilon_t = a_1\varepsilon_{t-1} + \eta_t$, where $a_1 \in \{-0.9, -0.5, -0.25, 0.25, 0.5, 0.9\}$, η_t are i.i.d. standard normal and the AR order $p^* = 1$ is treated as known. To simulate data under the null H_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 different non-constant trend functions which are specified below.
To implement our multiscale test $\mathcal{T}_{\mathrm{MS}}$, we choose K to be an Epanechnikov kernel and set $\mathcal{G}_T = U_T \times H_T$, where $U_T = \{u : u = 5k/T \text{ for some } 1 \leq k \leq T/5\}$ and $H_T = \{h : h = 5\ell/T \text{ for some } 0 \leq \ell \leq T/20\}$. We thus take into account all rescaled time points u on an equidistant grid U_T with step length $5/T$. For the bandwidth $h = 5\ell/T$ and any $u \in [h, 1-h]$, the kernel window is $[u-h, u+h]$ and contains 10ℓ observations. Hence, the bandwidths $h \in H_T$ correspond to effective sample sizes of $10, 20, 30, \ldots$ up to approximately $T/4$ data points. As a robustness check, we have re-run the simulations for a number of other grids. As the results are very similar, we do however not report them here. To estimate the long-run error variance σ^2, we apply the procedure from in Section 4, where we set $q = 25$ and $\overline{r} = 10$. As already discussed in Section 4, $q = 25$ should be an appropriate value for AR(1) errors that are not too strongly correlated, in particular, for $a_1 \in \{-0.5, -0.25, 0.25, 0.5\}$. When the errors are expected to be very strongly correlated, larger values of q are required to produce precise estimates of σ^2. In the case of AR errors with $a_1 \in \{-0.9, 0.9\}$, we thus set $q = 50$. The dependence of our long-run variance estimator on the tuning parameters q and $\overline{r}$ is explored more systematically in Section 5.4. To compute the critical values of the multiscale test $\mathcal{T}_{\mathrm{MS}}$, we simulate 1000 values of the statistic Φ_T defined in Section 3.2 and compute their empirical $(1-\alpha)$ quantile $q_T(\alpha)$. The versions $\mathcal{T}_{\mathrm{UC}}$ and $\mathcal{T}_{\mathrm{RW}}$ of our multiscale test are implemented analogously. The SiZer test is implemented as described in Park et al. (2009). The details are summarized in Section S.3 of the Supplementary Material. To compute our simulation results, we generate $S = 1000$ samples for each model specification and carry out the tests $\mathcal{T}_{\mathrm{MS}}$, $\mathcal{T}_{\mathrm{UC}}$, $\mathcal{T}_{\mathrm{RW}}$ and $\mathcal{T}_{\mathrm{SiZer}}$ for each sample.

Table 1: Size of $\mathcal{T}_{\text{MS}}$ for the AR parameters $a_1\{-0.5, -0.25, 0.25, 0.5\}$.

	$a_1 = -0.5$			$a_1 = -0.25$			$a_1 = 0.25$			$a_1 = 0.5$		
	nominal size α			nominal size α			nominal size α			nominal size α		
	0.01	0.05	0.1	0.01	0.05	0.1	0.01	0.05	0.1	0.01	0.05	0.1
$T = 250$	0.015	0.049	0.117	0.009	0.052	0.111	0.013	0.047	0.110	0.017	0.048	0.109
$T = 500$	0.008	0.056	0.128	0.017	0.047	0.112	0.014	0.050	0.100	0.014	0.048	0.107
$T = 1000$	0.007	0.053	0.090	0.014	0.059	0.106	0.014	0.056	0.099	0.017	0.057	0.099

Table 2: Size of $\mathcal{T}_{\text{MS}}$ for the AR parameters $a_1 \in \{-0.9, 0.9\}$.

	$a_1 = -0.9$							$a_1 = 0.9$						
	sample size T							sample size T						
	250	500	1000	2000	3000	4000	5000	250	500	1000	2000	3000	4000	5000
$\alpha = 0.01$	0.057	0.033	0.032	0.014	0.011	0.018	0.011	0.004	0.012	0.023	0.015	0.019	0.019	0.020
$\alpha = 0.05$	0.134	0.098	0.091	0.055	0.046	0.057	0.049	0.016	0.040	0.076	0.062	0.050	0.057	0.060
$\alpha = 0.1$	0.233	0.193	0.156	0.109	0.103	0.108	0.095	0.039	0.079	0.103	0.120	0.116	0.111	0.113

5.1.1 Size properties of the multiscale test $\mathcal{T}_{\text{MS}}$

In the first part of our simulation study, we focus on the multiscale test $\mathcal{T}_{\text{MS}}$ and explore its size properties under the null that the trend m is constant. Table 1 presents the actual size for the AR parameters $a_1 \in \{-0.5, -0.25, 0.25, 0.5\}$. The size entries in this and the following tables are computed as the number of simulations in which the test rejects divided by the total number of simulations. The long-run variance σ^2 needed to compute the test statistic of $\mathcal{T}_{\text{MS}}$ is estimated by the procedure from Section 4, where we set $q = 25$ and $\overline{r} = 10$. As a robustness check, we have re-run the simulations for other choices of q and $\overline{r}$, which yields very similar results. The dependence of the estimators of our long-run variance estimator on q and $\overline{r}$ is further explored in Section 5.4. Inspecting Table 11, the actual size of the test can be seen to be fairly close to the nominal target α for all the considered AR parameters and sample sizes. Hence, the test has approximately the correct size.

In Table 1, we have explored the size of $\mathcal{T}_{\text{MS}}$ when the error terms are moderately autocorrelated. The case of very strong autocorrelated errors is investigated in Table 2, where we consider AR errors with $a_1 \in \{-0.9, 0.9\}$. Inspecting the table, the size numbers can be seen to stabilize around their target α for sample sizes $T \geq 2000$. For smaller sample sizes, there are considerable size distortions which are worst for $T = 250$ and slowly disappear as T increases. Hence, in the case of strongly autocorrelated errors, our multiscale test has good size properties only for sufficiently large sample sizes. This is not very surprising: Statistical inference in the presence of strongly autocorrelated data is a very difficult problem in general and satisfying results can only be expected for fairly large sample sizes.

Table 3: Comparison of size performance of $\mathcal{T}_{\mathrm{MS}}$, $\mathcal{T}_{\mathrm{UC}}$, $\mathcal{T}_{\mathrm{RW}}$ and $\mathcal{T}_{\mathrm{SiZer}}$

	$a_1 = -0.5$				$a_1 = 0.5$			
	$\mathcal{T}_{\mathrm{MS}}$	$\mathcal{T}_{\mathrm{UC}}$	$\mathcal{T}_{\mathrm{RW}}$	$\mathcal{T}_{\mathrm{SiZer}}$	$\mathcal{T}_{\mathrm{MS}}$	$\mathcal{T}_{\mathrm{UC}}$	$\mathcal{T}_{\mathrm{RW}}$	$\mathcal{T}_{\mathrm{SiZer}}$
$T = 250$	0.064	0.080	0.305	0.379	0.050	0.037	0.121	0.310
$T = 500$	0.055	0.072	0.351	0.445	0.047	0.042	0.180	0.394
$T = 1000$	0.059	0.076	0.413	0.552	0.046	0.042	0.232	0.491

Insert figure.

Figure 2: Parallel coordinate plots for row-wise size comparisons.

5.1.2 Size comparisons

We now compare the multiscale test $\mathcal{T}_{\mathrm{MS}}$ with the versions $\mathcal{T}_{\mathrm{UC}}$ and $\mathcal{T}_{\mathrm{RW}}$ and the SiZer test $\mathcal{T}_{\mathrm{SiZer}}$ in terms of size. To keep the comparison study to a reasonable length, we restrict attention to a subset of the parameters T, α and a_1. In particular, we focus on the sample sizes $T \in \{250, 500, 1000\}$, the significance level $\alpha = 0.05$ and the AR parameters $a_1 \in \{-0.5, 0.5\}$. To simplify the implementation of the SiZer test $\mathcal{T}_{\mathrm{SiZer}}$, we assume that the autocovariance function $\gamma_\varepsilon(\cdot)$ of the error process and thus the long-run error variance σ^2 is known. To keep the comparison fair, we treat σ^2 as known also when implementing $\mathcal{T}_{\mathrm{MS}}$, $\mathcal{T}_{\mathrm{UC}}$ and $\mathcal{T}_{\mathrm{RW}}$. Moreover, we use exactly the same grid $\mathcal{G}_T$ for all four methods. To achieve this, we start off with the grid $\mathcal{G}_T = U_T \times H_T$ with U_T and H_T defined above. We then follow Rondonotti et al. (2007) and Park et al. (2009) and restrict attention to those points $(u, h) \in \mathcal{G}_T$ for which the effective sample size $\mathrm{ESS}^*(u, h)$ for correlated data is not smaller than 5. This yields the grid $\mathcal{G}_T^* = \{(u, h) \in \mathcal{G}_T : \mathrm{ESS}^*(u, h) \geq 5\}$. A definition of the effective sample size $\mathrm{ESS}^*(u, h)$ is given in Section S.3 of the Supplement.

Table 3 reports the actual size of the four test procedures $\mathcal{T}_{\mathrm{MS}}$, $\mathcal{T}_{\mathrm{UC}}$, $\mathcal{T}_{\mathrm{RW}}$ and $\mathcal{T}_{\mathrm{SiZer}}$. As can be seen, the size produced by our multiscale test $\mathcal{T}_{\mathrm{MS}}$ and its uncorrected version $\mathcal{T}_{\mathrm{UC}}$ is fairly close to the target $\alpha = 0.05$. The two row-wise procedures $\mathcal{T}_{\mathrm{RW}}$ and $\mathcal{T}_{\mathrm{SiZer}}$, in contrast, are much too liberal, having an actual size much larger than thr target $\alpha = 0.05$. This is of course not surprising: Both $\mathcal{T}_{\mathrm{RW}}$ and $\mathcal{T}_{\mathrm{SiZer}}$ are not rigorous level-α-tests of the overall null H_0. In particular, they are not designed to control the size α uniformly over both locations $u \in U$ and scales $h \in H_T$. Rather, they are calibrated to control the size α separately for each scale $h \in H_T$. This is illustrated by Figure ?? which is produced as follows: For each test $\mathcal{T}_j$ with $j \in \{\mathrm{MS}, \mathrm{UC}, \mathrm{RW}, \mathrm{SiZer}\}$, we compute the actual size separately for each scale $h \in H_T$. In particular, for each $h \in H_T$, we calculate the percentage of simulation runs in which $\mathcal{T}_j$ rejects $H_0(u, h)$ for some $u \in U_T$. We thus obtain a curve for each $\mathcal{T}_j$ which specifies the actual size as a function of h. The resulting curves are plotted in Figure ?? for $a_1 = -0.5$ in panel (a) and for $a_1 = 0.5$ in panel (b). As one can see, the row-wise (scale-wise) version $\mathcal{T}_{\mathrm{RW}}$ of

Table 4: Global power comparison for the bump signal

	$a_1 = -0.5$				$a_1 = 0.5$			
	$\mathcal{T}_{\text{MS}}$	$\mathcal{T}_{\text{UC}}$	$\mathcal{T}_{\text{RW}}$	$\mathcal{T}_{\text{SiZer}}$	$\mathcal{T}_{\text{MS}}$	$\mathcal{T}_{\text{UC}}$	$\mathcal{T}_{\text{RW}}$	$\mathcal{T}_{\text{SiZer}}$
Power	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
Spurious power	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000

our test holds the size quite accurately across scales h. The size of the row-wise SiZer test $\mathcal{T}_{\text{SiZer}}$ is also reasonably close to the target across cales h. However, it appears that $\mathcal{T}_{SiZer}$ is a bit more liberal than $\mathcal{T}_{\text{RW}}$, its size lying somewhat above the target of $\alpha = 0.05$. In contrast to the two row-wise procedures, our multiscale test $\mathcal{T}_{\text{MS}}$ and $\mathcal{T}_{\text{UC}}$ have a row-wise size much smaller than $\alpha = 0.05$. This is compeltely nmatural as they control the size globally, that is, simultaneously across scales h.

5.1.3 Power comparisons

In the final part of our simulation study, we compare the multiscale test $\mathcal{T}_{\text{MS}}$ with the versions $\mathcal{T}_{\text{UC}}$ and $\mathcal{T}_{\text{RW}}$ and the SiZer test $\mathcal{T}_{\text{SiZer}}$ in terms of power. As for the size comparisons, we focus on a subset of the parameters T, α and a_1. In particular, we fix $T = 500$ and $\alpha = 0.05$ and consider the AR parameters $a_1 \in \{-0.5, 0.5\}$. We first consider a very simple trend function m, which allows us to make systematic power comparisons. The function is given by $m(u) = 0.5 \cdot 1(u \in [0.45, 0.55]) \cdot (1 - \{\frac{(u-0.5}{0.05}\}^2)^2$. A graphical illustrateion is given in ??. As one can see, the function is constantly eqaul to zero on $[0, 1]$ except for a small region around $u = 0.5$, where it is a sharp bump. We make the following power comparisons for this signal: For each test, we compute the

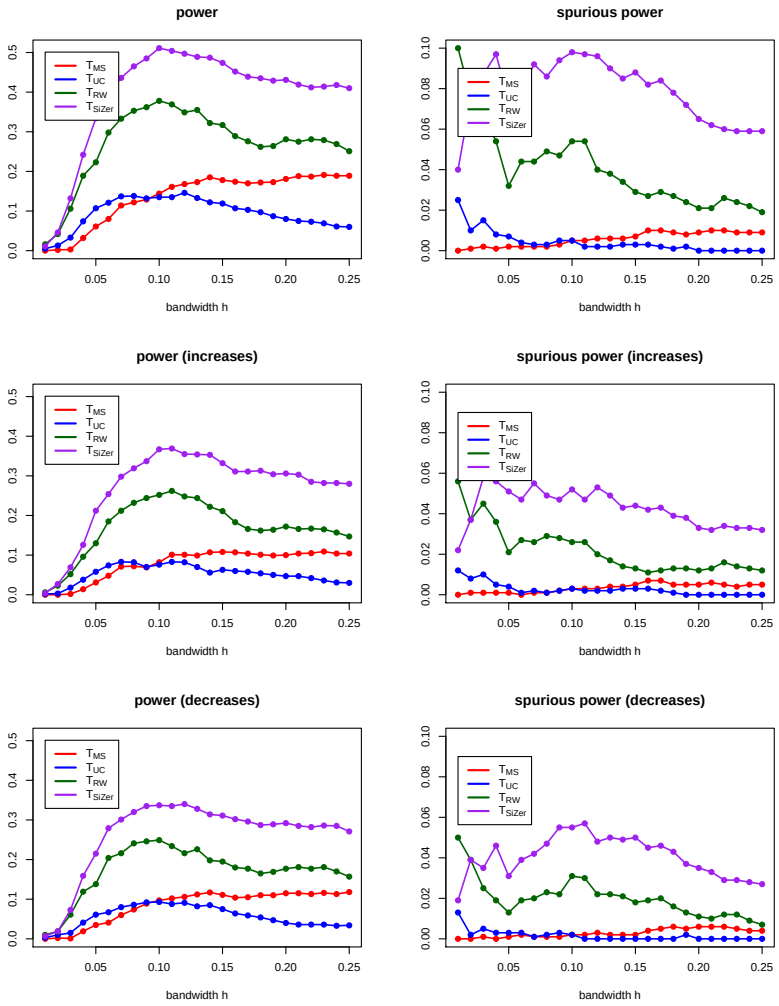


Figure 3: Rowwise power comparison for bump signal.

Tables 11 and 12 report the simulation results for the sample sizes $T = 250, 350, 500$ 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 6. To produce our simulation results, we generate $S = 1000$ samples for each model specification and carry out the multiscale test for each sample. The entries of Tables 11 and 12 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 11, the actual size of the test is fairly close to the nominal target α for all the considered AR specifications and sample sizes. Hence, the test has approximately the correct size. Inspecting Table 12, one can further see that the test has reasonable power properties. For all the considered AR specifications, the power increases quickly (i) as the sample size gets larger and (ii) as we move away from the null by increasing the slope parameter β. The power is of course quite different across the various AR specifications. In particular, it is much lower for positive than for negative values of a_1 in the AR(1) case, the lowest power numbers being obtained for the largest positive value $a_1 = 0.5$ under consideration. This reflects the fact that it

is more difficult to detect a trend when there is strong positive autocorrelation in the data. For the AR(2) specification of the errors, the sample size $T = 350$ and the slopes $\beta = 2.0$ and $\beta = 2.5$, which yield the two model specifications that resemble the real-life data in Section 6 the most, the power of the test is above 92% for the significance levels $\alpha = 0.05$ and $\alpha = 0.1$ and above 75% for $\alpha = 0.01$. Hence, our method has substantial power in the two simulation scenarios which are closest to the situation in the application.

(1) Size properties of our multiscale test $\mathcal{T}_{\mathrm{MS}}$.

Table 5: Size of $\mathcal{T}_{\mathrm{MS}}$ for different AR parameters a_1, sample sizes T and nominal sizes α for the estimated long-run variance.

	$a_1 = -0.5$			$a_1 = -0.25$			$a_1 = 0.25$			$a_1 = 0.5$		
	nominal size α			nominal size α			nominal size α			nominal size α		
	0.01	0.05	0.1	0.01	0.05	0.1	0.01	0.05	0.1	0.01	0.05	0.1
$T = 250$	0.015	0.049	0.117	0.009	0.052	0.111	0.013	0.047	0.110	0.017	0.048	0.109
$T = 500$	0.008	0.056	0.128	0.017	0.047	0.112	0.014	0.050	0.100	0.014	0.048	0.107
$T = 1000$	0.007	0.053	0.090	0.014	0.059	0.106	0.014	0.056	0.099	0.017	0.057	0.099

Table 6: Size of $\mathcal{T}_{\mathrm{MS}}$ for different AR parameters a_1, sample sizes T and nominal sizes α for the true long-run variance.

	$a_1 = -0.5$			$a_1 = -0.25$			$a_1 = 0.25$			$a_1 = 0.5$		
	nominal size α			nominal size α			nominal size α			nominal size α		
	0.01	0.05	0.1	0.01	0.05	0.1	0.01	0.05	0.1	0.01	0.05	0.1
$T = 250$	0.014	0.053	0.107	0.013	0.049	0.104	0.014	0.046	0.093	0.011	0.043	0.082
$T = 500$	0.014	0.049	0.103	0.014	0.049	0.099	0.014	0.047	0.096	0.013	0.044	0.091
$T = 1000$	0.010	0.050	0.105	0.010	0.050	0.107	0.010	0.044	0.105	0.009	0.043	0.097

Table 7: Size of $\mathcal{T}_{\mathrm{MS}}$ for different AR parameters a_1, sample sizes T and nominal sizes α for the estimated long-run variance.

	$a_1 = -0.9$							$a_1 = 0.9$						
	sample size T							sample size T						
	250	500	1000	2000	3000	4000	5000	250	500	1000	2000	3000	4000	5000
$\alpha = 0.01$	0.057	0.033	0.032	0.014	0.011	0.018	0.011	0.004	0.012	0.023	0.015	0.019	0.019	0.020
$\alpha = 0.05$	0.134	0.098	0.091	0.055	0.046	0.057	0.049	0.016	0.040	0.076	0.062	0.050	0.057	0.060
$\alpha = 0.1$	0.233	0.193	0.156	0.109	0.103	0.108	0.095	0.039	0.079	0.103	0.120	0.116	0.111	0.113

(2) Comparison of the multiscale test $\mathcal{T}_{\mathrm{MS}}$ with its uncorrected version $\mathcal{T}_{\mathrm{UC}}$, its row-wise version $\mathcal{T}_{\mathrm{RW}}$ and (row-wise) SiZer $\mathcal{T}_{\mathrm{SiZer}}$.

For the comparison, we focus on the significance level $\alpha = 0.05$ and on the AR parameters $a_1 \in \{-0.5, 0.5\}$. (Maybe report results for other AR parameters and significance levels in supplement?)

Table 8: Size of $\mathcal{T}_{\text{MS}}$ for different AR parameters a_1, sample sizes T and nominal sizes α for the true long-run variance.

	$a_1 = -0.9$							$a_1 = 0.9$						
	sample size T							sample size T						
	250	500	1000	2000	3000	4000	5000	250	500	1000	2000	3000	4000	5000
$\alpha = 0.01$	0.058	0.036	0.027	0.014	0.010	0.016	0.011	0.003	0.005	0.005	0.008	0.009	0.008	0.009
$\alpha = 0.05$	0.136	0.092	0.090	0.056	0.048	0.056	0.048	0.010	0.019	0.028	0.036	0.042	0.046	0.040
$\alpha = 0.1$	0.231	0.196	0.154	0.109	0.107	0.107	0.103	0.024	0.044	0.053	0.068	0.087	0.093	0.090

Table 9: Size of $\mathcal{T}_{\text{MS}}$ for different AR parameters a_1, sample sizes T and nominal sizes α for the true long-run variance.

	$a_1 = -0.5$				$a_1 = 0.5$			
	$\mathcal{T}_{\text{MS}}$	$\mathcal{T}_{\text{uncor}}$	$\mathcal{T}_{\text{rows}}$	$\mathcal{T}_{\text{SiZer}}$	$\mathcal{T}_{\text{MS}}$	$\mathcal{T}_{\text{uncor}}$	$\mathcal{T}_{\text{rows}}$	$\mathcal{T}_{\text{SiZer}}$
$T = 250$	0.064	0.080	0.305	0.379	0.050	0.037	0.121	0.310
$T = 500$	0.055	0.072	0.351	0.445	0.047	0.042	0.180	0.394
$T = 1000$	0.059	0.076	0.413	0.552	0.046	0.042	0.232	0.491

(a) Size comparisons:

- Comparison of global size (as in Figure 6 of Hannig & Marron (2006)).
- Comparison of row-wise size (as in Figure 7 of Hannig & Marron (2006)). (Three plots, each corresponding to one sample size $T = 250, 500, 1000$. Or one plot for all three sample sizes? Too packed?)

(b) Power comparisons:

- We first consider a version of the bump signal that we already had in the old Section 7.2 (Comparison with SiZer): $m(u) = 0.5 \cdot 1(u \in [0.45, 0.55]) \cdot (1 - \{\frac{(u-0.5}{0.05}\}^2)^2$. (This is the same bump function as in Section 7.2, but more localized and smaller, so that it is harder to detect.)

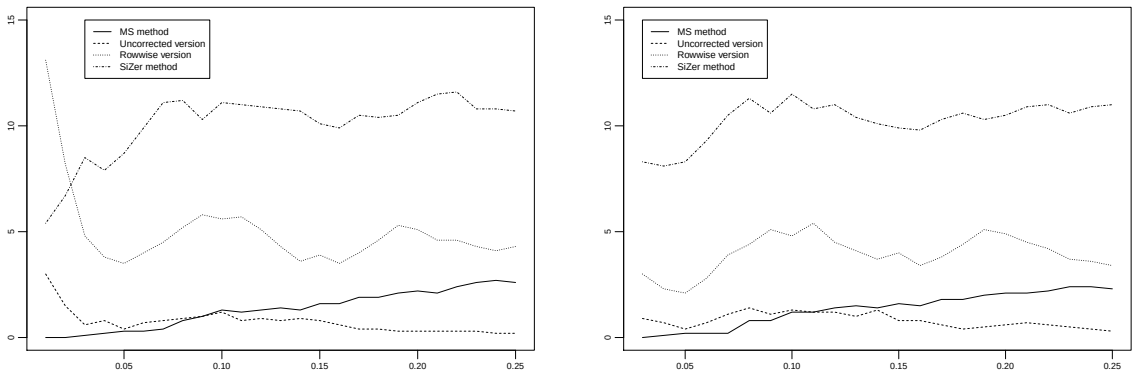
  Set $T = 500$. We make the following power comparisons for the bump signal:

  (i) comparison of global power and global spurious power reported in a table of the same form as Table 10.

Table 10: Global power comparison for the bump signal

	$a_1 = -0.5$				$a_1 = 0.5$			
	$\mathcal{T}_{\text{MS}}$	$\mathcal{T}_{\text{UC}}$	$\mathcal{T}_{\text{RW}}$	$\mathcal{T}_{\text{SiZer}}$	$\mathcal{T}_{\text{MS}}$	$\mathcal{T}_{\text{UC}}$	$\mathcal{T}_{\text{RW}}$	$\mathcal{T}_{\text{SiZer}}$
Power	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
Spurious power	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000

(ii) comparison of rowwise power reported in a plot of the same format as Figure 7 of Hannig & Marron (2006) / Figure 5 below.



(a) $a_1 = -0.50$ (b) $a_1 = 0.50$

Figure 4: Rowwise power comparison

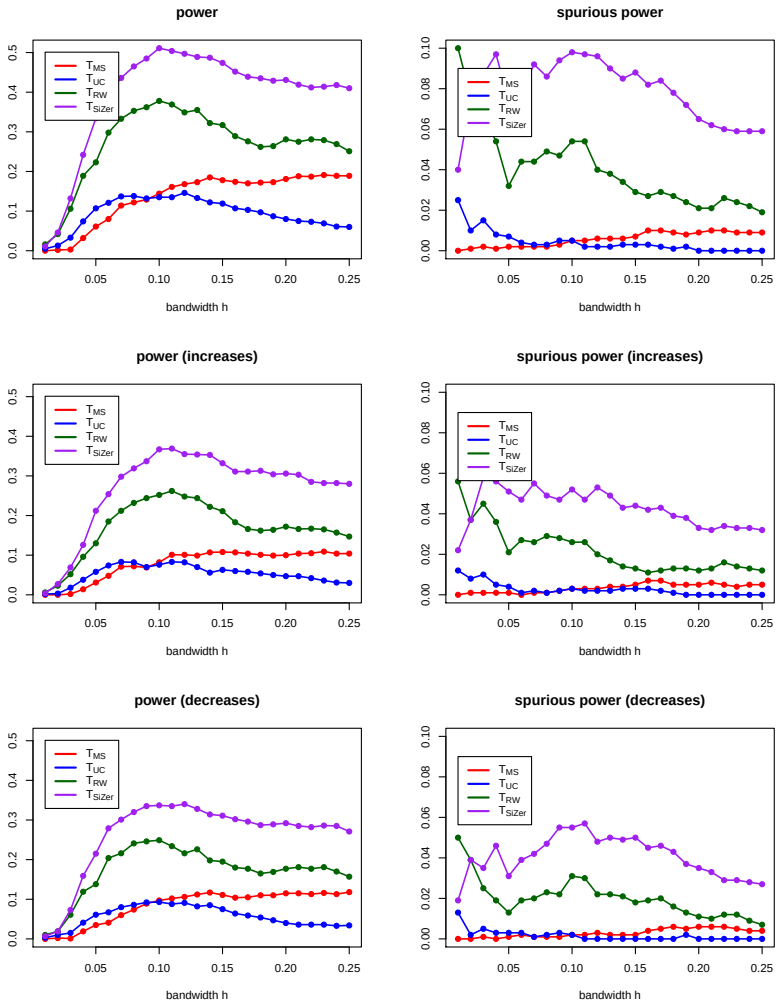


Figure 5: Rowwise power comparison for bump signal.

(c) Power comparisons by means of "more interesting" signals (e.g. the blocks signal of Donoho): plot of SiZer maps and minimal intervals.

5.2 Size and power properties of the multiscale test

Our simulation design mimics the situation in the application example of Section 6. We generate data from the model $Y_{t,T} = m(t/T) + \varepsilon_t$ for different trend functions m, error processes $\{\varepsilon_t\}$ and time series lengths T. The error terms are supposed to have the AR(1) structure $\varepsilon_t = a_1\varepsilon_{t-1} + \eta_t$, where $a_1 \in \{-0.5, -0.25, 0.25, 0.5\}$ and η_t are i.i.d. standard normal. In addition, we consider the AR(2) specification $\varepsilon_t = a_1\varepsilon_{t-1} + a_2\varepsilon_{t-2} + \eta_t$, where η_t are normally distributed with $\mathbb{E}[\eta_t] = 0$ and $\mathbb{E}[\eta_t^2] = \nu^2$. We set $a_1 = 0.167$, $a_2 = 0.178$ and $\nu^2 = 0.322$, thus matching the estimated values obtained in the application of Section 6. To simulate data under the null hypothesis, 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 $m(u) = \beta(u - 0.5) \cdot 1(0.5 \leq u \leq 1)$ with $\beta = 1.5, 2.0, 2.5$. These functions are broken lines with a kink at $u = 0.5$ and different slopes β. Their shape roughly resembles the trend estimates in the application of Section 6. The slope parameter β corresponds to a trend with the value $m(1) = 0.5\beta$ at the right endpoint $u = 1$. We thus consider broken lines with the values $m(1) = 0.75, 1.0, 1.25$. Inspecting the middle panel of Figure 11, the broken lines with the endpoints $m(1) = 1.0$ and $m(1) = 1.25$ (that is, with $\beta = 2.0$ and $\beta = 2.5$) can be seen to resemble the local linear trend estimates in the real-data example the most (where we neglect the nonlinearities of the local linear fits at the beginning of the observation period). The broken line with $\beta = 1.5$ is closer to the null, making it harder for our test to detect this alternative.[6]

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

$$\begin{aligned}\mathcal{G}_T = \big\{(u,h) : u = 5k/T \text{ for some } 1 \leq k \leq T/5 \text{ and} \\ h = (3+5\ell)/T \text{ for some } 0 \leq \ell \leq T/20\big\}.\end{aligned} \tag{5.1}$$

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 kernel weights $K(h^{-1}\{t/T - u\})$ 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 $T/4$ data points. As a robustness check, we have re-run the simulations for a number of other grids. As the results are very similar, we do however not report them here. The long-run error variance σ^2 is estimated by the procedures from Section **??**: We first compute the estimator $\widehat{\boldsymbol{a}}$ of the AR parameter(s), where we use $\overline{r} = 10$ and the pilot estimator $\widetilde{\boldsymbol{a}}_q$ with $q = 25$. Based on $\widehat{\boldsymbol{a}}$, we then compute the estimator

[6]The broken lines m are obviously non-differentiable at the kink point. We could replace them by slightly smoothed versions to satisfy the differentiability assumption that is imposed in the theoretical part of the paper. However, as this leaves the simulation results essentially unchanged but only creates additional notation, we stick to the broken lines.

Table 11: Size of our multiscale test for different AR parameters a_1 and a_2, sample sizes T and nominal sizes α.

	$a_1 = -0.5$			$a_1 = -0.25$			$a_1 = 0.25$			$a_1 = 0.5$			$(a_1, a_2) = (0.167, 0.178)$		
	nominal size α			nominal size α			nominal size α			nominal size α			nominal size α		
	0.01	0.05	0.1	0.01	0.05	0.1	0.01	0.05	0.1	0.01	0.05	0.1	0.01	0.05	0.1
$T = 250$	0.015	0.050	0.127	0.014	0.057	0.120	0.011	0.046	0.116	0.013	0.042	0.108	0.011	0.052	0.117
$T = 350$	0.009	0.067	0.120	0.010	0.055	0.095	0.009	0.055	0.096	0.010	0.049	0.090	0.010	0.059	0.114
$T = 500$	0.015	0.053	0.128	0.015	0.047	0.100	0.018	0.048	0.101	0.015	0.042	0.106	0.015	0.056	0.107

Table 12: Power of our multiscale test for different AR parameters a_1 and a_2, sample sizes T and nominal sizes α. The three panels (a)–(c) corresponds to different slope parameters β of the broken line m.

(a) $\beta = 1.5$

	$a_1 = -0.5$			$a_1 = -0.25$			$a_1 = 0.25$			$a_1 = 0.5$			$(a_1, a_2) = (0.167, 0.178)$		
	nominal size α			nominal size α			nominal size α			nominal size α			nominal size α		
	0.01	0.05	0.1	0.01	0.05	0.1	0.01	0.05	0.1	0.01	0.05	0.1	0.01	0.05	0.1
$T = 250$	0.484	0.726	0.853	0.319	0.548	0.702	0.077	0.177	0.324	0.036	0.097	0.181	0.269	0.460	0.612
$T = 350$	0.735	0.913	0.955	0.463	0.753	0.834	0.116	0.273	0.385	0.050	0.141	0.221	0.390	0.654	0.770
$T = 500$	0.945	0.988	0.997	0.775	0.925	0.972	0.195	0.389	0.551	0.060	0.162	0.285	0.623	0.815	0.907

(b) $\beta = 2.0$

	$a_1 = -0.5$			$a_1 = -0.25$			$a_1 = 0.25$			$a_1 = 0.5$			$(a_1, a_2) = (0.167, 0.178)$		
	nominal size α			nominal size α			nominal size α			nominal size α			nominal size α		
	0.01	0.05	0.1	0.01	0.05	0.1	0.01	0.05	0.1	0.01	0.05	0.1	0.01	0.05	0.1
$T = 250$	0.869	0.961	0.985	0.663	0.846	0.916	0.164	0.340	0.520	0.062	0.143	0.259	0.549	0.724	0.851
$T = 350$	0.979	0.997	1.000	0.863	0.969	0.986	0.262	0.483	0.615	0.092	0.231	0.334	0.759	0.922	0.958
$T = 500$	1.000	1.000	1.000	0.983	0.997	0.999	0.469	0.716	0.821	0.137	0.309	0.451	0.933	0.983	0.994

(c) $\beta = 2.5$

	$a_1 = -0.5$			$a_1 = -0.25$			$a_1 = 0.25$			$a_1 = 0.5$			$(a_1, a_2) = (0.167, 0.178)$		
	nominal size α			nominal size α			nominal size α			nominal size α			nominal size α		
	0.01	0.05	0.1	0.01	0.05	0.1	0.01	0.05	0.1	0.01	0.05	0.1	0.01	0.05	0.1
$T = 250$	0.989	1.000	1.000	0.901	0.971	0.993	0.322	0.543	0.703	0.100	0.224	0.367	0.804	0.918	0.958
$T = 350$	1.000	1.000	1.000	0.990	1.000	1.000	0.470	0.737	0.833	0.162	0.361	0.481	0.950	0.988	0.997
$T = 500$	1.000	1.000	1.000	0.999	1.000	1.000	0.773	0.919	0.968	0.285	0.473	0.649	0.994	0.999	1.000

$\widehat{\sigma}^2$ of the long-run error variance σ^2. As a further robustness check, we have re-run the simulations for other choices of the parameters q and $\overline{r}$, which yields very similar results. The dependence of the estimators $\widehat{\boldsymbol{a}}$ and $\widehat{\sigma}^2$ on q and $\overline{r}$ is further explored in Section **??**. To compute the critical values of the multiscale test, we simulate 1000 values of the statistic Φ_T defined in Section 3.2 and compute their empirical $(1-\alpha)$ quantile $q_T(\alpha)$.

Tables 11 and 12 report the simulation results for the sample sizes $T = 250, 350, 500$ 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 6. To produce our simulation results, we generate $S = 1000$ samples for each model specification and carry out the multiscale test for each sample. The entries of Tables 11 and 12 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 11, the actual size of the test is fairly close to the nominal target α for all the considered AR specifications and sample sizes. Hence, the test has approximately the correct size. Inspecting Table 12, one can further see that the test has reasonable power properties. For all the considered AR specifications, the power increases quickly (i) as the sample size gets larger and (ii) as we move away from the null by increasing the slope parameter β. The power is of course quite different across the various AR specifications. In particular, it is much lower for positive than for negative values of a_1 in the AR(1) case, the lowest power numbers being obtained for the largest positive value $a_1 = 0.5$ under consideration. This reflects the fact that it is more difficult to detect a trend when there is strong positive autocorrelation in the data. For the AR(2) specification of the errors, the sample size $T = 350$ and the slopes $\beta = 2.0$ and $\beta = 2.5$, which yield the two model specifications that resemble the real-life data in Section 6 the most, the power of the test is above 92% for the significance levels $\alpha = 0.05$ and $\alpha = 0.1$ and above 75% for $\alpha = 0.01$. Hence, our method has substantial power in the two simulation scenarios which are closest to the situation in the application.

5.3 Comparison with SiZer

We now compare our multiscale test to SiZer for times series which was developed in Park et al. (2004), Rondonotti et al. (2007) and Park et al. (2009). Roughly speaking, the SiZer method proceeds as follows: For each location u and bandwidth h in a pre-specified set, SiZer computes an estimator $\widehat{m}'_h(u)$ of the derivative $m'(u)$ and a corresponding confidence interval. For each (u, h), it then checks whether the confidence interval includes the value 0. The set Π_T^{SiZer} of points (u, h) for which the confidence interval does not include 0 corresponds to the set of intervals $\Pi_T^{\pm}$ for which our multiscale test finds an increase/decrease in the trend m. In order to explore how our test performs in comparison to SiZer, we compare the two sets $\Pi_T^{\pm}$ and Π_T^{SiZer} in different

ways to each other in what follows.

In order to implement SiZer for time series, we follow the exposition in Park et al. (2009).[7] The details are given in Section S.3 in the Supplementary Material. To simplify the implementation of SiZer, we assume that the autocovariance function $\gamma_\varepsilon(\cdot)$ of the error process and thus the long-run error variance σ^2 is known. Our multiscale test is implemented in the same way as in Section 5.2. To keep the comparison fair, we treat σ^2 as known also when implementing our method. Moreover, we use the same grid $\mathcal{G}_T$ of points (u, h) for both methods. To achieve this, we start off with the grid $\mathcal{G}_T$ from (5.1). We then follow Rondonotti et al. (2007) and Park et al. (2009) and restrict attention to those points $(u, h) \in \mathcal{G}_T$ for which the effective sample size $\mathrm{ESS}^*(u, h)$ for correlated data is not smaller than 5. This yields the grid $\mathcal{G}_T^* = \{(u, h) \in \mathcal{G}_T : \mathrm{ESS}^*(u, h) \geq 5\}$. A detailed discussion of the effective sample size $\mathrm{ESS}^*(u, h)$ for correlated data can be found in Rondonotti et al. (2007).

In the first part of the comparison study, we analyse the size and power of the two methods. To do so, we treat SiZer as a rigorous statistical test of the null hypothesis H_0 that m is constant on all intervals $[u - h, u + h]$ with $(u, h) \in \mathcal{G}_T^*$. In particular, we let SiZer reject the null if the set Π_T^{SiZer} is non-empty, that is, if the value 0 is not included in the confidence interval for at least one point $(u, h) \in \mathcal{G}_T^*$. We simulate data from the model $Y_{t,T} = m(t/T) + \varepsilon_t$ with different AR(1) error processes and different trends m. In particular, we let $\{\varepsilon_t\}$ be an AR(1) process of the form $\varepsilon_t = a_1\varepsilon_{t-1} + \eta_t$ with $a_1 \in \{-0.25, 0.25\}$ and i.i.d. standard normal innovations η_t. To simulate data under the null, we set $m = 0$ as in the previous section. To generate data under the alternative, we consider the linear trends $m(u) = \beta(u - 0.5)$ with different slopes β. As it is more difficult to detect a trend m in the data when the error terms are positively autocorrelated, we choose the slopes β larger in the AR(1) case with $a_1 = 0.25$ than in the case with $a_1 = -0.25$. In particular, we let $\beta \in \{1.0, 1.25, 1.5\}$ when $a_1 = -0.25$ and $\beta \in \{2.0, 2.25, 2.5\}$ when $a_1 = 0.25$. Further model specifications with nonlinear trends are considered in the second part of the comparison study. To produce our simulation results, we generate $S = 1000$ samples for each model specification and carry out the two methods for each sample.

The simulation results are reported in Tables 13 and 14. Both for our multiscale test and SiZer, the entries in the tables are computed as the number of simulations in which the respective method rejects the null hypothesis H_0 divided by the total number of simulations. As can be seen from Table 13, our test has approximately correct size in all of the considered settings, whereas SiZer is very liberal and rejects the null way too often. Examining Table 14, one can further see that our procedure has reasonable power against the considered alternatives. The power numbers are of course higher for

[7] We have also examined the somewhat different implementation from Rondonotti et al. (2007). As this yields worse simulation results than the procedure from Park et al. (2009), we however do not report them here.

Table 13: Size of our multiscale test (MT) and SiZer for different model specifications.

	$a_1 = -0.25$						$a_1 = 0.25$					
	$\alpha = 0.01$		$\alpha = 0.05$		$\alpha = 0.1$		$\alpha = 0.01$		$\alpha = 0.05$		$\alpha = 0.1$	
	MT	SiZer	MT	SiZer	MT	SiZer	MT	SiZer	MT	SiZer	MT	SiZer
$T = 250$	0.018	0.112	0.040	0.374	0.104	0.575	0.017	0.106	0.034	0.347	0.092	0.522
$T = 350$	0.012	0.140	0.058	0.426	0.080	0.621	0.012	0.130	0.046	0.399	0.074	0.578
$T = 500$	0.005	0.140	0.041	0.489	0.097	0.680	0.006	0.136	0.039	0.452	0.097	0.639

Table 14: Power of our multiscale test (MT) and SiZer for different model specifications. The three panels (a)–(c) corresponds to different slope parameters β of the linear tend m.

(a) $\beta = 1.0$ for negative a_1 and $\beta = 2.0$ for positive a_1

	$a_1 = -0.25$						$a_1 = 0.25$					
	$\alpha = 0.01$		$\alpha = 0.05$		$\alpha = 0.1$		$\alpha = 0.01$		$\alpha = 0.05$		$\alpha = 0.1$	
	MT	SiZer	MT	SiZer	MT	SiZer	MT	SiZer	MT	SiZer	MT	SiZer
$T = 250$	0.218	0.544	0.454	0.869	0.664	0.949	0.359	0.717	0.653	0.947	0.829	0.989
$T = 350$	0.385	0.707	0.665	0.958	0.753	0.986	0.599	0.888	0.864	0.995	0.913	0.998
$T = 500$	0.581	0.899	0.862	0.993	0.949	0.999	0.851	0.981	0.983	1.000	0.999	1.000

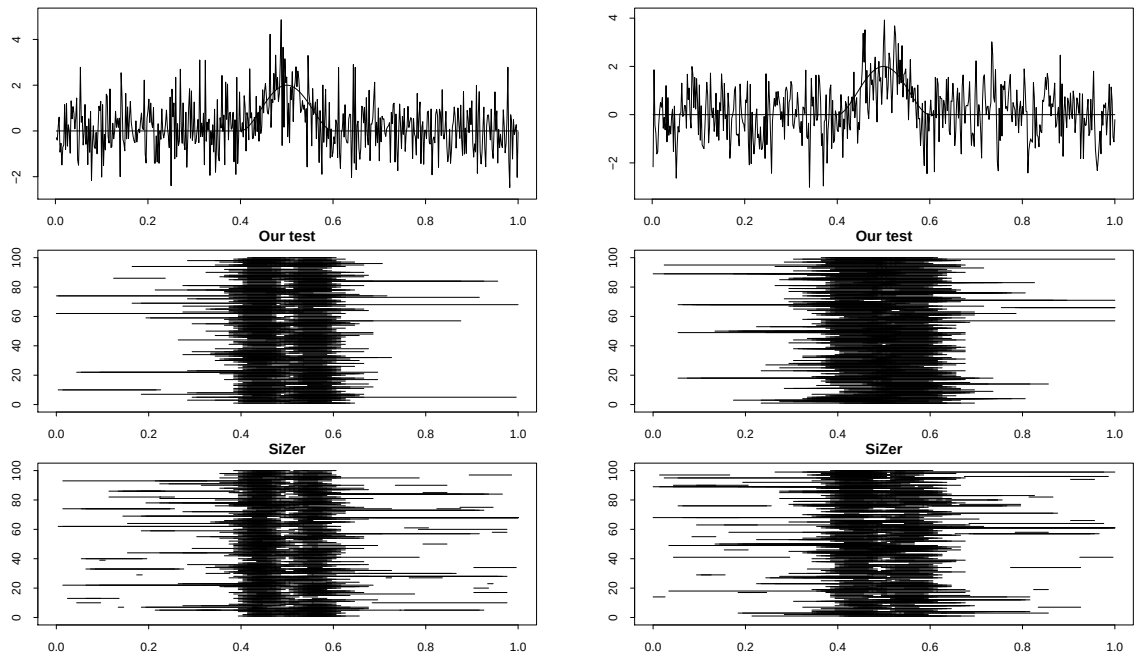
(b) $\beta = 1.25$ for negative a_1 and $\beta = 2.25$ for positive a_1

	$a_1 = -0.25$						$a_1 = 0.25$					
	$\alpha = 0.01$		$\alpha = 0.05$		$\alpha = 0.1$		$\alpha = 0.01$		$\alpha = 0.05$		$\alpha = 0.1$	
	MT	SiZer	MT	SiZer	MT	SiZer	MT	SiZer	MT	SiZer	MT	SiZer
$T = 250$	0.426	0.771	0.705	0.969	0.878	0.996	0.537	0.861	0.791	0.987	0.932	0.999
$T = 350$	0.645	0.912	0.882	0.993	0.954	1.000	0.773	0.955	0.948	0.999	0.985	1.000
$T = 500$	0.915	0.994	0.993	1.000	0.998	1.000	0.962	0.999	1.000	1.000	0.999	1.000

(c) $\beta = 1.5$ for negative a_1 and $\beta = 2.5$ for positive a_1

	$a_1 = -0.25$						$a_1 = 0.25$					
	$\alpha = 0.01$		$\alpha = 0.05$		$\alpha = 0.1$		$\alpha = 0.01$		$\alpha = 0.05$		$\alpha = 0.1$	
	MT	SiZer	MT	SiZer	MT	SiZer	MT	SiZer	MT	SiZer	MT	SiZer
$T = 250$	0.701	0.942	0.911	0.992	0.972	1.000	0.698	0.941	0.908	0.993	0.970	1.000
$T = 350$	0.895	0.994	0.981	1.000	0.996	1.000	0.893	0.993	0.980	1.000	0.996	1.000
$T = 500$	0.995	1.000	1.000	1.000	1.000	1.000	0.995	1.000	1.000	1.000	1.000	1.000

SiZer, which is a trivial consequence of the fact that SiZer is extremely liberal. These numbers should thus be treated with caution. All in all, the simulations suggest that SiZer can hardly be regarded as a rigorous statistical test of the null hypothesis H_0 that m is constant on all intervals $[u-h, u+h]$ with $(u, h) \in \mathcal{G}_T^*$. This is not very surprising as SiZer is not designed to be such a test but to produce informative SiZer maps. In particular, the confidence intervals of SiZer are not constructed to control the level α under H_0. In what follows, we thus attempt to compare the two methods in a different way which goes beyond mere size and power comparisons.



(a) $a_1 = -0.25$ (b) $a_1 = 0.25$

Figure 6: Comparison of the regions $\mathcal{R}_T^{\pm}$ and $\mathcal{R}_T^{\text{SiZer}}$. Subfigure (a) corresponds to the model setting with the AR parameter $a_1 = -0.25$, subfigure (b) to the setting with $a_1 = 0.25$. The upper panel of each subfigure shows a simulated time series path together with the underlying trend function m. The middle panel depicts the regions $\mathcal{R}_T^{\pm}$ produced by our multiscale test for 100 simulation runs. The lower panel presents the regions $\mathcal{R}_T^{\text{SiZer}}$ produced by SiZer.

Both our method and SiZer can be regarded as statistical tools to identify time regions where the curve m is increasing/decreasing.[8] Suppose that m is increasing/decreasing in the time region $\mathcal{R} \subset [0,1]$ but constant otherwise, that is, $m'(u) \neq 0$ for all $u \in \mathcal{R}$ and $m'(u) = 0$ for all $u \notin \mathcal{R}$. A natural question is the following: How well can the two methods identify the time region $\mathcal{R}$? In our framework, information on the region $\mathcal{R}$ is contained in the minimal intervals of the set $\Pi_T^{\pm}$. In particular, the union $\mathcal{R}_T^{\pm}$ of the minimal intervals in $\Pi_T^{\pm}$ can be regarded as an estimate of $\mathcal{R}$. This follows from the results in Propositions 3.2 and 3.3. Let $\mathcal{R}_T^{\text{SiZer}}$ be the union of the minimal intervals in Π_T^{SiZer}. In what follows, we compare $\mathcal{R}_T^{\pm}$ and $\mathcal{R}_T^{\text{SiZer}}$ to the region $\mathcal{R}$. This gives us information on how well the two methods approximate the true region where m is increasing/decreasing.[9]

We consider the same simulation setup as in the first part of the comparison study, only the trend function m is different. We let m be defined as $m(u) = 2 \cdot 1(u \in [0.4, 0.6]) \cdot (1 - 100\{u - 0.5\}^2)^2$, which implies that $\mathcal{R} = (0.4, 0.5) \cup (0.5, 0.6)$. The function m is plotted in the two upper panels of Figure 6. We set the significance level to

[8] More precisely speaking, SiZer is usually interpreted as investigating the curve m, viewed at different levels of resolution, rather than the curve m itself. Put differently, the underlying object of interest is a family of smoothed versions of m rather than m itself.

[9] The same exercise could of course also be carried out separately for the time region where the trend m increases and the region where it decreases.

$\alpha = 0.05$ and the sample size to $T = 500$. For each AR parameter $a_1 \in \{-0.25, 0.25\}$, we simulate $S = 100$ samples and compute $\mathcal{R}_T^{\pm}$ and $\mathcal{R}_T^{\text{SiZer}}$ for each sample. The simulation results are depicted in Figure 6, the two subfigures (a) and (b) corresponding to different AR parameters. The upper panel of each subfigure displays the time series path of a representative simulation together with the trend function m. The middle panel shows the regions $\mathcal{R}_T^{\pm}$ produced by our multiscale approach for the 100 simulation runs: On the y-axis, the simulation runs i are enumerated for $1 \leq i \leq 100$, and the black line at y-level i represents $\mathcal{R}_T^{\pm}$ for the i-th simulation. Finally, the lower panel of each subfigure depicts the regions $\mathcal{R}_T^{\text{SiZer}}$ in an analogous way.

Inspecting Figure 6, our multiscale method can be seen to approximate the region $\mathcal{R}$ fairly well in both simulation scenarios under consideration. In particular, $\mathcal{R}_T^{\pm}$ gives a good approximation to the region $\mathcal{R}$ for most simulations. Only in some simulation runs, $\mathcal{R}_T^{\pm}$ is too large compared to $\mathcal{R}$, which means that our method is not able to locate the region $\mathcal{R}$ sufficiently precisely. Overall, the SiZer method also produces quite satisfactory results. However, the SiZer estimates of $\mathcal{R}$ are not as precise as ours. In particular, SiZer spuriously finds regions of decrease/increase outside the interval $\mathcal{R}$ much more often than our method. It thus frequently mistakes fluctuations in the time series which are due to the dependence in the error terms for increases/decreases in the trend m.

To sum up, our multiscale test exhibits good size and power properties in the simulations, and the minimal intervals produced by it identify the time regions where m increases/decreases in a quite reliable way. SiZer performs clearly worse in these respects. Nevertheless, it may still produce informative SiZer plots. All in all, we would like to regard the two methods as complementary rather than direct competitors. SiZer is an explorative tool which aims to give an overview of the increases/decreases in m by means of a SiZer plot. Our method, in contrast, is tailored to be a rigorous statistical test of the hypothesis H_0. In particular, it allows to make rigorous confidence statements about the time regions where the trend m increases/decreases.

5.4 Small sample properties of the long-run variance estimator

In the final part of the simulation study, we examine the estimators of the AR parameters and the long-run error variance from Section **??**. We simulate data from the model $Y_{t,T} = m(t/T) + \varepsilon_t$, where $\{\varepsilon_t\}$ is an AR(1) process of the form $\varepsilon_t = a_1\varepsilon_{t-1} + \eta_t$. We consider the AR parameters $a_1 \in \{-0.95, -0.75, -0.5, -0.25, 0.25, 0.5, 0.75, 0.95\}$ and let η_t be i.i.d. standard normal innovation terms. We report our findings for a specific sample size T, in particular for $T = 500$, as the results for other sample sizes are very similar. For simplicity, m is chosen to be a linear function of the form $m(u) = \beta u$ with the slope parameter β. For each value of a_1, we consider two different slopes β, one corresponding to a moderate and one to a pronounced trend m. In particular,

we let $\beta = s_\beta \sqrt{\text{Var}(\varepsilon_t)}$ with $s_\beta \in \{1, 10\}$. When $s_\beta = 1$, the slope β is equal to the standard deviation $\sqrt{\text{Var}(\varepsilon_t)}$ of the error process, which yields a moderate trend m. When $s_\beta = 10$, in contrast, the slope β is 10 times as large as $\sqrt{\text{Var}(\varepsilon_t)}$, which results in a quite pronounced trend m.
For each model specification, we generate $S = 1000$ data samples and compute the following quantities for each simulated sample:

(i) the pilot estimator $\widetilde{a}_q$ from (4.9) with the tuning parameter q.

(ii) the estimator $\widehat{a}$ from (4.12) with the tuning parameter $\overline{r}$ as well as the long-run variance estimator $\widehat{\sigma}^2$ from (4.13).

(iii) the estimators of a_1 and σ^2 from Hall and Van Keilegom (2003), which are denoted by $\widehat{a}_{\text{HvK}}$ and $\widehat{\sigma}^2_{\text{HvK}}$ for ease of reference. The estimator $\widehat{a}_{\text{HvK}}$ is computed as described in Section 2.2 of Hall and Van Keilegom (2003) and $\widehat{\sigma}^2_{\text{HvK}}$ as defined at the bottom of p.447 in Section 2.3. The estimator $\widehat{a}_{\text{HvK}}$ (as well as $\widehat{\sigma}^2_{\text{HvK}}$) depends on two tuning parameters which we denote by m_1 and m_2 as in Hall and Van Keilegom (2003).

(iv) oracle estimators $\widehat{a}_{\text{oracle}}$ and $\widehat{\sigma}^2_{\text{oracle}}$ of a_1 and σ^2, which are constructed under the assumption that the error process $\{\varepsilon_t\}$ is observed. For each simulation run, we compute $\widehat{a}_{\text{oracle}}$ as the maximum likelihood estimator of a_1 from the time series of simulated error terms $\varepsilon_1, \ldots, \varepsilon_T$. We then calculate the residuals $r_t = \varepsilon_t - \widehat{a}_{\text{oracle}}\, \varepsilon_{t-1}$ and estimate the innovation variance $\nu^2 = \mathbb{E}[\eta_t^2]$ by $\widehat{\nu}^2_{\text{oracle}} = (T-1)^{-1} \sum_{t=2}^{T} r_t^2$. Finally, we set $\widehat{\sigma}^2_{\text{oracle}} = \widehat{\nu}^2_{\text{oracle}} / (1 - \widehat{a}_{\text{oracle}})^2$.

Throughout the section, we set $q = 25$, $\overline{r} = 10$ and $(m_1, m_2) = (20, 30)$. We in particular choose q to be in the middle of m_1 and m_2 to make the tuning parameters of the estimators $\widetilde{a}_q$ and $\widehat{a}_{\text{HvK}}$ more or less comparable. In order to assess how sensitive our estimators are to the choice of q and $\overline{r}$, we carry out a number of robustness checks, considering a range of different values for q and $\overline{r}$. In addition, we vary the tuning parameters m_1 and m_2 of the estimators from Hall and Van Keilegom (2003) in order to make sure that the results of our comparison study are not driven by the particular choice of any of the involved tuning parameters. The results of our robustness checks are reported in Section S.3 of the Supplementary Material. They show that the results of our comparison study are robust to different choices of the parameters q, $\overline{r}$ and (m_1, m_2). Moreover, they indicate that our estimators are rather insensitive to the choice of tuning parameters.
For each estimator $\widehat{a}$, $\widehat{a}_{\text{HvK}}$, $\widehat{a}_{\text{oracle}}$ and $\widehat{\sigma}^2$, $\widehat{\sigma}^2_{\text{HvK}}$, $\widehat{\sigma}^2_{\text{oracle}}$ and for each model specification, the simulation output consists in a vector of length $S = 1000$ which contains the 1000 simulated values of the respective estimator. Figures 7 and 8 report the mean squared error (MSE) of these 1000 simulated values for each estimator. On the x-axis of each

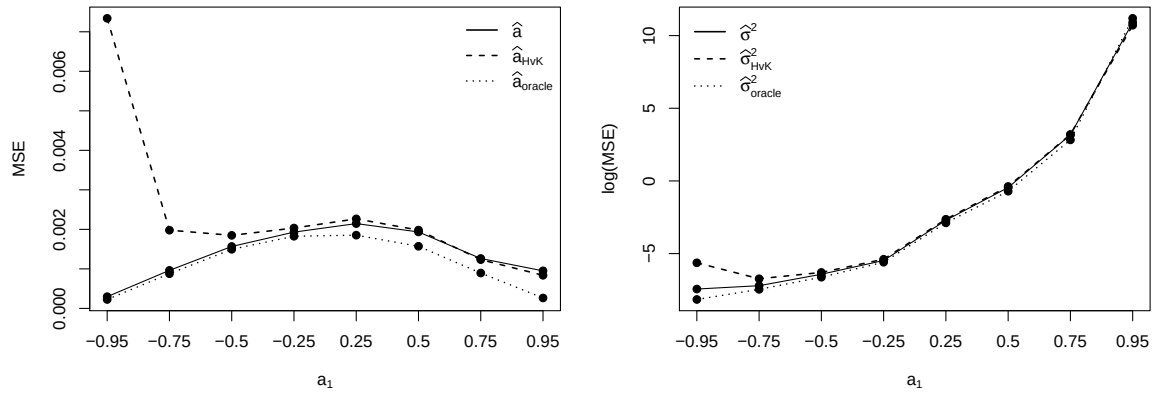


Figure 7: MSE values for the estimators $\widehat{a}$, $\widehat{a}_{\mathrm{HvK}}$, $\widehat{a}_{\mathrm{oracle}}$ and $\widehat{\sigma}^2$, $\widehat{\sigma}^2_{\mathrm{HvK}}$, $\widehat{\sigma}^2_{\mathrm{oracle}}$ in the simulation scenarios with a moderate trend ($s_\beta = 1$).

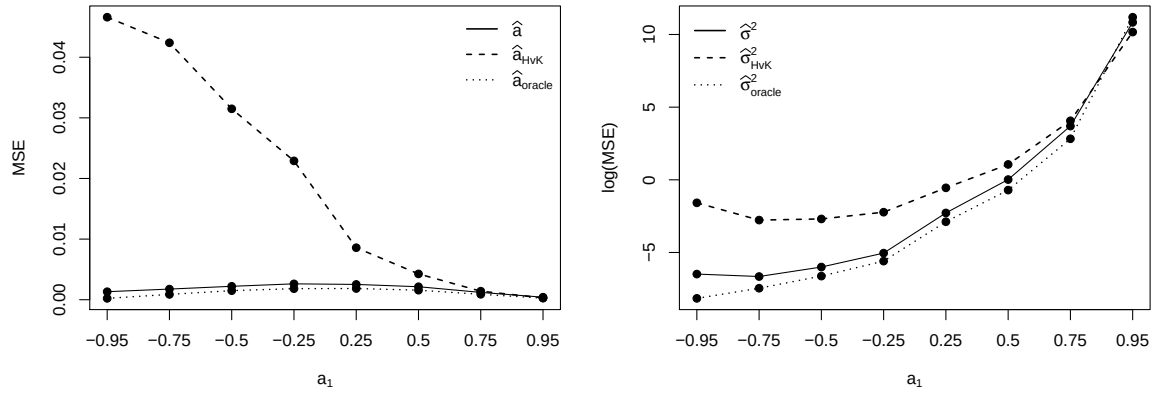


Figure 8: MSE values for the estimators $\widehat{a}$, $\widehat{a}_{\mathrm{HvK}}$, $\widehat{a}_{\mathrm{oracle}}$ and $\widehat{\sigma}^2$, $\widehat{\sigma}^2_{\mathrm{HvK}}$, $\widehat{\sigma}^2_{\mathrm{oracle}}$ in the simulation scenarios with a pronounced trend ($s_\beta = 10$).

plot, the various values of the AR parameter a_1 are listed which are considered. The solid line in each plot gives the MSE values of our estimators. The dashed and dotted lines specify the MSE values of the HvK and the oracle estimators, respectively. Note that for the long-run variance estimators, the plots report the logarithm of the MSE rather than the MSE itself since the MSE values are too different across simulation scenarios to obtain a reasonable graphical presentation. In addition to the MSE values presented in Figures 7 and 8, we depict histograms of the 1000 simulated values produced by the estimators $\widehat{a}$, $\widehat{a}_{\mathrm{HvK}}$, $\widehat{a}_{\mathrm{oracle}}$ and $\widehat{\sigma}^2$, $\widehat{\sigma}^2_{\mathrm{HvK}}$, $\widehat{\sigma}^2_{\mathrm{oracle}}$ for two specific simulation scenarios in Figures 9 and 10. The main findings can be summarized as follows:

(a) In the simulation scenarios with a moderate trend ($s_\beta = 1$), the estimators $\widehat{a}_{\mathrm{HvK}}$ and $\widehat{\sigma}^2_{\mathrm{HvK}}$ of Hall and Van Keilegom (2003) exhibit a similar performance as our estimators $\widehat{a}$ and $\widehat{\sigma}^2$ as long as the AR parameter a_1 is not too close to -1. For strongly negative values of a_1 (in particular for $a_1 = -0.75$ and $a_1 = -0.95$), the estimators perform much worse than ours. This can be clearly seen from the much larger MSE values of the estimators $\widehat{a}_{\mathrm{HvK}}$ and $\widehat{\sigma}^2_{\mathrm{HvK}}$ for $a_1 = -0.75$ and

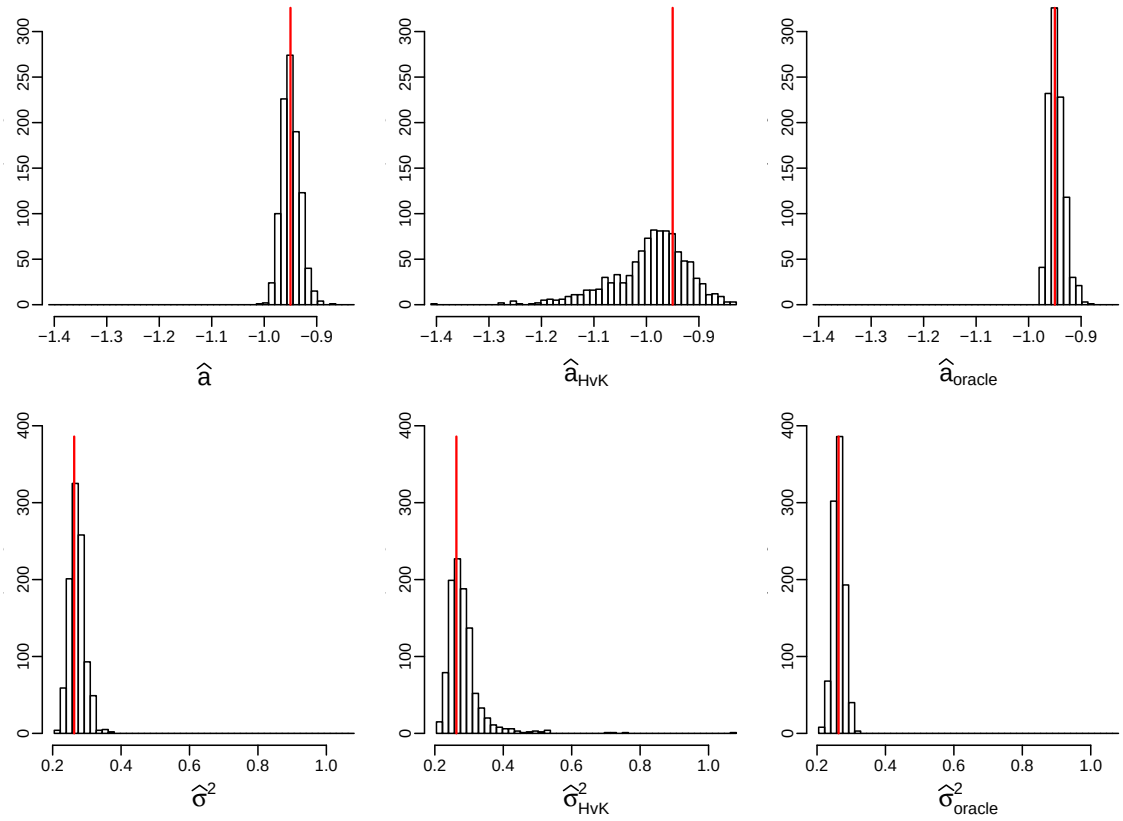


Figure 9: Histograms of the simulated values produced by the estimators $\widehat{a}$, $\widehat{a}_{\text{HvK}}$, $\widehat{a}_{\text{oracle}}$ and $\widehat{\sigma}^2$, $\widehat{\sigma}^2_{\text{HvK}}$, $\widehat{\sigma}^2_{\text{oracle}}$ in the scenario with $a_1 = -0.95$ and $s_\beta = 1$. The vertical red lines indicate the true values of a_1 and σ^2.

$a_1 = -0.95$ in Figure 7. Figure 9 gives some further insights into what is happening here. It shows the histograms of the simulated values produced by the estimators $\widehat{a}$, $\widehat{a}_{\text{HvK}}$, $\widehat{a}_{\text{oracle}}$ and the corresponding long-run variance estimators in the scenario with $a_1 = -0.95$ and $s_\beta = 1$. As can be seen, the estimator $\widehat{a}_{\text{HvK}}$ does not obey the causality restriction $|a_1| \leq 1$ but frequently takes values substantially smaller than -1. This results in a very large spread of the histogram and thus in a disastrous performance of the estimator.[10] A similar point applies to the histogram of the long-run variance estimator $\widehat{\sigma}^2_{\text{HvK}}$. Our estimators $\widehat{a}$ and $\widehat{\sigma}^2$, in contrast, exhibit a stable behaviour in this case.

Interestingly, the estimator $\widehat{a}_{\text{HvK}}$ (as well as the corresponding long-run variance estimator $\widehat{\sigma}^2_{\text{HvK}}$) performs much worse than ours for large negative values but not for large positive values of a_1. This can be explained as follows: In the special case of an AR(1) process, the estimator $\widehat{a}_{\text{HvK}}$ may produce estimates smaller than -1 but it cannot become larger than 1. This can be easily seen upon inspecting the

[10] One could of course set $\widehat{a}_{\text{HvK}}$ to $-(1-\delta)$ for some small $\delta > 0$ whenever it takes a value smaller than -1. This modified estimator, however, is still far from performing in a satisfying way when a_1 is close to -1.

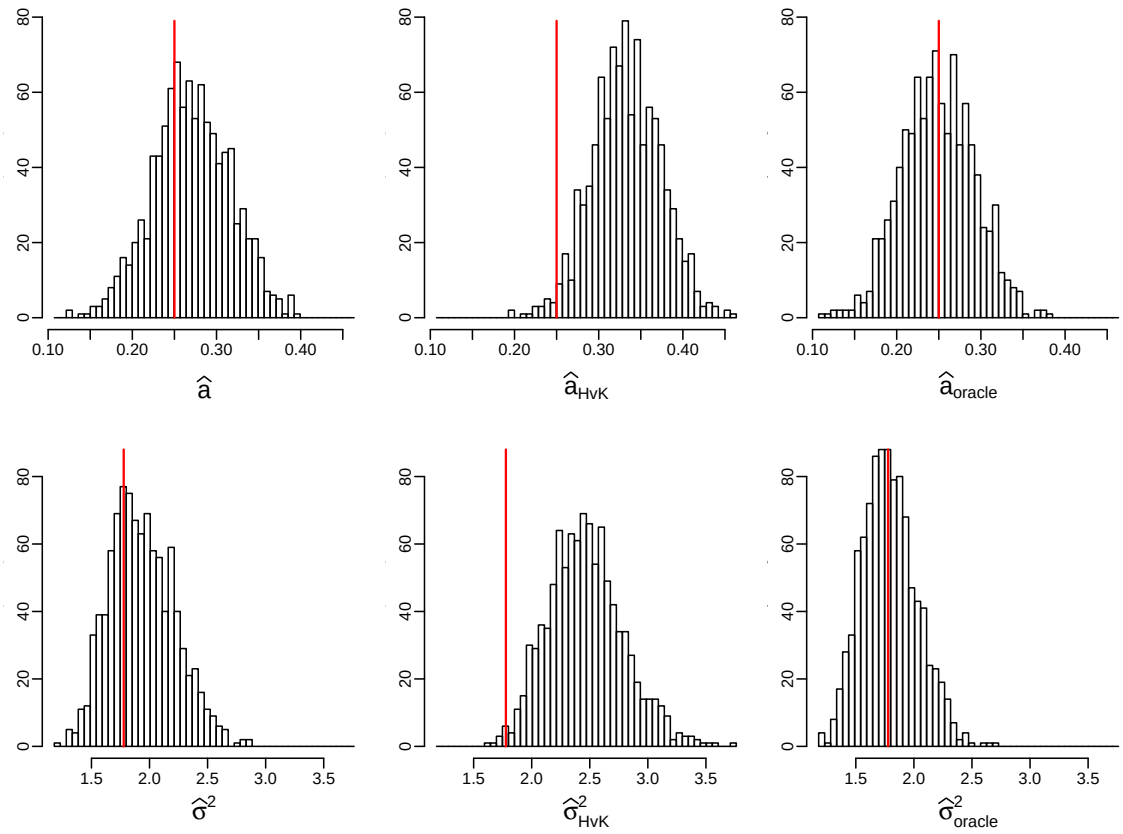


Figure 10: Histograms of the simulated values produced by the estimators $\widehat{a}$, $\widehat{a}_{\text{HvK}}$, $\widehat{a}_{\text{oracle}}$ and $\widehat{\sigma}^2$, $\widehat{\sigma}^2_{\text{HvK}}$, $\widehat{\sigma}^2_{\text{oracle}}$ in the scenario with $a_1 = 0.25$ and $s_\beta = 10$. The vertical red lines indicate the true values of a_1 and σ^2.

definition of the estimator. Hence, for large positive values of a_1, the estimator $\widehat{a}_{\text{HvK}}$ performs well as it satisfies the causality restriction that the estimated AR parameter should be smaller than 1.

(b) In the simulation scenarios with a pronounced trend ($s_\beta = 10$), the estimators of Hall and Van Keilegom (2003) are clearly outperformed by ours for most of the AR parameters a_1 under consideration. In particular, their MSE values reported in Figure 8 are much larger than the values produced by our estimators for most parameter values a_1. The reason is the following: The HvK estimators have a strong bias since the pronounced trend with $s_\beta = 10$ is not eliminated appropriately by the underlying differencing methods. This point is illustrated by Figure 10 which shows histograms of the simulated values for the estimators $\widehat{a}$, $\widehat{a}_{\text{HvK}}$, $\widehat{a}_{\text{oracle}}$ and the corresponding long-run variance estimators in the scenario with $a_1 = 0.25$ and $s_\beta = 10$. As can be seen, the histogram produced by our estimator $\widehat{a}$ is approximately centred around the true value $a_1 = 0.25$, whereas that of the estimator $\widehat{a}_{\text{HvK}}$ is strongly biased upwards. A similar picture arises for the long-run variance estimators $\widehat{\sigma}^2$ and $\widehat{\sigma}^2_{\text{HvK}}$.

Whereas the methods of Hall and Van Keilegom (2003) perform much worse than

ours for negative and moderately positive values of a_1, the performance (in terms of MSE) is fairly similar for large values of a_1. This can be explained as follows: When the trend m is not eliminated appropriately by taking differences, this creates spurious persistence in the data. Hence, the estimator $\widehat{a}_{\mathrm{HvK}}$ tends to overestimate the AR parameter a_1, that is, $\widehat{a}_{\mathrm{HvK}}$ tends to be larger in absolute value than a_1. Very loosely speaking, when the parameter a_1 is close to 1, say $a_1 = 0.95$, there is not much room for overestimation since $\widehat{a}_{\mathrm{HvK}}$ cannot become larger than 1. Consequently, the effect of not eliminating the trend appropriately has a much smaller impact on $\widehat{a}_{\mathrm{HvK}}$ for large positive values of a_1.

6 Application

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 $Y_{t,T} = m(t/T) + \varepsilon_t$, where m is the unknown time trend of interest. The error process $\{\varepsilon_t\}$ is supposed to have the AR(p) structure $\varepsilon_t = \sum_{j=1}^{p} a_j \varepsilon_{t-j} + \eta_t$, where η_t are i.i.d. innovations with mean 0 and variance ν^2. As pointed out in Mudelsee (2010) among others, this is the most widely used error model for discrete climate time series. To select the AR order p, we proceed as follows: We estimate the AR parameters and the corresponding variance of the innovation terms for different AR orders by our methods from Section **??** and choose p to be the minimizer of the Bayesian information criterion (BIC). This yields the AR order $p = 2$. We then estimate the parameters $\boldsymbol{a} = (a_1, a_2)$ and the long-run error variance σ^2 by the estimators $\widehat{\boldsymbol{a}} = (\widehat{a}_1, \widehat{a}_2)$ and $\widehat{\sigma}^2$, which gives the values $\widehat{a}_1 = 0.167$, $\widehat{a}_2 = 0.178$ and $\widehat{\sigma}^2 = 0.749$. To select the AR order p and to produce the estimators $\widehat{\boldsymbol{a}}$ and $\widehat{\sigma}^2$, we set $q = 25$ and $\overline{r} = 10$ as in the simulation study of Section 5.2.[11]

With the help of our multiscale method from Section 3, we now test the null hypothesis H_0 that m is constant on all intervals $[u - h, u + h]$ with $(u, h) \in \mathcal{G}_T$, where we use the grid $\mathcal{G}_T$ defined in (5.1). To do so, we set the significance level to $\alpha = 0.05$ and implement the test in exactly the same way as in the simulations of Section 5.2. The

[11] As a robustness check, we have repeated the process of order selection and parameter estimation for other values of q and $\overline{r}$ as well as for other criteria such as FPE, AIC and AICC, which gave similar results.

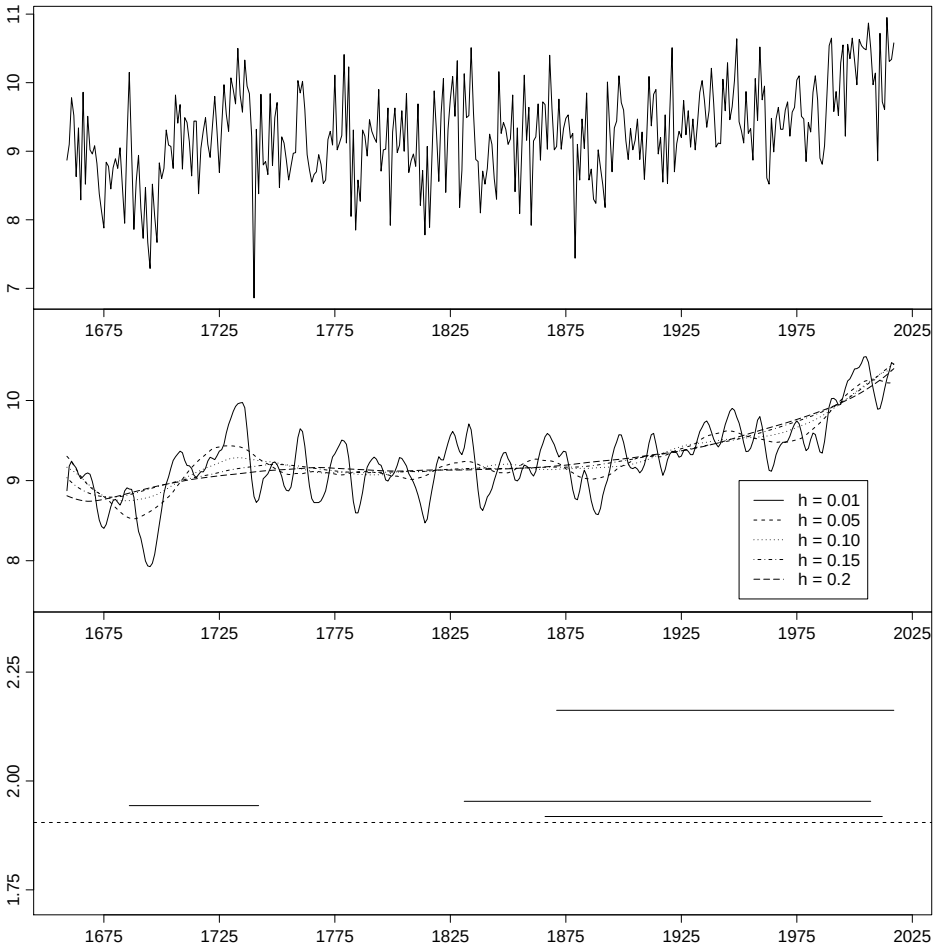


Figure 11: Summary of the application results. The upper panel shows the Central England mean temperature time series. The middle panel depicts local linear kernel estimates of the time trend for a number of different bandwidths h. The lower panel presents the minimal intervals in the set Π_T^+ produced by the multiscale test. These are $[1686, 1742]$, $[1831, 2007]$, $[1866, 2012]$ and $[1871, 2017]$.

results are presented in Figure 11. 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 11. The plot depicts the minimal intervals contained in the set Π_T^+, which is defined in Section 3.3. 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 3.3, 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 1870 onwards. Hence, our test in particular provides evidence that there has been some warming trend in the period over approximately the last 150 years. On the other hand, as the set Π_T^- is empty, there is no evidence of any downward movement of the trend.

References

Benner, T. C. (1999). Central england temperatures: long-term variability and teleconnections. *International Journal of Climatology*, **19** 391–403.

Berkes, I., Liu, W. and Wu, W. B. (2014). Komlós-Major-Tusnády approximation under dependence. *Annals of Probability*, **42** 794–817.

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

Chaudhuri, P. and Marron, J. S. (2000). Scale space view of curve estimation. *Annals of Statistics*, **28** 408–428.

Chernozhukov, V., Chetverikov, D. and Kato, K. (2014). Gaussian approximation of suprema of empirical processes. *Annals of Statistics*, **42** 1564–1597.

Chernozhukov, V., Chetverikov, D. and Kato, K. (2015). Comparison and anti-concentration bounds for maxima of Gaussian random vectors. *Probability Theory and Related Fields*, **162** 47–70.

Chernozhukov, V., Chetverikov, D. and Kato, K. (2017). Central limit theorems and bootstrap in high dimensions. *Annals of Probability*, **45** 2309–2352.

Cho, H. and Fryzlewicz, P. (2012). Multiscale and multilevel technique for consistent segmentation of nonstationary time series. *Statistica Sinica*, **22** 207–229.

Donoho, D., Johnstone, I., Kerkyacharian, G. and Picard, D. (1995). Wavelet shrinkage: Asymptopia? *Journal of the Royal Statistical Society: Series B*, **57** 301–369.

Dümbgen, L. (2002). Application of local rank tests to nonparametric regression. *Journal of Nonparametric Statistics*, **14** 511–537.

Dümbgen, L. and Spokoiny, V. G. (2001). Multiscale testing of qualitative hypotheses. *Annals of Statistics*, **29** 124–152.

Dümbgen, L. and Walther, G. (2008). Multiscale inference about a density. *Annals of Statistics*, **36** 1758–1785.

Eckle, K., Bissantz, N. and Dette, H. (2017). Multiscale inference for multivariate

deconvolution. *Electronic Journal of Statistics*, **11** 4179–4219.

Hall, P. and Heckman, N. E. (2000). Testing for monotonicity of a regression mean by calibrating for linear functions. *Annals of Statistics*, **28** 20–39.

Hall, P. and Van Keilegom, I. (2003). Using difference-based methods for inference in nonparametric regression with time series errors. *Journal of the Royal Statistical Society: Series B*, **65** 443–456.

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

Herrmann, E., Gasser, T. and Kneip, A. (1992). Choice of bandwidth for kernel regression when residuals are correlated. *Biometrika*, **79** 783–795.

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

Müller, H.-G. and Stadtmüller, U. (1988). Detecting dependencies in smooth regression models. *Biometrika*, **75** 639–650.

Park, C., , Hannig, J. and Kang, K.-H. (2009). Improved SiZer for time series. *Statistica Sinica*, **19** 1511–1530.

Park, C., Marron, J. S. and Rondonotti, V. (2004). Dependent SiZer: goodness-of-fit tests for time series models. *Journal of Applied Statistics*, **31** 999–1017.

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.

Proksch, K., Werner, F. and Munk, A. (2018). Multiscale scanning in inverse problems. *Forthcoming in Annals of Statistics*.

Qiu, D., Shao, Q. and Yang, L. (2013). Efficient inference for autoregressive coefficients in the presence of trends. *Journal of Multivariate Analysis*, **114** 40–53.

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

Rohde, A. (2008). Adaptive goodness-of-fit tests based on signed ranks. *Annals of Statistics*, **36** 1346–1374.

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.

Rufibach, K. and Walther, G. (2010). The block criterion for multiscale inference about a density, with applications to other multiscale problems. *Journal of Computational and Graphical Statistics*, **19** 175–190.

Schmidt-Hieber, J., Munk, A. and Dümbgen, L. (2013). Multiscale methods for shape constraints in deconvolution: confidence statements for qualitative features. *Annals of Statistics*, **41** 1299–1328.

Shao, Q. and Yang, L. J. (2011). Autoregressive coefficient estimation in nonparametric analysis. *Journal of Time Series Analysis*, **32** 587–597.

Tecuapetla-Gómez, I. and Munk, A. (2017). Autocovariance estimation in regression with a discontinuous signal and m-dependent errors: a difference-based approach. *Scandinavian Journal of Statistics*, **44** 346–368.

Truong, Y. K. (1991). Nonparametric curve estimation with time series errors. *Journal of Statistical Planning and Inference*, **28** 167–183.

Von Sachs, R. and MacGibbon, B. (2000). Non-parametric curve estimation by Wavelet thresholding with locally stationary errors. *Scandinavian Journal of Statistics*, **27** 475–499.

Wu, W. B. (2005). Nonlinear system theory: another look at dependence. *Proc. Natn. Acad. Sci. USA*, **102** 14150–14154.

Wu, W. B. and Shao, X. (2004). Limit theorems for iterated random functions. *Journal of Applied Probability* 425–436.