Khismatullina, Marina; Vogt, Michael

Multiscale inference and long-run variance estimation in non-parametric regression with time series errors

Journal of the Royal Statistical Society: Series B (Statistical Methodology)

Provided in Cooperation with:
John Wiley & Sons

Suggested Citation: Khismatullina, Marina; Vogt, Michael (2019) : Multiscale inference and long-run variance estimation in non-parametric regression with time series errors, Journal of the Royal Statistical Society: Series B (Statistical Methodology), ISSN 1467-9868, Wiley, Hoboken, NJ, Vol. 82, Iss. 1, pp. 5-37,
https://doi.org/10.1111/rssb.12347

This Version is available at:
https://hdl.handle.net/10419/230027

ROYAL STATISTICAL SOCIETY
DATA | EVIDENCE | DECISIONS

Journal of the Royal Statistical Society
Statistical Methodology
Series B

J. R. Statist. Soc. B (2020)
82, *Part* 1, *pp.* 5–37

Multiscale inference and long-run variance estimation in non-parametric regression with time series errors

Marina Khismatullina and Michael Vogt

University of Bonn, Germany

[Received November 2018. Revised October 2019]

Summary. We develop new multiscale methods to test qualitative hypotheses about the function m in the non-parametric regression model $Y_{t,T} = m(t/T) + \varepsilon_t$ with time series errors ε_t. In time series applications, m represents a non-parametric 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 or decreasing in certain time intervals. Our multiscale methods enable us to test for such shape properties of the trend m. To perform the methods, we require an estimator of the long-run error variance $\sigma^2 = \sum_{l=-\infty}^{\infty} \mathrm{cov}(\varepsilon_0, \varepsilon_l)$. We propose a new difference-based estimator of σ^2 for the case that $\{\varepsilon_t\}$ belongs to the class of auto-regressive 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.

Keywords: Anticoncentration bounds; Long-run variance; Multiscale statistics; Non-parametric regression; Shape constraints; Strong approximations; Time series errors

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 or decreasing in certain time intervals? Can one identify the intervals of increase and decrease? As an example, consider the time series plotted in Fig. 1 which shows the yearly mean temperature in central England from 1659 to 2017. Climatologists are very much interested in learning about the trending behaviour of temperature time series like this; see for example 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 non-parametric time trend. We in particular construct a multiscale test which enables us to identify local increases and decreases of the trend function. We develop our test in the context of the following model setting: we observe a time series $\{Y_{t,T} : 1 \leqslant t \leqslant T\}$ of the form

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

Address for correspondence: Michael Vogt, Department of Economics and Hausdorff Center for Mathematics, University of Bonn, 53113 Bonn, Germany.
E-mail: michael.vogt@uni-bonn.de

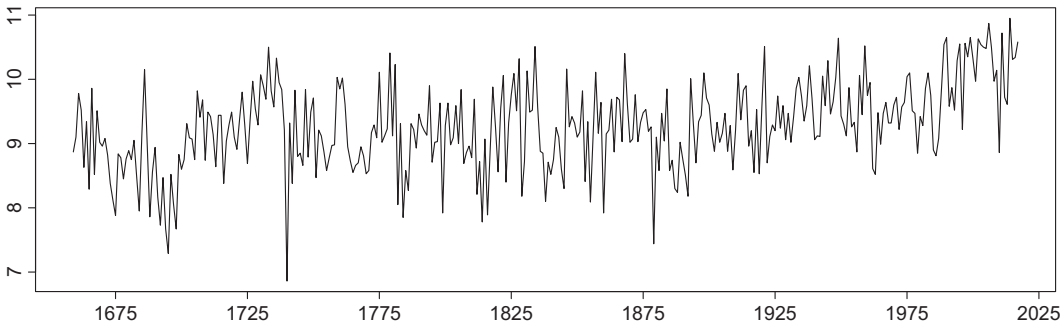

Fig. 1. Yearly mean temperature in central England from 1659 to 2017 measured in degrees centigrade

for $1 \leqslant t \leqslant T$, where $m : [0,1] \to \mathbb{R}$ is an unknown non-parametric 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 non-parametric time trend. As usual in non-parametric regression, we let the function m depend on rescaled time t/T rather than on realtime 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 $\hat{\varphi}_T(u,h)$ for the hypothesis $H_0(u,h)$. In a second step, we aggregate the statistics $\hat{\varphi}_T(u,h)$ for a large number of time windows $[u-h, u+h]$. We thereby construct a multiscale statistic which enables us 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 anticoncentration 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 level of significance $\alpha \in (0,1)$, the test enables us to make simultaneous confidence statements of the following form: we can claim, with statistical confidence $1-\alpha$, that there is an increase or 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 enables us to identify, with a prespecified statistical confidence, time intervals where the trend m is increasing or decreasing.

For independent data, multiscale tests have been developed in a variety of 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 for example 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 non-parametric 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) introduced SiZer methods for dependent data which can be used to find local increases or 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 enables us to make simultaneous confidence statements about the time intervals where the trend m is increasing or decreasing. Some theoretical results for dependent SiZer methods were derived in Park *et al.* (2009), but only under quite a 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 0 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, enables us to consider simultaneously scales h of fixed size and scales h that converge to 0 at various rates. We can thus take into account time windows of many sizes. In Section 3.4, we compare our approach with SiZer methods for dependent data in more detail.

Our multiscale approach is also related to wavelet-based methods: similar to the wavelet-based methods, it takes into account different locations u and resolution levels or scales h simultaneously. However, whereas our multiscale approach is designed to test for local increases and decreases of a non-parametric trend, wavelet methods are commonly used for other purposes. Among other things, they are employed for estimating or reconstructing non-parametric regression curves (see for example Donoho *et al.* (1995) or Von Sachs and MacGibbon (2000)) and for change point detection (see for example Cho and Fryzlewicz (2012)).

The test statistic of our multiscale method depends on the long-run error variance $\sigma^2 = \sum_{l=-\infty}^{\infty} \mathrm{cov}(\varepsilon_0, \varepsilon_l)$, 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 considerable 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 introduce a new difference-based estimator of σ^2 for the case that $\{\varepsilon_t\}$ belongs to the class of auto-regressive 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 two empirical applications in Section 6. In the simulation study, we examine the finite sample properties of our multiscale test and compare it with the dependent SiZer methods that were introduced in Park *et al.* (2004) and Rondonotti *et al.* (2007). Moreover, we investigate the small sample performance of our estimator of σ^2 and compare it with the estimator of Hall and Van Keilegom (2003). In Section 6, we use our methods to analyse the temperature data from Fig. 1 as well as a sample of global temperature data. The data that are analysed in the paper and the computer code that was used for the simulations and the analysis of the empirical data can be obtained from

`https://rss.onlinelibrary.wiley.com/hub/journal/14679868/series-b-datasets`.

2. The model

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

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

for $1 \leqslant t \leqslant T$. Here, m is an unknown non-parametric function defined on $[0,1]$ and $\{\varepsilon_t : 1 \leqslant t \leqslant T\}$ 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.

Condition 1. 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 independent and identically distributed (IID) random variables and $G : \mathbb{R}^{\mathbb{Z}} \to \mathbb{R}$ is a measurable function.

Condition 2. 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 IID copy of $\{\eta_t\}$. In particular, we make the following assumption.

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

Conditions 1–3 are fulfilled by a wide range of stationary processes $\{\varepsilon_t\}$. As a first example, consider linear processes of the form $\varepsilon_t = \Sigma_{i=0}^{\infty} c_i \eta_{t-i}$ with $\|\varepsilon_t\|_q < \infty$, where c_i are absolutely summable coefficients and η_t are IID innovations with $\mathbb{E}[\eta_t] = 0$ and $\|\eta_t\|_q < \infty$. Trivially, conditions 1 and 2 are fulfilled in this case. Moreover, if $|c_i| = O(\rho^i)$ for some $\rho \in (0,1)$, then condition 3 is easily seen to be satisfied as well. As a special case, consider an auto-regressive moving average (ARMA) process $\{\varepsilon_t\}$ of the form $\varepsilon_t - \Sigma_{i=1}^{p} a_i \varepsilon_{t-i} = \eta_t + \Sigma_{j=1}^{r} b_j \eta_{t-j}$ 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 IID innovations with $\mathbb{E}[\eta_t] = 0$ and $\|\eta_t\|_q < \infty$. Moreover, as usual, we suppose that the complex polynomials $A(z) = 1 - \Sigma_{j=1}^{p} a_j z^j$ and $B(z) = 1 + \Sigma_{j=1}^{r} b_j z^j$ 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 $\varepsilon_t = \Sigma_{i=0}^{\infty} c_i \eta_{t-i}$ with $|c_i| = O(\rho^i)$ for some $\rho \in (0,1)$, implying that conditions 1–3 are fulfilled. The results in Wu and Shao (2004) show that condition 3 (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 and 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

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

where m' is the first derivative of m. We want to test the hypothesis $H_0(u,h)$ not just for a single interval $[u-h, u+h]$ but simultaneously for many 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. 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, however, we 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

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

where $w_{t,T}(u,h)$ is a kernel weight and h is the bandwidth. 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

$$\Lambda_{t,T}(u,h) = K\left(\frac{t/T-u}{h}\right)\left\{S_{T,0}(u,h)\left(\frac{t/T-u}{h}\right) - S_{T,1}(u,h)\right\},$$

$$S_{T,l}(u,h) = (Th)^{-1}\sum_{t=1}^{T} K\left(\frac{t/T-u}{h}\right)\left(\frac{t/T-u}{h}\right)^{l}$$

for $l = 0, 1, 2$ and K is a kernel function with the following properties.

Condition 4. The kernel K is non-negative, symmetric about zero and integrates to 1. Moreover, it has compact support $[-1, 1]$ and is Lipschitz continuous, i.e. $|K(v) - K(w)| \leqslant C|v-w|$ for any $v, w \in \mathbb{R}$ and some constant $C > 0$.

The kernel average $\hat{\psi}_T(u,h)$ is nothing other than a rescaled local linear estimator of the derivative $m'(u)$ with bandwidth h. Alternatively to the local linear weights defined in equation (3.1), we could work with the weights $w_{t,T}(u,h) = K'\{(t/T-u)/h\}/[\Sigma_{t=1}^{T} K'\{(t/T-u)/h\}^2]^{1/2}$, where the kernel function K is assumed to be differentiable and K' is its derivative. However, we prefer to use local linear weights as these have superior theoretical properties at the boundary.

A test statistic for the hypothesis $H_0(u,h)$ is given by the normalized kernel average $\hat{\psi}_T(u,h)/\hat{\sigma}$, where $\hat{\sigma}^2$ is an estimator of the long-run variance $\sigma^2 = \Sigma_{l=-\infty}^{\infty} \mathrm{cov}(\varepsilon_0, \varepsilon_l)$ 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 $\hat{\sigma}^2$ is an estimator with reasonable theoretical properties. Specifically, we assume that $\hat{\sigma}^2 = \sigma^2 + o_p(\rho_T)$ with $\rho_T = o\{1/\log(T)\}$. This is a fairly weak condition which is in particular satisfied by the estimator 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, $\mathrm{var}\{\hat{\psi}_T(u,h)\} = \sigma^2$ 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, $\mathrm{var}\{\hat{\psi}_T(u,h)\} = \sigma^2 + o(1)$ for any u and h under consideration. Hence, for sufficiently large sample sizes T, the test statistic $\hat{\psi}_T(u,h)/\hat{\sigma}$ has approximately unit variance.

We now combine the test statistics $\hat{\psi}_T(u,h)/\hat{\sigma}$ for a wide range of 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

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

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 $\hat{\Psi}_T$ does not simply aggregate the individual statistics $\hat{\psi}_T(u,h)/\hat{\sigma}$ by taking the supremum over all points $(u,h)\in\mathcal{G}_T$ as in more traditional multiscale approaches. We rather calibrate the statistics $\hat{\psi}_T(u,h)/\hat{\sigma}$ 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 for example 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

$$\hat{\Psi}_{T,\text{uncorrected}} = \max_{(u,h)\in\mathcal{G}_T} \left| \frac{\hat{\psi}_T(u,h)}{\hat{\sigma}} \right| \tag{3.3}$$

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 IID normally distributed and neglect the estimation error in $\hat{\sigma}$, i.e. set $\hat{\sigma}=\sigma$. Moreover, suppose that the set $\mathcal{G}_T$ consists of only the points $(u_k,h_l)=((2k-1)h_l,h_l)$ with $k=1,\dots,\lfloor 1/(2h_l)\rfloor$ and $l=1,\dots,L$. In this case, we can write

$$\hat{\Psi}_{T,\text{uncorrected}} = \max_{1\leqslant l\leqslant L} \max_{1\leqslant k\leqslant \lfloor 1/(2h_l)\rfloor} \left| \frac{\hat{\psi}_T(u_k,h_l)}{\sigma} \right|.$$

Under our simplifying assumptions, the statistics $\hat{\psi}_T(u_k,h_l)/\sigma$ with $k=1,\dots,\lfloor 1/(2h_l)\rfloor$ 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 $\max_k \hat{\psi}_T(u_k,h_l)/\sigma$ is approximately of size $\lambda(h_l)$ for small bandwidths h_l. As $\lambda(h)\to\infty$ for $h\to 0$, this implies that $\max_k \hat{\psi}_T(u_k,h_l)/\sigma$ tends to be much larger for small than for large bandwidths h_l. As a result, the stochastic behaviour of the uncorrected statistic $\hat{\Psi}_{T,\text{uncorrected}}$ tends to be dominated by the statistics $\hat{\psi}_T(u_k,h_l)$ corresponding to small bandwidths h_l. The additively corrected statistic $\hat{\Psi}_T$, in contrast, puts the statistics $\hat{\psi}_T(u_k,h_l)$ corresponding to different bandwidths h_l on a more equal footing, thus counteracting the dominance of small bandwidth values.

The multiscale statistic $\hat{\Psi}_T$ 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 $\mathcal{G}_T^{\text{full}} = \{(u,h): u=t/T \text{ for some } 1\leqslant t\leqslant T \text{ and } h\in[h_{\min},h_{\max}]\}$, 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.

Condition 5. $|\mathcal{G}_T| = O(T^\theta)$ for some arbitrarily large but fixed constant $\theta>0$, where $|\mathcal{G}_T|$ denotes the cardinality of $\mathcal{G}_T$.

Condition 6. $h_{\min} \gg T^{-(1-2/q)}\log(T)$, i.e. $h_{\min}/\{T^{-(1-2/q)}\log(T)\}\to\infty$ with $q>4$ defined in condition 2 and $h_{\max}<\frac{1}{2}$.

According to condition 5, 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 with the sample size T. For example, we may work with the set $\mathcal{G}_T = \{(u,h) : u = t/T \text{ for some } 1 \leqslant t \leqslant T \text{ and } h \in [h_{\min}, h_{\max}] \text{ with } h = t/T \text{ for some } 1 \leqslant t \leqslant T\}$, which contains more than enough points (u,h) for most practical applications. Condition 6 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 condition 2 which specifies the number of existing moments for the error terms ε_t. As we 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 1. The above construction of the multiscale statistic can be easily adapted to hypotheses other than H_0. To do so, we simply need to replace the kernel weights $w_{t,T}(u,h)$ that are defined in equation (3.1) by appropriate versions which are suited to test the hypothesis of interest. For example, if we want to test for local convexity or concavity of m, we may define the kernel weights $w_{t,T}(u,h)$ such that the kernel average $\hat{\psi}_T(u,h)$ is a (rescaled) estimator of the second derivative of m at the location u with bandwidth h.

3.2. The test procedure

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

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

where $\phi_T(u,h) = \Sigma_{t=1}^T w_{t,T}(u,h)\sigma Z_t$ and Z_t are independent standard normal random variables. The statistic Φ_T can be regarded as a Gaussian version of the test statistic $\hat{\Psi}_T$ 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 is now defined as follows: for a given level of significance $\alpha \in (0,1)$, we reject the overall null hypothesis H_0 if $\hat{\Psi}_T > q_T(\alpha)$. In particular, for any $(u,h) \in \mathcal{G}_T$, we reject $H_0(u,h)$ if the (corrected) test statistic $|\hat{\psi}_T(u,h)/\hat{\sigma}| - \lambda(h)$ lies above the critical value $q_T(\alpha)$, i.e. if $|\hat{\psi}_T(u,h)/\hat{\sigma}| > q_T(\alpha) + \lambda(h)$.

3.3. Theoretical properties of the test

To examine the theoretical properties of our multiscale test, we introduce the auxiliary multiscale statistic

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

with $\hat{\phi}_T(u,h) = \hat{\psi}_T(u,h) - \mathbb{E}[\hat{\psi}_T(u,h)] = \Sigma_{t=1}^T w_{t,T}(u,h)\varepsilon_t$. 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 that was defined in Section 3.2 can be used as a proxy for the $(1-\alpha)$-quantile of the multiscale statistic $\hat{\Phi}_T$.

Theorem 1. Let conditions 1–6 be fulfilled and assume that $\hat{\sigma}^2 = \sigma^2 + o_p(\rho_T)$ with $\rho_T = o\{1/\log(T)\}$. Then

$$\mathbb{P}\{\hat{\Phi}_T \leqslant q_T(\alpha)\} = 1 - \alpha + o(1).$$

A full proof of theorem 1 is given in the on-line supplementary material. Here we briefly outline the proof strategy, which splits up into two main steps. In the first, we replace the statistic $\hat{\Phi}_T$ for each $T \geqslant 1$ by a statistic $\tilde{\Phi}_T$ with the same distribution as $\hat{\Phi}_T$ and the property that

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

where $\delta_T = o(1)$ and the Gaussian statistic Φ_T is defined in Section 3.2. We thus replace the statistic $\hat{\Phi}_T$ 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

$$\sup_{x \in \mathbb{R}} |\mathbb{P}(\tilde{\Phi}_T \leqslant x) - \mathbb{P}(\Phi_T \leqslant x)| = o(1), \tag{3.7}$$

which immediately implies the statement of theorem 1. Importantly, the convergence result (3.6) is not sufficient for establishing the result in equation (3.7). Put differently, the fact that $\tilde{\Phi}_T$ can be approximated by Φ_T in the sense that $\tilde{\Phi}_T - \Phi_T = o_p(\delta_T)$ does not imply that the distribution of $\tilde{\Phi}_T$ is close to that of Φ_T in the sense of equation (3.7). For equation (3.7) to hold, we additionally require that the distribution of Φ_T has some sort of continuity property. Specifically, we prove that

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

which says that Φ_T does not concentrate too strongly in small regions of the form $[x - \delta_T, x + \delta_T]$. Anticoncentration bounds for Gaussian random vectors as derived in Chernozhukov *et al.* (2015) are the main tool for verifying the result in equation (3.8). The claim (3.7) can be proved by using equation (3.6) together with equation (3.8), which in turn yields theorem 1.

The main idea of our proof strategy is to combine strong approximation theory with anticoncentration bounds for Gaussian random vectors to show that the quantiles of the multiscale statistic $\hat{\Phi}_T$ 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 for example Schmidt-Hieber *et al.* (2013). However, it has not been combined with anticoncentration 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 that were 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 1. The first result is an immediate consequence of theorem 1. It says that the test has the correct (asymptotic) size.

Proposition 1. Let the conditions of theorem 1 be satisfied. Under the null hypothesis H_0, it holds that

$$\mathbb{P}\{\hat{\Psi}_T \leqslant q_T(\alpha)\} = 1 - \alpha + o(1).$$

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

$$m'_T(w) \geqslant c_T \sqrt{\left\{ \frac{\log(T)}{Th^3} \right\}} \qquad \text{for all } w \in [u-h, u+h], \tag{3.9}$$

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

Proposition 2. Let the conditions of theorem 1 be satisfied and consider any sequence of functions m_T with the property (3.9). Then

$$\mathbb{P}\{\hat{\Psi}_T \leqslant q_T(\alpha)\} = o(1).$$

According to proposition 2, our test has asymptotic power 1 against local alternatives of the form (3.9). The proof can be found in the on-line supplementary material.

The next result formally shows that we can make simultaneous confidence statements about the time intervals where the trend m is increasing or decreasing. To formulate it, we define

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

where

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

The object $\Pi_T^{\pm}$ can be interpreted as follows: our multiscale test rejects the null hypothesis $H_0(u,h)$ if $|\hat{\psi}_T(u,h)/\hat{\sigma}| > q_T(\alpha) + \lambda(h)$. 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 $H_0(u,h)$. The objects Π_T^{+} and Π_T^{-} can be interpreted analogously: if $\hat{\psi}_T(u,h)/\hat{\sigma} > q_T(\alpha) + \lambda(h)$, i.e., if $(u,h) \in \mathcal{A}_T^{+}$, then our test rejects $H_0(u,h)$ and indicates an increase in the trend m on the interval $I_{u,h}$, taking into account the positive sign of the statistic $\hat{\psi}_T(u,h)/\hat{\sigma}$. Hence, Π_T^{+} is the collection of time intervals $I_{u,h}$ for which our test indicates an increase in the trend m. Likewise, Π_T^{-} is the collection of intervals for which the test indicates a decrease. Note that $\Pi_T^{\pm}$ (as well as Π_T^{+} and Π_T^{-}) 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. Having defined the objects $\Pi_T^{\pm}$, Π_T^{+} and Π_T^{-}, we now consider the events

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

$E_T^{\pm}$ (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 on-line supplement.

Proposition 3. Let the conditions of theorem 1 be fulfilled. Then for $l \in \{\pm, +, -\}$, it holds that

$$\mathbb{P}(E_T^l) \geqslant 1 - \alpha + o(1).$$

According to proposition 3, we can make simultaneous confidence statements of the following form: with (asymptotic) probability $1-\alpha$ or greater, the trend function m is non-constant (increasing, decreasing) on each interval $I_{u,h} \in \Pi_T^{\pm}$ (Π_T^+, Π_T^-). Hence, our multiscale procedure enables us to identify, with prespecified confidence, time intervals where there is an increase or decrease in the trend m.

Remark 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, i.e. 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 enable us to say, with asymptotic confidence $1-\alpha$ or greater, that $m'(v) \neq 0$ for some $v \in I_{u,h}$. However, we cannot say whether $m'(v) > 0$ or $m'(v) < 0$, i.e. 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. The statement of proposition 3 suggests that we graphically present the results of our multiscale test by plotting the intervals $I_{u,h} \in \Pi_T^l$ for $l \in \{\pm, +, -\}$, i.e. by plotting the intervals where (with asymptotic confidence $1-\alpha$ or greater) 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 $\Pi_T^{l,\min}$ which is constructed as follows: as in Dümbgen (2002), we call an interval $I_{u,h} \in \Pi_T^l$ minimal if there is no other interval $I_{u',h'} \in \Pi_T^l$ with $I_{u',h'} \subset I_{u,h}$. Let $\Pi_T^{l,\min}$ be the set of all minimal intervals in Π_T^l for $l \in \{\pm, +, -\}$ and define the events

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

It is easily seen that $E_T^l = E_T^{l,\min}$ for $l \in \{\pm, +, -\}$. Hence, by proposition 3, it holds that

$$\mathbb{P}(E_T^{l,\min}) \geqslant 1 - \alpha + o(1)$$

for $l \in \{\pm, +, -\}$. This suggests that we plot the minimal intervals in $\Pi_T^{l,\min}$ 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 enables us to make confidence statements for a fixed level of significance $\alpha \in (0,1)$. In some situations, we may be interested in letting $\alpha = \alpha_T \in (0,1) \to 0$ as $T \to \infty$. This situation is considered in the following corollary to proposition 3, whose proof can be found in the on-line supplementary material.

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

Corollary 1 can be interpreted as a consistency result: if we let the level of significance $\alpha = \alpha_T$ go to 0, then the event $E_T^{\pm}$ (E_T^+, E_T^-) occurs with probability tending to 1, i.e. the trend m is non-constant (increasing, decreasing) on each interval $I_{u,h} \in \Pi_T^{\pm}$ (Π_T^+, Π_T^-) with probability tending to 1.

3.4. Comparison with SiZer methods

As already mentioned in Section 1, 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 and decreases of a non-parametric 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, in this section, we let $\mathcal{G}_T = U_T \times H_T$ with U_T being the set of locations and H_T the set of bandwidths or scales. In what follows, we compare our approach with dependent SiZer and point out the most important differences.

Dependent SiZer is based on the statistics $\hat{s}_T(u, h) = \hat{m}'(u, h)/\widehat{\mathrm{sd}}\{\hat{m}'(u, h)\}$, where $\hat{m}'(u, h)$ is a local linear kernel estimator of $m'(u)$ with bandwidth h and $\widehat{\mathrm{sd}}\{\hat{m}'(u, h)\}$ is an estimator of its standard deviation. The statistic $\hat{s}_T(u, h)$ parallels the statistic $\hat{\psi}_T(u, h)/\hat{\sigma}$ 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, as follows.

(a) The global version aggregates the individual statistics $\hat{s}_T(u, h)$ into the overall statistic $\hat{S}_T = \max_{h \in H_T} \hat{S}_T(h)$, where $\hat{S}_T(h) = \max_{u \in U_T} |\hat{s}_T(u, h)|$. The statistic $\hat{S}_T$ is the counterpart to the multiscale statistic $\hat{\Psi}_T$ in our approach.

(b) The rowwise version considers each scale $h \in H_T$ separately. In particular, for each bandwidth $h \in H_T$, a test is carried out based on the statistic $\hat{S}_T(h)$. A rowwise analogue of our approach would be obtained by carrying out a test for each scale $h \in H_T$ separately based on the statistic $\hat{\Psi}_T(h) = \max_{u \in U_T} |\hat{\psi}_T(u, h)/\hat{\sigma}|$. 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.

In practice, SiZer is commonly implemented in its rowwise 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: rowwise SiZer carries out a test *separately* for each scale $h \in H_T$, 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. The main differences are as follows.

(a) Theory for SiZer is derived under the assumption that $H_T \subseteq H$ for all T, where H is a compact subset of $(0, \infty)$. As already pointed out in Chaudhuri and Marron (2000) on page 420, this is quite a severe restriction: only bandwidths h are taken into account that remain bounded away from 0 as the sample size T increases. Bandwidths h that converge to 0 are excluded. Our theory, in contrast, enables us to consider simultaneously bandwidths h of fixed size and bandwidths h that converge to 0 at 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 IID data case and in Park *et al.* (2009) for the dependent data case, $\hat{S}_T$ weakly converges to some limit 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 with this, our proof strategy (which combines

strong approximation theory with anticoncentration bounds as outlined in Section 3.3) does not even require the statistic $\hat{\Psi}_T$ to have a weak limit and is thus not restricted by the limitations of classic weak convergence theory.

(b) There are different ways to combine the test statistics $\hat{S}_T(h) = \max_{u \in U_T} |\hat{s}_T(u,h)|$ for different scales $h \in H_T$. One way is to take their maximum, which leads to the SiZer statistic $\hat{S}_T = \max_{h \in H_T} \hat{S}_T(h)$. We could proceed analogously and consider the statistic $\hat{\Psi}_{T,\text{uncorrected}} = \max_{h \in H_T} \hat{\Psi}_T(h) = \max_{(u,h) \in U_T \times H_T} |\hat{\psi}_T(u,h)/\hat{\sigma}|$. 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_T contains scales h of many rates. Following the lead of Dümbgen and Spokoiny (2001), we consider the test statistic $\hat{\Psi}_T = \max_{(u,h) \in U_T \times H_T} \{|\hat{\psi}_T(u,h)/\hat{\sigma}| - \lambda(h)\}$ with the additive correction terms $\lambda(h)$. Hence, even though related, our multiscale test statistic $\hat{\Psi}_T$ differs from the SiZer statistic $\hat{S}_T$ in important ways.

(c) The main complication in carrying out both our multiscale test and SiZer is to determine the critical values, i.e. the quantiles of the test statistics $\hat{\Psi}_T$ and $\hat{S}_T$ under H_0. To approximate the quantiles, we proceed quite differently from the SiZer literature. The quantiles of the SiZer statistic $\hat{S}_T$ can be approximated by those of the weak limit S. Usually, however, the quantiles of S cannot be determined analytically but must be approximated themselves (e.g. by the bootstrap procedures of Chaudhuri and Marron (1999, 2000)). Alternatively, the quantiles of $\hat{S}_T$ 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 $\hat{\Psi}_T$ under H_0 are approximated by those of a suitably constructed Gaussian analogue of $\hat{\Psi}_T$. 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 that were proposed in Chaudhuri and Marron (1999, 2000). Both procedures can of course be regarded as resampling methods. However, the resampling is done in quite a 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 $\sigma^2 = \sum_{l=-\infty}^{\infty} \text{cov}(\varepsilon_0, \varepsilon_l)$ 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 $\hat{\varepsilon}_t = Y_{t,T} - \hat{m}_h(t/T)$, where $\hat{m}_h$ is a non-parametric estimator of m with the bandwidth or smoothing parameter h. Difference-based methods proceed by estimating σ^2 from the lth differences $Y_{t,T} - Y_{t-l,T}$ 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 non-parametric estimator of the function m and thus do not require us 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 fulfils the weak dependence condition 3. 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 estimates of σ^2 that are produced can be expected to be fairly reliable even though they are a little biased.

Estimators of the long-run error variance σ^2 in model (2.1) have been developed for different kinds of error models. Various researchers have analysed the case of moving average 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 quite a 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 IID with $\mathbb{E}[\eta_t] = 0$ and $\mathbb{E}[\eta_t^2] = \nu^2$. We assume that $A(z) := 1 - \Sigma_{j=1}^{\infty} a_j z^j \neq 0$ for all complex numbers $|z| \leqslant 1 + \delta$ with some small $\delta > 0$, which has the following implications:

(a) $\{\varepsilon_t\}$ is stationary and causal;
(b) the coefficients a_j decay to 0 exponentially fast, i.e. $|a_j| \leqslant C\xi^j$ with some $C > 0$ and $\xi \in (0, 1)$;
(c) $\{\varepsilon_t\}$ has an MA(∞) representation of the form $\varepsilon_t = \Sigma_{k=0}^{\infty} c_k \eta_{t-k}$.

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 $b_0 = 1$ and $b_k = 0$ for $k > 0$. Moreover, they decay to 0 exponentially fast, i.e. $|c_k| \leqslant C\xi^k$ with some $C > 0$ 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 the process (4.1) as

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

where the AR order p^* is treated as unknown.

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 do so, we shall fit AR(p)-type models to $\{\varepsilon_t\}$, where we distinguish between the following two cases (which are referred to as case A and case B).

(a) We do not know the precise AR order p^* but we know an upper bound p on it. In this case, p is a fixed natural number with $p \geqslant p^*$ (case A).
(b) We neither know p^* nor an upper bound on it. In this case, we let $p = p_T \to \infty$ as $T \to \infty$, where formal conditions on the growth of $p = p_T$ are specified later (case B).

To simplify the notation, we let $\Delta_l Z_t = Z_t - Z_{t-l}$ denote the lth 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

$$\Delta_q Y_{t,T} = \varepsilon_t - \varepsilon_{t-q} + m\left(\frac{t}{T}\right) - m\left(\frac{t-q}{T}\right) = \Delta_q \varepsilon_t + O\left(\frac{q}{T}\right). \tag{4.5}$$

Taken together, equations (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 method.

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

$$\gamma_q(l) - \sum_{j=1}^{p^*} a_j \gamma_q(l-j) = \begin{cases} -\nu^2 c_{q-l} & \text{for } 1 \leqslant l < q+1, \\ 0 & \text{for } l \geqslant q+1, \end{cases} \tag{4.6}$$

where $\gamma_q(l) = \operatorname{cov}(\Delta_q \varepsilon_t,\ \Delta_q \varepsilon_{t-l})$ and c_k are the coefficients from the MA(∞) expansion of $\{\varepsilon_t\}$. Combining equations (4.6) for $l = 1, \ldots, p$, we obtain that

$$\mathbf{\Gamma}_q \mathbf{a} = \boldsymbol{\gamma}_q + \nu^2 \mathbf{c}_q - \boldsymbol{\rho}_q, \tag{4.7}$$

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

$$\tilde{\mathbf{a}}_q = \hat{\mathbf{\Gamma}}_q^{-1} \hat{\boldsymbol{\gamma}}_q, \tag{4.8}$$

where $\hat{\mathbf{\Gamma}}_q$ and $\hat{\boldsymbol{\gamma}}_q$ are defined analogously to $\mathbf{\Gamma}_q$ and $\boldsymbol{\gamma}_q$ with $\gamma_q(l)$ replaced by the sample autocovariances $\hat{\gamma}_q(l) = (T-q)^{-1} \sum_{t=q+l+1}^{T} \Delta_q Y_{t,T} \Delta_q Y_{t-l,T}$ and $q = q_T \to \infty$ as $T \to \infty$. For our theory to work, we require that $q/p \to \infty$, i.e. q needs to grow faster than p. Formal conditions on the growth of q are given later.

The estimator $\tilde{\mathbf{a}}_q$ depends on the tuning parameter q, i.e. 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.

(a) q should be chosen sufficiently large to ensure that the vector $\mathbf{c}_q = (c_{q-1}, \ldots, c_{q-p})^{\mathrm{T}}$ is close to zero. As we have already seen, the constants c_k decay to 0 exponentially fast and can be computed from the recursive equations (4.2) for given parameters $a_1, a_2, a_3, \ldots$. In the special case of an AR(1) process, for example, one can readily calculate that $c_k \leqslant 0.0035$ for any $k \geqslant 20$ and any $|a_1| \leqslant 0.75$. Hence, if we have an AR(1) model for the errors ε_t and the error process is not too persistent, choosing $q \geqslant 20$ should make sure that $\mathbf{c}_q$ is close to 0. Generally speaking, the recursive equations (4.2) can be used to obtain some idea for which values of q the vector $\mathbf{c}_q$ can be expected to be approximately 0.

(b) q should not be chosen too large to ensure that the trend m is appropriately eliminated by taking qth differences.

As long as the trend m is not very strong, the two requirements (a) and (b) can be fulfilled without much difficulty. For example, by choosing $q = 20$ in the AR(1) case just discussed, we not only take care of point (a) 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 $\tilde{\mathbf{a}}_q$ will have a strong bias. To reduce this bias, we refine our estimation procedure as follows: by solving the recursive equations (4.2) with $\mathbf{a}$ replaced by $\tilde{\mathbf{a}}_q$, we can compute estimators $\tilde{c}_k$ of the coefficients c_k and thus estimators $\tilde{\mathbf{c}}_r$ of the vectors $\mathbf{c}_r$ for any $r \geqslant 1$. Moreover, the innovation variance ν^2 can be estimated by $\tilde{\nu}^2 = (2T)^{-1}\Sigma_{t=p+2}^{T}\tilde{r}_{t,T}^2$, where $\tilde{r}_{t,T} = \Delta_1 Y_{t,T} - \Sigma_{j=1}^{p}\tilde{a}_j \Delta_1 Y_{t-j,T}$ and $\tilde{a}_j$ is the jth entry of the vector $\tilde{\mathbf{a}}_q$. Plugging the expressions $\hat{\mathbf{\Gamma}}_r$, $\hat{\gamma}_r$, $\tilde{\mathbf{c}}_r$ and $\tilde{\nu}^2$ into equation (4.7), we can estimate $\mathbf{a}$ by

$$\hat{\mathbf{a}}_r = \hat{\mathbf{\Gamma}}_r^{-1}(\hat{\gamma}_r + \tilde{\nu}^2 \tilde{\mathbf{c}}_r), \tag{4.9}$$

where r is a much smaller differencing order than q. Specifically, in case A, we can choose r to be any fixed number $r \geqslant 1$. Unlike q, the parameter r thus remains bounded as T increases. In case B, our theory enables us to choose any number r with $r \geqslant (1+\delta)p$ for some small $\delta > 0$. Since $q/p \to \infty$, it holds that $q/r \to \infty$ as well, which means that r is of smaller order than q. Hence, in both case A and case B, the estimator $\hat{\mathbf{a}}_r$ is based on a differencing order r that is much smaller than q; only the pilot estimator $\tilde{\mathbf{a}}_q$ relies on differences of the larger order q. As a consequence, $\hat{\mathbf{a}}_r$ should eliminate the trend m more appropriately and should thus be less biased than the pilot estimator $\tilde{\mathbf{a}}_q$. To make the method more robust against estimation errors in $\tilde{\mathbf{c}}_r$, we finally average the estimators $\hat{\mathbf{a}}_r$ for a few values of r. In particular, we define

$$\hat{\mathbf{a}} = \frac{1}{\bar{r} - \underline{r} + 1}\sum_{r=\underline{r}}^{\bar{r}} \hat{\mathbf{a}}_r, \tag{4.10}$$

where $\underline{r}$ and $\bar{r}$ are chosen as follows: in case A, we let $\underline{r}$ and $\bar{r}$ be small natural numbers. In case B, we set $\underline{r} = (1-\delta)p$ for some small $\delta > 0$ and choose $\bar{r}$ such that $\bar{r} - \underline{r}$ remains bounded. For ease of notation, we suppress the dependence of $\hat{\mathbf{a}}$ on the parameters $\underline{r}$ and $\bar{r}$. Once $\hat{\mathbf{a}} = (\hat{a}_1, \ldots, \hat{a}_p)^{\mathrm{T}}$ has been computed, the long-run variance σ^2 can be estimated by

$$\hat{\sigma}^2 = \frac{\hat{\nu}^2}{(1 - \sum_{j=1}^{p}\hat{a}_j)^2}, \tag{4.11}$$

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

We briefly compare the estimator $\hat{\mathbf{a}}$ with competing methods. Presumably closest to our approach is that of Hall and Van Keilegom (2003) which was designed for AR(p^*) processes of known finite order p^*. For comparing the two methods, we thus assume that p^* is known and set $p = p^*$. 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, i.e. even when the AR polynomial $A(z) = 1 - \Sigma_{j=1}^{p^*}a_j z^j$ has a root that is 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 $\tilde{\mathbf{a}}_q = (\tilde{a}_1, \ldots, \tilde{a}_{p^*})^{\mathrm{T}}$ has the property that the estimated AR polynomial $\tilde{A}(z) = 1 - \Sigma_{j=1}^{p^*}\tilde{a}_j z^j$ has no root inside the unit disc, i.e. $\tilde{A}(z) \neq 0$ for all complex numbers z with $|z| \leqslant 1$. (More precisely, $\tilde{A}(z) \neq 0$ for all z with $|z| \leqslant 1$, whenever the covariance matrix $(\hat{\gamma}_q(i-j) : 1 \leqslant i, j \leqslant p^*+1)$ is non-singular. Moreover, $(\hat{\gamma}_q(i-j) : 1 \leqslant i, j \leqslant p^*+1)$ is non-singular whenever $\hat{\gamma}_q(0) > 0$, which is the generic case.) Hence, the fitted AR model

with the coefficients $\tilde{\mathbf{a}}_q$ 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 $\tilde{\mathbf{a}}_q$: 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 reliable behaviour of the estimator $\hat{\mathbf{a}}$ 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 5.2. A striking example is Fig. 6 there, which presents the simulation results for the case of an AR(1) process $\varepsilon_t = a_1 \varepsilon_{t-1} + \eta_t$ with $a_1 = -0.95$ and clearly shows the much better performance of our method.

(b) Both our pilot estimator $\tilde{\mathbf{a}}_q$ and the estimator of Hall and Van Keilegom (2003) tend to have a substantial bias when the trend m is pronounced. Our estimator $\hat{\mathbf{a}}$ reduces this bias considerably as demonstrated in the simulations of Section 5.2. 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 $\tilde{\mathbf{a}}_q$, $\hat{\mathbf{a}}$ and $\hat{\sigma}^2$. To formulate the following result, we use the shorthand $v_T \ll w_T$ which means that $v_T / w_T \to 0$ as $T \to \infty$.

Proposition 4. Let m be Lipschitz continuous and suppose that $\{\varepsilon_t\}$ is an AR(p^*) process of the form (4.3) with the following properties: $A(z) \neq 0$ for all $|z| \leqslant 1 + \delta$ with some small $\delta > 0$ and the innovations η_t have a finite fourth moment. Assume that p, q, $\underline{r}$ and $\bar{r}$ satisfy the following conditions. In case A, p, $\underline{r}$ and $\bar{r}$ are fixed natural numbers and $\log(T) \ll q \ll \sqrt{T}$. In case B, $C \log(T) \leqslant p \ll \min\{T^{1/5}, q\}$ for some sufficiently large C, $q \ll \sqrt{T}$, $\underline{r} = (1+\delta)p$ for some small $\delta > 0$ and $\bar{r} - \underline{r}$ remains bounded. Under these conditions, $\tilde{\mathbf{a}}_q - \mathbf{a} = O_p\{\sqrt{(p/T)}\}$ as well as $\hat{\mathbf{a}} - \mathbf{a} = O_p\{\sqrt{(p^3/T)}\}$ and $\hat{\sigma}^2 - \sigma^2 = O_p\{\sqrt{(p^4/T)}\}$.

The proof is provided in the on-line supplementary material. As we can see, the convergence rate of the second-step estimator $\hat{\mathbf{a}}$ is somewhat slower than that of the pilot estimator $\tilde{\mathbf{a}}_q$. Hence, from an asymptotic perspective, there is no gain from using the second-step estimator. Nevertheless, in finite samples, the estimator $\hat{\mathbf{a}}$ vastly outperforms $\tilde{\mathbf{a}}_q$ since it considerably reduces the bias of the latter.

5. Simulations

5.1. Small sample properties of the multiscale test

In what follows, we investigate the performance of our multiscale test and compare it with the dependent SiZer methods from Park *et al.* (2004, 2009) and Rondonotti *et al.* (2007). We consider the following versions of our multiscale test and SiZer.

(a) $\mathcal{T}_{\mathrm{MS}}$ is our multiscale test with the statistic $\hat{\Psi}_T = \max_{h \in H_T}\{\hat{\Psi}_T(h) - \lambda(h)\}$, where $\hat{\Psi}_T(h) = \max_{u \in U_T} |\hat{\psi}_T(u,h)/\hat{\sigma}|$. 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.

(b) $\mathcal{T}_{\mathrm{UC}}$ is the uncorrected version of our multiscale test with the test statistic $\hat{\Psi}_{T,\mathrm{uncorrected}} = \max_{h \in H_T} \hat{\Psi}_T(h)$, which has already been introduced in equation (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.

(c) $\mathcal{T}_{\mathrm{RW}}$ is a rowwise (or scalewise) version of our multiscale test as briefly mentioned in Section 3.4. This version carries out a test scalewise, i.e. separately for each scale $h \in H_T$ based on the statistic $\hat{\Psi}_T(h)$. Note the following.
   (i) For each $h \in H_T$, the test based on $\hat{\Psi}_T(h)$ 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 $\hat{\Psi}_T(h)$ by subtracting $\lambda(h)$ or not, since $\lambda(h)$ acts as a fixed constant when only one bandwidth h is taken into account.

(d) $\mathcal{T}_{\mathrm{SiZer}}$ is the rowwise version of dependent SiZer from Park *et al.* (2004, 2009) and Rondonotti *et al.* (2007). We do not consider a global version of dependent SiZer as such a version was not fully developed in Park *et al.* (2004, 2009) and Rondonotti *et al.* (2007).

The simulation set-up 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 IID standard normal and the AR order $p^* = 1$ is treated as known. To simulate data under the null, 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. For each model specification, we simulate $S = 1000$ data samples and carry out the tests $\mathcal{T}_{\mathrm{MS}}$, $\mathcal{T}_{\mathrm{UC}}$, $\mathcal{T}_{\mathrm{RW}}$ and $\mathcal{T}_{\mathrm{SiZer}}$ for each simulated sample.

To implement our multiscale test $\mathcal{T}_{\mathrm{MS}}$, we choose K to be an Epanechnikov kernel and let $\mathcal{G}_T = U_T \times H_T$ with

$$U_T = \{u \in [0,1] : u = 5t/T \text{ for some } t \in \mathbb{N}\},$$

$$H_T = \{h \in [\log(T)/T, \tfrac{1}{4}] : h = 5l/T \text{ for some } l \in \mathbb{N}\}.$$

We thus take into account all locations u on an equidistant grid U_T with step length $5/T$ and all bandwidths $h = 5/T, 10/T, 15/T, \ldots$ with $\log(T)/T \leqslant h \leqslant \frac{1}{4}$. Note that the lower bound $\log(T)/T$ is motivated by condition 6 which requires that $\log(T)/T \ll h_{\min}$ (given that all moments of ε_t exist). As a robustness check, we have rerun the simulations for other grids. As the results are very similar, we do not, however, report them here. To estimate the long-run error variance σ^2, we apply the procedure from Section 4 with $\underline{r} = 1$ and $\bar{r} = 10$ and the following choices of q: for $a_1 \in \{-0.5, -0.25, 0.25, 0.5\}$, we set $q = 25$. As already discussed in Section 4, this should be an appropriate choice 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 very strongly correlated, larger values of q are required to produce precise estimates of σ^2. In the case of AR(1) 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, $\underline{r}$ and $\bar{r}$ is explored more systematically in Section 5.2. To compute the critical values of the multiscale test $\mathcal{T}_{\mathrm{MS}}$, we simulate 5000 values of the statistic Φ_T defined in Section 3.2 and compute their empirical $(1-\alpha)$-quantile $q_T(\alpha)$. The uncorrected and rowwise 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 on-line supplementary material.

5.1.1. Size simulations

The first part of our simulation study investigates the size properties of the four tests $\mathcal{T}_{\mathrm{MS}}$, $\mathcal{T}_{\mathrm{UC}}$, $\mathcal{T}_{\mathrm{RW}}$ and $\mathcal{T}_{\mathrm{SiZer}}$ under the null that the trend m is constant. To start with, we focus on

the multiscale test $\mathcal{T}_{\mathrm{MS}}$. Table 1 reports the actual size of $\mathcal{T}_{\mathrm{MS}}$ for the AR parameters $a_1 \in \{-0.5, -0.25, 0.25, 0.5\}$, which is computed as the number of simulations in which $\mathcal{T}_{\mathrm{MS}}$ rejects the null divided by the total number of simulations. As can be seen, the actual size of the multiscale test $\mathcal{T}_{\mathrm{MS}}$ is fairly close to the nominal target α for all the AR parameters and sample sizes considered. Hence, the test has approximately the correct size.

In Table 1, we have explored the size of $\mathcal{T}_{\mathrm{MS}}$ when the errors are moderately auto-correlated. The case of strongly auto-correlated errors is investigated in Table 2, where we consider AR(1) errors with $a_1 \in \{-0.9, 0.9\}$. We first discuss the results for the positive AR parameter $a_1 = 0.9$. As can be seen, the size numbers are substantially downwardly biased for small sample sizes, in particular, for $T = 250$ and $T = 500$. As the sample size increases, this downward bias diminishes and the size numbers stabilize around their target α. In particular, for $T \geqslant 1000$, the size numbers give a decent approximation to α. An analogous picture arises for the negative AR parameter $a_1 = -0.9$. The size numbers, however, are upwardly rather than downwardly biased for small sample sizes T and the size distortions appear to vanish a little more slowly as T increases. To summarize, in the case of strongly auto-correlated 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 auto-correlated data is a very difficult problem in general and satisfying results can only be expected for fairly large sample sizes.

We next compare our multiscale test $\mathcal{T}_{\mathrm{MS}}$ with $\mathcal{T}_{\mathrm{UC}}$, $\mathcal{T}_{\mathrm{RW}}$ and $\mathcal{T}_{\mathrm{SiZer}}$ in terms of size. There is an important difference between $\mathcal{T}_{\mathrm{MS}}$ and $\mathcal{T}_{\mathrm{UC}}$ on the one hand and $\mathcal{T}_{\mathrm{RW}}$ and $\mathcal{T}_{\mathrm{SiZer}}$ on the other. $\mathcal{T}_{\mathrm{MS}}$ and its uncorrected version $\mathcal{T}_{\mathrm{UC}}$ are *global* test procedures: they test $H_0(u,h)$ simultaneously for all locations $u \in U_T$ and scales $h \in H_T$. Hence, they control the size simultaneously over both locations u and scales h. The methods $\mathcal{T}_{\mathrm{RW}}$ and $\mathcal{T}_{\mathrm{SiZer}}$, in contrast, are *rowwise* (or *scalewise*) in nature: they test the hypothesis $H_0(u,h)$ simultaneously for all $u \in U_T$ but separately for each scale $h \in H_T$. Hence, they control the size for each scale $h \in H_T$ separately.

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

T	*Results for $a_1 = -0.5$ and the following nominal sizes α:*			*Results for $a_1 = -0.25$ and the following nominal sizes α:*			*Results for $a_1 = 0.25$ and the following nominal sizes α:*			*Results for $a_1 = 0.5$ and the following nominal sizes 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*
250	0.013	0.040	0.086	0.016	0.054	0.106	0.009	0.045	0.094	0.014	0.058	0.106
500	0.013	0.044	0.102	0.008	0.041	0.089	0.013	0.057	0.107	0.014	0.056	0.101
1000	0.011	0.052	0.090	0.007	0.057	0.114	0.011	0.049	0.106	0.007	0.050	0.098

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

α	*Results for $a_1 = -0.9$ and the following sample sizes T:*					*Results for $a_1 = 0.9$ and the following sample sizes T:*				
	250	*500*	*1000*	*2000*	*3000*	*250*	*500*	*1000*	*2000*	*3000*
0.01	0.040	0.032	0.017	0.009	0.012	0.003	0.016	0.015	0.021	0.017
0.05	0.137	0.093	0.067	0.061	0.047	0.017	0.038	0.055	0.059	0.057
0.1	0.218	0.160	0.124	0.108	0.098	0.040	0.054	0.095	0.096	0.106

Table 3. Global size comparisons for the level of significance $\alpha = 0.05$

T	*Results for* $a_1 = -0.5$				*Results for* $a_1 = 0.5$			
	$\mathcal{T}_{\mathrm{MS}}$	$\mathcal{T}_{\mathrm{UC}}$	$\mathcal{T}_{\mathrm{RW}}$	$\mathcal{T}_{\mathrm{SiZer}}$	$\mathcal{T}_{\mathrm{MS}}$	$\mathcal{T}_{\mathrm{UC}}$	$\mathcal{T}_{\mathrm{RW}}$	$\mathcal{T}_{\mathrm{SiZer}}$
250	0.069	0.065	0.230	0.333	0.049	0.048	0.143	0.289
500	0.054	0.065	0.288	0.448	0.042	0.026	0.187	0.397
1000	0.046	0.051	0.318	0.522	0.052	0.049	0.276	0.509

We conduct some simulation exercises to illustrate this important distinction. To keep the simulation study to a reasonable length, we restrict attention to the level of significance $\alpha = 0.05$ and the AR parameters $a_1 \in \{-0.5, 0.5\}$. To simplify the implementation of $\mathcal{T}_{\mathrm{SiZer}}$, we assume that the autocovariance function of the error process and thus the long-run error variance σ^2 are 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 location–scale grid 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 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 $\mathcal{G}_T^* = \{(u, h) \in \mathcal{G}_T : \mathrm{ESS}^*(u, h) \geqslant 5\}$. A definition of $\mathrm{ESS}^*(u, h)$ is given in section S.3 of the on-line supplement.

For our simulation exercises, we distinguish between global and rowwise (or scalewise) size: global size is defined as the percentage of simulations in which the test under consideration rejects $H_0(u, h)$ for some $(u, h) \in \mathcal{G}_T^*$. Hence, it is identical to the size as computed in Tables 1 and 2. Rowwise size for scale $h^* \in H_T$, in contrast, is the percentage of simulations in which the test rejects $H_0(u, h^*)$ for some $(u, h^*) \in \mathcal{G}_T^*$. Table 3 reports the global size of the four tests. As can be seen, the size numbers of our multiscale test $\mathcal{T}_{\mathrm{MS}}$ and its uncorrected version $\mathcal{T}_{\mathrm{UC}}$ are reasonably close to the target $\alpha = 0.05$. The global size numbers of the rowwise methods $\mathcal{T}_{\mathrm{RW}}$ and $\mathcal{T}_{\mathrm{SiZer}}$, in contrast, are much larger than the target $\alpha = 0.05$. Since the number of scales h in the grid $\mathcal{G}_T^*$ increases with T, they even move away from α as the sample size T increases. To summarize, as expected, the global tests $\mathcal{T}_{\mathrm{MS}}$ and $\mathcal{T}_{\mathrm{UC}}$ hold the size reasonably well, whereas the rowwise methods $\mathcal{T}_{\mathrm{RW}}$ and $\mathcal{T}_{\mathrm{SiZer}}$ are much too liberal.

Fig. 2 reports the rowwise size of the four tests by so-called parallel co-ordinate plots (Inselberg, 1985) for the sample size $T = 1000$. Each curve in Fig. 2 specifies the rowwise size of one of the tests for the scales h under consideration. As can be seen, the rowwise version $\mathcal{T}_{\mathrm{RW}}$ of our multiscale test holds the size quite accurately across scales. The rowwise size of $\mathcal{T}_{\mathrm{SiZer}}$ also gives an acceptable approximation to the target $\alpha = 5\%$, even though the size numbers are upwardly biased quite considerably. The global tests $\mathcal{T}_{\mathrm{MS}}$ and $\mathcal{T}_{\mathrm{UC}}$, in contrast, have a rowwise size that is much smaller than the target $\alpha = 5\%$, which reflects the fact that they control global rather than rowwise size.

5.1.2 Power comparisons

In the second part of our simulation study, we compare the tests $\mathcal{T}_{\mathrm{MS}}$, $\mathcal{T}_{\mathrm{UC}}$, $\mathcal{T}_{\mathrm{RW}}$ and $\mathcal{T}_{\mathrm{SiZer}}$ in terms of power. As above, we use the location–scale grid $\mathcal{G}_T^*$ and treat the autocovariance function of the error terms as known when implementing the tests. Moreover, we restrict attention to the level of significance $\alpha = 0.05$ and the AR parameters $a_1 \in \{-0.5, 0.5\}$. Our simulation exercises investigate the ability of the four tests to detect local increases in the trend m. (The same could of

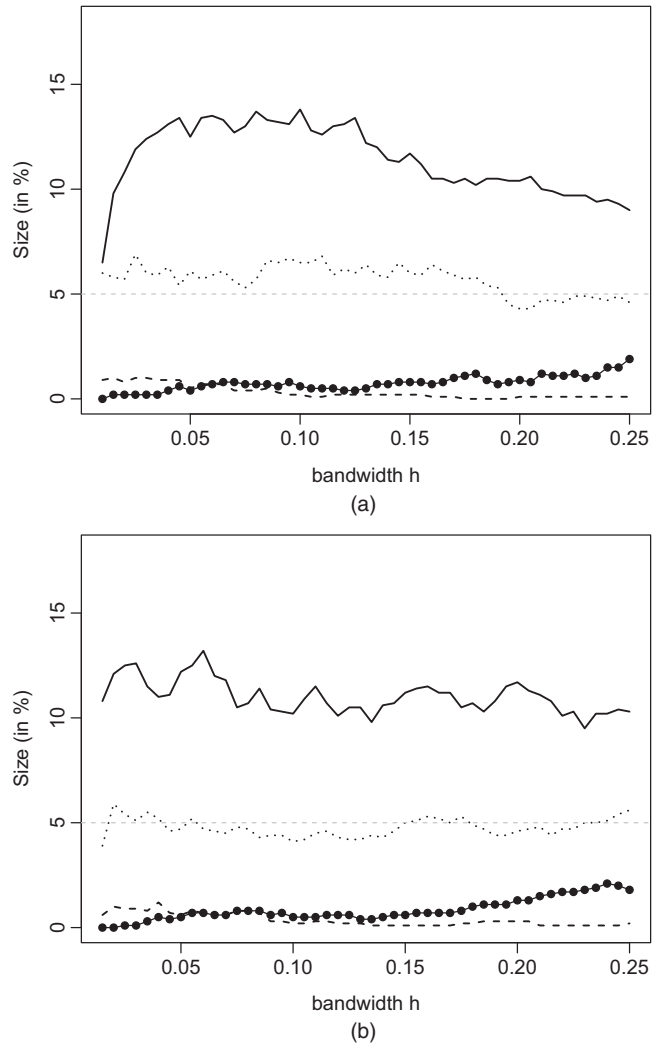

Fig. 2. Rowwise size comparisons for level of significance $\alpha = 5\%$ and sample size $T = 1000$ (each curve shows the rowwise size as a function of the bandwidth h for one of the four tests $\mathcal{T}_{\text{MS}}$ (●), $\mathcal{T}_{\text{UC}}$ (- - - - -), $\mathcal{T}_{\text{RW}}$ (· · · · · ·) and $\mathcal{T}_{\text{SiZer}}$ (———)): (a) $a_1 = -0.5$; (b) $a_1 = 0.5$

course be done for decreases.) The tests indicate a local increase in m according to the following decision rules: for each $(u,h) \in \mathcal{G}_T^*$,

$$\begin{aligned}
\mathcal{T}_{\text{MS}} \text{ indicates an increase on } [u-h,u+h] &\Leftrightarrow \hat{\psi}_T(u,h)/\hat{\sigma} > q_T(\alpha) + \lambda(h),\\
\mathcal{T}_{\text{UC}} \text{ indicates an increase on } [u-h,u+h] &\Leftrightarrow \hat{\psi}_T(u,h)/\hat{\sigma} > q_T^{\text{UC}}(\alpha),\\
\mathcal{T}_{\text{RW}} \text{ indicates an increase on } [u-h,u+h] &\Leftrightarrow \hat{\psi}_T(u,h)/\hat{\sigma} > q_T^{\text{RW}}(\alpha,h),\\
\mathcal{T}_{\text{SiZer}} \text{ indicates an increase on } [u-h,u+h] &\Leftrightarrow \hat{s}_T(u,h) > q_T^{\text{SiZer}}(\alpha,h),
\end{aligned}$$

where $q_T^{\text{UC}}(\alpha)$, $q_T^{\text{RW}}(\alpha,h)$ and $q_T^{\text{SiZer}}(\alpha,h)$ are the critical values of $\mathcal{T}_{\text{UC}}$, $\mathcal{T}_{\text{RW}}$ and $\mathcal{T}_{\text{SiZer}}$ respectively. Note that the critical values of $\mathcal{T}_{\text{RW}}$ and $\mathcal{T}_{\text{SiZer}}$ depend on the scale h as these are rowwise procedures.

To be able to make systematic power comparisons, we consider a very simple trend function m. More complicated signals m are analysed in section S.3 of the on-line supplementary material. The trend function that we are considering here is defined as $m(u) = c\mathbf{1}(u \in [0.45, 0.55])[1 - \{(u - 0.5)/0.05\}^2]^2$, where $c = 0.45$ in the AR case with $a_1 = -0.5$ and $c = 1.3$ in the case with $a_1 = 0.5$. The function m is increasing on $I^+ = (0.45, 0.5)$, decreasing on $I^- = (0.5, 0.55)$ and constant elsewhere. Figs 3(a) and 3(d) give a graphical illustration of m, where the grey line in the background is the time series path of a representative simulated data sample. As can be seen, m is a small bump around $u = 0.5$, where c determines the height of the bump. The constant c is chosen such that the bump is difficult but not impossible to detect for the four tests. We distinguish between the following types of power for the tests $\mathcal{T}_j$ with $j \in \{\text{MS}, \text{UC}, \text{RW}, \text{SiZer}\}$, where we restrict attention to increases in m:

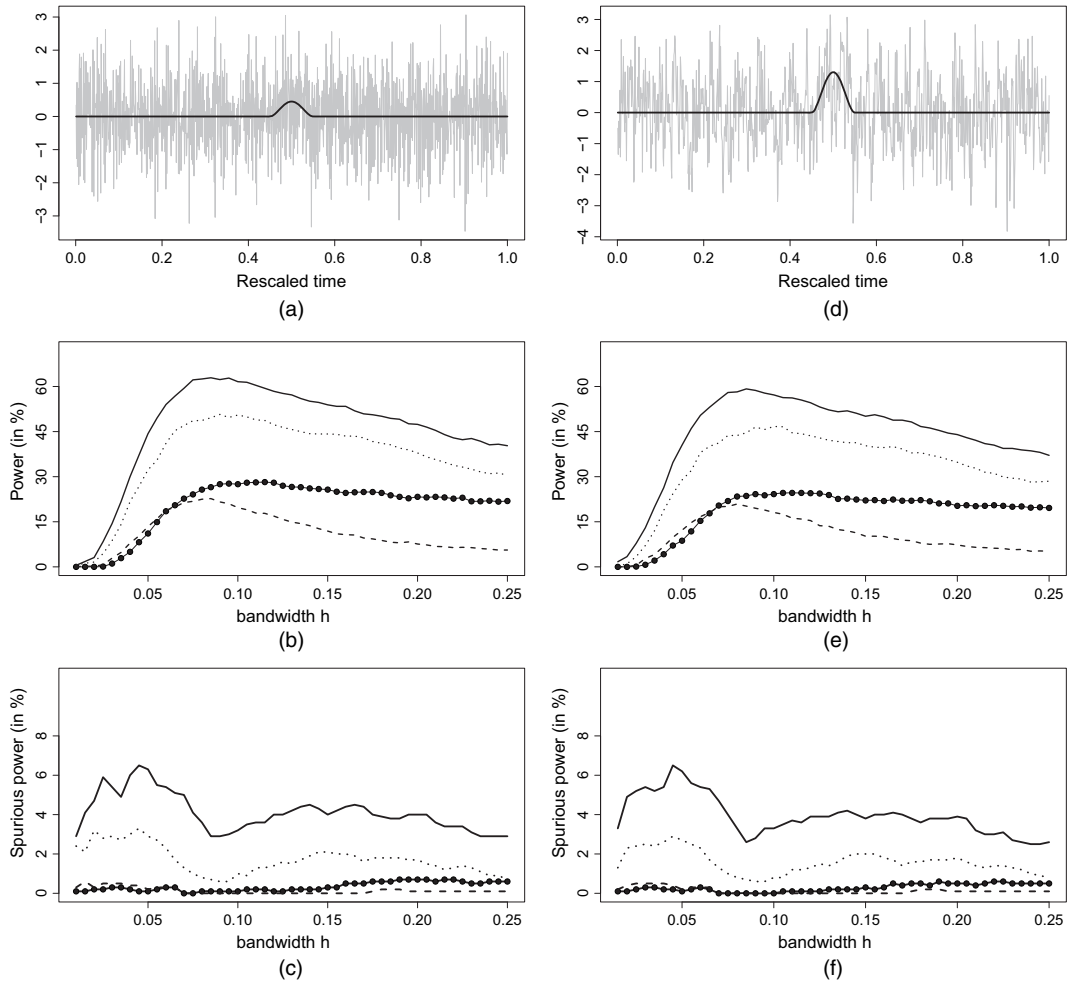

Fig. 3. Rowwise power and rowwise spurious power comparisons for $\alpha = 5\%$ and $T = 1000$: (a)–(c) $a_1 = -0.5$; (d)–(f) $a_1 = 0.5$; (a), (d) bump function m with a representative data sample in the background; (b), (e) parallel co-ordinate plot reporting rowwise power (in particular, each curve shows the rowwise power as a function of the bandwidth h for one of the four tests $\mathcal{T}_{\text{MS}}$ (●), $\mathcal{T}_{\text{UC}}$ (- - - - -), $\mathcal{T}_{\text{RW}}$ (· · · · · ·) and $\mathcal{T}_{\text{SiZer}}$ (———)); (c), (f) parallel co-ordinate plot reporting rowwise spurious power in an analogous fashion

(a) global power; the percentage of simulation runs in which the test $\mathcal{T}_j$ indicates an increase on some interval $I_{u,h} = [u-h, u+h]$ where m is indeed increasing, i.e. on some $I_{u,h}$ with $I_{u,h} \cap I^+ \neq \emptyset$;
(b) spurious global power; the percentage of simulation runs in which the test $\mathcal{T}_j$ indicates an increase on some interval $I_{u,h} = [u-h, u+h]$ where m is not increasing, i.e. on some $I_{u,h}$ with $I_{u,h} \cap I^+ = \emptyset$;
(c) rowwise power on scale h^*; the percentage of simulation runs in which the test $\mathcal{T}_j$ indicates an increase on some interval $I_{u,h^*} = [u-h^*, u+h^*]$ where m is indeed increasing, i.e. on some I_{u,h^*} with $I_{u,h^*} \cap I^+ \neq \emptyset$;
(d) spurious rowwise power on scale h^*; the percentage of simulation runs in which the test $\mathcal{T}_j$ indicates an increase on some interval $I_{u,h^*} = [u-h^*, u+h^*]$ where m is not increasing, i.e. on some I_{u,h^*} with $I_{u,h^*} \cap I^+ = \emptyset$.

Table 4 reports the global power and global spurious power of the four tests. As can be seen, our multiscale test $\mathcal{T}_{\mathrm{MS}}$ has higher power than the uncorrected version $\mathcal{T}_{\mathrm{UC}}$. This confirms the theoretical optimality theory in Dümbgen and Spokoiny (2001) (see also Dümbgen and Walther (2008) and Rufibach and Walther (2010)) according to which the aggregation scheme of $\mathcal{T}_{\mathrm{MS}}$ with its additive correction term should yield better power properties than the simpler scheme of $\mathcal{T}_{\mathrm{UC}}$. As expected, the rowwise methods $\mathcal{T}_{\mathrm{RW}}$ and $\mathcal{T}_{\mathrm{SiZer}}$ have substantially more power than the global tests. Indeed, $\mathcal{T}_{\mathrm{SiZer}}$ is even a little more powerful than $\mathcal{T}_{\mathrm{RW}}$, which is presumably because it is somewhat too liberal in terms of rowwise size as observed in Fig. 2. The higher power of the rowwise procedures comes at some cost: their spurious global power is much higher than that of the global tests. For the sample size $T = 1000$ and the AR parameter $a_1 = -0.5$, for example, $\mathcal{T}_{\mathrm{SiZer}}$ spuriously finds an increase in the trend m in more than 28% of the simulations, and $\mathcal{T}_{\mathrm{RW}}$ in more than 15%. The multiscale test $\mathcal{T}_{\mathrm{MS}}$ (as well as its uncorrected version $\mathcal{T}_{\mathrm{UC}}$), in contrast, controls the probability of finding a spurious increase. In particular, as implied by proposition 3, its spurious global power is below $100\alpha\% = 5\%$.

Fig. 3 gives a more detailed picture of the power properties of the four tests for the sample size $T = 1000$. The parallel co-ordinate plots of Fig. 3 show how power and spurious power are distributed across scales h. Let us first have a look at the rowwise methods. As can be seen, $\mathcal{T}_{\mathrm{SiZer}}$ is more powerful than $\mathcal{T}_{\mathrm{RW}}$ on all scales under consideration. As already mentioned when discussing the global power results, this is presumably because $\mathcal{T}_{\mathrm{SiZer}}$ is a little too liberal in terms of rowwise size. Comparing the power curves of the two global methods gives an interesting insight: our multiscale test $\mathcal{T}_{\mathrm{MS}}$ has substantially more power than the uncorrected version $\mathcal{T}_{\mathrm{UC}}$ on medium and large scales. On small scales, in contrast, it is slightly less powerful than $\mathcal{T}_{\mathrm{UC}}$.

Table 4. Global power and global spurious power comparisons for $\alpha = 0.05$

T		*Results for* $a_1 = -0.5$				*Results for* $a_1 = 0.5$			
		$\mathcal{T}_{\mathrm{MS}}$	$\mathcal{T}_{\mathrm{UC}}$	$\mathcal{T}_{\mathrm{RW}}$	$\mathcal{T}_{\mathrm{SiZer}}$	$\mathcal{T}_{\mathrm{MS}}$	$\mathcal{T}_{\mathrm{UC}}$	$\mathcal{T}_{\mathrm{RW}}$	$\mathcal{T}_{\mathrm{SiZer}}$
250	Power	0.102	0.086	0.228	0.328	0.096	0.079	0.190	0.295
	Spurious power	0.021	0.032	0.109	0.166	0.012	0.017	0.054	0.131
500	Power	0.212	0.166	0.464	0.617	0.186	0.160	0.406	0.587
	Spurious power	0.020	0.024	0.137	0.212	0.016	0.016	0.082	0.192
1000	Power	0.575	0.425	0.817	0.901	0.526	0.394	0.780	0.884
	Spurious power	0.023	0.024	0.158	0.283	0.020	0.019	0.123	0.252

This again illustrates the theoretical optimality theory in Dümbgen and Spokoiny (2001) which suggests that, asymptotically, the multiscale test $\mathcal{T}_{\mathrm{MS}}$ should be as powerful as $\mathcal{T}_{\mathrm{UC}}$ on small scales but more powerful on large scales. This is essentially what we see in Figs 3(b) and 3(e). Of course, $\mathcal{T}_{\mathrm{MS}}$ does not have exactly as much power as $\mathcal{T}_{\mathrm{UC}}$ on fine scales. However, the loss of power on fine scales is very small compared with the gain of power on larger scales (which is also reflected by the fact that $\mathcal{T}_{\mathrm{MS}}$ has more global power than $\mathcal{T}_{\mathrm{UC}}$).

The main findings of our simulation exercises can be summarized as follows: if we are interested in an exploratory data tool for finding local increases or decreases of a trend, the rowwise methods $\mathcal{T}_{\mathrm{RW}}$ and $\mathcal{T}_{\mathrm{SiZer}}$ both do a good job. However, if we want to make rigorous statistical inference simultaneously across locations and scales, we need to opt for a global method. Our simulation exercises have demonstrated that our multiscale test $\mathcal{T}_{\mathrm{MS}}$ is a global method which enjoys good size and power properties. In particular, as predicted by the theory, it is a more effective test than the uncorrected version $\mathcal{T}_{\mathrm{UC}}$.

5.2. *Small sample properties of the long-run variance estimator*

In the final part of our simulation study, we analyse the estimators of the AR parameters and of the long-run error variance from Section 4 and compare them with the estimators of Hall and Van Keilegom (2003). 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 IID standard normal innovation terms. Throughout the simulation study, the AR order $p^* = 1$ is treated as known. We report our findings for the sample size $T = 500$; 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 slopes β: one corresponding to a moderate and one to a pronounced trend m. In particular, we let $\beta = s_\beta\sqrt{\mathrm{var}(\varepsilon_t)}$ with $s_\beta \in \{1, 10\}$. When $s_\beta = 1$, the slope β is equal to the standard deviation $\sqrt{\mathrm{var}(\varepsilon_t)}$ 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 $\sqrt{\mathrm{var}(\varepsilon_t)}$, which results in quite a pronounced trend m.

For each model specification, we generate $S = 1000$ data samples and compute the following quantities for each simulated sample:

(a) the pilot estimator $\tilde{a}_q$ from equation (4.8) with the tuning parameter q, the estimator $\hat{a}$ from equation (4.10) with the tuning parameters $(\underline{r}, \bar{r})$ and the long-run variance estimator $\hat{\sigma}^2$ from equation (4.11);
(b) the estimators of a_1 and σ^2 from Hall and Van Keilegom (2003), which are denoted by $\hat{a}_{\mathrm{HvK}}$ and $\hat{\sigma}^2_{\mathrm{HvK}}$ (the estimator $\hat{a}_{\mathrm{HvK}}$ is computed as described in Section 2.2 of Hall and Van Keilegom (2003) and $\hat{\sigma}^2_{\mathrm{HvK}}$ as defined at the bottom of page 447 in section 2.3 there; the estimator $\hat{a}_{\mathrm{HvK}}$ (as well as $\hat{\sigma}^2_{\mathrm{HvK}}$) depends on two tuning parameters which we denote by m_1 and m_2 as in Hall and Van Keilegom (2003));
(c) oracle estimators $\hat{a}_{\mathrm{oracle}}$ and $\hat{\sigma}^2_{\mathrm{oracle}}$ 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 $\hat{a}_{\mathrm{oracle}}$ 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 $r_t = \varepsilon_t - \hat{a}_{\mathrm{oracle}}\varepsilon_{t-1}$ and estimate the innovation variance $\nu^2 = \mathbb{E}[\eta_t^2]$ by $\hat{\nu}^2_{\mathrm{oracle}} = (T-1)^{-1}\Sigma_{t=2}^T r_t^2$. Finally, we set $\hat{\sigma}^2_{\mathrm{oracle}} = \hat{\nu}^2_{\mathrm{oracle}}/(1-\hat{a}_{\mathrm{oracle}})^2$.

Throughout this section, we set $q = 25$, $(\underline{r}, \bar{r}) = (1, 10)$ 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 $\tilde{a}_q$ and $\hat{a}_{\mathrm{HvK}}$ comparable. To assess how sensitive our estimators are to the choice of q and $(\underline{r}, \bar{r})$,

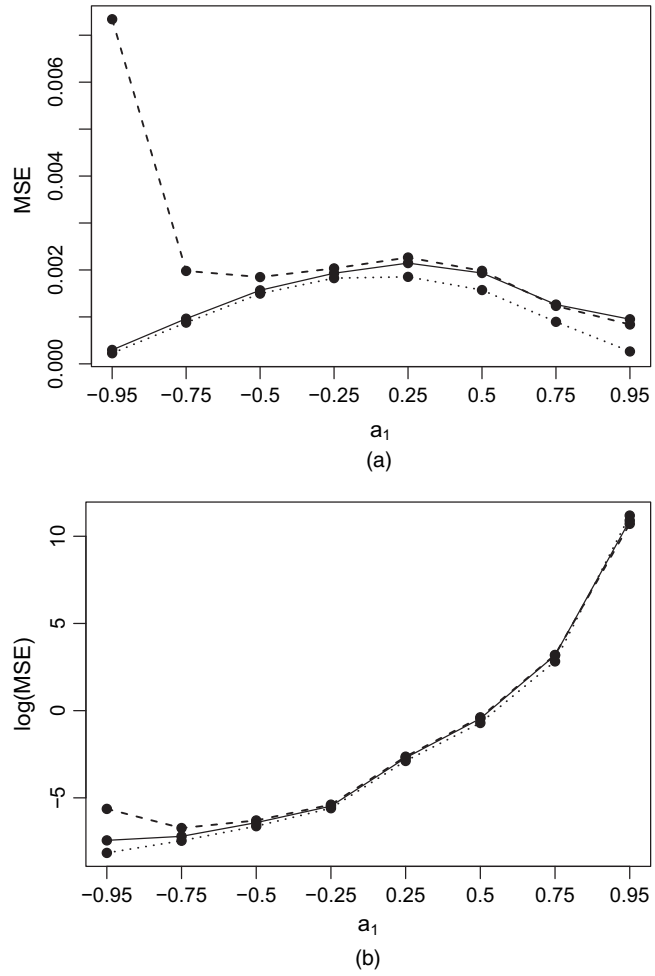

Fig. 4. MSE values for the estimators (a) $\hat{a}$ (———), $\hat{a}_{\text{HvK}}$ (- - - - -) and $\hat{a}_{\text{oracle}}$ (· · · · · ·) and (b) $\hat{\sigma}^2$ (———), $\hat{\sigma}^2_{\text{HvK}}$ (- - - - -) and $\hat{\sigma}^2_{\text{oracle}}$ (· · · · · ·) in the simulation scenarios with a moderate trend ($s_\beta = 1$)

we carry out robustness checks, considering a range of values for q and $(\underline{r}, \bar{r})$. In addition, we vary the tuning parameters m_1 and m_2 of the estimators from Hall and Van Keilegom (2003) to make sure that the results of our comparison study are not driven by the particular choice of any of the tuning parameters involved. The results of our robustness checks are reported in section S.3 of the on-line supplement. They show that the results of our comparison study are robust to different choices of the parameters q, $(\underline{r}, \bar{r})$ and (m_1, m_2).

For each estimator $\hat{a}$, $\hat{a}_{\text{HvK}}$, $\hat{a}_{\text{oracle}}$, $\hat{\sigma}^2$, $\hat{\sigma}^2_{\text{HvK}}$ and $\hat{\sigma}^2_{\text{oracle}}$, and for each model specification, the simulation output consists of a vector of length $S = 1000$ which contains the 1000 simulated values of the respective estimator. Figs 4 and 5 report the mean-squared error (MSE) of these 1000 simulated values for each estimator. On the x-axis of each plot, the various values of the AR parameter a_1 are listed which are considered. The full curve in each plot gives the MSE values of our estimators. The broken and dotted curves specify the MSE values of the Hall and Van Keilegom 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

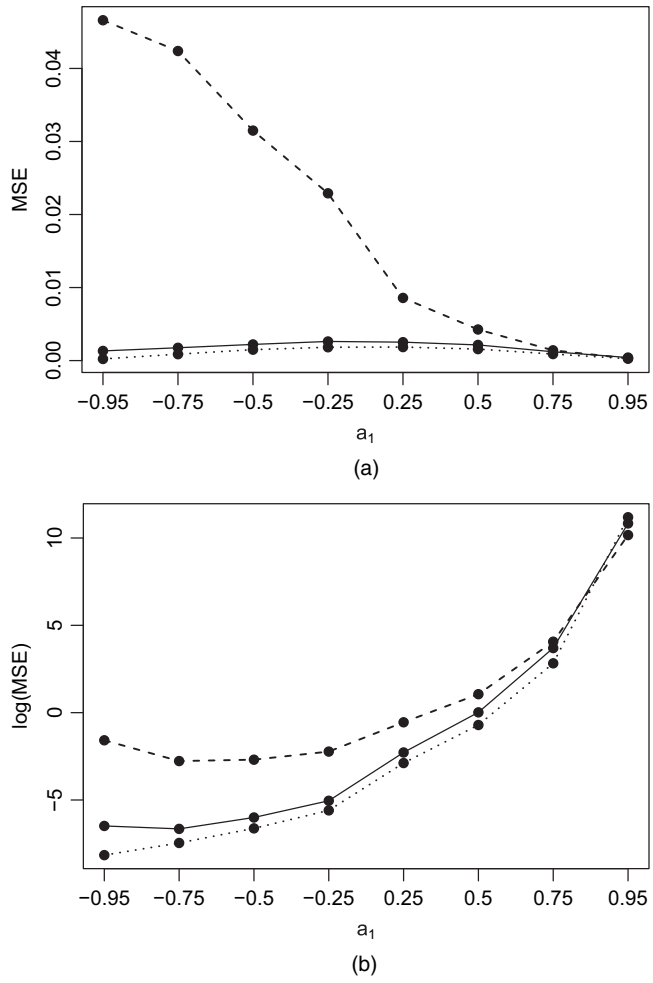

Fig. 5. MSE values for the estimators (a) $\hat{a}$ (———), $\hat{a}_{\text{HvK}}$ (- - - - -) and $\hat{a}_{\text{oracle}}$ (· · · · · ·) and (b) $\hat{\sigma}^2$ (———), $\hat{\sigma}^2_{\text{HvK}}$ (- - - - -) and $\hat{\sigma}^2_{\text{oracle}}$ (· · · · · ·) in the simulation scenarios with a pronounced trend ($s_\beta = 10$)

different across simulation scenarios to obtain a reasonable graphical presentation. In addition to the MSE values that are presented in Figs 4 and 5, we depict histograms of the 1000 simulated values that are produced by the estimators $\hat{a}$, $\hat{a}_{\text{HvK}}$, $\hat{a}_{\text{oracle}}$, $\hat{\sigma}^2$, $\hat{\sigma}^2_{\text{HvK}}$ and $\hat{\sigma}^2_{\text{oracle}}$ for two specific simulation scenarios in Figs 6 and 7. The main findings can be summarized as follows.

(a) In the simulation scenarios with a moderate trend ($s_\beta = 1$), the estimators $\hat{a}_{\text{HvK}}$ and $\hat{\sigma}^2_{\text{HvK}}$ of Hall and Van Keilegom (2003) exhibit a similar performance to that of our estimators $\hat{a}$ and $\hat{\sigma}^2$ 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 $\hat{a}_{\text{HvK}}$ and $\hat{\sigma}^2_{\text{HvK}}$ for $a_1 = -0.75$ and $a_1 = -0.95$ in Fig. 4. Fig. 6 gives some further insights into what is happening here. It shows the histograms of the simulated values that are produced by the estimators $\hat{a}$, $\hat{a}_{\text{HvK}}$ and $\hat{a}_{\text{oracle}}$ 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 $\hat{a}_{\text{HvK}}$

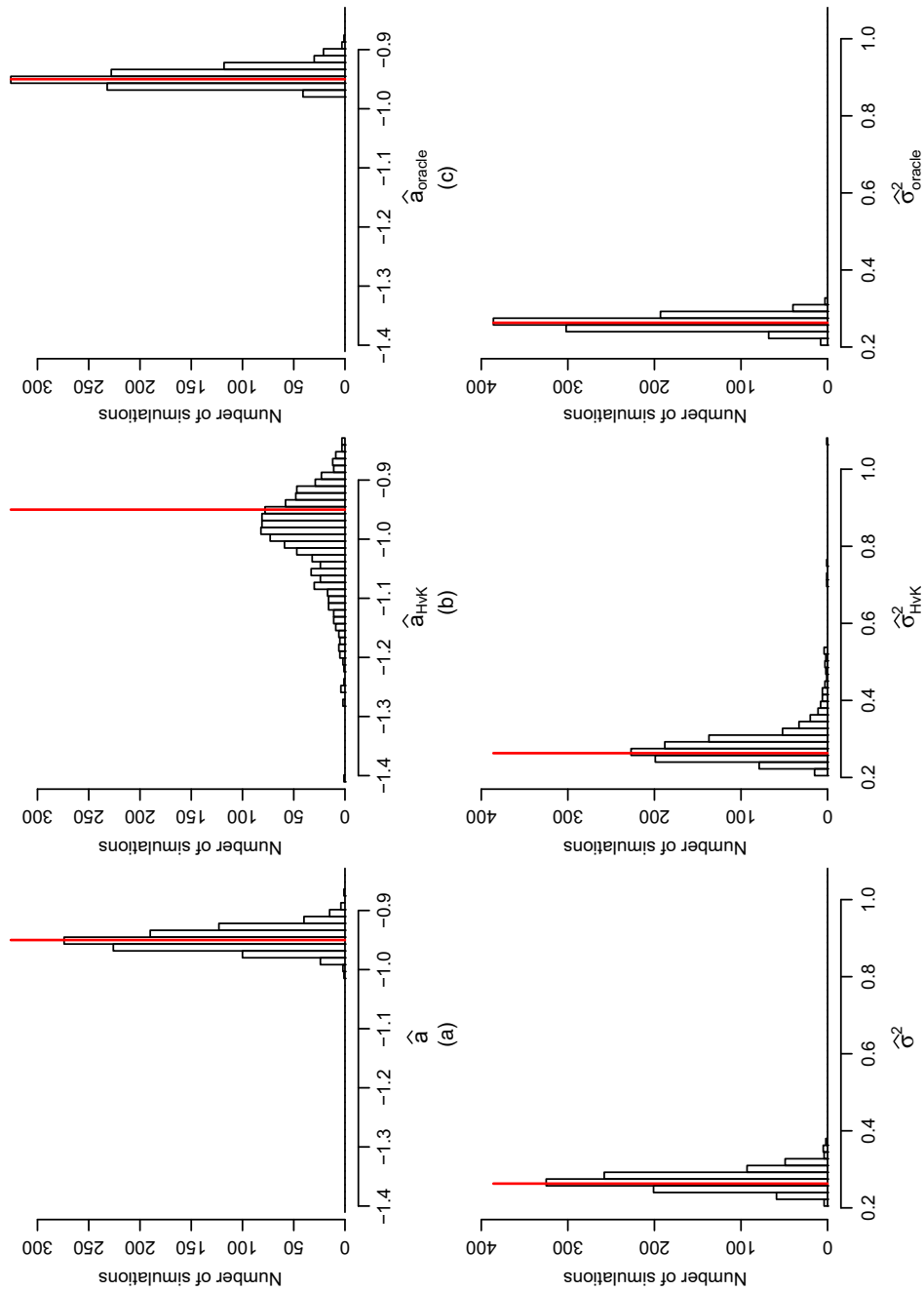

Fig. 6. Histograms of the simulated values produced by the estimators (a) $\hat{a}$, (b) $\hat{a}_{HvK}$, (c) $\hat{a}_{oracle}$, (d) $\hat{\sigma}^2$, (e) $\hat{\sigma}^2_{HvK}$ and (f) $\hat{\sigma}^2_{oracle}$ in the scenario with $a_1 = -0.95$ and $s_\beta = 1$: the vertical lines indicate the true values of a_1 and σ^2

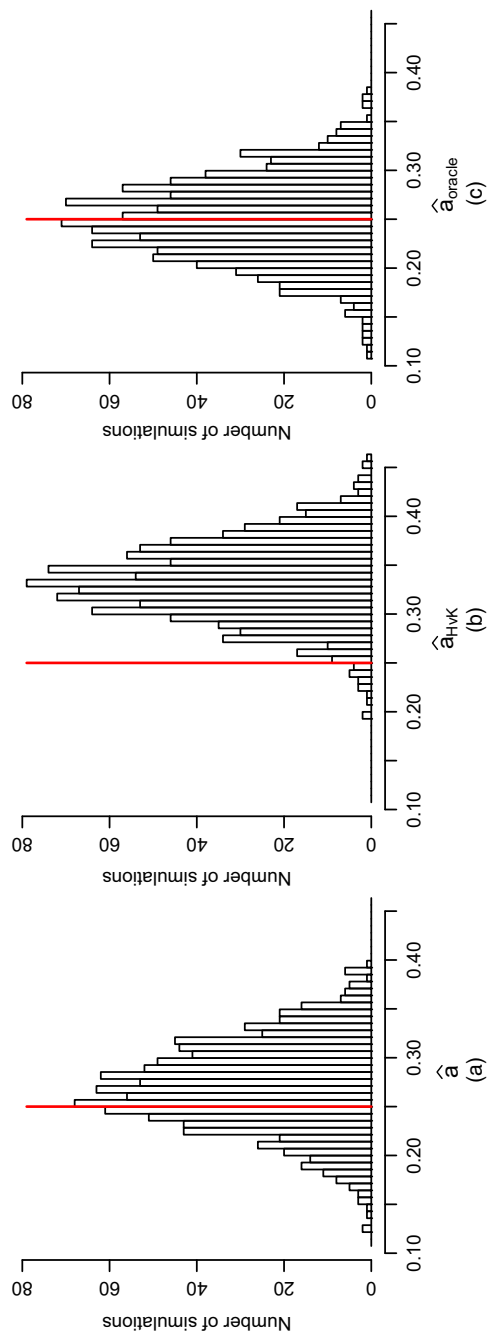

Fig. 7. Histograms of the simulated values produced by the estimators (a) $\hat{a}$, (b) $\hat{a}_{HvK}$, (c) $\hat{a}_{oracle}$, (d) $\hat{\sigma}^2$, (e) $\hat{\sigma}^2_{HvK}$ and (f) $\hat{\sigma}^2_{oracle}$ in the scenario with $a_1 = 0.25$ and $s_\beta = 10$: the vertical lines indicate the true values of a_1 and σ^2

does not obey the causality restriction $|a_1| < 1$ but frequently takes values that are substantially smaller than -1. This results in a very large spread of the histogram and thus in a disastrous performance of the estimator. A similar point applies to the histogram of the long-run variance estimator $\hat{\sigma}^2_{\mathrm{HvK}}$. Our estimators $\hat{a}$ and $\hat{\sigma}^2$, in contrast, exhibit stable behaviour in this case.

Interestingly, the estimator $\hat{a}_{\mathrm{HvK}}$ (as well as the corresponding long-run variance estimator $\hat{\sigma}^2_{\mathrm{HvK}}$) 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 $\hat{a}_{\mathrm{HvK}}$ may produce estimates that are smaller than -1 but it cannot become larger than 1. This can be easily seen on inspecting the definition of the estimator. Hence, for large positive values of a_1, the estimator $\hat{a}_{\mathrm{HvK}}$ 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 Fig. 5 are much larger than the values that are produced by our estimators for most parameter values a_1. The reason is as follows: the Hall and Van Keilegom 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 Fig. 7 which shows histograms of the simulated values for the estimators $\hat{a}$, $\hat{a}_{\mathrm{HvK}}$ and $\hat{a}_{\mathrm{oracle}}$ 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 that is produced by our estimator $\hat{a}$ is centred near the true value $a_1 = 0.25$, whereas that of $\hat{a}_{\mathrm{HvK}}$ is strongly biased upwards. A similar picture arises for the long-run variance estimators $\hat{\sigma}^2$ and $\hat{\sigma}^2_{\mathrm{HvK}}$.

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 $\hat{a}_{\mathrm{HvK}}$ tends to overestimate the AR parameter a_1, i.e. $\hat{a}_{\mathrm{HvK}}$ 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 $\hat{a}_{\mathrm{HvK}}$ cannot become larger than 1. Consequently, the effect of not eliminating the trend appropriately has a much smaller effect on $\hat{a}_{\mathrm{HvK}}$ 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 to understand long-term climate variability better. In what follows, we use our multiscale test $\mathcal{T}_{\mathrm{MS}}$ to analyse two long-term temperature records. Throughout the section, we set the level of significance to $\alpha = 0.05$ and implement the multiscale test in exactly the same way as in the simulation study of Section 5.

6.1. Analysis of the central England temperature record

The central England temperature record is the longest instrumental temperature time series in the world. The data are publicly available on the web page 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 data set of yearly mean temperatures which consists of $T = 359$ observations $Y_{t,T}$ covering the years from 1659 to 2017. A plot of the time series is given in Fig. 8(a). We assume that the temperature

data $Y_{t,T}$ follow the non-parametric 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 $\varepsilon_t = \Sigma_{j=1}^{p^*} a_j \varepsilon_{t-j} + \eta_t$, where η_t are IID 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. We select the AR order p^* by the Bayesian information criterion, which yields $p^* = 2$. More precisely, we proceed as follows: we estimate the AR parameters and the corresponding variance of the innovation terms for different AR orders by the methods from Section 4 and then choose p^* as the minimizer of the BIC. As a check of robustness, we have repeated this procedure for a wide range of the tuning parameters q and $(\underline{r}, \bar{r})$, which produces the value $p^* = 2$ throughout. Moreover, we have considered other information criteria such as the final prediction error criterion, the Akaike information criterion and a bias-corrected version thereof, which give the AR order $p^* = 2$ for almost all values of q and $(\underline{r}, \bar{r})$. Given the AR order $p^* = 2$, we estimate the AR(2) parameters $\mathbf{a} = (a_1, a_2)$ and the long-run error variance σ^2 by the procedures from

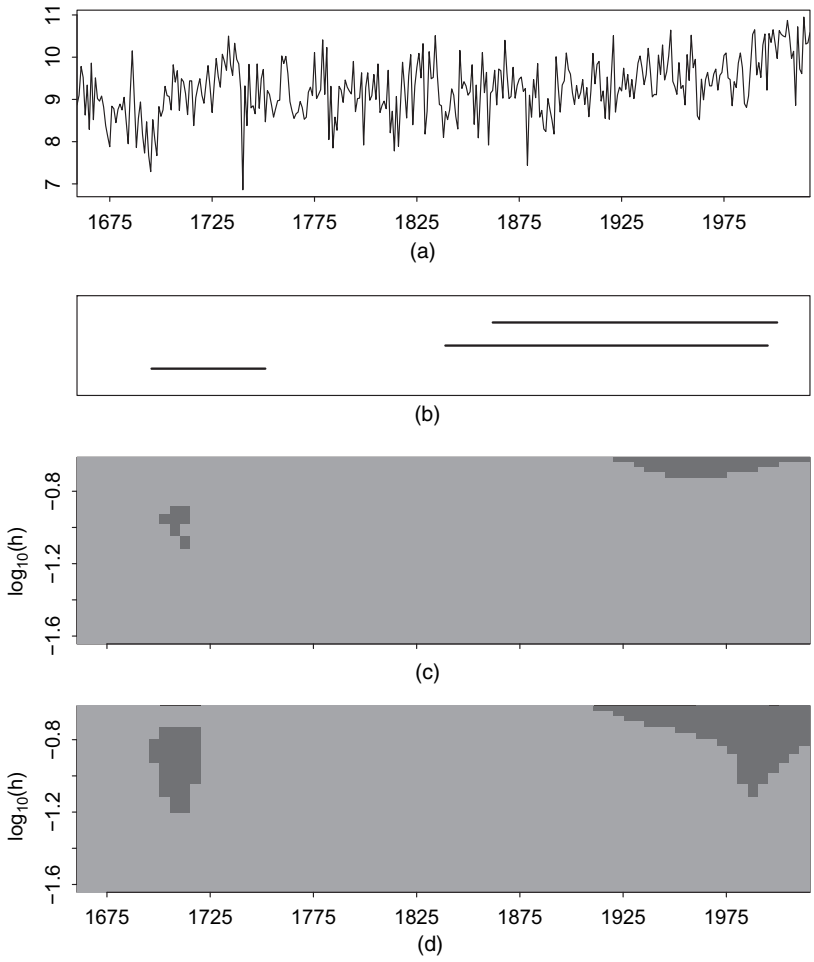

Fig. 8. Summary of the results for the central England temperature record: (a) observed temperature time series (in degrees centigrade); (b) minimal intervals in the set $\Pi_T^{\pm}$ produced by our multiscale test (these are [1684, 1744], [1839, 2009] and [1864, 2014]); (c) SiZer map produced by our multiscale test $\mathcal{T}_{\mathrm{MS}}$; (d) SiZer map produced by $\mathcal{T}_{\mathrm{SiZer}}$

Section 4 with $q = 25$ and $(\underline{r}, \bar{r}) = (1, 10)$. This gives the estimators $\hat{a}_1 = 0.164$, $\hat{a}_2 = 0.175$ and $\hat{\sigma}^2 = 0.737$.

With the help of our multiscale method, 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 the grid $\mathcal{G}_T^*$ is defined in the same way as in Section 5. The results are presented in Fig. 8. Fig. 8(b) depicts the minimal intervals in the set Π_T^+ which is produced by our multiscale test $\mathcal{T}_{\mathrm{MS}}$. The set of intervals Π_T^- is empty in the present case. According to proposition 3, we can make the following simultaneous confidence statement about the collection of minimal intervals plotted in Fig. 8(b). We can claim, with confidence of about 95%, that the trend 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. In contrast, as the set Π_T^- is empty, there is no evidence of any downward movement of the trend.

Fig. 8(c) presents the SiZer map that was produced by our multiscale test $\mathcal{T}_{\mathrm{MS}}$. For comparison, the SiZer map of the dependent SiZer test $\mathcal{T}_{\mathrm{SiZer}}$ is shown in Fig. 8(d). To produce Fig. 8(d), we have implemented SiZer as described in section S.3 of the on-line supplement, where the autocovariance function of the errors $\{\varepsilon_t\}$ is estimated with the help of our procedures from Section 4 under the assumption that $\{\varepsilon_t\}$ is an AR(2) process. The SiZer maps of Figs 8(c) and 8(d) are to be read as follows: each pixel of the map corresponds to a location–scale point (u,h) or, put differently, to a time interval $[u-h, u+h]$. The pixel (u,h) is coloured dark grey if the test indicates an increase in the trend m on the interval $[u-h, u+h]$, white if the test indicates a decrease and light grey if the test does not reject the null hypothesis that m is constant on $[u-h, u+h]$. As can be seen, the two SiZer maps in Figs 8(c) and 8(d) have a similar structure. Both our multiscale test and SiZer indicate increases in the trend m during a short time period around 1700 and towards the end of the sample. However, in contrast with SiZer, our method enables us to make formal confidence statements about the regions of dark pixels in the SiZer map. In particular, as the set of dark grey pixels in Fig. 8(c) exactly corresponds to the collection of intervals Π_T^+, we can claim, with confidence of about 95%, that the trend m has an increase on each time interval represented by a dark grey pixel in Fig. 8(c).

6.2. Analysis of global temperature data

We next analyse a data set which consists of annual global temperature anomalies from 1850 onwards. The data are plotted in Fig. 9(a) and are described in detail in Morice *et al.* (2012). They are publicly available on the web page `https://cdiac.ess-dive.lbl.gov/trends/temp/jonescru/jones.html`. As before, we assume that the data come from the model $Y_{t,T} = m(t/T) + \varepsilon_t$, where m is the trend and $\{\varepsilon_t\}$ the noise process. We apply our multiscale methods to test the null hypothesis H_0 that m is constant on all time intervals $[u-h, u+h]$ with $(u,h) \in \mathcal{G}_T$, where the grid $\mathcal{G}_T$ is defined as in Section 5. We compare our results with those obtained by Wu *et al.* (2001) who developed a method for testing the hypothesis that m is constant on $[0,1]$ against the alternative that m is an arbitrary monotonic function. For comparability, we use exactly the same data as in Wu *et al.* (2001): in particular, the yearly temperature anomalies from 1856 to 1998. Moreover, we use their estimate of the long-run error variance σ^2 which amounts to 0.01558. As we do not have an estimate available from Wu *et al.* (2001) for the autocovariance function of the error process, we do not consider dependent SiZer in the application example at hand.

The results produced by our multiscale test are reported in Fig. 9. Fig. 9(b) shows the minimal intervals in Π_T^+ and Fig. 9(c) the SiZer map of the test. As can be clearly seen from both

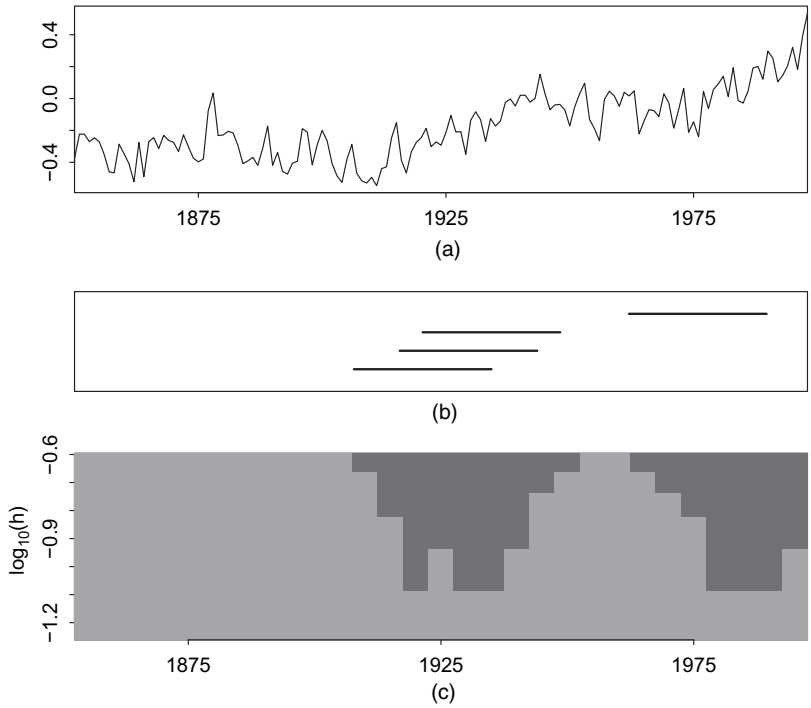

Fig. 9. Summary of the results for the global temperature anomalies: (a) observed temperature time series (in degrees centigrade); (b) minimal intervals in the set Π_T^+ produced by the multiscale test (these are [1905, 1935], [1915, 1945], [1920, 1950] and [1965, 1995]); (c) SiZer map of our test

Fig. 9(b) and Fig. 9(c), the test indicates an increase in the trend m during the first half of the 20th century followed by another increase during the second half. These findings are in line with those in Wu *et al.* (2001) who rejected the null hypothesis that m is constant. In contrast with the test of Wu *et al.* (2001), however, our multiscale method not only enables us to test whether the null is violated. It also enables us to make formal confidence statements about where violations occur, i.e. about where the trend m is increasing. In particular, we can claim, with confidence of about 95%, that the trend has an increase on each interval plotted in Fig. 9(b).

Acknowledgements

We thank the Joint Editor, the Associate Editor and two referees for their constructive and helpful comments on an earlier version of the paper.

Michael Vogt was funded by the Deutsche Forschungsgemeinschaft (German Research Foundation) under Germany's 'Excellence strategy'—GZ 2047/1, project 390685813.

References

Benner, T. C. (1999) Central England temperatures: long-term variability and teleconnections. *Int. J. Clim.*, **19**, 391–403.

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

Chaudhuri, P. and Marron, J. S. (1999) SiZer for the exploration of structures in curves. *J. Am. Statist. Ass.*, **94**, 807–823.

Chaudhuri, P. and Marron, J. S. (2000) Scale space view of curve estimation. *Ann. Statist.*, **28** 408–428.
Chernozhukov, V., Chetverikov, D. and Kato, K. (2014) Gaussian approximation of suprema of empirical processes. *Ann. Statist.*, **42**, 1564–1597.
Chernozhukov, V., Chetverikov, D. and Kato, K. (2015) Comparison and anti-concentration bounds for maxima of Gaussian random vectors. *Probab. Theory Reltd Flds*, **162**, 47–70.
Chernozhukov, V., Chetverikov, D. and Kato, K. (2017) Central limit theorems and bootstrap in high dimensions. *Ann. Probab.*, **45**, 2309–2352.
Cho, H. and Fryzlewicz, P. (2012) Multiscale and multilevel technique for consistent segmentation of nonstationary time series. *Statist. Sin.*, **22**, 207–229.
Donoho, D. L., Johnstone, I. M., Kerkyacharian, G. and Picard, D. (1995) Wavelet shrinkage: asymptopia (with discussion)? *J. R. Statist. Soc.* B, **57**, 301–369.
Dümbgen, L. (2002) Application of local rank tests to nonparametric regression. *J. Nonparam. Statist.*, **14**, 511–537.
Dümbgen, L. and Spokoiny, V. G. (2001) Multiscale testing of qualitative hypotheses. *Ann. Statist.*, **29**, 124–152.
Dümbgen, L. and Walther, G. (2008) Multiscale inference about a density. *Ann. Statist.*, **36**, 1758–1785.
Eckle, K., Bissantz, N. and Dette, H. (2017) Multiscale inference for multivariate deconvolution. *Electron. J. Statist.*, **11**, 4179–4219.
Hall, P. and Heckman, N. E. (2000) Testing for monotonicity of a regression mean by calibrating for linear functions. *Ann. Statist.*, **28**, 20–39.
Hall, P. and Van Keilegom, I. (2003) Using difference-based methods for inference in nonparametric regression with time series errors. *J. R. Statist. Soc.* B, **65**, 443–456.
Hannig, J. and Marron, J. S. (2006) Advanced distribution theory for SiZer. *J. Am. Statist. Ass.*, **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.
Inselberg, A. (1985) The plane with parallel coordinates. *Visl Comput.*, **1**, 69–91.
Morice, C. P., Kennedy, J. J., Rayner, N. A. and Jones, P. D. (2012) Quantifying uncertainties in global and regional temperature change using an ensemble of observational estimates: the HadCRUT4 data set. *J. Geophys. Res.*, **117**.
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. *Statist. Sin.*, **19**, 1511–1530.
Park, C., Marron, J. S. and Rondonotti, V. (2004) Dependent SiZer: goodness-of-fit tests for time series models. *J. Appl. Statist.*, **31**, 999–1017.
Parker, D. E., Legg, T. P. and Folland, C. K. (1992) A new daily central England temperature series, 1772-1991. *Int. J. Clim.*, **12**, 317–342.
Proksch, K., Werner, F. and Munk, A. (2018) Multiscale scanning in inverse problems. *Ann. Statist.*, **46**, 3569–3602.
Qiu, D., Shao, Q. and Yang, L. (2013) Efficient inference for autoregressive coefficients in the presence of trends. *J. Multiv. Anal.*, **114**, 40–53.
Rahmstorf, S., Foster, G. and Cahill, N. (2017) Global temperature evolution: recent trends and some pitfalls. *Environ. Res. Lett.*, **12**, article 054001.
Rohde, A. (2008) Adaptive goodness-of-fit tests based on signed ranks. *Ann. Statist.*, **36**, 1346–1374.
Rondonotti, V., Marron, J. S. and Park, C. (2007) SiZer for time series: a new approach to the analysis of trends. *Electron. J. Statist.*, **1**, 268–289.
Rufibach, K. and Walther, G. (2010) The block criterion for multiscale inference about a density, with applications to other multiscale problems. *J. Computnl Graph. Statist.*, **19**, 175–190.
Schmidt-Hieber, J., Munk, A. and Dümbgen, L. (2013) Multiscale methods for shape constraints in deconvolution: confidence statements for qualitative features. *Ann. Statist.*, **41**, 1299–1328.
Shao, Q. and Yang, L. J. (2011) Autoregressive coefficient estimation in nonparametric analysis. *J. Time Ser. Anal.*, **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. *Scand. J. Statist.*, **44**, 346–368.
Truong, Y. K. (1991) Nonparametric curve estimation with time series errors. *J. Statist. Planng Inf.*, **28**, 167–183.
Von Sachs, R. and MacGibbon, B. (2000) Non-parametric curve estimation by Wavelet thresholding with locally stationary errors. *Scand. J. Statist.*, **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. *J. Appl. Probab.*, **41**, 425–436.
Wu, W. B., Woodroofe, M. and Mentz, G. (2001) Isotonic regression: another look at the changepoint problem. *Biometrika*, **88**, 793–804.

Supporting information
Additional 'supporting information' may be found in the on-line version of this article:

'Supplement to "Multiscale inference and long-run variance estimation in nonparametric regression with time series errors"'.