The validity of bootstrap testing for threshold autoregression

The validity of bootstrap testing in the threshold framework

Simone Giannerini[1], Greta Goracci[1,2], and Anders Rahbek[3]

[1]Department of Statistical Sciences, University of Bologna, Italy
[2]Faculty of Economics and Management, Free University of Bozen-Bolzano, Italy
[3]Department of Economics, University of Copenhagen, Denmark

January 10, 2023

Abstract

We consider bootstrap-based testing for threshold effects in non-linear threshold autoregressive (TAR) models. It is well-known that classic tests based on asymptotic theory tend to be biased in case of small, or even moderate sample sizes, especially when the estimated parameters indicate non-stationarity, or in presence of heteroskedasticity, as often witnessed in the analysis of financial or climate data. To address the issue we propose a supremum Lagrange Multiplier test statistic (sLM), where the null hypothesis specifies a linear autoregressive (AR) model against the alternative of a TAR model. We consider both the classical recursive residual i.i.d. bootstrap (sLMi) and a wild bootstrap (sLMw), applied to the sLM statistic, and establish their validity under the null hypothesis. The framework is new, and requires the proof of non-standard results for bootstrap analysis in time series models; this includes a uniform bootstrap law of large numbers and a bootstrap functional central limit theorem. The Monte Carlo evidence shows that the bootstrap tests have correct empirical size even for small samples; the wild bootstrap version (sLMw) is also robust against the presence of heteroskedasticity. Moreover, there is no loss of empirical power when compared to the asymptotic test and the size of the tests is not affected if the order of the tested model is selected through AIC. Finally, we use our results to analyse the time series of the Greenland ice sheet mass balance. We find a significant threshold effect and an appropriate specification that manages to reproduce the main nonlinear features of the series, such as the asymmetric seasonal cycle, the main periodicities and the multimodality of the probability density function.

1 Introduction

The problem of testing for a linear time series model versus its threshold extension has attracted considerable attention for a number of reasons. First and foremost, threshold autoregressive models (TAR) are among the simplest nonlinear specifications and retain a good interpretability. Second, they can encompass many complex features such as jumps, limit-cycles, time irreversibility and chaos, see e.g. Tong [1990, 2011]. Petruccelli [1992] proved that TAR models approximate a wide range of nonlinear autoregressive processes. Moreover, they have been proven to describe successfully many real-world phenomena in economics and finance, see e.g. Chan et al. [2017], Hansen [2011], Tong [2017].

Seminal works on asymptotic quasi-likelihood ratio tests for threshold autoregression include Chan [1990], Chan and Tong [1990], Chan [1991]. Other contributions include those of Petruccelli and Davies [1986], Su and Chan [2017] and that of Tsay [1998] for the multivariate case. Tests based upon Lagrange Multipliers were proposed in Luukkonen et al. [1988] for the smooth transition case and in Wong and Li [1997, 2000] for TAR models with conditional heteroscedasticity, see also Tong [2011] for a review.

The main theoretical problem associated with testing for threshold autoregression is the nuisance parameter (the threshold) being present only under the alternative hypothesis, as adduced in Davies [1977, 1987] and Andrews [1993]. In the present context, one solution is to derive the test statistic as a random function of the nuisance parameter. Then, the overall test statistic is the supremum (or some other convenient function) of the statistic over the grid of values of the nuisance parameter. The derivation of the null distribution of the overall test statistic requires proving the stochastic equicontinuity (tightness) of the sequence of random functions, see e.g. van der Vaart [1998], and this is often the most challenging task.

One key issue with asymptotic tests is the sample size requirement to deliver a good performance. Typically, the rate of convergence towards the asymptotic null distribution depends upon the true parameters values of the data generating process and might produce a size bias that can be severe, for instance when these are close to non-stationarity and/or non-invertibility, see e.g. Goracci et al. [2023]. Furthermore, the null distribution, which has no closed analytical form, depends both upon the threshold range and the number of tested parameters, so that one has to make use of simulated critical values for each combination of the threshold grid and number of parameters, see e.g. Andrews [2003]. Also, the presence of unaccounted heteroskedasticity can produce a substantial bias that increases with sample size. One way to overcome the aforementioned problems is to resort to resampling methods. Hansen [1996] proposes tests based on a randomization device where the score function is perturbed through an auxiliary random variable. The same approach has been deployed in Li and Li [2011] to test a linear model against its threshold ARMA extension by means of a quasi likelihood ratio statistic. More recently, Hill [2021] adopts a similar approach to introduce robust conditional moment tests of omitted nonlinearity. To the best of our knowledge, to date, there are no available results on the validity of

the classical bootstrap (both parametric and nonparametric) for testing a linear AR model against a TAR model.

In this paper we fill this gap and provide a proof of the validity of the tests (under H_0) based upon two residual bootstraps combined with a supremum Lagrange Multiplier test statistic where the null hypothesis specifies a linear AR(p) model against the alternative of a TAR(p) model. The first bootstrap test, denoted with sLMi, is valid under the hypothesis of i.i.d. errors, whereas the wild bootstrap test (sLMw) is valid under the error process having a martingale difference structure. One of the main advantages of Lagrange multiplier tests over likelihood ratio tests is that the former only need estimating the model under the null hypothesis and avoid direct estimation of the TAR model.

We prove that, under the null hypothesis, the bootstrap distributions of the test statistics sLMi and sLMw converge to the corresponding asymptotic distributions, namely, a functional of a centered Gaussian process. Note that, as also shown in Hansen [1996], the Wald, the supLM and the likelihood-ratio test statistics share the same asymptotic distribution. The inherent difficulties associated with working in the bootstrap framework, i.e. simultaneously coping with the two kinds of randomness (the first one is the sampling variability and the second one is the bootstrap variability) are amplified by the discontinuity of the threshold function and the absence of the nuisance parameter under the null hypothesis. To this end, we provide a *uniform* bootstrap law of large numbers and a functional bootstrap central limit theorem that represent a general theoretical framework that can be adapted to other situations.

We carry out a simulation study where we compare the finite sample performance of our bootstrap tests, the asymptotic test and Hansen's test Hansen [1996], both under i.i.d. errors and under GARCH errors. The results show that the bootstrap tests have a correct empirical size for a series' length as small as 50, whereas both the asymptotic test and Hansen's test tend to be undersized and this impinges negatively upon their power. Additional simulations also show that the tests are well behaved in case of near-integrated processes or local unit-roots, see Section B of the Supplementary Material. The good performance of the bootstrap tests in small samples makes them suitable to many situations where either data collection/production is expensive, as in the experimental context, or longer series are simply not available. Moreover, the wild bootstrap test is robust against the presence of heteroskedasticity, which induces a non-negligible size bias both on the asymptotic test and on the i.i.d. bootstrap test. The simulations also show that the size of the tests is not affected if the order of the autoregression is treated as unknown and is selected by means of the AIC.

In the last section we apply our methodology to the monthly time series of the Greenland ice sheet mass balance [Sasgen et al., 2020a]. The sheet covers around the 85% of the island and the ice loss in the last decades is one of the main contributors to the rise of sea level at a global scale. For this reason, its study is of key importance for monitoring climate change. The series is characterised by a declining trend, an asymmetric seasonal cycle, a multimodal density, and complex patterns of spectral modes and resonances. We find a

significant threshold effect and an appropriate two-regime specification that manages to reproduce the main nonlinear features of the series. The lower regime corresponds to the slow, rising phase of the cycle, whereas the upper regime is associated to the fast declining phase occurring in the summer season.

The rest of the paper is organised as follows. Section 2 introduces the problem and describes the theory behind the standard asymptotic sLM test. In Section 3 we present the two bootstrap versions of the test, together with the main results on their validity. Section 4 shows the finite sample behaviour of the tests where our bootstrap tests (sLMi and sLMw) are compared to the asymptotic test (sLMa) and Hansen's test (sLMh). Section 5 is devoted to the application to the Greenland ice sheet mass balance. All the proofs are detailed in A. The Supplementary Material contains technical lemmas, further results from the simulation study and from the analysis of the Greenland ice sheet mass balance. The tests introduced in this work are implemented in routines included in the forthcoming R package `tseriesTARMA`.

1.1 Notation

We write $P^*(\cdot)$, $E^*[\cdot]$ to indicate probability and expectation conditional on the data, respectively; $\xrightarrow[n\to\infty]{w^*}{}_p$ denotes the weak convergence in probability and $Y_n^* \xrightarrow[n\to\infty]{p^*}{}_p Y$ or, equivalently, $Y_n^* - Y = o_{p^*}(1)$, means that, for any $\delta > 0$, $P^*(\|Y_n^* - Y\| > \delta) \xrightarrow[n\to\infty]{p} 0$; lastly, $Y_n^* = O_{p^*}(1)$ means that, for any $\delta > 0$, there exists $M > 0$ such that $P(P^*(\|Y_n^*\| > M) < \delta)$ is arbitrarily close to one for sufficiently large n. Here, $\|\cdot\|$ is the $\mathcal{L}^2$ matrix norm (the Frobenius' norm, i.e. $\|A\| = \sqrt{\sum_{i=1}^n \sum_{j=1}^m |a_{ij}|^2}$, where A is a $n \times m$ matrix); $\|A\| = (E[A]^r)^{1/r}$ is the $\mathcal{L}^r$ norm of a random matrix. Moreover, let $\mathcal{D}_{\mathbb{R}}(a, b)$, $a < b$ be the space of functions from (a, b) to $\mathbb{R}$ that are right continuous with left-hand limits.

2 Preliminaries

Let the time series $\{X_t\}$ follow the threshold autoregressive TAR(p) model defined by the difference equation:

$$X_t = \phi_0 + \sum_{i=1}^{p} \phi_i X_{t-i} + \left(\Psi_0 + \sum_{i=1}^{p} \Psi_i X_{t-i} \right) I(X_{t-d} \leq r) + \varepsilon_t. \qquad (1)$$

The positive integers p and d are the autoregressive order and the delay parameter, respectively; we assume p and d to be known. Moreover $I(\cdot)$ indicates the indicator function and $r \in \mathbb{R}$ is the threshold parameter. $\{\varepsilon_t\}$ is the innovation process, we assume $E[\varepsilon_t] = 0$ and, for each t, ε_t is independent of X_{t-1}, X_{t-2}, ... Further assumptions on $\{\varepsilon_t\}$ will be detailed in Section 3.1.

Clearly, Eq. (1) specifies a regime-switching process where each regime follows a linear autoregressive process. The parameters are given by

$$\begin{aligned}
&\boldsymbol{\phi} = (\phi_0, \phi_1, \ldots, \phi_p)^\intercal \in \Theta_\phi;\\
&\boldsymbol{\Psi} = (\Psi_0, \Psi_1, \ldots, \Psi_p)^\intercal \in \Theta_\Psi;\\
&\boldsymbol{\eta} = (\boldsymbol{\phi}^\intercal, \boldsymbol{\Psi}^\intercal, \sigma^2)^\intercal \in \Theta = \Theta_\phi \times \Theta_\Psi \times (0, +\infty),
\end{aligned}$$

with Θ_ϕ and Θ_Ψ being subsets of $\mathbb{R}^{p+1}$. We use $\boldsymbol{\eta} = (\boldsymbol{\phi}^\intercal, \boldsymbol{\Psi}^\intercal, \sigma^2)^\intercal$ to refer to unknown parameters, whereas the true parameters are indicated by

$$\boldsymbol{\eta}_0 = (\boldsymbol{\phi}_0^\intercal, \boldsymbol{\Psi}_0^\intercal, \sigma_0^2)^\intercal = (\phi_{0,0}, \phi_{0,1}, \ldots, \phi_{0,p}, \Psi_{0,0}, \Psi_{0,1}, \ldots, \Psi_{0,p}, \sigma_0^2)^\intercal.$$

We test whether the TAR model fits the data significantly better than its linear counterpart. As $\boldsymbol{\Psi}$ contains the differences of the autoregressive parameters in the two regimes, the system of hypotheses reduces to

$$\begin{cases} H_0 & : \boldsymbol{\Psi} = \mathbf{0} \\ H_1 & : \boldsymbol{\Psi} \neq \mathbf{0}, \end{cases}$$

where $\mathbf{0}$ is the vector of zeros. Suppose we observe $\{X_t, t = 1, \ldots, n\}$. We develop the Lagrange Multiplier (hereafter LM) test based on the quasi Gaussian log-likelihood conditional on the initial values $X_0, X_{-1}, \ldots, X_{-p+1}$:

$$\ell_n(\boldsymbol{\eta}, r) = -\frac{1}{2\sigma^2} \sum_{t=1}^{n} \varepsilon_t^2(\boldsymbol{\eta}, r), \tag{2}$$

where

$$\varepsilon_t(\boldsymbol{\eta}, r) = X_t - \left\{ \phi_0 + \sum_{i=1}^{p} \phi_i X_{t-i} \right\} - \left\{ \Psi_0 + \sum_{i=1}^{p} \Psi_i X_{t-i} \right\} I\left(X_{t-d} \leq r\right). \tag{3}$$

Under the null hypothesis, model (1) reduces to an AR(p) model:

$$X_t = \phi_0 + \sum_{i=1}^{p} \phi_i X_{t-i} + \varepsilon_t, \tag{4}$$

and let $\tilde{\boldsymbol{\phi}} = (\tilde{\phi}_0, \tilde{\phi}_1, \ldots, \tilde{\phi}_p)^\intercal$ be the Maximum Likelihood Estimator (hereafter MLE) of the autoregressive parameters in Eq. (4) based upon the Gaussian likelihood, i.e.:

$$\tilde{\boldsymbol{\phi}} = \left(\tilde{\phi}_0, \tilde{\phi}_1, \ldots, \tilde{\phi}_p\right)^\intercal = \underset{\boldsymbol{\phi} \in \Theta_\phi}{\operatorname{argmin}} \left[-\frac{1}{2\sigma^2} \sum_{t=1}^{n} \varepsilon_t^2((\boldsymbol{\phi}, \mathbf{0}, \sigma^2), r) \right].$$

Note that $\tilde{\boldsymbol{\phi}}$ coincides with the least squares estimator. The associated residuals (*restricted* residuals) are

$$\tilde{\varepsilon}_t = X_t - \tilde{\phi}_0 - \sum_{i=1}^{p} \tilde{\phi}_i X_{t-i} = (\phi_{0,0} - \tilde{\phi}_0) + \sum_{i=1}^{p} (\phi_{0,i} - \tilde{\phi}_i) X_{t-i} + \varepsilon_t. \tag{5}$$

Moreover, σ^2 is estimated by

$$\tilde{\sigma}^2 = \frac{1}{n-p-1} \sum_{t=1}^{n-p-1} \tilde{\varepsilon}_t^2. \tag{6}$$

Lastly, define $\tilde{\boldsymbol{\eta}} = (\tilde{\boldsymbol{\phi}}^\intercal, \mathbf{0}^\intercal, \tilde{\sigma}^2)$, i.e. $\tilde{\boldsymbol{\eta}}$ is the restricted MLE under the null hypothesis.

In order to test the null hypothesis define:

$$\frac{\partial \tilde{\ell}_n}{\partial \boldsymbol{\Psi}}(r) = \left.\frac{\partial \ell_n(\boldsymbol{\eta}, r)}{\partial \boldsymbol{\Psi}}\right|_{\boldsymbol{\eta}=\tilde{\boldsymbol{\eta}}}$$

i.e. the first derivative with respect to the tested parameters evaluated at the restricted estimators. Moreover, let

$$\begin{aligned} D_t(r) = (&-1, -X_t, \ldots, -X_{t-p+1}, -I(X_{t-d+1} \leq r), \\ &-X_t I(X_{t-d+1} \leq r), \ldots, -X_{t-p+1} I(X_{t-d+1} \leq r))^\intercal . \end{aligned} \tag{7}$$

Hence, the Fisher information matrix is as follows:

$$\begin{aligned} I_n(r) &= \begin{pmatrix} I_{n,11} & I_{n,12}(r) \\ I_{n,21}(r) & I_{n,22}(r) \end{pmatrix} := \begin{pmatrix} -\frac{\partial^2 \ell_n(\boldsymbol{\eta}, r)}{\partial \boldsymbol{\phi} \partial \boldsymbol{\phi}^\intercal} & -\frac{\partial^2 \ell_n(\boldsymbol{\eta}, r)}{\partial \boldsymbol{\phi} \partial \boldsymbol{\Psi}^\intercal} \\ -\frac{\partial^2 \ell_n(\boldsymbol{\eta}, r)}{\partial \boldsymbol{\Psi} \partial \boldsymbol{\phi}^\intercal} & -\frac{\partial^2 \ell_n(\boldsymbol{\eta}, r)}{\partial \boldsymbol{\Psi} \partial \boldsymbol{\Psi}^\intercal} \end{pmatrix} \\ &= \frac{1}{\sigma^2} \sum_{t=1}^{n} D_{t-1}(r) D_{t-1}^\intercal(r). \end{aligned} \tag{8}$$

The supremum Lagrange Multipliers test statistic (hereafter supLM) is

$$T_n = \sup_{r \in [r_L, r_U]} T_n(r), \tag{9}$$

where

$$T_n(r) = \left(\frac{\partial \tilde{\ell}_n}{\partial \boldsymbol{\Psi}}(r)\right)^\intercal \left(I_{n,22}(r) - I_{n,21}(r) I_{n,11}^{-1} I_{n,12}(r)\right)^{-1} \frac{\partial \tilde{\ell}_n}{\partial \boldsymbol{\Psi}}(r) \tag{10}$$

with $[r_L, r_U]$ being a data driven interval, e.g. r_L and r_U can be some percentiles of the observed data.

In order to define the asymptotic null distribution of T_n, define

$$I_\infty(r) = \begin{pmatrix} I_{\infty,11} & I_{\infty,12}(r) \\ I_{\infty,21}(r) & I_{\infty,22}(r) \end{pmatrix} \tag{11}$$

where $I_{\infty,22}(r) = I_{\infty,12}(r) = I_{\infty,21}^\intercal(r)$ are $(p+1) \times (p+1)$ symmetric matrices whose $(i+1, j+1)$th element is

$$\begin{aligned} &E[I(X_{t-d} \leq r)], && \text{if } i = 0, j = 0 \\ &E[X_{t-j} I(X_{t-d} \leq r)], && \text{if } i = 0, j \neq 0 \\ &E[X_{t-i} X_{t-j} I(X_{t-d} \leq r)], && \text{if } i \neq 0, j \neq 0 \end{aligned}$$

and $I_{\infty,11} = I_{\infty,22}(\infty)$. Here and in the following, $P(\cdot)$ and $E[\cdot]$ are, respectively, the probability and expectation taken under the true probability distribution for which the null hypothesis holds. Note that

$$I_\infty(r) = E\left[D_{t-1}(r)D_{t-1}(r)^\intercal\right]. \tag{12}$$

The null distribution of the supLM test statistic is a functional of the centered Gaussian process $\{\xi(r),\ r \in \mathbb{R}\}$ with covariance kernel

$$\Xi(r_1, r_2) = E\left[\varepsilon_t^2 D_{t-1}(r_1)D_{t-1}^\intercal(r_2)\right], \tag{13}$$

where $a_1 \wedge a_2 = \min\{a_1, a_2\}$, for any $a_1, a_2 \in \mathbb{R}$. Under standard regularity conditions as in Li and Li [2011], it holds that

$$T_n \xrightarrow[n\to\infty]{w} \sup_{r\in[r_L,r_U]} \xi(r)^\intercal \Sigma(r)\xi(r) := T_\infty \tag{14}$$

where $\xrightarrow[n\to\infty]{w}$ means the convergence in distribution as the sample size n increases, and

$$\Sigma(r) = \sigma^{-2}\left(-I_{\infty,21}(r)I_{\infty,11}^{-1}, \mathrm{I}_{p+1}\right)^\intercal \\ \left(I_{\infty,22}(r) - I_{\infty,21}(r)I_{\infty,11}^{-1}I_{\infty,12}(r)\right)^{-1}\left(-I_{\infty,21}(r)I_{\infty,11}^{-1}, \mathrm{I}_{p+1}\right)$$

with I_{p+1} being the $p+1$ identity matrix.

Remark 1. For any fixed r, if $\{\varepsilon_t\}$ is an i.i.d. process then T_∞ reduces to a χ^2 random variable with $p+1$ degrees of freedom. This is not the case if the innovation process presents heteroskedasticity.

Remark 2. Note that if H_1 is true and X_t follows a stationary TAR(p) DGP, then there can be a local unit-root regime but X_t is still globally stationary. This is does not entail additional complications and is encompassed within our testing framework. On the other hand, testing for X_t being a $I(1)$ process under H_0 and a stationary TAR(p), possibly with local unit-root regimes, under H_1 requires a different asymptotic theory based upon the limit theory of stochastic integrals, see e.g. Chan et al. [2024] and references therein.

3 The bootstrap

We focus on the following residual-based bootstrap approaches. Consider the recursively defined bootstrap process generated by the bootstrap parameters $\boldsymbol{\phi}^* = (\phi_0^*, \phi_1^*, \ldots, \phi_p^*)^\intercal$:

$$X_t^* = \phi_0^* + \sum_{i=1}^{p} \phi_i^* X_{t-i}^* + \varepsilon_t^*, \tag{15}$$

where the initial values $X_0^*, X_1^*, \ldots, X_{-p+1}^*$, are equal to their sample counterpart. We consider the case where the bootstrap parameters are the restricted

MLE, i.e. $\boldsymbol{\phi}^* = \tilde{\boldsymbol{\phi}}$; therefore the process defined in Eq. (15) equals:

$$X_t^* = \tilde{\phi}_0 + \sum_{i=1}^{p} \tilde{\phi}_i X_{t-i}^* + \varepsilon_t^*, \tag{16}$$

which is an example of the so-called restricted bootstrap, see Cavaliere and Rahbek [2021]. Different types of bootstrap can be introduced according to the definition of the bootstrap innovation process $\{\varepsilon_t^*\}$. In this work we focus upon the following two cases:

(B.iid) i.i.d. bootstrap: for each t, ε_t^* is sampled with replacement from the re-centred residuals $\tilde{\varepsilon}_t^c := \tilde{\varepsilon}_t - n^{-1}\sum_{t=1}^{n}\tilde{\varepsilon}_t$, where $\tilde{\varepsilon}_t$ are defined in Eq. (5).

(B.wild) wild bootstrap: for each t, $\varepsilon_t^* = \tilde{\varepsilon}_t \cdot \upsilon_t$, $\tilde{\varepsilon}_t$ are defined in Eq. (5) and $\{\upsilon_t\}$ is a sequence of i.i.d. random variables with zero mean and unit variance. Typically, $\{\upsilon_t\}$ follows either the standard normal or the Rademacher random variable.

Remark 3. Conditionally upon the data, under the i.i.d. bootstrap scheme $\{\varepsilon_t^*\}$ is a sequence of i.i.d. random variables, whereas in the wild bootstrap scheme it is a sequence of independent random variables with $E[\varepsilon_t^{*2}] = \tilde{\varepsilon}_t^2$.

Given $\{X_t^*, t = 1, \ldots, n\}$, the bootstrap sample from Eq. (16), the bootstrap log-likelihood function results:

$$\ell_n^*(\boldsymbol{\eta}, r) = -\frac{1}{2\sigma^2}\sum_{t=1}^{n}\varepsilon_t^{*2}(\boldsymbol{\eta}, r), \tag{17}$$

where $\varepsilon_t^*(\boldsymbol{\eta}, r)$ is defined as in Eq. (3) with X being replaced by X^*:

$$\varepsilon_t^*(\boldsymbol{\eta}, r) = X_t^* - \left\{\phi_0 + \sum_{i=1}^{p}\phi_i X_{t-i}^*\right\} - \left\{\Psi_0 + \sum_{i=1}^{p}\Psi_i X_{t-i}^*\right\} I\left(X_{t-d}^* \leq r\right). \tag{18}$$

In analogy with the standard asymptotic case, we present the methodology and the related theory for the two bootstrap schemes. Let

$$\tilde{\boldsymbol{\phi}}^* = (\tilde{\phi}_0^*, \tilde{\phi}_1^*, \ldots, \tilde{\phi}_p^*)$$

be the MLE computed upon $\{X_t^*, t = 1, \ldots, n\}$ defined in Eq. (15) and $\tilde{\sigma}^{*2} = (n-p-1)^{-1}\sum_{t=1}^{n-p-1}\tilde{\varepsilon}_t^{*2}$, with $\tilde{\varepsilon}_t^{*2}$ being the corresponding bootstrap restricted residuals. Hence, we define $\tilde{\boldsymbol{\eta}}^* = (\tilde{\boldsymbol{\phi}}^{*\intercal}, \mathbf{0}^\intercal, \tilde{\sigma}^{*2})$ to be the bootstrap estimator maximising the bootstrap log-likelihood function in Eq. (17). Let

$$\frac{\partial \ell_n^*}{\partial \boldsymbol{\phi}} = \frac{\partial \ell_n^*(\boldsymbol{\eta}, r)}{\partial \boldsymbol{\phi}}\bigg|_{\boldsymbol{\eta}=\tilde{\boldsymbol{\eta}}}, \quad \frac{\partial \ell_n^*}{\partial \boldsymbol{\Psi}}(r) = \frac{\partial \ell_n^*(\boldsymbol{\eta}, r)}{\partial \boldsymbol{\Psi}}\bigg|_{\boldsymbol{\eta}=\tilde{\boldsymbol{\eta}}}$$

and

$$\nabla_n^*(r) = \left(\left(\frac{\partial \ell_n^*}{\partial \boldsymbol{\phi}}\right)^\intercal, \left(\frac{\partial \ell_n^*}{\partial \boldsymbol{\Psi}}(r)\right)^\intercal\right)^\intercal, \tag{19}$$

thereby

$$\{\nabla_n^*(r)\} = \{\nabla_n^*(r), r_L \leq r \leq r_U\} \tag{20}$$

is the bootstrap score process as a function of r, evaluated in $\boldsymbol{\eta} = \tilde{\boldsymbol{\eta}}$. In order to build the bootstrap LM test statistic, we also need the first derivative of the bootstrap log-likelihood function with respect to $\boldsymbol{\Psi}$ evaluated in $\tilde{\boldsymbol{\eta}}^*$, i.e:

$$\frac{\partial \tilde{\ell}_n^*}{\partial \boldsymbol{\Psi}}(r) = \left.\frac{\partial \ell_n^*(\boldsymbol{\eta}, r)}{\partial \boldsymbol{\Psi}}\right|_{\boldsymbol{\eta}=\tilde{\boldsymbol{\eta}}^*}. \tag{21}$$

Moreover, let $D_t^*(r)$ denote the first-order derivative of $\varepsilon_{t+1}^*(\boldsymbol{\eta}, r)$ with respect to $\boldsymbol{\eta}$. It follows that:

$$\begin{aligned} D_t^*(r) = \big(&-1, -X_t^*, \ldots, -X_{t-p+1}^*, -I(X_{t-d+1}^* \leq r), \\ &-X_t^* I(X_{t-d+1}^* \leq r), \ldots, -X_{t-p+1}^* I(X_{t-d+1}^* \leq r)\big)^\intercal. \end{aligned} \tag{22}$$

Similar to Eq. (8), the bootstrap observed information matrix is defined by:

$$\begin{aligned} I_n^*(r) &= \begin{pmatrix} I_{n,11}^* & I_{n,12}^*(r) \\ I_{n,21}^*(r) & I_{n,22}^*(r) \end{pmatrix} \\ &= \begin{pmatrix} -\frac{\partial^2 \ell_n^*(\boldsymbol{\eta},r)}{\partial \boldsymbol{\phi} \partial \boldsymbol{\phi}^\intercal} & -\frac{\partial^2 \ell_n^*(\boldsymbol{\eta},r)}{\partial \boldsymbol{\phi} \partial \boldsymbol{\Psi}^\intercal} \\ -\frac{\partial^2 \ell_n^*(\boldsymbol{\eta},r)}{\partial \boldsymbol{\Psi} \partial \boldsymbol{\phi}^\intercal} & -\frac{\partial^2 \ell_n^*(\boldsymbol{\eta},r)}{\partial \boldsymbol{\Psi} \partial \boldsymbol{\Psi}^\intercal} \end{pmatrix} = \frac{1}{\sigma^{*2}} \sum_{t=1}^{n} D_{t-1}^*(r) D_{t-1}^{*\intercal}(r). \end{aligned} \tag{23}$$

We compute the bootstrap supLM statistic T_n^* as:

$$T_n^* = \sup_{r \in [r_L, r_U]} T_n^*(r); \tag{24}$$

$$T_n^*(r) = \left(\frac{\partial \tilde{\ell}_n^*}{\partial \boldsymbol{\Psi}}(r)\right)^\intercal \left(I_{n,22}^*(r) - I_{n,21}^*(r)(I_{n,11}^*)^{-1} I_{n,12}^*(r)\right)^{-1} \frac{\partial \tilde{\ell}_n^*}{\partial \boldsymbol{\Psi}}(r). \tag{25}$$

Finally, the bootstrap p-value is given by

$$B^{-1} \sum_{b=1}^{B} I(T_n^{*b} \geq T_n),$$

where T_n^{*b}, $b = 1, \ldots B$ is the bootstrap test statistics and T_n is the value of the supLM statistic computed on the original sample, defined in Eq. (9).

3.1 Bootstrap asymptotic theory

In order to derive the bootstrap asymptotic theory we rely on some assumptions involving both the process $\{X_t\}$ and the innovation process $\{\varepsilon_t\}$.

Assumption 4. *$\{X_t\}$ is stationary and ergodic under the null hypothesis. Moreover, ε_t has a continuous and strictly positive density on the real line and $E[\varepsilon_t^4] = \kappa < \infty$.*

This is a standard assumption since, under H_0, $\{X_t\}$ reduces to a AR(p) process. The existence of the density of the innovation process $\{\varepsilon_t\}$ implies the existence of the density of $\{X_t\}$. Moreover, we consider two different sets of additional assumptions on the innovation process.

(A.iid) $\{\varepsilon_t\}$ is a sequence of independent and identically distributed random variables with $E[\varepsilon_t] = 0$, $E[\varepsilon_t^2] = \sigma^2 < \infty$.

(A.wild) $\varepsilon_t = h_t \cdot z_t$, with $\{z_t\}$ being a sequence of i.i.d. random variables with zero mean and unit variance. $h_t > 0$ is a strictly stationary and ergodic random process bounded away from zero with probability one and with finite fourth moments. Also it is measurable with respect to the filtration $\mathcal{F}_t = \sigma\{z_t, z_{t-1}, \dots\}$.

In Theorem 12, under Assumption (A.iid), ((A.wild), respectively) we prove the validity of the i.i.d. bootstrap, (wild bootstrap) by showing that T_n^* converges weakly in probability to T_∞. To this aim, in Proposition 5 we derive a *new* uniform bootstrap law of large numbers (hereafter UBLLN) that allows us to (i) verify that $n^{-1}I_n^*(r)$ converges in probability (in probability) to $\sigma^{-2}I_\infty(r)$ uniformly on r (Proposition 7) and (ii) derive an approximation of $\partial\tilde{\ell}_n^*/\partial\mathbf{\Psi}(r)$ in terms of $\nabla_n^*(r)$ (Proposition 8). We next state the UBLLN in Proposition 5 that establishes a new result which is of independent interest since it is the first proof of the validity of the bootstrap when testing for a regime switching mechanism where a nuisance parameter is absent under the null hypothesis. The main difficulty here resides in the indicator function $I(y \leq r)$ being not differentiable. Hence, standard methods based upon Taylor's expansion cannot be applied. Notice that in Hansen [1996], the problem is circumvented by adopting a stochastic perturbation of the score vector for which no UBLLN is required. Our proof of the bootstrap validity also extends the approach of Chan et al. [2020]. One of the key aspects of our approach is the approximation of the step function with a parameterised sequence of continuous and differentiable functions. The results can be used as the building block for proving the validity of bootstrap testing in all those situations where additive nonlinearity is tested within a (auto)regression framework. These include threshold regression, regime-switching processes with exogenous threshold variables, testing for structural breaks. Establishing the validity of bootstrap testing in such cases requires checking additional assumptions on a case by case basis and will be studied in future works.

Proposition 5. ***(UBLLN)*** *Let $\{X_t\}$ and $\{X_t^*\}$ be defined in Eq. (4) and Eq. (16), respectively. Assume that Assumption 4 holds and one of the two following conditions is satisfied:*

(i) *$\{\varepsilon_t^*\}$ is defined as in (B.iid) and Assumption (A.iid) is fulfilled;*

(w) *$\{\varepsilon_t^*\}$ is defined as in (B.wild) and Assumption (A.wild) is fulfilled.*

Under the null hypothesis, it follows that:

1. *if* $E[|X_t|^u] < \infty$, *for* $u \geq 0$, *then:*

$$\sup_{r\in[r_L,r_U]} \left| \frac{1}{n}\sum_{t=1}^{n} X_t^{*u} I(X_t^* \leq r) - E[X_t^u I(X_t \leq r)] \right| \xrightarrow[n\to\infty]{p^*}{}_p\ 0. \tag{26}$$

2. *if* $E[|X_t|^u] < \infty$, *for* $u = 1, 2$, *then, for every* i, j, d*:*

$$\sup_{r\in[r_L,r_U]} \left| \frac{1}{n}\sum_{t=1}^{n} X_{t-i}^* X_{t-j}^* I(X_{t-d}^* \leq r) - E[X_{t-i}X_{t-j}I(X_{t-d} \leq r)] \right| \xrightarrow[n\to\infty]{p^*}{}_p\ 0. \tag{27}$$

Remark 6. Under the null hypothesis, $E[\varepsilon_t] = 0$ and $E[\varepsilon_t] = \sigma^2 < \infty$ imply that $E[|X_t|^u] < \infty$, for $u = 1, 2$.

In the next proposition we prove that $n^{-1}I_n^*(r)$ converges in probability (in probability) to $\sigma^{-2}I_\infty(r)$ uniformly on r.

Proposition 7. *Let* $\{X_t\}$ *and* $\{X_t^*\}$ *be defined in (4) and Eq. (16), respectively. Assume that Assumption 4 holds and if one of the two following conditions is satisfied:*

(i) $\{\varepsilon_t^*\}$ *is defined as in (B.iid) and Assumption (A.iid) is fulfilled;*

(w) $\{\varepsilon_t^*\}$ *is defined as in (B.wild) and Assumption (A.wild) is fulfilled.*

Under the null hypothesis, it follows that:

$$\sup_{r\in[r_L,r_U]} \left| n^{-1}I_n^*(r) - \sigma^{-2}I_\infty(r) \right| \xrightarrow[n\to\infty]{p^*}{}_p\ 0,$$

with $I_n^*(r)$ *and* $I_\infty(r)$ *being defined in Eq. (23) and Eq. (12), respectively.*

Now, we derive an approximation of $\partial\tilde{\ell}_n^*/\partial\boldsymbol{\Psi}(r)$ in terms of $\nabla_n^*(r)$.

Proposition 8. *Let* $\{X_t\}$ *and* $\{X_t^*\}$ *be defined in Eq. (4) and Eq. (16), respectively. Assume that Assumption 4 holds and one of the two following conditions is satisfied:*

(i) $\{\varepsilon_t^*\}$ *is defined as in (B.iid) and Assumption (A.iid) is fulfilled;*

(w) $\{\varepsilon_t^*\}$ *is defined as in (B.wild) and Assumption (A.wild) is fulfilled.*

Under the null hypothesis, it holds that the bootstrap score defined in Eq. (21) satisfies:

$$\frac{\partial\tilde{\ell}_n^*}{\partial\boldsymbol{\Psi}}(r) = \frac{\partial\ell_n^*}{\partial\boldsymbol{\Psi}}(r) - I_{n,21}(r)I_{n,11}^{-1}\frac{\partial\ell_n^*}{\partial\boldsymbol{\phi}}.$$

Remark 9. By analogy with standard, non-bootstrap, asymptotics [Chan, 1990, Ling and Tong, 2005, Goracci et al., 2023], thanks to Proposition 8 the asymptotic null distribution of T_n^* is predominantly determined by the asymptotic behaviour of $\{\nabla_n^*(r)\}$ rather than $\left\{\partial\tilde{\ell}_n^*/\partial\boldsymbol{\Psi}(r)\right\}$ and this simplifies substantially the derivations.

In Proposition 11 we prove a bootstrap central limit theorem (hereafter BCLT) for $\{\nabla_n^*(r)\}$, the bootstrap score process defined in Eq. (20). The BCLT is the key result to obtain the validity of the bootstrap schemes and requires proving the following UBLLN.

Proposition 10. ***(UBLLN)*** *Let $\{X_t\}$ and $\{X_t^*\}$ be defined in Eq. (4) and Eq. (16), respectively. Assume that Assumption 4 holds and one of the two following conditions is satisfied:*

(i) *$\{\varepsilon_t^*\}$ is defined as in (B.iid) and Assumption (A.iid) is fulfilled;*

(w) *$\{\varepsilon_t^*\}$ is defined as in (B.wild) and Assumption (A.wild) is fulfilled.*

Under the null hypothesis, it follows that:

1. *if $E[|X_t|^u] < \infty$, for $u = 1, 2$, then, for every i, d:*

$$\sup_{r\in[r_L,r_U]} \left| \frac{1}{n}\sum_{t=1}^{n} \varepsilon_t^{*2} X_{t-i}^* I(X_{t-d}^* \leq r) - E[\varepsilon_t^2 X_{t-i} I(X_{t-d} \leq r)] \right| \xrightarrow[n\to\infty]{p^*}{}_p \; 0. \quad (28)$$

2. *if $E[|X_t|^u] < \infty$, for $u = 1, 2$, then, for every i, j, d:*

$$\sup_{r\in[r_L,r_U]} \left| \frac{1}{n}\sum_{t=1}^{n} \varepsilon_t^{*2} X_{t-i}^* X_{t-j}^* I(X_{t-d}^* \leq r) - E[\varepsilon_t^2 X_{t-i} X_{t-j} I(X_{t-d} \leq r)] \right| \xrightarrow[n\to\infty]{p^*}{}_p \; 0. \quad (29)$$

Proposition 11. ***(BCLT)*** *Assume that Assumption 4 holds and one of the following two conditions is satisfied:*

(i) *$\{\varepsilon_t^*\}$ is defined as in (B.iid) and Assumption (A.iid) is fulfilled;*

(w) *$\{\varepsilon_t^*\}$ is defined as in (B.wild) and Assumption (A.wild) is fulfilled.*

Under the null hypothesis, for any fixed r, it holds that

$$\frac{1}{\sqrt{n}} \nabla_n^*(r) \xrightarrow[n\to\infty]{w^*}{}_p \; \xi(r),$$

where $\nabla_n^*(r)$ is defined in Eq. (19), $\xi(r)$ is a Gaussian distributed $2(p+1)$-dimensional random vector with zero-mean and variance-covariance matrix $\Xi(r, r)$ defined in Eq. (13).

Next, we present the main result that guarantees the validity of the proposed bootstrap methods.

Theorem 12. ***(BFCLT)*** *Let T_n^* be the supLM statistic defined in Eq. (24). Assume that Assumption 4 holds and one of the following two conditions is satisfied:*

(i) *$\{\varepsilon_t^*\}$ is defined as in (B.iid) and Assumption (A.iid) is fulfilled;*

(w) $\{\varepsilon_t^*\}$ *is defined as in (B.wild) and Assumption (A.wild) is fulfilled.*

Then, under the null hypothesis it holds that

$$T_n^* \xrightarrow[n\to\infty]{w^*}{}_p\, T_\infty,$$

where T_∞ *is defined in Eq. (14).*

4 Finite sample performance

In this section we investigate the finite sample performance of the bootstrap sLM tests. We adopt the following notation: sLMi is the residual i.i.d. bootstrap test, sLMw is the residual wild bootstrap test. We compare them with the asymptotic test (sLMa) and with Hansen's bootstrap test (sLMh) [Hansen, 1996]. The length of the series is $n = 50, 100, 200$. These are small to moderate sample sizes that are quite common in many fields, especially when the cost of producing the data is not negligible. The nominal size is $\alpha = 5\%$ and the number of Monte Carlo replications is 10000. For the asymptotic test we have used the tabulated values of Andrews [2003], whereas the bootstrap p-values are based on $B = 999$ resamples. The threshold is searched from percentile 25th to 75th of the sample distribution. In Section 4.1 we study the size and the power of the tests when the innovation process is i.i.d. Also, in Section 4.2 we assess the behaviour of the tests when the order of the AR process tested is treated as unknown and is selected through the AIC. Then, in Section 4.3 we study the performance of the tests when the innovation process presents conditional heteroskedasticity. Section B of the Supplementary Material contains additional simulations. In particular, in Section B.1 we study the size of the tests when the DGP is close to non-stationarity, whereas Section B.2 contains additional results on the power of the tests, covering the case of unit root regimes and white noise regimes. All the results are presented as percentages as to enhance the readability of the tables.

4.1 i.i.d. innovations

Throughout this section the innovation process $\{\varepsilon_t\}$ is simulated from a i.i.d. $N(0,1)$ random process. To study the size of the tests, we generate time series from the following AR(1) model:

$$X_t = \phi_0 + \phi_1 X_{t-1} + \varepsilon_t \tag{30}$$

where $\phi_1 = 0, \pm 0.3, \pm 0.6, \pm 0.9$. Also, we set $\phi_0 = 0$ since preliminary investigations showed that its value has no impact upon the results. Figure 1 shows the boxplots of the rejection percentages for the four tests. Each boxplot groups together the size for all the 7 values of ϕ_1. The bootstrap tests sLMi and sLMb present a size very close to the nominal 5% even for a sample size as small as 50 and are not influenced by the value of the autoregressive parameter close

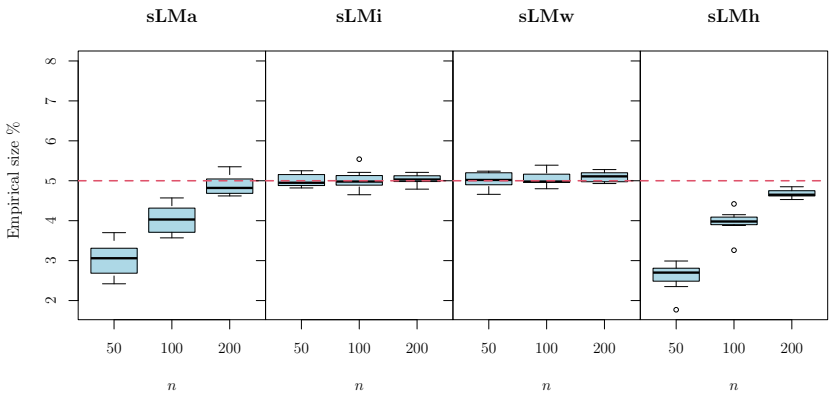

Figure 1: Empirical size in percentage, at nominal 5% level, of the four sLM tests for varying sample size n. sLMa: asymptotic test; sLMi: residual i.i.d. bootstrap test, sLMw: residual wild bootstrap test; sLMh: Hansen's test. Each boxplot groups together the size for the 7 cases.

to non-stationarity, see also Section B.1 of the Supplementary Material. On the contrary, both the asymptotic test sLMa and Hansen's test sLMh result undersized, especially for small sample sizes. As we will show, this impinges negatively upon the power of the tests.

As concerns power, we simulate from the following TAR(1) model:

$$X_t = \phi_0 + \phi_1 X_{t-1} + (\Psi + \Psi X_{t-1})\, I(X_{t-1} \leq 0) + \varepsilon_t. \tag{31}$$

where $(\phi_0, \phi_1) = (-0.1, -0.8)$ (M1) and $(0.8, 0.2)$ (M2). Also, $\Psi \in \{0.0, 0.3, 0.6, 0.9\}$ as to obtain 8 different parameter settings. Note that the parameter Ψ represents the departure from the null hypothesis and in all the simulations below we take sequences of increasing distance from H_0. The case $\Psi = 0$ corresponds to H_0. Table 1 presents the empirical power (in percentage) at nominal level 5% where, for each sample size, the first row corresponds to the null hypothesis and reflects the size of the tests, while the subsequent three rows represent increasing departures from H_0 and reflect the power of the tests. As expected, the bootstrap tests sLMi and sLMw have superior power with respect to Hansen's test sLMh. The asymptotic test sLMa is also slightly inferior in terms of power. As the sample size increases the four tests tend to have comparable power but the bootstrap tests retain a small advantage. Additional simulations show that the tests have power also in case of local unit-roots or in the local white noise case, see Section B.2 of the Supplementary Material.

4.2 The impact of order selection

In practical situations, the order of the autoregressive model to be tested is unknown and has to be estimated. This can impinge on the performance of the tests so that we assess the impact of treating the order p of the AR model as

		M1				M2			
n	Ψ	sLMa	sLMi	sLMw	sLMh	sLMa	sLMi	sLMw	sLMh
50	0.0	3.1	4.8	4.7	2.6	3.5	5.2	5.4	3.3
	0.3	8.8	12.0	11.9	7.4	4.1	6.1	6.2	3.8
	0.6	29.5	35.4	34.4	24.6	6.8	9.5	9.2	5.6
	0.9	60.5	66.5	65.3	51.2	12.8	16.6	16.1	9.7
100	0.0	4.1	5.0	5.0	4.0	4.3	5.1	5.3	4.3
	0.3	20.1	22.6	22.3	18.2	6.4	7.4	7.3	6.0
	0.6	64.1	66.6	66.2	59.4	15.1	16.8	16.2	13.0
	0.9	94.1	94.9	94.6	91.8	33.2	35.4	34.5	29.2
200	0.0	4.6	4.7	4.7	4.4	4.8	4.9	5.0	4.6
	0.3	40.8	41.1	40.8	38.7	10.2	10.4	9.9	9.0
	0.6	94.2	94.2	94.2	92.9	34.4	34.4	33.5	31.4
	0.9	99.9	99.9	99.9	99.9	67.6	67.7	66.6	64.1

Table 1: Empirical power (%) at nominal level $\alpha = 5\%$ for the TAR(1) process of Eq. (31) and sample size $n = 50, 100, 200$. The parameter Ψ represents the departure from the null hypothesis.

unknown and selecting it by means of the AIC. We study the impact on the size of the tests by simulating from the following AR(2) model.

$$X_t = \phi_0 + \phi_1 X_{t-1} + \phi_2 X_{t-2} + \varepsilon_t \tag{32}$$

where $\phi_0 = 0$, whereas ϕ_1 and ϕ_2 are as follows:

ϕ_1	-0.65	0.80	-0.35	1.15	0.45	0.45	-0.90
ϕ_2	0.25	0.10	-0.45	-0.55	0.25	-0.55	-0.25

The results are presented in the boxplots of Figure 2 where in the upper panel we have used the true AR order, whereas in the lower panel the order is selected using the AIC. Each boxplot groups together the size for all the 7 values of (ϕ_1, ϕ_2). The results confirm that treating the order of the autoregression as unknown has negligible impact upon the size of the four tests.

We study the impact of model selection upon the power of the tests by simulating from the following TAR(2) process:

$$X_t = -0.35 X_{t-1} - 0.45 X_{t-2} + (\Psi + \Psi X_{t-1} + \Psi X_{t-2})\, I(X_{t-1} \leq 0) + \varepsilon_t. \tag{33}$$

as before, $\Psi \in \{0.0, 0.3, 0.6, 0.9\}$ represents the level of departure from H_0. The rejection percentages are shown in Table 2, where in the left panel the true order of the autoregression is used, whereas in the right panel it is selected using the AIC. Contrarily to size, the procedure of order selection does have a negative impact upon power for $n = 50, 100$, much less so for $n = 200$. This can

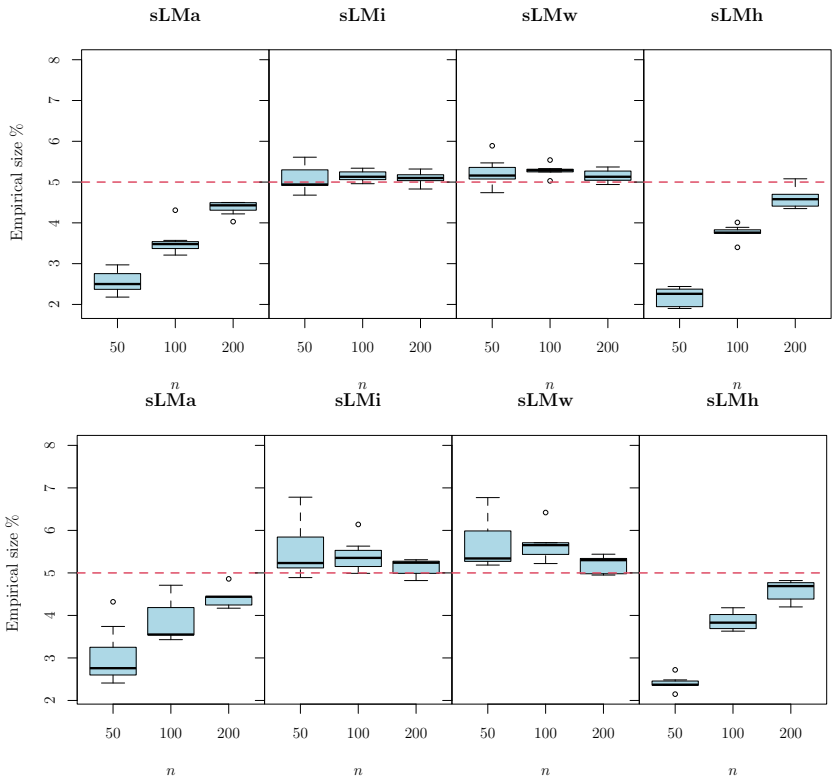

Figure 2: Empirical size in percentage, at nominal 5% level, of the four sLM tests for varying sample size n for the AR(2) process of Eq. (32). sLMa: asymptotic test; sLMi: residual i.i.d. bootstrap test, sLMw: residual wild bootstrap test; sLMh: Hansen's test. Upper panel: the true AR order is used in the test. Lower panel: the AR order is treated as unknown and selected using the AIC. Each boxplot groups together the size for the 7 cases.

		true				AIC			
n	Ψ	sLMa	sLMi	sLMw	sLMh	sLMa	sLMi	sLMw	sLMh
50	0.0	2.7	5.0	5.1	2.2	2.7	4.9	5.0	2.1
	0.3	6.5	11.3	11.1	5.3	5.6	10.0	10.1	4.4
	0.6	27.3	37.5	36.7	22.7	18.4	25.3	25.1	14.5
	0.9	68.6	77.2	76.0	60.5	37.9	45.1	44.1	30.6
100	0.0	3.3	4.7	4.9	3.5	3.2	4.6	4.8	3.3
	0.3	17.8	22.2	22.3	17.2	16.3	20.3	20.2	15.2
	0.6	72.6	77.0	76.5	70.8	54.8	58.9	58.3	52.1
	0.9	98.8	99.1	99.0	98.3	77.9	79.5	78.5	75.5
200	0.0	4.2	4.9	5.1	4.5	4.1	5.0	5.1	4.3
	0.3	43.5	46.1	45.7	42.8	40.4	43.2	42.9	39.7
	0.6	98.7	98.8	98.7	98.4	92.5	92.7	92.4	91.5
	0.9	100.0	100.0	100.0	100.0	97.8	97.8	97.5	97.4

Table 2: Empirical power (in percentage) at nominal level $\alpha = 5\%$ for the TAR(2) process of Eq. (33). In the left panel the true order of the autoregression is used, whereas in the right panel it is selected through AIC.

be somehow expected since the AIC is prone to overfitting and leads to using critical values larger than the correct ones. This suggests that other model selection criteria might be more appropriate in this context, such as the BIC or, as advocated in Goracci et al. [2023], the Hannan-Rissanen criterion.

4.3 Heteroskedastic innovations

In this section we investigate the performance of the tests when the innovation process $\{\varepsilon_t\}$ presents conditional heteroskedasticity. As for the size, we generate time series from $7 \times 3 = 21$ different simulation settings of the following AR(1)-GARCH(1,1) model:

$$\begin{aligned} X_t &= \phi_1 X_{t-1} + \varepsilon_t \\ \varepsilon_t &= \sigma_t z_t \\ \sigma_t^2 &= 1 + a_1 \varepsilon_{t-1}^2 + b_1 \sigma_{t-1}^2 \end{aligned} \tag{34}$$

where $z_t \sim$ i.i.d. $N(0,1)$ and, as before, $\phi_1 = 0, \pm 0.3, \pm 0.6, \pm 0.9$. We combine these with the following parameters for the GARCH specification: $(a_1, b_1) = (0.04, 0.95)$ (case A), $(0.3, 0.0)$ (case B), $(0.4, 0.4)$ (case C). Note that case B corresponds to an ARCH(1) process. Also, all three cases fulfil the condition of finite fourth moments. The boxplots of the size of the four tests are presented in Figure 3, where we have grouped together the three cases since they behave similarly. Hence, each boxplot contains the size for $7 \cdot 3 = 21$ parameters'

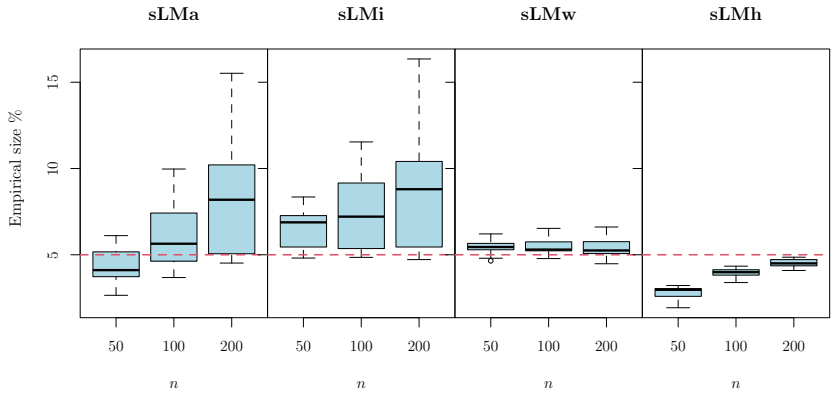

Figure 3: Empirical size (%) in presence of heteroskedasticity, at nominal 5% level, of the four sLM tests for varying sample size n. sLMa: asymptotic test; sLMi: residual i.i.d. bootstrap test, sLMw: residual wild bootstrap test; sLMh: Hansen's test. Each boxplot groups together the size for the 21 cases.

combinations. Clearly, both the asymptotic test sLMa and the i.i.d. bootstrap test sLMi present a size bias that increases with sample size, whereas the wild bootstrap test sLMw behaves as expected even for $n = 50$ and its size tends to the nominal 5% as the sample size increases. Hansen's sLMh test is undersized also in this scenario and approaches the nominal size from below as the sample size increases. Note that in Hansen [1996] p. 422, Corollary 2, the innovations are assumed to be i.i.d. to guarantee absolute regularity of the TAR process, but the author remarks that the condition could be relaxed to a martingale difference sequence. Our results confirm that this is indeed the case. Finally, we have replicated the same experiment when the DGP is near to non stationarity and found no significant differences from the above results. These latter findings are reported in Section B.1 of the Supplementary Material.

In order to study the power of the tests we simulate from the following TAR(1)-GARCH(1, 1) model:

$$\begin{aligned} X_t &= \phi_0 + \phi_1 X_{t-1} + (\Psi + \Psi X_{t-1})\, I(X_{t-1} \leq 0) + \varepsilon_t. \\ \varepsilon_t &= \sigma_t z_t \\ \sigma_t^2 &= 1 + a_1 \varepsilon_{t-1}^2 + b_1 \sigma_{t-1}^2 \end{aligned} \tag{35}$$

where, as above, $(\phi_0, \phi_1) = (-0.1, -0.8)$ (M1) and $(0.8, 0.2)$ (M2), $\Psi \in \{0.0, 0.3, 0.6, 0.9\}$, and $(a_1, b_1) = (0.04, 0.95)$ (case A), $(0.3, 0.0)$ (case B), $(0.4, 0.4)$ (case C). We test the following system of hypotheses:

$$\begin{cases} H_0 & : \boldsymbol{\Psi} = \mathbf{0} \\ H_1 & : \boldsymbol{\Psi} \neq \mathbf{0}, \end{cases}$$

where $\boldsymbol{\Psi} = (\Psi, \Psi)$, so that the parameter Ψ represents the departure from the null hypothesis. The empirical power (rejection percentages) is shown in

		$n = 50$				$n = 100$				$n = 200$			
	Ψ	sLMa	sLMi	sLMw	sLMh	sLMa	sLMi	sLMw	sLMh	sLMa	sLMi	sLMw	sLMh
M1-A	0.0	3.2	5.0	5.1	2.7	4.0	4.8	4.7	3.8	5.0	5.1	5.0	4.7
	0.3	5.8	8.1	8.0	5.0	12.4	13.9	13.7	11.0	23.9	24.3	23.4	21.4
	0.6	13.7	17.6	16.9	11.1	31.9	34.3	32.9	28.4	64.6	64.5	63.0	59.9
	0.9	25.6	30.7	29.5	20.9	57.1	59.7	57.7	52.1	89.7	89.7	88.9	87.3
M2-A	0.0	3.8	5.4	5.5	3.2	4.4	5.1	5.1	4.2	5.1	5.3	4.9	4.4
	0.3	4.6	6.5	6.8	4.3	8.4	9.8	9.7	8.0	14.8	15.1	14.7	13.6
	0.6	9.6	13.5	13.2	9.1	22.6	25.4	24.8	22.0	50.8	51.3	49.6	47.3
	0.9	20.3	27.1	26.3	18.5	52.0	56.0	54.7	50.1	88.8	89.4	88.8	87.5
M1-B	0.0	4.5	6.2	5.2	2.7	5.7	6.8	5.1	3.9	7.0	7.1	5.0	4.2
	0.3	10.3	13.7	11.5	6.7	21.7	23.6	20.0	15.0	41.8	42.4	35.5	31.1
	0.6	28.9	34.8	31.5	19.9	60.0	62.5	57.9	48.6	89.9	89.9	86.8	81.2
	0.9	55.2	60.5	57.6	41.7	89.9	91.0	89.0	81.5	99.7	99.7	99.3	98.0
M2-B	0.0	5.2	7.3	6.0	3.3	6.9	8.1	5.4	3.8	9.8	10.1	5.9	4.7
	0.3	6.5	8.8	7.4	4.3	11.0	12.4	9.1	7.0	18.1	18.5	12.2	10.2
	0.6	9.6	12.9	11.1	6.7	21.9	24.0	19.4	15.2	43.6	44.3	35.7	31.4
	0.9	17.3	22.1	19.4	12.6	41.7	44.1	39.1	33.2	75.6	76.1	70.3	65.8
M1-C	0.0	5.6	7.6	5.4	2.8	7.9	8.8	4.8	3.6	12.2	12.8	5.1	4.2
	0.3	11.6	14.4	11.2	6.4	21.9	24.1	16.2	12.1	38.6	39.4	25.3	20.8
	0.6	24.3	28.8	23.7	14.8	48.6	51.0	41.0	32.6	78.3	78.9	65.8	57.5
	0.9	41.6	47.4	42.0	28.8	74.7	76.8	69.0	58.9	96.2	96.3	90.8	85.3
M2-C	0.0	5.7	8.0	5.9	3.3	10.0	11.6	6.4	4.2	15.2	16.0	6.5	4.3
	0.3	7.0	9.8	7.7	4.7	14.3	16.5	10.4	7.6	24.3	25.6	13.3	10.5
	0.6	11.7	15.4	12.8	8.4	27.3	29.9	22.5	17.7	52.3	53.6	39.0	33.9
	0.9	21.8	27.2	24.1	17.0	49.9	53.0	46.0	40.2	82.9	83.7	74.4	69.2

Table 3: Empirical power (%) in presence of heteroskedasticity, at nominal level $\alpha = 5\%$ for the TAR(1)-GARCH(1,1) process of Eq. (35) and sample size $n = 50, 100, 200$. The parameter Ψ represents the departure from the null hypothesis.

Table 3, where each block of four rows represents one of the six combinations above. Note that the first values of each block correspond to $\Psi = 0.0$ and reflect the size for that specific parameterization. In general, when the size of the four tests is comparable, also their power is. However, Hansen's test sLMh tends to be undersized also in this case and this causes its ensuing inferior power. As also shown above, the tests sLMa and sLMi are oversized in a number of situations and their power reflects this. The comparison of these results with the case of i.i.d. innovations of Table 1 shows that, the presence of heteroskedastic innovations interacts non-trivially with the discriminating power of the tests, depending also upon the TAR parameters. For instance, case M1 with i.i.d. innovations presents higher power than cases M1-A, M1-B, M1-C with heteroskedastic innovations, while the reverse holds for case M2.

5 An application: the Greenland ice sheet mass balance

We apply our methodology to the time series of the Greenland ice sheet mass balance. The Greenland ice sheet covers around 85% of the whole island and can reach more than 3 km of height. After decades of relative stability, in 1998 the ice sheet began to lose mass every year and marked records of ice loss in 2012 and 2019. A decline is observed when the mass gain from snow accumulation is exceeded by mass loss due to several factors: surface meltwater runoff, marine-terminating glacier calving, submarine melting, basal melting. This has been recognized as one of the main contributors to the rise of global sea-level and its study is of fundamental importance for monitoring and understanding climate change. In particular, the contribution from the Arctic, primarily Greenland, Alaska, and Arctic Canada, is approximately 350 billion tons/yr, which corresponds to an increase of 1 mm of the sea level each year.

It is now recognised that the behaviour of the Greenland ice sheet mass subject to global warming is governed by feedback mechanisms and coupling [Zeitz et al., 2022]. As argued in Wouters et al. [2013], ice sheets can be affected by processes with very different timescales: from daily variations in incoming radiation and transient stormy events to low-frequency atmospheric (e.g. the El Niño–Southern Oscillation, the North Atlantic Oscillation, the Southern Annular Mode) and oceanic fluctuations (e.g., changes in ocean circulation leading to warm water intruding into glacier fjords). In particular, the major driving forces determining mass loss are identified as increasing temperatures and decreasing reflectivity of surface snow. However, the latter is also influenced by temperature, so that warming is recognised as the most important factor [Scambos et al., 2021]. Notably, the causes of such an increase are to be found in changes in the seasonal pattern with greater influence of meridional heat advection during the summer. Indeed, years with dominant meridional circulation have become more frequent since the end of the 1990s [Sasgen et al., 2022]. For this reason, even if the majority of past studies focused on the long-term trend, recent works highlighted the importance of investigating both the interannual variability and seasonal patterns [Zhang et al., 2020]. In particular, the increase in mass loss of the last decades was also accompanied by an increase in variability both at the seasonal level and in the interannual component. This was also noted in Sasgen et al. [2020a], that found out that the variability observed in the rates of mass loss were due to anomalous seasonal behaviour. Moreover, there is increasing consensus on the fact that the dynamics underlying the ice sheet mass balance contain a stochastic component [Wouters et al., 2013, Zhang et al., 2020]. For instance, Zhang et al. [2020] propose a specification containing a polynomial deterministic trend, a trigonometric, non-random, seasonality and a linear autoregressive model, estimated through a Kalman filter, for the interannual component.

Given the importance of the matter, many different projects are devoted to estimating mass balance of the Greenland ice sheet and three methods can be

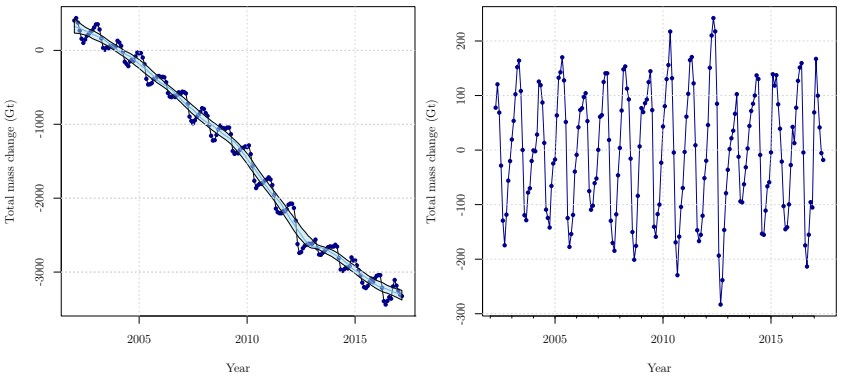

Figure 4: Monthly time series of the Greenland ice sheet mass balance (in Gigatons) from April 2002 to June 2017. Left: time series with the estimated local linear trend and robust confidence bands. Right: detrended series.

identified, each one with its pros and cons. The first method measures changes in gravity, the second one measures changes in volume, and the third method is based upon an input-output model. In Mankoff et al. [2021] these are reviewed and it is shown that they produce data that are in general agreement. Here we focus on the data originating from the Gravity Recovery and Climate Experiment (GRACE), a joint mission of NASA and the German Aerospace Center that monitors ice-sheet dynamics through the measurement of anomalies in the Earth's gravitational field. The monthly time series comes from Sasgen et al. [2020a] and is available at Sasgen et al. [2020b]; in October 2017 the mission ended and was replaced by a follow-on mission (GRACE-FO) launched in May 2018. This produced a gap of almost one year in the data so that we focused on the GRACE data that ranges from April 2002 to June 2017 ($n = 183$). By using spline interpolation, we imputed the missing values and rendered the series regularly spaced. The time plot of the resulting series is presented in Figure 4(left). A local linear trend and associated robust wild bootstrap confidence bands at 95% are superimposed in light blue. The bandwidth for the trend spans approximately two years of data, whereas the robust confidence bands have been derived according to Friedrich et al. [2020] and are valid under serial dependence and heteroskedasticity. Since 2002, there has been a 281 Gigatons/years average mass loss with a partial slowdown from 2013 onwards. The detrended series is shown in Figure 4(right). It is characterised by a strongly asymmetric yearly cyclical component where the slow-rising phase of the cycle lasts 8 months, from October to May, and is followed by a fast decline from June to September, see Figure 5(left). The lag plot of Figure 5(right) hints at the presence of a nonlinear oscillatory mechanism. The months associated to the rising/declining phase of the cycle are coloured in orange/blue and are clearly separated in the state space. The power spectral density of the detrended series is shown in Figure 6, where the frequencies associated to the dominant peaks are annotated

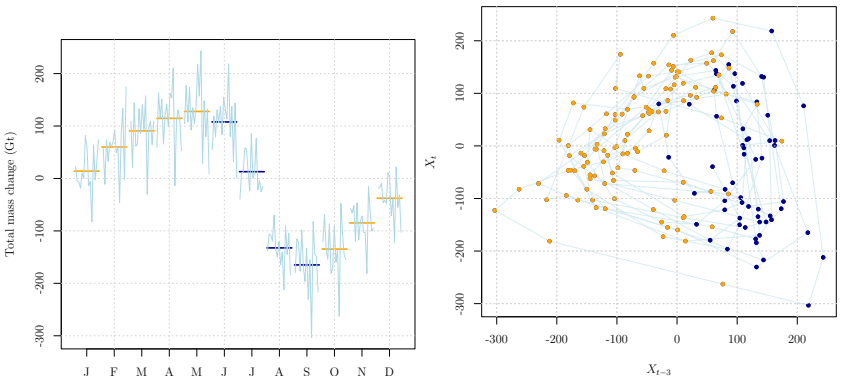

Figure 5: Monthly time series of the Greenland ice sheet mass balance (in Gigatons) from April 2002 to June 2017. Left: month plot with monthly averages. Right: lag plot of the series X_t vs X_{t-3}. Months corresponding to the rising phase of the cycle are coloured in orange/light gray, whereas those associated to the declining phase are coloured in blue/dark gray).

as fractions to facilitate the identification of the corresponding periodicities (in months). Besides the yearly cycle and its harmonics, the spectrum has a complex pattern of peaks that hints at the presence of resonances induced by forcing mechanisms, responsible of the ice melting, interacting with the nonlinear dynamics. Resonances, also known as phase-lockings, are produced by nonlinear systems subject to external forcing frequencies that lock into a resonant periodic response possessing a rational frequency ratio [Cartwright et al., 1999, 2001]. In particular, a non-dynamic nonlinearity forced quasiperiodically by two frequencies ω_1 and ω_2 generates a resonance at frequency ω_R, which is a solution of the equation $p\omega_1 + q\omega_2 + \omega_R = 0$, where p, q are integer numbers. This appears to be the case for the Greenland ice sheet as it is immediate to show that $1/12 + 1/27.5 \approx 1/8.4$ and $1/12 - 1/27.5 \approx 1/21.3$. Three-frequency resonances can also be identified. A full analysis of the resonances of the Greenland ice sheet mass balance dynamics is beyond the scope of the present work and will be pursued in future projects. The above exploratory analysis indicates that threshold autoregressive models appear as natural candidates to describe the Greenland ice sheet dynamics, especially due to their ability to reproduce the nonlinear features of the series. Indeed, they are the discrete-time version of differential equations [see Tong, 1990], which are commonly used in climate models. The direct modelling of the asymmetric limit cycle as a nonlinear stochastic oscillator can provide additional insights with respect to existing approaches. Hence, we test for the presence of a significant threshold effect by applying our methodology.

Table 4 shows the results of the application of the sLM tests upon the detrended series, for different values of the delay parameter $d = 1, \ldots, 7$. In the first four columns the AR(1) specification is tested against the alternative of

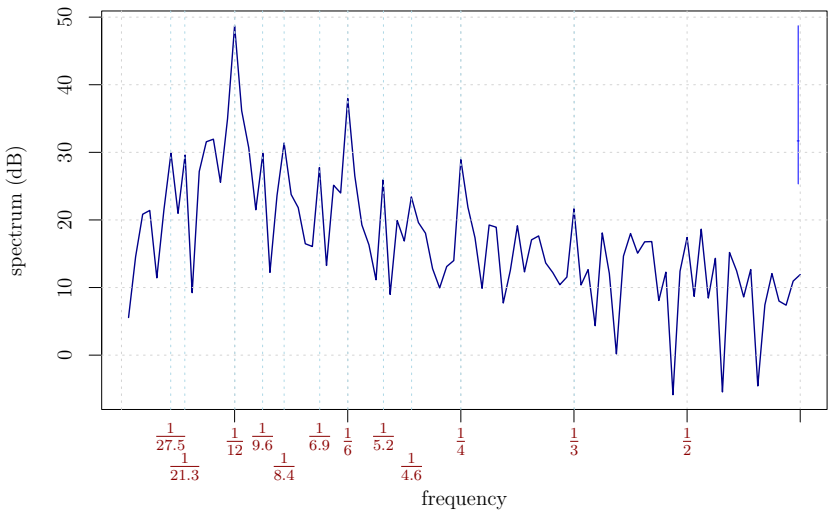

Figure 6: Power spectral density of the detrended time series of the Greenland ice sheet mass balance. The frequencies corresponding to the main peaks are annotated as fractions to facilitate the identification of the associated periods (in months).

a TAR(1), whereas in the last four columns we used AR order 13, which was selected by means of the AIC. The columns denoted by sLMa report the value of the test statistic, whereas the other columns contain the bootstrap p-values of the tests based upon $B = 9999$ resamples. Our bootstrap tests clearly reject the null hypothesis for values of d greater than one, both for order 1 and order 13. Note that the rejection is less neat for Hansen's test when the AR order tested is 13, and this probably reflects its inferior power. Overall, there is a strong indication of a threshold effect and this justifies the adoption of a threshold autoregressive specification. The series is likely to be affected by measurement error so that a threshold ARMA specification is more appropriate than the TAR model (see Chan et al. [2024] for the theoretical justification). Typically, the MA parameters greatly enhance the flexibility of the model, while retaining parsimony Goracci [2020b,a]. We adopt a TARMA model where the MA component is fixed across regimes. As also discussed in Goracci et al. [2023], the asymptotic behaviour of the test statistics is not affected by the presence of the MA part, if it is not tested, but depends only upon the number of tested parameters. The proposed TARMA specification is presented in Eq. (36), where

	AR order = 1				AR order = 13			
d	sLMa	sLMi	sLMw	sLMh	sLMa	sLMi	sLMw	sLMh
1	4.2	0.576	0.589	0.685	33.5	0.042	0.102	0.236
2	70.1	0.000	0.000	0.000	46.5	0.001	0.003	0.034
3	94.5	0.000	0.000	0.000	48.9	0.000	0.001	0.026
4	85.1	0.000	0.000	0.000	45.4	0.001	0.004	0.037
5	68.9	0.000	0.000	0.000	45.3	0.001	0.003	0.024
6	58.7	0.000	0.000	0.000	62.5	0.000	0.000	0.001
7	28.2	0.000	0.000	0.002	47.6	0.000	0.002	0.020

Table 4: Results of the application of the sLM tests to the detrended time series for delay $d = 1, \ldots, 7$. sLMa columns report the value of the sLM test statistic, whereas remaining columns contain the p-values of the bootstrap tests. The AR order 13 was selected through AIC.

the standard errors are reported in parenthesis below the estimates.

$$X_t = \underset{(0.05)}{0.46}\, X_{t-1} + \underset{(0.07)}{0.75}\, \varepsilon_{t-1} + \tag{36}$$
$$\begin{cases} \underset{(8.85)}{56.39} + \underset{(0.08)}{0.21}\, X_{t-6} + \underset{(0.05)}{0.30}\, X_{t-11} + \varepsilon_t, & \text{if } X_{t-6} \leq -5.86 \\ \underset{(6.41)}{-24.14} + \underset{(0.11)}{0.21}\, X_{t-12} - \underset{(0.07)}{0.26}\, X_{t-13} + \underset{(0.08)}{0.37}\, X_{t-24} & \\ \quad \underset{(0.04)}{-0.15}\, X_{t-27} + \underset{(0.09)}{0.07}\, X_{t-36} + \varepsilon_t, & \text{if } X_{t-6} > -5.86 \end{cases}$$

The model includes a common ARMA(1,1) effect plus two regimes separated by a threshold close to zero, which is consistent with the state space partition suggested in Figure 5(right). Hence, the lower regime corresponds to the slow, rising phase of the cycle, whereas the upper regime is associated to the fast declining phase and depends upon seasonal lags, multiple of the dominant yearly period. The delay $d = 6$ also seems to imply a seasonal dependence from forcing factors. In Figure 7(left) we show a simulated path from the deterministic skeleton of the fitted model, i.e. the fitted model without noise components: it is a stable limit cycle with an ascending phase of 8 months and a descending phase of 4 months that matches the observed asymmetric cycle. Also, Figure 7(right) presents the histogram of the series with its kernel density estimate (red line). We have superimposed in blue the estimated density of a realization of 100k observations from the fitted model. The density of the estimated model manages to reproduce in part the complex multimodal asymmetric density of the series. The main observed periodicities are also captured by the fitted model. This is shown in Figure 8(left) that contains the power spectral density of the simulated series, where we have reported the main observed frequencies derived from Figure 6. The plot of the observed series (light blue) and the superimposed fitted series (dark blue) is shown in Figure 8(right). The diagnostic analysis performed both on the residuals and on the squared residuals of the

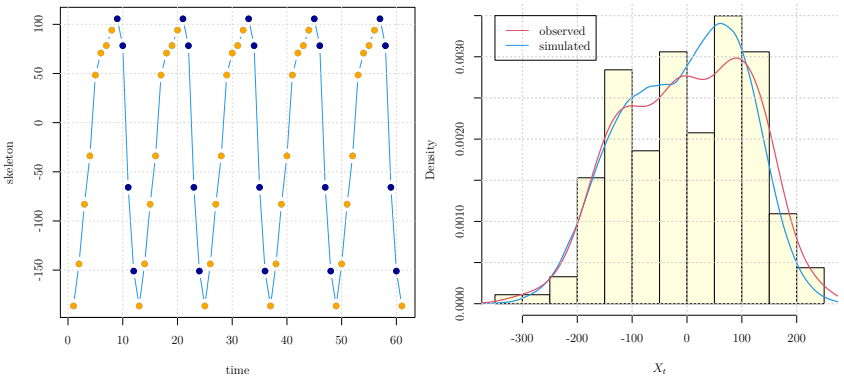

Figure 7: Left: simulated trajectory from the deterministic skeleton of the fitted model of Eq. (36). Right: histogram with superimposed kernel density estimate of the detrended series (red/light gray line) and of a simulated series from the fitted model (blue/dark gray line).

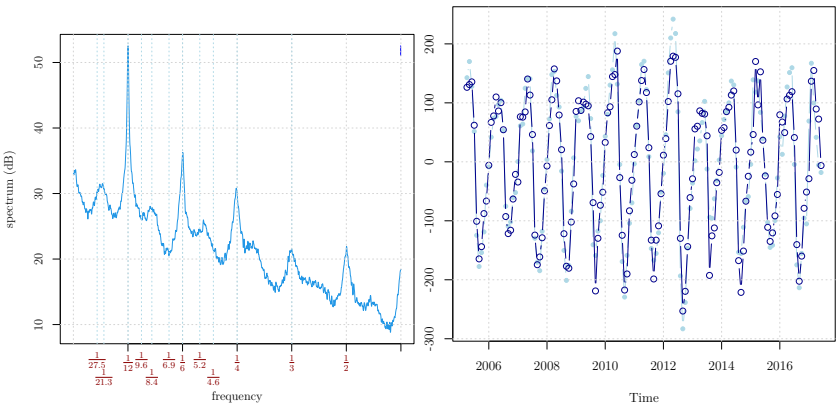

Figure 8: (Left): power spectral density of the simulated time series of 100k observations from the model fit of Eq. (36). The frequencies corresponding to the main peaks derived from Figure 6 are also annotated. (Right): detrended series (light blue, filled points) with superimposed fitted model from Eq. (36) (dark blue, empty points).

fitted models does not show any unaccounted dependence, see Figures 10, 11 of the Supplementary Material. Finally, the Shapiro-Wilk test applied to the residuals does not show departures from normality (see also Figure 12 of the Supplementary Material). A more parsimonious model that reproduces the observed asymmetric limit cycle is presented in Eq. (37).

$$X_t = \underset{(0.05)}{0.61}\, X_{t-1} + \underset{(0.07)}{0.58}\,\varepsilon_{t-1} + \begin{cases} \underset{(7.06)}{24.66} + \underset{(0.05)}{0.22}\, X_{t-11} + \varepsilon_t, & \text{if } X_{t-6} \leq -3.85 \\ \underset{(6.66)}{-41.05} - \underset{(0.04)}{0.34}\, X_{t-27} + \varepsilon_t, & \text{if } X_{t-6} > -3.85 \end{cases} \tag{37}$$

Also in this case, the associated deterministic map of the fitted model (skeleton) converges to a stable limit cycle with an ascending phase of 8 months and a fast 4-month declining phase.

The nonlinearity in the seasonal cycle is expected since it is heavily influenced by the atmospheric dynamics, which is known to be highly nonlinear. An analysis of the association between temperature and ice loss is performed in Hanna et al. [2021]. They found that the Greenland Ice Sheet dynamics is highly sensitive to global warming and anomalies in seasonal temperature. They also highlight a high positive association between coastal Greenland mean surface air temperatures and the Greenland Blocking Index (GBI), which measures the tendency of large anticyclones to develop and persist for several days or longer in the vicinity of southern Greenland. The limitation of existing studies that typically assume linear dependence is pointed out in Woollings et al. [2010], which suggest that the frequency of Greenland blocking occurrence determines the two-regime behaviour observed in the North Atlantic Oscillation. To the best of our knowledge this is the first time a threshold autoregressive model is applied in this field and we believe it complements existing approaches by taking into account different aspects of the modelling: first, the endogeneity of the threshold variable automatically accounts for the unknown/unobservable, possibly multifactorial, cause of the regime switching behaviour; moreover, the TARMA fit manages to reproduce the main nonlinear features of the observed series, such as its dominant spectral modes and hints at the presence of complex resonances, whose study can prompt further research and help elucidating the forcing mechanisms behind the dynamics of the ice sheet mass balance.

Acknowledgements

This work was supported by a STSM Grant from COST Action CA17120 (European Cooperation in Science and Technology). Greta Goracci acknowledges the support of Libera Università di Bolzano, Grant WW202G (TARMAECON). Simone Giannerini Acknowledges the support of University of Bologna, ALMArie CURIE 2021 project. Anders Rahbek gratefully acknowledges financial support from the Independent Research Fund Denmark (Grant no. 0133-00162B).

We thank the editor (Serena Ng), the associate editor and two anonymous referees for the comments that helped improving the paper. We are grateful

to Diego Luis Gonzalez, Julyan Cartwright and Giuseppe Cavaliere for useful discussions. We also thank all the participants of the VieCo 2022 workshop.

Appendix

This appendix contains the proofs of theorems, lemmas, and propositions. The online supplement contains technical lemmas, further results from the simulation study and from the analysis of the Greenland ice sheet mass balance.

A Proofs

A.1 Proof of Proposition 5

PART 1. The proof is divided in two parts: first, we show Eq. (26) for a given r and then we prove that the result holds also uniformly for $r \in [r_L, r_U]$.

Pointwise convergence. We assume r to be fixed and, for each $\eta > 0$, we show that

$$P^*\left(\left|\frac{1}{n}\sum_{t=1}^{n} X_t^{*u} I(X_t^* \leq r) - E[X_t^u I(X_t \leq r)]\right| > 2\eta\right) \xrightarrow[n\to\infty]{p} 0. \tag{38}$$

Since the indicator function $I(y \leq r)$ is not differentiable, standard methods based upon Taylor's expansion cannot be applied. We exploit the fact that the function is discontinuous only at r. By extending the approach used in Chan et al. [2020], we approximate the step function with a sequence of continuous and differentiable functions $G_\alpha(y)$, parameterized by $\alpha \geq 0$:

$$G_\alpha(y) = \begin{cases} \frac{1}{2} + \frac{1}{\pi}\arctan(\frac{r-y}{\alpha}) & \text{if } y \neq r \\ \frac{1}{2} & \text{if } y = r \end{cases} \tag{39}$$

In Figure 9 we show the plot of $G_\alpha(y)$ for three values of α, together with the limit value $\alpha = 0$ for which $G_\alpha(y) = I(y \leq r)$ almost surely. For each $\delta > 0$, define the interval

$$[L_{\alpha,\delta}, U_{\alpha,\delta}] := r \pm q_{\alpha,\delta}, \tag{40}$$

where $q_{\alpha,\delta} = \alpha \tan(\pi(\delta - 1/2))$. This implies:

$$|I(y \leq r) - G_\alpha(y)| < \delta \text{ if } y \notin [L_{\alpha,\delta}, U_{\alpha,\delta}]; \tag{41}$$

$$|I(y \leq r) - G_\alpha(y)| < 1 \text{ if } y \in [L_{\alpha,\delta}, U_{\alpha,\delta}]. \tag{42}$$

Conditions Eq. (40)–(42) assure that, when α and δ approach zero, the interval $[L_{\alpha,\delta}, U_{\alpha,\delta}]$ collapses on r and the distance between $G_\alpha(\cdot)$ and $I(\cdot \leq r)$, which

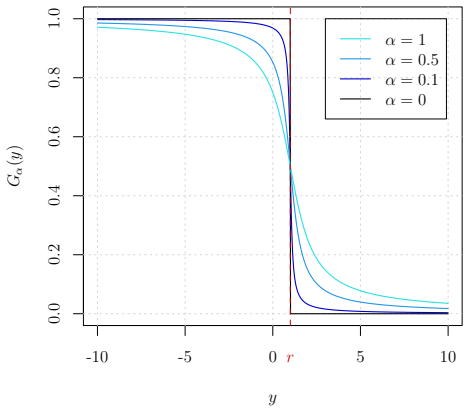

Figure 9: $G_\alpha(y)$ for $\alpha = 1, 0.5, 0.1$, together with the limit value $\alpha = 0$ for which $G_\alpha(y) = I(y \leq r)$ almost surely.

is bounded by δ, vanishes. Now, it holds that:

$$
\begin{aligned}
& P^* \left(\left| \frac{1}{n} \sum_{t=1}^{n} X_t^{*u} I(X_t^* \leq r) - E[X_t^u I(X_t \leq r)] \right| > 2\eta \right) \\
& \leq P^* \left(\left| \frac{1}{n} \sum_{t=1}^{n} X_t^{*u} I(X_t^* \leq r) - \frac{1}{n} \sum_{t=1}^{n} X_t^{*u} G_\alpha(X_t^*) \right. \right. \\
& \qquad \left. \left. - E[X_t^u I(X_t \leq r)] + E[X_t^u G_\alpha(X_t)] \right| > \eta \right) \qquad (43) \\
& + P^* \left(\left| \frac{1}{n} \sum_{t=1}^{n} X_t^{*u} G_\alpha(X_t^*) - E[X_t^u G_\alpha(X_t)] \right| > \eta \right). \qquad (44)
\end{aligned}
$$

Markov's inequality implies that, in order to prove that Eq. (43) is $o_{p^*}(1)$, it suffices to show that the following two expectations vanish in probability:

$$
E^* \left[\left| E[X_t^u I(X_t \leq r)] - E[X_t^u G_\alpha(X_t)] \right| \right], \qquad (45)
$$

$$
E^* \left[\left| \frac{1}{n} \sum_{t=1}^{n} X_t^{*u} I(X_t^* \leq r) - \frac{1}{n} \sum_{t=1}^{n} X_t^{*u} G_\alpha(X_t^*) \right| \right]. \qquad (46)
$$

As for Eq. (45): let $f_X(\cdot)$ be the stationary probability density function of the AR(p) process $\{X_t\}$. Since it is continuous [see e.g. Anděl and Hrach, 2000, theorem 1.3] and $E[|X_t|^u] < \infty$, by using the same argument developed in Ling and Tong [2005], it is possible to show that there exists a positive finite constant,

say M, such that $\sup_{x\in\mathbb{R}} |x|^u f_X(x) < M$. It holds that:

$$
\begin{aligned}
&E^*\left[|E[X_t^u I(X_t \le r)] - E[X_t^u G_\alpha(X_t)]|\right] = |E[X_t^u I(X_t \le r)] - E[X_t^u G_\alpha(X_t)]| \\
&\le E[|X_t^u| \cdot |I(X_t \le r) - G_\alpha(X_t)|] \\
&= E[|X_t^u| \cdot |I(X_t \le r) - G_\alpha(X_t)| I(X_t \notin [L_{\alpha,\delta}, U_{\alpha,\delta}])] \\
&+ E[|X_t^u| \cdot |I(X_t \le r) - G_\alpha(X_t)| I(X_t \in [L_{\alpha,\delta}, U_{\alpha,\delta}])] \\
&\le \delta E[|X_t|^u] + M(U_{\alpha,\delta} - L_{\alpha,\delta}),
\end{aligned}
$$

where the last inequality follows from Eq. (40), Eq. (41) and Eq. (42). Hence, Eq. (45) can be made arbitrarily small in probability by choosing α and δ sufficiently small. A similar argument handles Eq. (46):

$$
\begin{aligned}
&E^*\left[\left|\frac{1}{n}\sum_{t=1}^n X_t^{*u} I(X_t^* \le r) - \frac{1}{n}\sum_{t=1}^n X_t^{*u} G_\alpha(X_t^*)\right|\right] \\
&\le E^*\left[\frac{1}{n}\sum_{t=1}^n |X_t^{*u}| \cdot |I(X_t^* \le r) - G_\alpha(X_t^*)|\right] \\
&= E^*\left[\frac{1}{n}\sum_{t=1}^n |X_t^{*u}| \cdot |I(X_t^* \le r) - G_\alpha(X_t^*)| I(X_t^* \notin [L_{\alpha,\delta}, U_{\alpha,\delta}])\right] \\
&+ E^*\left[\frac{1}{n}\sum_{t=1}^n |X_t^{*u}| \cdot |I(X_t^* \le r) - G_\alpha(X_t^*)| I(X_t^* \in [L_{\alpha,\delta}, U_{\alpha,\delta}])\right] \\
&\le \delta \frac{1}{n}\sum_{t=1}^n E^*[|X_t^{*u}|] + \mathcal{M}\frac{1}{n}\sum_{t=1}^n P^*(X_t^* \in [L_{\alpha,\delta}, U_{\alpha,\delta}]),
\end{aligned}
$$

where $\mathcal{M} = \max\{|L_{\alpha,\delta}|^u, |U_{\alpha,\delta}|^u, 1\}$. Lemma 13 implies that $n^{-1}\sum_{t=1}^n P^*(X_t^* \in [L_{\alpha,\delta}, U_{\alpha,\delta}]) \le n^{-1}$ with probability 1, hence Eq. (46) is $o_{p^*}(1)$. Lastly, in order to show that also Eq. (44) is $o_{p^*}(1)$, we use the following two expansions:

$$
\begin{aligned}
G_\alpha(X_t^*) &= G_\alpha(X_t) + g_\alpha(Y_t^*)(X_t^* - X_t); &(47)\\
G_\alpha(X_t) &= G_\alpha(q_{\alpha,\delta} + 2r) + g_\alpha(Y_t)(X_t - q_{\alpha,\delta} - 2r) &(48)
\end{aligned}
$$

where $q_{\alpha,\delta}$ is defined in Eq. (40), $Y_t^* = \lambda_{1,t} X_t^* + (1-\lambda_{1,t})X_t$ and $Y_t = \lambda_{2,t} X_t + (1-\lambda_{2,t})(q_\delta + 2r)$ for some $\lambda_{j,t}$ with $0 \le \lambda_{j,t} \le 1$ and $j = 1, 2$; moreover,

$$
g_\alpha(y) = \frac{\partial G_\alpha(y)}{\partial y} = \begin{cases} -\frac{\alpha}{\pi(\alpha^2 + (r-y)^2)} & \text{if } y \neq r \\ 0 & \text{if } y = r. \end{cases}
$$

Note that $g_\alpha(y) \xrightarrow[\alpha\to 0]{} 0$ for each y. Since the ergodicity of $\{X_t\}$ implies that

$$
\frac{1}{n}\sum_{t=1}^n X_t^u G_\alpha(X_t) \xrightarrow[n\to\infty]{p} E[X_t^u G_\alpha(X_t)],
$$

it suffices to prove that

$$P^*\left(\left|\frac{1}{n}\sum_{t=1}^{n} X_t^{*u} G_\alpha(X_t^*) - \frac{1}{n}\sum_{t=1}^{n} X_t^{u} G_\alpha(X_t)\right| > \eta/2\right) \xrightarrow[n\to\infty]{p} 0, \tag{49}$$

which can be achieved by using Markov's inequality. Indeed, by using Eq. (47), Eq. (48) and since $G_\alpha(q_\delta + 2r) = \delta$ we have

$$\begin{aligned}
&E^*\left[\left|\frac{1}{n}\sum_{t=1}^{n} X_t^{*u} G_\alpha(X_t^*) - \frac{1}{n}\sum_{t=1}^{n} X_t^{u} G_\alpha(X_t)\right|\right] \\
&\leq E^*\left[\left|\frac{1}{n}\sum_{t=1}^{n} (X_t^{*u} - X_t^u)\left\{\delta + g_\alpha(Y_t)(X_t - q_\delta - 2r)\right\}\right|\right] \\
&+ E^*\left[\left|\frac{1}{n}\sum_{t=1}^{n} X_t^{*u} g_\alpha(Y_t^*)(X_t^* - X_t)\right|\right]
\end{aligned}$$

which can be made arbitrarily small in probability by taking α and δ sufficiently small and this completes the proof.

Uniform convergence. By deploying arguments similar to Cavaliere et al. [2017], we show that for each $\eta > 0$

$$P^*\left(\sup_{r\in[r_L,r_U]} |\Delta_n^*(r)| > 2\eta\right) \xrightarrow[n\to\infty]{p} 0, \tag{50}$$

where $\Delta_n^*(r) = n^{-1}\sum_{t=1}^{n} X_t^{*u} I(X_t^* \leq r) - E[X_t^u I(X_t \leq r)]$. Since $[r_L, r_U]$ is a compact subset of $\mathbb{R}$, for any $c > 0$, there exists a finite coverage $\{[r_{i-1}, r_i]; i = 1, \ldots, m\}$, with m being a constant, such that $r_L = r_0 < r_1 < \ldots < r_{m-1} < r_m = r_U$ ad $r_i - r_{i-1} \leq c$, for each $i = 1, \ldots, m$. Therefore, it holds that

$$\sup_{r\in[r_L,r_U]} |\Delta_n^*(r)| \leq \max_{i=0,\ldots,m} |\Delta_n^*(r_i)| + \max_{i=1,\ldots,m} \sup_{r\in[r_{i-1},r_i]} |\Delta_n^*(r) - \Delta_n^*(r_{i-1})|$$

which implies:

$$\begin{aligned}
&P^*\left(\sup_{r\in[r_L,r_U]} |\Delta_n^*(r)| > 2\eta\right) \\
&\leq P^*\left(\max_{i=0,\ldots,m} |\Delta_n^*(r_i)| > \eta\right) + P^*\left(\max_{i=1,\ldots,m} \sup_{r\in[r_{i-1},r_i]} |\Delta_n^*(r) - \Delta_n^*(r_{i-1})| > \eta\right)
\end{aligned} \tag{51}$$

By combining Bonferroni's inequality, the pointwise convergence and the finiteness of m, we have that

$$P^*\left(\max_{i=1,\ldots,m} |\Delta_n^*(r_i)| > \eta\right) \leq \sum_{i=1}^{m} P^*\left(|\Delta_n^*(r_i)| > \eta\right) \xrightarrow[n\to\infty]{p} 0.$$

It remains to show that the second term of the RHS of Eq. (51) converges to zero in probability (in probability), which is the case because:

$$
\begin{aligned}
&E^*\left[\sup_{r\in[r_{i-1},r_i]}|\Delta_n^*(r)-\Delta_n^*(r_{i-1})|\right]\\
&\leq E^*\left[\sup_{r\in[r_{i-1},r_i]}\frac{1}{n}\sum_{t=1}^{n}|X_t^{*u}|I(r_{i-1}<X_t^*\leq r)+\sup_{r\in[r_{i-1},r_i]}E[|X_t|^uI(r_{i-1}<X_t\leq r)]\right]\\
&\leq E^*\left[\frac{1}{n}\sum_{t=1}^{n}|X_t^{*u}|I(r_{i-1}-c<X_t^*\leq r_{i-1}+c)+E[|X_t|^uI(r_{i-1}<X_t\leq r_i)]\right]\\
&\leq \frac{\mathcal{M}_1}{n}\sum_{t=1}^{n}P^*(r_{i-1}-c<X_t^*\leq r_{i-1}+c)+\mathcal{M}_2P(r_{i-1}<X_t\leq r),
\end{aligned}
$$

with $\mathcal{M}_1=\max\{|r_{i-1}-c|^u,|r_{i-1}+c|^u,1\}$ and $\mathcal{M}_2=\max\{|r_{i-1}|^u,|r_i|^u,1\}$. By combining Lemma 13 and Markov's inequality, the proof is completed since c can be chosen arbitrarily small.

PART 2. The proof follows via the same arguments used in 1. and, hence, it is omitted.

A.2 Proof of Proposition 7

By routine algebra it holds that $I^*_{n,22}(r)=I^*_{n,12}(r)=I^{*\intercal}_{n,21}(r)$ are $(p+1)\times(p+1)$ symmetric matrices whose $(i+1,j+1)$th element is

$$
\begin{aligned}
&\frac{1}{\sigma^{*2}}\sum_{t=1}^{n}I(X^*_{t-d}\leq r), && \text{if } i=0, j=0\\
&\frac{1}{\sigma^{*2}}\sum_{t=1}^{n}X^*_{t-j}I(X^*_{t-d}\leq r), && \text{if } i=0, j\neq 0\\
&\frac{1}{\sigma^{*2}}\sum_{t=1}^{n}X^*_{t-i}X^*_{t-j}I(X^*_{t-d}\leq r), && \text{if } i\neq 0, j\neq 0
\end{aligned}
$$

and $I^*_{n,11}=I^*_{n,22}(\infty)$. The results readily follows by combining Proposition 5 with $u=0,1,2$ for point 1 and standard results of bootstrap asymptotic analysis.

A.3 Proof of Proposition 8

The proof is based upon verifying the following two equalities:

$$
\sqrt{n}(\tilde{\boldsymbol{\phi}}-\tilde{\boldsymbol{\phi}}^*)=-\left(\frac{I^*_{n,11}}{n}\right)^{-1}\frac{1}{\sqrt{n}}\frac{\partial\ell^*_n}{\partial\boldsymbol{\phi}} \tag{52}
$$

$$
\frac{1}{\sqrt{n}}\frac{\partial\tilde{\ell}^*_n}{\partial\boldsymbol{\Psi}}(r)=\frac{1}{\sqrt{n}}\frac{\partial\ell^*_n}{\partial\boldsymbol{\Psi}}(r)+\frac{I^*_{n,21}(r)}{n}\sqrt{n}(\tilde{\boldsymbol{\phi}}-\tilde{\boldsymbol{\phi}}^*), \tag{53}
$$

where $\frac{\partial \ell_n^*}{\partial \boldsymbol{\phi}}$, $\frac{\partial \ell_n^*}{\partial \boldsymbol{\Psi}}(r)$ and $\frac{\partial \tilde{\ell}_n^*}{\partial \boldsymbol{\Psi}}(r)$ are defined in Eq. (19) and (21) whereas $I_{n,11}^*$ and $I_{n,21}^*(r)$ in Eq. (23). As previously stated we use $\boldsymbol{\phi}$ as to refer to a generic parameter and let $\frac{\partial \ell_n^*}{\partial \boldsymbol{\phi}}(\boldsymbol{\phi})$ be the partial derivative of the bootstrap log-likelihood computed under the null hypothesis, i.e.:

$$\frac{\partial \ell_n^*}{\partial \boldsymbol{\phi}}(\boldsymbol{\phi}) = \left.\frac{\partial \ell_n^*(\boldsymbol{\eta}, r)}{\partial \boldsymbol{\phi}}\right|_{\boldsymbol{\Psi}=\mathbf{0},\sigma^2=\tilde{\sigma}^2}.$$

Next, we derive two first order Taylor expansions of the function $\frac{\partial \ell_n^*}{\partial \boldsymbol{\phi}}(\boldsymbol{\phi})$: one at the true bootstrap value $\tilde{\boldsymbol{\phi}}$ ad the other at the bootstrap MLE $\tilde{\boldsymbol{\phi}}^*$. Note that, since the νth partial derivatives of $\ell_n^*(\boldsymbol{\eta}, r)$ are zero for $\nu > 2$, the Taylor expansion of $\frac{\partial \ell_n^*}{\partial \boldsymbol{\phi}}(\boldsymbol{\phi})$ coincides with its first-order Taylor polynomial; moreover, the Jacobian matrix of $\frac{\partial \ell^*}{\partial \boldsymbol{\phi}}(\boldsymbol{\phi})$ is $-I_{n,11}^*$, defined in Eq. (23), which does not depend on $\boldsymbol{\phi}$. Hence it results that:

$$\frac{\partial \ell_n^*}{\partial \boldsymbol{\phi}}(\boldsymbol{\phi}) = \frac{\partial {\ell_n}^*}{\partial \boldsymbol{\phi}} - I_{n,11}(\boldsymbol{\phi} - \tilde{\boldsymbol{\phi}}), \tag{54}$$

$$\frac{\partial \ell_n^*}{\partial \boldsymbol{\phi}}(\boldsymbol{\phi}) = \frac{\partial \tilde{\ell}_n^*}{\partial \boldsymbol{\phi}} - I_{n,11}(\boldsymbol{\phi} - \tilde{\boldsymbol{\phi}}^*) \tag{55}$$

with $\frac{\partial \tilde{\ell}_n^*}{\partial \boldsymbol{\phi}}$ and $\frac{\partial \ell_n^*}{\partial \boldsymbol{\phi}}$ being defined in Eq. (19). By subtracting Eq. (55) from Eq. (54) and dividing by $\sqrt{n}$, we get

$$\frac{1}{\sqrt{n}}\frac{\partial \tilde{\ell}_n^*}{\partial \boldsymbol{\phi}} = \frac{1}{\sqrt{n}}\frac{\partial \ell_n^*}{\partial \boldsymbol{\phi}} + \frac{I_{n,11}^*}{n}\sqrt{n}(\tilde{\boldsymbol{\phi}} - \tilde{\boldsymbol{\phi}}^*). \tag{56}$$

Since $\tilde{\boldsymbol{\phi}}^*$ is the bootstrap MLE obtained under the null hypothesis, $\frac{\partial \tilde{\ell}_n^*}{\partial \boldsymbol{\phi}} = 0$ thence Eq. (56) implies

$$\frac{I_{n,11}^*}{n}\sqrt{n}(\tilde{\boldsymbol{\phi}}^* - \tilde{\boldsymbol{\phi}}) = -\frac{1}{\sqrt{n}}\frac{\partial \ell_n^*}{\partial \boldsymbol{\phi}}.$$

and hence Eq. (52) follows. We prove Eq. (53) componentwise. We detail below the argument only for the first component since it can be easily adapted to the other ones. Therefore, we show that:

$$\begin{aligned}\frac{1}{\sqrt{n}}\frac{\partial \tilde{\ell}_n^*}{\partial \Psi_0}(r) = \frac{1}{\sqrt{n}}\frac{\partial \ell_n^*}{\partial \Psi_0}(r) + \sqrt{n}(\tilde{\phi}_0 - \tilde{\phi}_0^*)\frac{1}{n}\sum_{t=1}^{n} I(X_{t-d}^* \le r) \\ + \sum_{i=1}^{p}\sqrt{n}(\tilde{\phi}_i - \tilde{\phi}_i^*)\frac{1}{n}\sum_{t=1}^{n} X_{t-i}^* I(X_{t-d}^* \le r).\end{aligned} \tag{57}$$

Let $\tilde{\varepsilon}_t^*$ be the residuals obtained from the ML fit upon the bootstrap sample $\{X_t^*, t = 1 \dots, n\}$, i.e.:

$$\tilde{\varepsilon}_t^* = X_t^* - \tilde{\phi}_0^* - \sum_{i=1}^{p}\tilde{\phi}_i^* X_{t-i}^* = (\tilde{\phi}_0 - \tilde{\phi}_0^*) + \sum_{i=1}^{p}(\tilde{\phi}_i - \tilde{\phi}_i^*)X_{t-i} + \varepsilon_t^*.$$

By definition:

$$\varepsilon_t^* - \tilde{\varepsilon}_t^* = (\tilde{\phi}_0^* - \tilde{\phi}_0) + \sum_{i=1}^{p} (\tilde{\phi}_i^* - \tilde{\phi}_i) X_{t-i}^*. \tag{58}$$

Note that $\varepsilon_t^*(\boldsymbol{\eta}, r)$, defined in Eq. (18), does not depend on σ^2 and $\tilde{\varepsilon}_t^*$ and ε_t^* correspond to the function $\varepsilon_t^*(\boldsymbol{\eta}, r)$ evaluated at $\boldsymbol{\phi} = \tilde{\boldsymbol{\phi}}^*$, $\boldsymbol{\Psi} = \mathbf{0}$ and $\boldsymbol{\phi} = \tilde{\boldsymbol{\phi}}$, $\boldsymbol{\Psi} = \mathbf{0}$, respectively. Consider the partial derivatives of the function $\varepsilon_t^*(\boldsymbol{\eta}, r)$ and denote:

$$D_{\Psi_0 t}^*(r) = \left.\frac{\partial \varepsilon_t^*(\boldsymbol{\eta}, r)}{\partial \Psi_0}\right|_{\boldsymbol{\phi}=\tilde{\boldsymbol{\phi}}, \boldsymbol{\Psi}=\mathbf{0}}, \quad \tilde{D}_{\Psi_0 t}^*(r) = \left.\frac{\partial \varepsilon_t^*(\boldsymbol{\eta}, r)}{\partial \Psi_0}\right|_{\boldsymbol{\phi}=\tilde{\boldsymbol{\phi}}^*, \boldsymbol{\Psi}=\mathbf{0}}.$$

Note that:

$$D_{\Psi_0 t}^*(r) = \tilde{D}_{\Psi_0 t}^*(r) = -I(X_{t-d}^* \leq r)$$

therefore, we get:

$$\begin{aligned}
\frac{1}{\sqrt{n}} \frac{\partial \tilde{\ell}_n^*}{\partial \Psi_0}(r) &= -\frac{1}{\sqrt{n}} \sum_{t=1}^{n} \tilde{\varepsilon}_t^* \tilde{D}_{\Psi_0 t}^*(r) = -\frac{1}{\sqrt{n}} \sum_{t=1}^{n} \tilde{\varepsilon}_t^* D_{\Psi_0 t}^*(r) \\
&= -\frac{1}{\sqrt{n}} \sum_{t=1}^{n} \tilde{\varepsilon}_t^* D_{\Psi_0 t}^*(r) - \frac{1}{\sqrt{n}} \sum_{t=1}^{n} \varepsilon_t^* D_{\Psi_0 t}^*(r) + \frac{1}{\sqrt{n}} \sum_{t=1}^{n} \varepsilon_t^* D_{\Psi_0 t}^*(r) \\
&= \frac{1}{\sqrt{n}} \frac{\partial \ell_n^*}{\partial \Psi_0}(r) + \frac{1}{\sqrt{n}} \sum_{t=1}^{n} (\varepsilon_t^* - \tilde{\varepsilon}_t^*) D_{\Psi_0 t}^*(r).
\end{aligned}$$

The expression of $(\varepsilon_t^* - \tilde{\varepsilon}_t^*)$ in Eq. (58) implies that

$$\begin{aligned}
\frac{1}{\sqrt{n}} \frac{\partial \tilde{\ell}_n^*}{\partial \Psi_0}(r) &= \frac{1}{\sqrt{n}} \frac{\partial \ell_n^*}{\partial \Psi_0}(r) + \frac{1}{\sqrt{n}} \sum_{t=1}^{n} \left\{ (\tilde{\phi}_0^* - \tilde{\phi}_0) + \sum_{i=1}^{p} (\tilde{\phi}_i^* - \tilde{\phi}_i) X_{t-i}^* \right\} D_{\Psi_0 t}^*(r) \\
&= \frac{1}{\sqrt{n}} \frac{\partial \ell_n^*}{\partial \Psi_0}(r) + \frac{1}{\sqrt{n}} \sum_{t=1}^{n} \left\{ (\tilde{\phi}_0^* - \tilde{\phi}_0) + \sum_{i=1}^{p} (\tilde{\phi}_i^* - \tilde{\phi}_i) X_{t-i}^* \right\} \{-I(X_{t-d}^* \leq r)\} \\
&= \frac{1}{\sqrt{n}} \frac{\partial \ell_n^*}{\partial \Psi_0}(r) + \sqrt{n}(\tilde{\phi}_0 - \tilde{\phi}_0^*) \frac{1}{n} \sum_{t=1}^{n} I(X_{t-d}^* \leq r) \\
&+ \sum_{i=1}^{p} \sqrt{n}(\tilde{\phi}_i - \tilde{\phi}_i^*) \frac{1}{n} \sum_{t=1}^{n} X_{t-i}^* I(X_{t-d}^* \leq r)
\end{aligned}$$

and this completes the proof.

A.4 Proof of Proposition 10

The proof follows by using the same arguments developed to prove Proposition 5 by deploying that $\{\varepsilon_t^*\}$ is a sequence of independent random variables in both bootstrap schemes.

A.5 Proof of Proposition 11

We start by proving the proposition for the i.i.d. bootstrap case. Since $-\sum_{t=1}^{n} \varepsilon_t^* D_{t-1}^*(r)$, with $D_t^*(r)$ being defined in Eq. (22), forms a sequence of martingale difference arrays with respect to the filtration $\mathcal{F}_{t-1}^* := \sigma\{X_{t-1}^*, X_{t-2}^*, \dots\}$, the result holds upon proving, uniformly on r, the following two conditions:

$$\frac{1}{n}\sum_{t=1}^{n} E^*\left[\varepsilon_t^{*2}(D_{t-1}^*(r))(D_{t-1}^*(r))^\intercal | \mathcal{F}_{t-1}^*\right] \xrightarrow[n\to\infty]{p^*}{}_p \; E\left[\varepsilon_t^2 D_{t-1}(r) D_{t-1}^\intercal(r)\right]; \quad (59)$$

$$\frac{1}{n}\sum_{t=1}^{n} E^*\left[\varepsilon_t^{*2}\Lambda_{t-1}^{*2}(r) I\left(\left|\varepsilon_t^*\Lambda_{t-1}^*(r)\right| > \eta\sqrt{n}\right) | \mathcal{F}_{t-1}^*\right] \xrightarrow[n\to\infty]{p^*}{}_p \; 0, \quad (60)$$

with $\Lambda_t^*(r) := (\lambda_1, \dots, \lambda_{2(p+1)}) \cdot D_t^*(r)$, with λ_i, $i = 1, \dots, 2(p+1)$, being real numbers. In order to prove Eq. (59) note that the independence between ε_t^* and X_{t-j}^*, $j \geq 1$ implies that

$$\frac{1}{n}\sum_{t=1}^{n} E^*\left[\varepsilon_t^{*2}(D_{t-1}^*(r))(D_{t-1}^*(r))^\intercal | \mathcal{F}_{t-1}^*\right] = E^*\left[\varepsilon_t^{*2}\right] \frac{1}{n}\sum_{t=1}^{n} D_{t-1}^*(r)(D_{t-1}^*(r))^\intercal,$$

which converges in probability (in probability) to $\sigma^2 I_\infty(r)$ uniformly on r by Lemma 14 (Supplementary Material) and Proposition 7. As for Eq. (60) first observe that, by using Jensen's inequality and Proposition 5, $n^{-1}\sum_{t=1}^{n}\Lambda_{t-1}^{*2}(r)$ is bounded by

$$\frac{2(p+1)}{n}\sum_{t=1}^{n}\left[\left(\lambda_1^2 + \lambda_{p+2}^2\right) + \sum_{i=2}^{p+1}\left(\lambda_i^2 + \lambda_{i+p+1}^2\right) X_{t-i+1}^{*2}\right] = O_{p^*}(1), \quad (61)$$

whereas $n^{-1}\sum_{t=1}^{n}\Lambda_{t-1}^{*4}(r)$ is bounded by

$$\frac{8(p+1)^3}{n}\sum_{t=1}^{n}\left[\left(\lambda_1^4 + \lambda_{p+2}^4\right) + \sum_{i=2}^{p+1}\left(\lambda_i^4 + \lambda_{i+p+1}^4\right) X_{t-i+1}^{*4}\right] = O_{p^*}(1). \quad (62)$$

Now, since $|xy| \leq x^2 + y^2$, it follows that

$$\begin{aligned}
&\frac{1}{n}\sum_{t=1}^{n} E^*\left[\varepsilon_t^{*2}\Lambda_{t-1}^{*2}(r) I\left(\left|\varepsilon_t^*\Lambda_{t-1}^*(r)\right| > \eta\sqrt{n}\right) | \mathcal{F}_{t-1}^*\right] \\
&\leq \frac{1}{n}\sum_{t=1}^{n} E^*\left[\varepsilon_t^{*2}\Lambda_{t-1}^{*2}(r) I\left(\Lambda_{t-1}^{*2}(r) > 2^{-1}\eta\sqrt{n}\right) | \mathcal{F}_{t-1}^*\right] \\
&+ \frac{1}{n}\sum_{t=1}^{n} E^*\left[\varepsilon_t^{*2}\Lambda_{t-1}^{*2}(r) I\left(\varepsilon_t^{*2} > 2^{-1}\eta\sqrt{n}\right) | \mathcal{F}_{t-1}^*\right] \\
&\leq \frac{2}{\eta\sqrt{n}}\left\{\frac{1}{n}\sum_{t=1}^{n}\Lambda_{t-1}^{*4}(r) E^*[\varepsilon_t^{*2}] + \frac{1}{n}\sum_{t=1}^{n}\Lambda_{t-1}^{*2}(r) E^*[\varepsilon_t^{*4}]\right\}
\end{aligned}$$

which is $o_{p^*}(1)$ by combining Eq. (61), Eq. (62) and Lemma 14 (Supplementary Material).

As for the wild bootstrap case, by exploiting the same arguments of Proposition 10 it can be shown that, for every i, j, d

$$\sup_{r\in[r_L,r_U]}\left|\frac{1}{n}\sum_{t=1}^{n}\varepsilon_t^2 X^*_{t-i}I(X^*_{t-d}\leq r)-E[\varepsilon_t^2 X_{t-i}I(X_{t-d}\leq r)]\right|\xrightarrow[n\to\infty]{p^*}{}_p\ 0, \tag{63}$$

and

$$\sup_{r\in[r_L,r_U]}\left|\frac{1}{n}\sum_{t=1}^{n}\varepsilon_t^2 X^*_{t-i}X^*_{t-j}I(X^*_{t-d}\leq r)-E[\varepsilon_t^2 X_{t-i}X_{t-j}I(X_{t-d}\leq r)]\right|\xrightarrow[n\to\infty]{p^*}{}_p\ 0. \tag{64}$$

Hence, Eq. (59) holds for the wild bootstrap case, since

$$\begin{aligned}&\frac{1}{n}\sum_{t=1}^{n}E^*\left[\varepsilon_t^{*2}(D^*_{t-1}(r))(D^*_{t-1}(r))^\intercal|\mathcal{F}^*_{t-1}\right]=\frac{1}{n}\sum_{t=1}^{n}\tilde{\varepsilon}_t^2(D^*_{t-1}(r))(D^*_{t-1}(r))^\intercal\\&=\frac{1}{n}\sum_{t=1}^{n}\varepsilon_t^2(D^*_{t-1}(r))(D^*_{t-1}(r))^\intercal+o_p(1)\xrightarrow[n\to\infty]{p^*}{}_p\ E\left[\varepsilon_t^2 D_{t-1}(r)D^\intercal_{t-1}(r)\right].\end{aligned}$$

We now prove the following (unconditional) Lindeberg condition:

$$\frac{1}{n}\sum_{t=1}^{n}E^*\left[\varepsilon_t^{*2}\Lambda^{*2}_{t-1}(r)I\left(\left|\varepsilon_t^*\Lambda^*_{t-1}(r)\right|>\eta\sqrt{n}\right)\right]\xrightarrow[n\to\infty]{p^*}{}_p\ 0. \tag{65}$$

Routine algebra implies

$$\begin{aligned}&\frac{1}{n}\sum_{t=1}^{n}E^*\left[\varepsilon_t^{*2}\Lambda^{*2}_{t-1}(r)I\left(\left|\varepsilon_t^*\Lambda^*_{t-1}(r)\right|>\eta\sqrt{n}\right)\right]\\&\leq\frac{1}{n}\sum_{t=1}^{n}E^*\left[\varepsilon_t^{*2}\Lambda^{*2}_{t-1}(r)I\left(\Lambda^{*2}_{t-1}(r)>2^{-1}\eta\sqrt{n}\right)\right]\\&+\frac{1}{n}\sum_{t=1}^{n}E^*\left[\varepsilon_t^{*2}\Lambda^{*2}_{t-1}(r)I\left(\varepsilon_t^{*2}>2^{-1}\eta\sqrt{n}\right)\right]\\&\leq\frac{2}{\eta\sqrt{n}}E^*\left[\frac{1}{n}\sum_{t=1}^{n}\varepsilon_t^{*2}\Lambda^{*4}_{t-1}(r)+\frac{1}{n}\sum_{t=1}^{n}\varepsilon_t^{*4}\Lambda^{*2}_{t-1}(r)\right]\end{aligned}$$

which is also $o_{p^*}(1)$, and this completes the proof.

A.6 Proof of Theorem 12

In view of Proposition 11 and Theorem 18.14, p. 261 of van der Vaart [1998], it suffices to prove the stochastic equicontinuity of $\nabla^*_n(r)=-\sum_{t=1}^{n}\varepsilon_t^* D^*_{t-1}(r)$,

where $D^*_{t-1}(r)$ is defined in Eq. (22). The envelope of $\varepsilon^*_t D^*_{t-1}(r)$ is $\mathcal{L}^2$ integrable in probability:

$$\frac{1}{n}\sum_{t=1}^{n} E^*\left[\sup_{r\in[r_L,r_U]} \|\varepsilon^*_t x^*_{t-1}(r)\|^2\right] = \frac{1}{n}\sum_{t=1}^{n} E^*\left[\sup_{r\in[r_L,r_U]} (\varepsilon^*_t x^*_{t-1}(r))^\intercal(\varepsilon^*_t x^*_{t-1}(r))\right]$$
$$= \frac{1}{n}\sum_{t=1}^{n} E^*\left[\sup_{r\in[r_L,r_U]} \varepsilon^{*2}_t\left\{1+\sum_{i=1}^{p} X^{*2}_{t-i} + I(X^*_{t-d}\le r) + \sum_{i=1}^{p} X^{*2}_{t-i} I(X^*_{t-d}\le r)\right\}^2\right]$$
$$\le \frac{2}{n}\sum_{t=1}^{n} E^*\left[\varepsilon^{*2}_t\left\{1+\sum_{i=1}^{p} X^{*2}_{t-i}\right\}\right] = O_{p^*}(1).$$

Define the norms:

$$\rho^*_n(r_1,r_2) = \left\|\frac{1}{\sqrt{n}}\left(\nabla^*_n(r_2)-\nabla^*_n(r_1)\right)\right\|_2 \text{ and } \rho(r_1,r_2) \quad = \|\varepsilon_t D_{t-1}(r_2) - \varepsilon_t D_{t-1}(r_1)\|_2\,,$$

where, in analogy with Eq. (22), $D_t(r)$ is the first-order derivative of the function $\varepsilon_t(\boldsymbol{\eta}, r)$ defined in Eq. (3), i.e.:

$$D_t(r) = (-1, -X_t, \ldots, -X_{t-p+1},$$
$$-I(X_{t-d+1}\le r), -X_t I(X_{t-d+1}\le r), \ldots, -X_{t-p+1} I(X_{t-d+1}\le r))^\intercal\,.$$

It holds that

$$\rho^{*2}_n(r_1,r_2) = E^*\left\|\frac{1}{\sqrt{n}}\left(\frac{\partial \ell^*}{\partial \boldsymbol{\eta}}(r_2) - \frac{\partial \ell^*}{\partial \boldsymbol{\eta}}(r_1)\right)\right\|^2$$
$$= E^*\left[\frac{1}{n}\sum_{t=1}^{n}\varepsilon^{*2}_t\left(I(r_1 < X^*_{t-d}\le r_2) + \sum_{i=1}^{p} X^{*2}_{t-i} I(r_1 < X^*_{t-d}\le r_2)\right)\right].$$

Proposition 10 implies that $\rho^*_n(r_1,r_2)$ converges uniformly to

$$\left\{E\left[\varepsilon^2_t I(r_1 < X_{t-d}\le r_2)\right] + \sum_{i=1}^{p} E[\varepsilon^2_t X^2_{t-i} I(r_1 < X_{t-d}\le r_2)]\right\} = \rho^2(r_1,r_2).$$

Thence the same argument of Theorem 2 of Hansen [1996] holds, and this completes the proof.

Supplement for: The validity of bootstrap testing in the threshold framework

Simone Giannerini, Greta Goracci, Anders Rahbek

Abstract

This Supplement has 3 sections. In Section A we present auxiliary technical lemmas used in the proofs. Section B contains supplementary results from the simulation study. Section C presents supplementary results on the analysis of larvae population dynamics under the effect of warming.

Abstract

This Supplement has 2 sections. In Section A we present auxiliary technical lemmas used in the proofs. Section B contains supplementary results from the simulation study. Section C presents supplementary results on the analysis of the Greenland ice sheet mass balance.

A Auxiliary Lemmas

Lemma 13. Let $\{X_t^*, t = 1, \dots, n\}$ be defined in Eq. (16) and assume $b_1(c)$ and $b_2(c)$ to be two continuous functions in c such that

$$\lim_{c\to 0} b_1(c) = \lim_{c\to 0} b_2(c) = \mathrm{C},$$

with C being a real number. Then, for each $\gamma > 0$ we can choose c sufficiently small such that

$$P^*(b_1(c) \le X_t^* \le b_2(c)) \le \frac{1}{n} + \gamma \tag{66}$$

with probability one.

Proof. From the definition of limit, for each $\gamma > 0$ we can choose c sufficiently small such that

$$P^*(b_1(c) \le X_t^* \le b_2(c)) \le P^*(X_t^* = \mathrm{C}) + \gamma;$$

hence it remains to show that $P^*(X_t^* = \mathrm{C}) \le 1/n$ in probability. Define $\mathcal{A}_t^*$ to be the set of values that X_t^* can assume conditionally to the data. By using the

fact that: (i) for any real number $\kappa \in \mathbb{R}$, $P^*(\varepsilon_t^* = \kappa) \leq 1/n$ with probability one and (ii) $\sum_{a \in \mathcal{A}_s^*} P^*(X_s^* = a) = 1$, for any integer s, it holds that

$$\begin{aligned} P^*(X_t^* = \mathrm{C}) &= \sum_{a \in \mathcal{A}_{t-1}^*} P^*(X_t^* = \mathrm{C} | X_{t-1}^* = a) P^*(X_{t-1}^* = a) \\ &= \sum_{a \in \mathcal{A}_{t-1}^*} P^*(\varepsilon_t^* = \mathrm{C} - \tilde{\phi}_0 - \tilde{\phi}_1 a) P^*(X_{t-1}^* = a) \\ &\leq \frac{1}{n} \sum_{a \in \mathcal{A}_{t-1}^*} P^*(X_{t-1}^* = a) = \frac{1}{n} \end{aligned}$$

and the proof is completed. □

Lemma 14. **(LLN)** Assume that Assumption 4 holds and the following condition is satisfied:

(i) $\{\varepsilon_t^*\}$ is defined as in (B.iid) and Assumption (A.iid) is fulfilled;

It holds that:

$$E^*[\varepsilon_t^{*2}] \xrightarrow[n \to \infty]{p} \sigma^2 \quad \text{and} \quad E^*[\varepsilon_t^{*4}] \xrightarrow[n \to \infty]{p} \kappa.$$

Proof. Since

$$\tilde{\varepsilon}_t = (\phi_{0,0} - \tilde{\phi}_0) + \sum_{i=1}^{p} (\phi_{i,0} - \tilde{\phi}_i) X_{t-i} + \varepsilon_t$$

and $(\tilde{\boldsymbol{\phi}} - \boldsymbol{\phi}_0) = O_p(n^{-1/2})$, it follows that

$$\bar{\tilde{\varepsilon}} := \frac{1}{n} \sum_{t=1}^{n} \tilde{\varepsilon}_t = \frac{1}{n} \sum_{t=1}^{n} \left[(\phi_{0,0} - \tilde{\phi}_0) + \sum_{i=1}^{p} (\phi_{i,0} - \tilde{\phi}_i) X_{t-i} + \varepsilon_t \right]$$

converges in probability to zero. Similarly, by routine algebra, it is possible to show that $E^*[\varepsilon_t^{*2}] = \frac{1}{n} \sum_{t=1}^{n} (\tilde{\varepsilon}_t - \bar{\tilde{\varepsilon}})^2$ and $E^*[\varepsilon_t^{*4}] = \frac{1}{n} \sum_{t=1}^{n} (\tilde{\varepsilon}_t - \bar{\tilde{\varepsilon}})^4$ converge in probability to σ^2 and κ respectively. □

B Supplementary Monte Carlo results

In Section B.1 we study the size of the tests when the DGP is close to non-stationarity, whereas Section B.2 contains additional results on the power of the tests, covering the case of unit root regimes and white noise regimes.

B.1 Near non-stationarity and heteroskedasticity

We study the size of the tests when the DGP is near to non-stationarity. We simulate from the AR(1)-GARCH(1,1) of Eq. 67

$$
\begin{aligned}
X_t &= \phi_1 X_{t-1} + \varepsilon_t \\
\varepsilon_t &= \sigma_t z_t \\
\sigma_t^2 &= 1 + a_1 \varepsilon_{t-1}^2 + b_1 \sigma_{t-1}^2
\end{aligned} \tag{67}
$$

where $z_t \sim$ i.i.d.$N(0,1)$. We choose $\phi_1 = \pm 0.95, \pm 0.99$ and combine these with the following parameters for the GARCH specification: $(a_1, b_1) = (0.00, 0.00)$ (case N), $(a_1, b_1) = (0.04, 0.95)$ (case A), $(0.3, 0.0)$ (case B), $(0.4, 0.4)$ (case C). The empirical size (rejection percentages) is presented in Tables 1, 2 and 3 for $n = 50, 100, 200$, respectively. The four tests appear well behaved and the nearly non-stationary DGP has negligible influence. Indeed, the results are consistent with those of Figure 1 and Figure 3 of the main article: in case of GARCH innovations the wild bootstrap test sLMw has a correct size for small n whereas both sLMa and sLMi tests are oversized and the bias increases with sample size. Hansen's sLMh test is undersized and approaches the nominal size from below as the sample size increases.

	ϕ_1	a_1	b_1	sLMa	sLMi	sLMw	sLMh
N	-0.99	0.00	0.00	5.3	5.6	5.8	1.9
	-0.95	0.00	0.00	3.9	5.0	4.9	2.2
	0.95	0.00	0.00	3.3	5.6	5.5	1.8
	0.99	0.00	0.00	3.8	6.1	5.9	1.8
A	-0.99	0.04	0.95	5.2	5.6	5.4	1.8
	-0.95	0.04	0.95	3.8	4.7	4.8	2.0
	0.95	0.04	0.95	3.4	5.7	5.6	2.0
	0.99	0.04	0.95	3.7	5.9	5.9	1.8
B	-0.99	0.30	0.00	6.5	6.7	5.3	1.5
	-0.95	0.30	0.00	5.3	6.6	5.4	2.2
	0.95	0.30	0.00	5.3	8.2	6.1	1.9
	0.99	0.30	0.00	6.6	9.2	7.2	2.1
C	-0.99	0.40	0.40	9.4	9.2	6.4	1.7
	-0.95	0.40	0.40	6.5	7.8	5.8	2.4
	0.95	0.40	0.40	6.8	10.0	7.1	2.1
	0.99	0.40	0.40	9.5	12.3	8.4	2.2

Table 1: Empirical size (in percentage) for a nearly-non stationary AR(1) with GARCH innovations at nominal level $\alpha = 5\%$. n =50

	ϕ_1	a_1	b_1	sLMa	sLMi	sLMw	sLMh
N	-0.99	0.00	0.00	6.3	5.2	5.3	3.4
	-0.95	0.00	0.00	4.8	5.0	4.9	3.7
	0.95	0.00	0.00	4.1	5.1	5.0	3.1
	0.99	0.00	0.00	5.5	6.0	5.9	3.1
A	-0.99	0.04	0.95	6.9	5.7	5.2	3.5
	-0.95	0.04	0.95	4.7	5.0	4.8	3.6
	0.95	0.04	0.95	4.6	5.4	5.3	3.2
	0.99	0.04	0.95	5.6	6.2	5.9	3.3
B	-0.99	0.30	0.00	8.0	6.7	5.4	3.5
	-0.95	0.30	0.00	5.8	6.0	4.9	3.5
	0.95	0.30	0.00	6.6	7.8	5.7	3.5
	0.99	0.30	0.00	8.9	9.3	6.8	3.8
C	-0.99	0.40	0.40	11.0	8.8	5.6	3.2
	-0.95	0.40	0.40	7.9	8.0	5.1	4.1
	0.95	0.40	0.40	8.8	10.2	6.2	3.6
	0.99	0.40	0.40	12.2	12.7	7.7	3.8

Table 2: Empirical size (in percentage) for a nearly-non stationary AR(1) with GARCH innovations at nominal level $\alpha = 5\%$. n =100

	ϕ_1	a_1	b_1	sLMa	sLMi	sLMw	sLMh
N	-0.99	0.00	0.00	6.9	4.8	4.9	4.6
	-0.95	0.00	0.00	5.6	5.2	5.2	4.6
	0.95	0.00	0.00	5.4	5.2	5.3	4.4
	0.99	0.00	0.00	6.6	5.7	5.8	4.4
A	-0.99	0.04	0.95	7.3	5.4	4.9	4.6
	-0.95	0.04	0.95	5.9	5.4	5.2	4.9
	0.95	0.04	0.95	5.9	5.8	5.4	4.3
	0.99	0.04	0.95	7.4	6.5	5.8	4.9
B	-0.99	0.30	0.00	8.2	6.3	5.2	4.7
	-0.95	0.30	0.00	6.7	6.4	5.1	4.6
	0.95	0.30	0.00	7.1	6.9	5.1	4.5
	0.99	0.30	0.00	8.4	7.4	5.4	4.6
C	-0.99	0.40	0.40	11.3	8.8	5.1	4.4
	-0.95	0.40	0.40	9.5	9.2	4.8	4.6
	0.95	0.40	0.40	10.0	10.1	5.4	4.3
	0.99	0.40	0.40	13.1	12.0	6.6	5.1

Table 3: Empirical size (in percentage) for a nearly-non stationary AR(1) with GARCH innovations at nominal level $\alpha = 5\%$. n =200

B.2 Power of the tests: further results

In this section we present further results regarding the empirical power of the tests, also covering the cases of unit root regimes and white noise regimes. We include 4 additional models (M3–M6), which are combined with the vector Ψ as to obtain the 16 combinations of parameters presented in Table 4. As before, for each of the models M3–M6, the parameters of the lower regime deviate from those of the upper regime due to $|\Psi|$, which represents the distance from H_0. Hence, the first rows correspond to the null hypothesis ($\Psi = 0.0$). The 16 parameters are meant to explore the range of dynamic features of threshold models and all of them lie in the stationarity region: in model M3, (M4) the null hypothesis entails a very weak positive (negative) linear dependence, i.e. $\phi_{1,1} = \phi_{2,1} = 0.1$ (-0.1) and approaches a unit root in the lower regime, i.e. $\phi_{1,1} = 1$ (-1) as $|\Psi|$ increases. Conversely, model M5 (M6) specifies a quasi-unit root in H_0, i.e. $\phi_{1,1} = \phi_{2,1} = 0.9$ (-0.9) and becomes a local white noise ($\phi_{1,1} = 0$) for $|\Psi| = 0.9$. The results for models M3 and M4 are shown in Table 5, whereas those for models M5, M6 are presented in Table 6. As a

		lower		upper	
	Ψ	$\phi_{1,0}$	$\phi_{1,1}$	$\phi_{2,0}$	$\phi_{2,1}$
M3	0.0	-0.3	0.1	-0.3	0.1
	0.3	0.0	0.4	-0.3	0.1
	0.6	0.3	0.7	-0.3	0.1
	0.9	0.6	1.0	-0.3	0.1
M4	0.0	0.3	-0.1	0.3	-0.1
	-0.3	0.0	-0.4	0.3	-0.1
	-0.6	-0.3	-0.7	0.3	-0.1
	-0.9	-0.6	-1.0	0.3	-0.1
M5	0.0	-0.5	0.9	-0.5	0.9
	-0.3	-0.8	0.6	-0.5	0.9
	-0.6	-1.1	0.3	-0.5	0.9
	-0.9	-1.4	0.0	-0.5	0.9
M6	0.0	0.5	-0.9	0.5	-0.9
	0.3	0.8	-0.6	0.5	-0.9
	0.6	1.1	-0.3	0.5	-0.9
	0.9	1.4	0.0	0.5	-0.9

Table 4: 16 combinations of TAR parameters used in the simulations. Ψ indicates the departure from H_0 so that, for each of the models M3–M6, the first row ($\Psi = 0$) corresponds to the null hypothesis.

general comment to both tables: the results confirm that both the asymptotic test sLMa and Hansen's test sLMh are undersized for sample sizes $n = 50, 100$ and this is associated to their inferior power. The bootstrap tests sLMi and

sLMw have correct size also for small sample sizes. Moreover, all the tests seem to have power as $|\Psi|$ increases, irrespective of the presence of either a local unit-root or a local white noise under H_1. Specifically, all the tests present higher power for M3 and M6 than for M4 and M5, respectively and this indicates that the power of the tests depends non trivially upon the combination of the slopes and intercepts of the two regimes.

		M3					M4			
n	Ψ	sLMa	sLMi	sLMw	sLMh	Ψ	sLMa	sLMi	sLMw	sLMh
50	0.0	2.9	4.8	4.9	2.9	0.0	2.8	4.3	4.6	2.6
	0.3	6.0	9.1	9.2	6.2	-0.3	5.1	7.6	7.4	4.5
	0.6	19.0	25.6	25.9	19.3	-0.6	13.9	18.4	17.2	10.7
	0.9	47.6	56.4	56.3	46.8	-0.9	33.6	39.7	38.0	26.3
100	0.0	3.9	4.8	4.9	4.0	0.0	4.1	4.7	5.0	4.0
	0.3	12.8	15.5	15.4	13.1	-0.3	9.0	10.5	10.3	8.8
	0.6	51.1	55.2	54.7	51.1	-0.6	35.2	38.2	36.8	32.2
	0.9	90.0	91.6	91.4	89.7	-0.9	73.0	75.4	73.6	67.6
200	0.0	4.6	4.8	4.9	4.6	0.0	4.7	4.8	5.0	4.6
	0.3	28.1	28.7	28.7	27.7	-0.3	18.7	18.9	18.6	17.4
	0.6	88.6	88.7	88.6	87.9	-0.6	70.1	70.4	69.3	67.3
	0.9	99.9	99.9	99.9	99.9	-0.9	98.2	98.2	98.1	97.4

Table 5: Empirical power (in percentage) for models M3 and M4, at nominal level $\alpha = 5\%$.

C Supplementary results from the real application: the Greenland ice sheet mass balance

Figure 10 shows the global and partial correlograms up to lag 32 for the residuals (upper row) and for the squared residuals (lower row) of the TARMA model of Eq. (36). The confidence bands at level 99% under the null hypothesis of no-correlation are reported as blue dashed lines.

		M5					M6			
n	Ψ	sLMa	sLMi	sLMw	sLMh	Ψ	sLMa	sLMi	sLMw	sLMh
50	0.0	2.8	5.0	5.1	1.8	0.0	3.2	4.8	4.8	2.4
	-0.3	2.8	5.1	5.2	2.9	0.3	10.5	13.6	13.4	7.6
	-0.6	8.6	13.6	13.7	9.0	0.6	29.6	35.6	35.2	22.2
	-0.9	27.0	34.4	33.8	25.9	0.9	59.5	65.9	64.9	47.6
100	0.0	4.2	5.4	5.4	3.5	0.0	4.0	4.6	4.7	3.8
	-0.3	6.2	8.1	8.1	6.8	0.3	22.8	24.9	25.0	20.1
	-0.6	28.6	32.4	32.5	29.5	0.6	66.1	68.9	68.3	60.4
	-0.9	68.5	71.3	71.0	67.7	0.9	93.9	94.6	94.4	90.4
200	0.0	4.8	5.0	5.0	4.6	0.0	5.2	5.1	5.1	4.9
	-0.3	15.6	16.5	16.3	15.8	0.3	46.4	46.5	46.6	43.5
	-0.6	63.3	64.6	64.2	62.9	0.6	95.4	95.4	95.3	93.8
	-0.9	95.8	95.8	95.8	95.4	0.9	100.0	100.0	100.0	99.9

Table 6: Empirical power (in percentage) for models M3 and M4, at nominal level $\alpha = 5\%$.

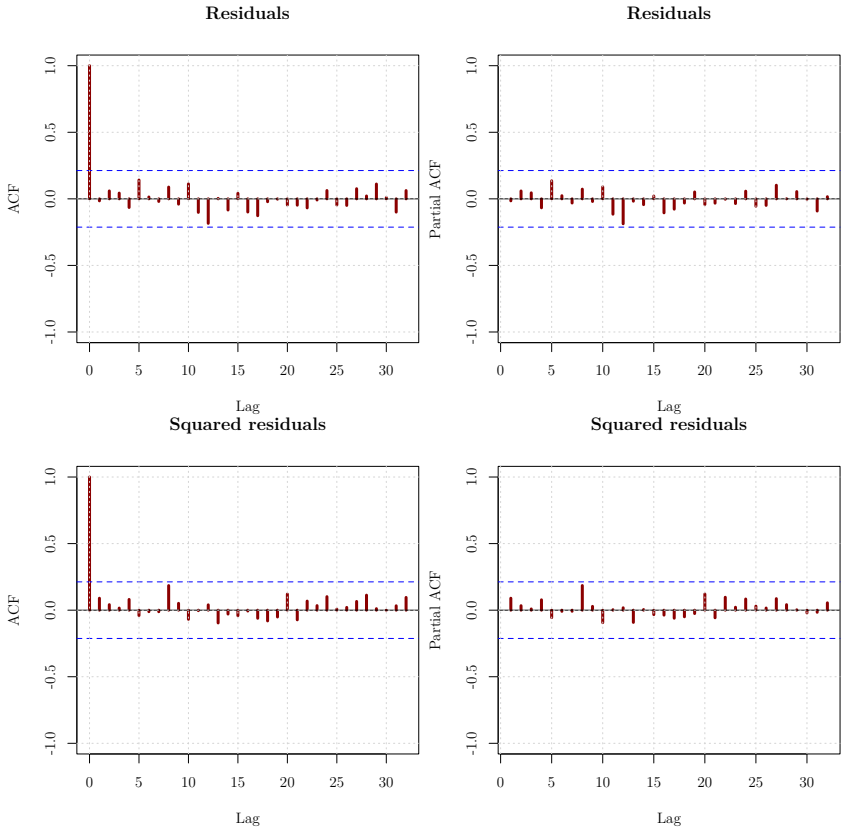

Figure 10: Correlograms of the residuals and squared residuals of the TARMA model of Eq. (36). Autocorrelation function (left) and partial autocorrelation function (right). The blue dashed lines indicate the rejection bands at 99% level.

Figure 11 reports the entropy based metric of serial dependence [Giannerini et al., 2015] up to lag 32 for the same residuals. The green and blue dashed lines indicate the confidence bands at levels 99% and 99.5% under the null hypothesis of serial independence. Figure 11 reports the histogram of the standardised residuals and the p-value of the Shapiro-Wilk test for normality. A standard Gaussian density is superimposed as a blue line.

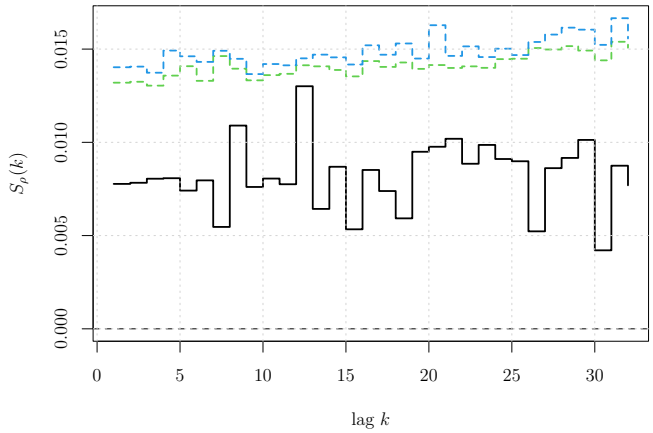

Figure 11: Entropy measure of serial dependence $S_\rho(k)$ computed on the residuals of the TARMA model of Eq. (36). The green and blue dashed lines indicate the rejection bands at levels 99% and 99.5%.

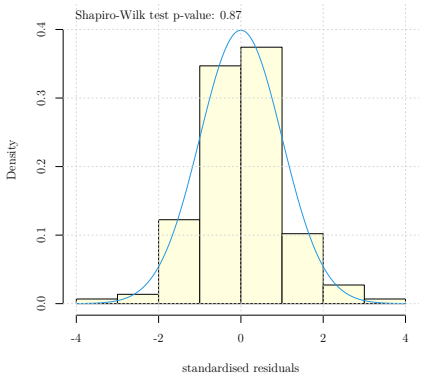

Figure 12: Histogram of the standardised residuals of the TARMA model of Eq. (36) with superimposed standard normal density in blue.

References

D.W.K. Andrews. Tests for parameter instability and structural change with unknown change point. *Econometrica*, 61(4):821–856, 1993. ISSN 00129682, 14680262. URL http://www.jstor.org/stable/2951764.

D.W.K. Andrews. Tests for parameter instability and structural change with unknown change point: A corrigendum. *Econometrica*, 71(1):395–397, 2003. doi: 10.1111/1468-0262.00405.

J. Anděl and K. Hrach. On calculation of stationary density of autoregressive processes. *Kybernetika*, 3:311–319, 01 2000.

J.H.E. Cartwright, D.L. González, and O. Piro. Universality in three-frequency resonances. *Phys. Rev. E*, 59:2902–2906, Mar 1999. doi: 10.1103/PhysRevE.59.2902. URL https://link.aps.org/doi/10.1103/PhysRevE.59.2902.

J.H.E. Cartwright, D.L. González, and O. Piro. Pitch perception: A dynamical-systems perspective. *Proceedings of the National Academy of Sciences*, 98(9):4855–4859, 2001. doi: 10.1073/pnas.081070998. URL https://www.pnas.org/doi/abs/10.1073/pnas.081070998.

G. Cavaliere and A. Rahbek. A primer on bootstrap testing of hypotheses in time series models: with an application to double autoregressive models. *Econometric Theory*, 37(1):1—-48, 2021. doi: 10.1017/S0266466620000067.

G. Cavaliere, H. B. Nielsen, and A. Rahbek. On the Consistency of Bootstrap Testing for a Parameter on the Boundary of the Parameter Space. *Journal of Time Series Analysis*, 38(4):513–534, 2017. doi: https://doi.org/10.1111/jtsa.12214. URL https://onlinelibrary.wiley.com/doi/abs/10.1111/jtsa.12214.

K.-S. Chan. Testing for threshold autoregression. *Ann. Statist.*, 18(4):1886–1894, 12 1990. doi: 10.1214/aos/1176347886. URL https://doi.org/10.1214/aos/1176347886.

K. S. Chan. Percentage points of likelihood ratio tests for threshold autoregression. *J. R. Stat. Soc. Ser. B. Stat. Methodol.*, 53(3):691–696, 1991. ISSN 00359246. URL http://www.jstor.org/stable/2345598.

K.-S. Chan and H. Tong. On likelihood ratio tests for threshold autoregression. *J. R. Stat. Soc. Ser. B. Stat. Methodol.*, 52(3):469–476, 1990. ISSN 00359246. URL http://www.jstor.org/stable/2345670.

K.-S. Chan, B.E. Hansen, and A. Timmermann. Guest editors' introduction: Regime switching and threshold models. *Journal of Business & Economic Statistics*, 35(2):159–161, 2017. doi: 10.1080/07350015.2017.1236521. URL http://dx.doi.org/10.1080/07350015.2017.1236521.

K.-S. Chan, S. Giannerini, G. Goracci, and H. Tong. Unit-root test within a threshold ARMA framework. Technical report, University of Iowa and University of Bologna, 2020. URL https://arxiv.org/abs/2002.09968v2.

K.-S. Chan, S. Giannerini, G. Goracci, and H. Tong. Testing for threshold regulation in presence of measurement error. *Statistica Sinica*, 34(3), 2024. URL https://doi.org/10.5705/ss.202022.0125.

R.B. Davies. Hypothesis testing when a nuisance parameter is present only under the alternative. *Biometrika*, 64(2):247–254, 1977. ISSN 00063444. URL http://www.jstor.org/stable/2335690.

R.B. Davies. Hypothesis testing when a nuisance parameter is present only under the alternatives. *Biometrika*, 74(1):33–43, 1987. ISSN 00063444. URL http://www.jstor.org/stable/2336019.

M. Friedrich, S. Smeekes, and J.-P. Urbain. Autoregressive wild bootstrap inference for nonparametric trends. *Journal of Econometrics*, 214(1):81 – 109, 2020. ISSN 0304-4076. doi: https://doi.org/10.1016/j.jeconom.2019.05.006. URL http://www.sciencedirect.com/science/article/pii/S0304407619301095. Annals Issue: Econometric Models of Climate Change.

S. Giannerini, E. Maasoumi, and E. Bee Dagum. Entropy testing for nonlinear serial dependence in time series. *Biometrika*, 102:661–675, 2015. URL http://biomet.oxfordjournals.org/content/102/3/661.abstract.

G. Goracci. Revisiting the canadian lynx time series analysis through TARMA models. *Statistica*, 80(4):357–394, 2020a. doi: 10.6092/issn.1973-2201/11478. URL https://rivista-statistica.unibo.it/article/view/11478.

G. Goracci. An empirical study on the parsimony and descriptive power of TARMA models. *Statistical Methods & Applications*, pages 109–137 , 02 2020b. doi: 10.1007/s10260-020-00516-8. URL https://doi.org/10.1007/s10260-020-00516-8.

G. Goracci, S. Giannerini, K.-S. Chan, and H. Tong. Testing for threshold effects in the TARMA framework. *Statistica Sinica*, 33(3), 2023. URL https://doi.org/10.5705/ss.202021.0120.

E. Hanna, J. Cappelen, X. Fettweis, S.H. Mernild, T.L. Mote, R. Mottram, K. Steffen, T.J. Ballinger, and R.J. Hall. Greenland surface air temperature changes from 1981 to 2019 and implications for ice-sheet melt and mass-balance change. *International Journal of Climatology*, 41(S1):E1336–E1352, 2021. doi: https://doi.org/10.1002/joc.6771. URL https://rmets.onlinelibrary.wiley.com/doi/abs/10.1002/joc.6771.

B.E. Hansen. Inference when a nuisance parameter is not identified under the null hypothesis. *Econometrica*, 64(2):413–430, 1996. ISSN 00129682, 14680262. URL http://www.jstor.org/stable/2171789.

B.E. Hansen. Threshold autoregression in economics. *Statistics and its Interface*, 4(2):123–127, 2011.

J.B. Hill. Weak-identification robust wild bootstrap applied to a consistent model specification test. *Econometric Theory*, 37(3):409–463, 2021. doi: 10.1017/S0266466620000201.

G. Li and W.K. Li. Testing a linear time series model against its threshold extension. *Biometrika*, 98(1):243–250, 02 2011. ISSN 0006-3444. doi: 10.1093/biomet/asq074.

S. Ling and H. Tong. Testing for a linear MA model against threshold MA models. *Ann. Statist.*, 33(6):2529–2552, 12 2005. URL https://doi.org/10.1214/009053605000000598.

R. Luukkonen, P. Saikkonen, and T. Teräsvirta. Testing linearity against smooth transition autoregressive models. *Biometrika*, 75(3):491–499, 1988. ISSN 00063444. URL http://www.jstor.org/stable/2336599.

K. D. Mankoff, X. Fettweis, P. L. Langen, M. Stendel, K. K. Kjeldsen, N. B. Karlsson, B. Noël, M. R. van den Broeke, A. Solgaard, W. Colgan, J. E. Box, S. B. Simonsen, M. D. King, A. P. Ahlstrøm, S. B. Andersen, and R. S. Fausto. Greenland ice sheet mass balance from 1840 through next week. *Earth System Science Data*, 13(10):5001–5025, 2021. doi: 10.5194/essd-13-5001-2021. URL https://essd.copernicus.org/articles/13/5001/2021/.

J. Petruccelli and N. Davies. A portmanteau test for self-exciting threshold autoregressive-type nonlinearity in time series. *Biometrika*, 73(3):687–694, 12 1986. ISSN 0006-3444. doi: 10.1093/biomet/73.3.687.

J.D. Petruccelli. On the approximation of time series by threshold autoregressive models. *Sankhyā: The Indian Journal of Statistics, Series B (1960-2002)*, 54 (1):106–113, 1992. URL http://www.jstor.org/stable/25052727.

I. Sasgen, B. Wouters, A. S. Gardner, M. D. King, M. Tedesco, F. W. Landerer, C. Dahle, H. Save, and X. Fettweis. Return to rapid ice loss in Greenland and record loss in 2019 detected by the GRACE-FO satellites. *Commun. Earth Environ.*, 1(1):8, 2020a. doi: 10.1038/s43247-020-0010-1.

I. Sasgen, B. Wouters, A. S. Gardner, M. D. King, M. Tedesco, F. W. Landerer, C. Dahle, H. Save, and X. Fettweis. Greenland Ice Sheet mass balance GRACE/GRACE-FO (2003-2019) and SMB-D (1948-2019), 2020b. URL https://doi.org/10.1594/PANGAEA.919670.

I. Sasgen, A. Salles, M. Wegmann, B. Wouters, X. Fettweis, B.P.Y. Noël, and C. Beck. Arctic glaciers record wavier circumpolar winds. *Nature Climate Change*, 12(3):249–255, Mar 2022. ISSN 1758-6798. doi: 10.1038/s41558-021-01275-4. URL https://doi.org/10.1038/s41558-021-01275-4.

T. Scambos, F. Straneo, and M. Tedesco. How fast is the greenland ice sheet melting? *Arctic, Antarctic, and Alpine Research*, 53(1):221–222, 2021. doi: 10.1080/15230430.2021.1946241. URL https://doi.org/10.1080/15230430.2021.1946241.

F. Su and K.-S. Chan. Testing for threshold diffusion. *J. Bus. Econom. Statist.*, 35:218–227, 04 2017. doi: 10.1080/07350015.2015.1073594.

H. Tong. *Non-linear Time Series: A Dynamical System Approach*. Clarendon Press, 1990.

H. Tong. Threshold models in time series analysis–30 years on. *Statistics and its Interface*, 4(2):107–118, 2011.

H. Tong. Threshold models in time series analysis—some reflections. *Journal of Econometrics*, 189(2):485 – 491, 2017. doi: https://doi.org/10.1016/j.jeconom.2015.03.039. URL http://www.sciencedirect.com/science/article/pii/S0304407615001177.

R.S. Tsay. Testing and modeling multivariate threshold models. *J. Amer. Statist. Assoc.*, 93(443):1188–1202, 1998. ISSN 01621459. URL http://www.jstor.org/stable/2669861.

A.W. van der Vaart. *Asymptotic statistics*. Cambridge series in statistical and probabilistic Mathematics, Cambridge University Press, 1998.

C. S. Wong and W.K. Li. Testing for threshold autoregression with conditional heteroscedasticity. *Biometrika*, 84(2):407–418, 1997. ISSN 00063444. URL http://www.jstor.org/stable/2337466.

C. S. Wong and W.K. Li. Testing for double threshold autoregressive conditional heteroscedastic model. *Statist. Sinica*, 10(1):173–189, 2000. ISSN 10170405, 19968507. URL http://www.jstor.org/stable/24306711.

T. Woollings, A. Hannachi, B. Hoskins, and A. Turner. A Regime View of the North Atlantic Oscillation and Its Response to Anthropogenic Forcing. *Journal of Climate*, 23(6):1291 – 1307, 2010. doi: 10.1175/2009JCLI3087.1. URL https://journals.ametsoc.org/view/journals/clim/23/6/2009jcli3087.1.xml.

B. Wouters, J.L. Bamber, M.R. van den Broeke, J.T.M. Lenaerts, and I. Sasgen. Limits in detecting acceleration of ice sheet mass loss due to climate variability. *Nature Geoscience*, 6(8):613–616, Aug 2013. ISSN 1752-0908. doi: 10.1038/ngeo1874. URL https://doi.org/10.1038/ngeo1874.

M. Zeitz, J. M. Haacker, J. F. Donges, T. Albrecht, and R. Winkelmann. Dynamic regimes of the greenland ice sheet emerging from interacting melt–elevation and glacial isostatic adjustment feedbacks. *Earth System Dynamics*, 13(3):1077–1096, 2022. doi: 10.5194/esd-13-1077-2022. URL https://esd.copernicus.org/articles/13/1077/2022/.

B. Zhang, L. Liu, Y. Yao, T. van Dam, and S.A. Khan. Improving the estimate of the secular variation of greenland ice mass in the recent decades by incorporating a stochastic process. *Earth and Planetary Science Letters*, 549:116518, 2020. ISSN 0012-821X. doi: https://doi.org/10.1016/j.epsl.2020.116518. URL https://www.sciencedirect.com/science/article/pii/S0012821X20304623.