Climate Dynamics (2022) 58:1349–1361
https://doi.org/10.1007/s00382-021-05974-8

How to determine the statistical significance of trends in seasonal records: application to Antarctic temperatures

Armin Bunde[1] · Josef Ludescher[2] · Hans Joachim Schellnhuber[2]

Received: 25 May 2021 / Accepted: 12 September 2021 / Published online: 5 October 2021

Abstract

We consider trends in the m seasonal subrecords of a record. To determine the statistical significance of the m trends, one usually determines the p value of each season either numerically or analytically and compares it with a significance level $\tilde{\alpha}$. We show in great detail for short- and long-term persistent records that this procedure, which is standard in climate science, is inadequate since it produces too many false positives (false discoveries). We specify, on the basis of the family wise error rate and by adapting ideas from multiple testing correction approaches, how the procedure must be changed to obtain more suitable significance criteria for the m trends. Our analysis is valid for data with all kinds of persistence. Specifically for long-term persistent data, we derive simple analytical expressions for the quantities of interest, which allow to determine easily the statistical significance of a trend in a seasonal record. As an application, we focus on 17 Antarctic station data. We show that only four trends in the seasonal temperature data are outside the bounds of natural variability, in marked contrast to earlier conclusions.

Keywords Seasonal records · Persistence · Trends · Statistical significance · Multiple testing

1 Introduction

In recent years, mainly due to climate change, the estimation of the statistical significance of climate trends has become an important issue, since the question whether a climate trend is of anthropogenic or natural origin is of great relevance for mitigation and adaptation measures alike (IPCC 2014). Of particular interest are, for instance, trends in temperature or precipitation records, river flows, and Arctic or Antarctic sea-ice-extent. Here we are particularly interested in the statistical significance of trends in seasonal records. Significant seasonal climate trends are of great importance, since they may affect considerably ecological systems, agricultural yields and human societies, this way creating major challenges for crop rotation management (Troost et al. 2015), river-borne transportation (Caldwell et al. 2002) and power generation (Rübbelke and Vögele 2013), as well as for the control of pests and vector-borne diseases (Rao et al. 2015).

Seasons are usually the four meteorological seasons winter, spring, summer and autumn, but also the 12 months, the 52 weeks or the 365 calendar days (without leap days) can be considered as generalized “seasons”. Also annual data can be considered as seasonal data where the season spans the whole year. The trends are usually obtained from a linear regression analysis. The relevant quantity is the positively defined relative trend $x \geq 0$ which is the ratio between the trend amplitude $|\Delta|$ and the standard deviation σ around the trend line (see, e.g., Ludescher et al. 2017).

For obtaining the p values of the trends in the m seasonal records, one needs to choose an appropriate model for the persistence of the record. Prominent examples are Gaussian white noise (for records without persistence), autoregressive processes of first order (AR(1)) for records with short-term memory, and scale-free processes for records with long-term memory (see Sect. 2). The surrogate data generated by these models allow to determine numerically, for each of the m trends, the probability $p_\nu^{(m)}(x)$ (p value) that in season ν, $\nu = 1, \ldots, m$, a relative trend above x is observed due to the record’s natural variability. Since no season is distinguished a priori from the others, there is no explicit dependence on

✉ Josef Ludescher
josef.ludescher@pik-potsdam.de

[1] Institute for Theoretical Physics, Justus Liebig University Giessen, 35392 Giessen, Germany

[2] Potsdam Institute for Climate Impact Research, 14473 Potsdam, Germany

ν and $p_\nu^{(m)}(x) \equiv p^{(m)}(x)$. This implies that the largest relative trend has the smallest p value. For white noise $p^{(m)}(x)$ is known exactly (Bronstein et al. 2004).

In climate science, one usually does not follow this route but instead assumes that the m seasonal records are independent and described by m different AR(1) processes with m detrended lag-1 autocorrelations $C_\nu(1), \nu = 1, \ldots, m$. Only in this approach, which we will refer to as the standard approximation, $p_\nu^{(m)}(x)$ depends explicitly on ν via $C_\nu(1)$ [(for details, see Sect. 2 and the discussions in (Mitchell et al. 1966; Santer et al. 2000)]. Here the largest relative trend does not have necessarily the smallest p value.

To decide whether one of the m relative trends x_ν is statistically significant, i.e., cannot be solely explained by the natural variability of the record, one usually compares its p-value with a certain significance level $\tilde{\alpha}$ (typically $\tilde{\alpha} = 0.05$ or 0.01). A trend $x_\nu, \nu = 1, \ldots, m$ is considered as significant, when the condition

$$p_\nu^{(m)}(x_\nu) \leq \tilde{\alpha}, \tag{1}$$

is met. For recent applications of this procedure in the important context of Antarctic temperatures, we refer to (Steig et al. 2009; O'Donnel et al. 2011; Bromwich et al. 2014; Jones et al. 2014; Chapman et al. 2007; Monaghan et al. 2008; Turner et al. 2016, 2019) and references therein.

By definition, when in a record the trend of at least *one* of the m seasons is found statistically significant, the record cannot be solely of natural origin. The question is how reliable the crucial significance criterion (1) is. Here we show in great detail that (1), in particular together with the standard approximation, is inadequate since it produces too many false positives (false discoveries, Type I errors). We explain in detail, on the basis of the family wise error rate and by adapting ideas from maximum statistics and multiple testing correction approaches (Bonferroni 1936; Holm 1979), how (1) must be changed to obtain appropriate significance criteria, this way confirming and considerably extending our recent work on this subject (Ludescher et al. 2017). Our analysis holds for Gaussian white noise, as well as for short-term persistent data (where the autocorrelation function $C(t)$ decays exponentially with time t) and long-term persistent data (where $C(t)$ decays algebraically), and also holds for a combination of both. Specifically for long-term persistent data, which are of great relevance in geoscience (Hurst 1951; Mandelbrot and Wallis 1968; Koscielny-Bunde et al. 1998; Malamud et al. 1999; Eichner et al. 2003; Blender et al. 2003; Vyushin et al. 2004; Cohn and Lins 2005; Santhanam and Kantz 2005; Kiraly et al. 2006; Livina et al. 2005; Varotsos and Kirk-Davidoff 2006; Lennartz et al. 2011; Mudelsee 2007; Yuan et al. 2010; Franzke 2010; Bogachev and Bunde 2012; Bunde et al. 2012; Lovejoy and Schertzer 2013; Dangendorf et al. 2014; Bunde et al. 2014; Yuan et al. 2015; Blender et al. 2015; Ludescher et al. 2016; Blesic et al. 2019; Blesic 2020; Ludescher et al. 2020), we derive simple analytical expressions for all quantities of interest, which allows to determine straightforwardly the significance of a trend in a seasonal record. As an important application, we focus on the significance of the observed trends in seasonal Antarctic temperature data in the past 50 years.

2 Statistical models and *p* values

We are interested in daily or monthly climate records $y(i)$, $i = 1, 2, \ldots, N$ and their m seasonal subrecords $y_\nu(j)$. The annual index j runs from 1 to L. For obtaining the p-values of the observed relative trends x_ν, one needs to choose an appropriate model for the persistence of the record.

Gaussian white noise: Here the data are independent and the m relative trends $x \equiv x_\nu$ follow (Bronstein et al. 2004; Mitchell et al. 1966; Santer et al. 2000) a Student's t distribution

$$P^{(m)}(t) \propto \left(1 + \frac{t^2}{\ell}\right)^{-\frac{\ell+1}{2}}, \tag{2}$$

with $t = x/a$ and the scaling factor

$$a = \sqrt{12/\ell} \;\; + \;\; O(L^{-3/2}). \tag{3}$$

Here, $\ell = L - 2$ is the number of degrees of freedom.

From the Student's t distribution one can easily obtain the p value $p^{(m)}(t) = \int_t^\infty dy P^{(m)}(y)$. Note that $P^{(m)}(t)$ and $p^{(m)}(t)$ do not depend explicitly on m since the data are uncorrelated. For $p^{(m)}(t) = 0.05$ and 0.01, the corresponding quantiles t_{05} and t_{01} are listed in mathematical tables for fixed ℓ. From t_{05} and t_{01} one obtains $x_{0.05} = at_{05}$ and $x_{0.01} = at_{01}$.

Short-term memory: In records with short-term memory, a measure for the strength of the correlation is the lag-1 autocorrelation $C(1)$. To eliminate effects of external trends, one often detrends the data by subtracting the linear trend and then uses the detrended data to determine the detrended $C(1)$ [see, e.g., (Santer et al. 2000)]. Unless otherwise stated, $C(1)$ refers here to the detrended lag-1 autocorrelation.

The simplest model that only requires the knowledge of $C(1)$ is an AR(1) process, where $y(i)$ and $y(i+1)$ are coupled by

$$y(i+1) = by(i) + \zeta(i), \quad i = 1, \ldots, N-1. \tag{4}$$

Here, b is the persistence parameter and $\zeta(i)$ is Gaussian white noise; in the limit of $N \to \infty$, b is identical to $C(1)$. For finite N, however, $C(1)$ is distributed around b. Accordingly, for instance, an observed $C(1) = 0.4$ may also arise from records with b below and above 0.4.

For obtaining surrogate data with fixed $C(1) = c_1$, one needs to simulate a large number of records with b between -1 and $+1$. In each record, one determines $C(1)$ and then selects only those records where $C(1)$ is in a narrow interval around c_1, this way obtaining a set of K surrogate records with the desired lag-1 autocorrelation c_1.

Next, we divide each of these K records into m seasonal subrecords and determine the relative seasonal trends x_ν. This way we obtain a set of $K_m = mK$ relative trend values. Next we order them in descending order, such that $x(1) < x(2) < \cdots < x(K_m)$. Then, by definition,

$$p^{(m)}\big(x = x(k)\big) = (K_m - k + 1)/K_m, \quad k = 1, \dots, K_m. \tag{5}$$

The relevant quantiles $x_{0.05}$ and $x_{0.01}$ are obtained from $k = 0.95K_m$ and $k = 0.99K_m$, respectively.

For annual data ($m = 1$), one can obtain a reasonable approximation for $p^{(1)}$ by replacing ℓ in (2) and (3) by $\ell^{\text{eff}} = L(1 - C(1))/(1 + C(1)) - 2$ (see, e.g., (Mitchell et al. 1966; Santer et al. 2000)). When assuming that the m seasons can be considered as independent, then this relation also holds for the m seasonal subrecords, i.e.,

$$\ell^{\text{eff}}_\nu = L\frac{1 - C_\nu(1)}{1 + C_\nu(1)} - 2, \quad \nu = 1, \dots, m. \tag{6}$$

Under the above assumptions, (2), (3), and (6) offer a simple way to obtain the desired p values and the relevant quantiles also in the presence of correlations. Since this approximation is very popular in climate science (see, e.g., (Turner et al. 2019)), we refer to it as the standard approximation.

Long-term memory: While in data with short-term memory the lag-t autocorrelation function $C(t)$ decays exponentially, $C(t)$ decays algebraically in data with long-term memory, in the limit of $N \to \infty$. For finite record length N (Lennartz and Bunde 2009b), the (undetrended) autocorrelation follows

$$C(t) \cong \frac{1}{1 - N^{-\gamma}}\left(\frac{(2-\gamma)(1-\gamma)}{2}t^{-\gamma} - N^{-\gamma}\right), \; t \geq 1, \tag{7}$$

where $0 < \gamma < 1$ denotes the correlation exponent. Such correlations are named "long-term" since the mean correlation time diverges in the limit of infinitely long series.

According to (7), $C(t)$ shows strong finite size effects such that the power-law dependence can only be seen for very small time lags t (for $\gamma = 0.4$ and the comparatively large record length $N = 16,000$ only for $t < 10$) (Lennartz and Bunde 2009b). In addition, C may depend on external trends. Since $C(t)$ is inappropriate to quantify the long-term memory in the relatively short climate records, one usually considers the Detrended Fluctuation Analysis 2 (DFA2) and its fluctuation function $F(s)$ (Kantelhardt et al. 2001). To obtain $F(s)$, one divides the seasonally detrended monthly (or daily) record $\{y(i)\}, i = 1, \dots, N$, into non-overlapping windows μ of lengths s. Then one focuses, in each segment μ, on the cumulated sum Y_i of the $\{y(i)\}$, and determines the variance $F^2_\mu(s)$ of the Y_i around the best polynomial fit of order 2. After averaging $F^2_\mu(s)$ over all segments μ and taking the square root, one arrives at the desired fluctuation function $F(s)$. One can show that in long-term persistent records (Kantelhardt et al. 2001)

$$F(s) \sim s^h, \quad 10 < s < N/4, \tag{8}$$

where the exponent h can be associated with the Hurst exponent (Hurst 1951; Mandelbrot and Wallis 1968) and is related to the correlation exponent γ by $h = 1 - \gamma/2$. By construction, h is not affected by external linear trends.

Long-term persistent records can also be characterized by their power spectral density $S(f)$ which decreases with frequency f as $S(f) \sim f^{-\beta}$ with $\beta = 1 - \gamma = 2h - 1$ for large record lengths N. This relation can be used to generate a set of long-term correlated surrogate data with length $N = mL$ and fixed DFA2 exponent h (for detailed descriptions of the method, we refer to (Lennartz and Bunde 2009a; Tamazian et al. 2015). Then following the procedure described above, one can easily determine numerically, for each m of interest, the relevant $p^{(m)}$ values of the m observed relative trends.

For annual data, analytical descriptions of $p^{(1)}(x)$ and the quantiles are available. For example, for monthly data with $L = 50$ years, the quantiles $x_{0.05}$ and $x_{0.01}$ follow the relations $x_{0.05} \simeq 0.2 + h^2|\ln(0.05/1.6)|$ and $x_{0.01} \simeq 0.2 + h^2|\ln(0.01/1.6)|$ (Lennartz and Bunde 2009a).

3 Fraction of statistically significant records and family wise error rate

When the respective p values have been obtained as described in Sect. 2, the significance test (1) can be applied to all kinds of seasonal data, from $m = 1$ (annual data) until $m = 365$ (where each calendar day is one season). The tests are well calibrated and consistent with each other when for a large number K of surrogate records (where by definition no external trends are present), at most $K\alpha$ (usually $\alpha = 0.05$ resp. 0.01) of them are found significant. In other words, it is required that the family wise error rate $F(m, \tilde{\alpha})$, which denotes the probability that in a record with m seasons at least one of them has a p-value below $\tilde{\alpha}$, satisfies

$$F(m, \tilde{\alpha}) \leq \alpha. \tag{9}$$

The question is how $\tilde{\alpha}$ in (1) must be chosen, in dependence of m and α, such that Eq. (9) is satisfied.

Before considering $F(m, \tilde{\alpha})$ in greater detail, we first put the standard approximation to a direct test. We have used Eq. (4) to generate 1000 short-term correlated daily records with $L = 50$ years. In each record, b has been chosen such that the detrended lag-1 autocorrelation between successive months was 0.4, which is typical for temperature data sets. In each record, we determined the p values of the trends in (i) the annual record ($m = 1$), (ii) the 4 records for the 4 meteorological seasons, (iii) the 12 records for the 12 months, (iv) the 52 records for the 52 weeks, and finally (v) the 365 records for the 365 calendar days.

In each of these sets of m subrecords we first focus on the trend with the lowest p value. By averaging over all 1,000 data sets, we obtain the mean lowest p value and the standard deviation from this average. Then we do the same for the 2nd, 3rd and 4th lowest p value in each data set. The results are shown in Table 1: already for $m = 4$ there is a good chance (20.2%) that the trend with the minimum p value is significant, and for m above 52 the chances are high that even 2 trends or more are highly significant. Accordingly, when testing whether a record is of natural origin or not, the standard method together with (1) produces a large fraction of false positives (false alarms) for $m \geq 4$ and thus is not applicable.

To obtain a more quantitative picture, we now focus directly on the family wise error rate $F(m, \tilde{\alpha})$. Figure 1 shows $F(m, \tilde{\alpha})$ with $\tilde{\alpha} = 0.05$ for (a) short-term persistent data where the p values have been determined both numerically and by the standard approximation, (b) white noise records, and (c) long-term correlated records ($h = 0.75$) where $p^{(m)}(x)$ has been obtained numerically. For each kind of data set, we considered 20,000 records, with length 50 years.

The figure shows that in all cases, $F(m, \tilde{\alpha})$ increases monotonously with m. For m above 52, most of the 20,000 records are falsely indicated as significant, irrespective of their persistence properties.

For Gaussian white noise, we can estimate $F(m, \tilde{\alpha})$ analytically: Since the m seasonal records are independent of each other, the probability that none of them has a p value below $\tilde{\alpha}$ is $(1 - \tilde{\alpha})^m$. Therefore,

$$F(m, \tilde{\alpha}) = 1 - (1 - \tilde{\alpha})^m. \tag{10}$$

Equations (9) and (10) imply $\tilde{\alpha} \leq 1 - (1 - \alpha)^{1/m}$. Together with (1), this yields

$$p^{(m)}(x_\nu) \leq 1 - (1 - \alpha)^{1/m}, \quad \nu = 1, \ldots, m. \tag{11}$$

Accordingly, a trend x_ν is significant when its p value is below $1 - (1 - \alpha)^{1/m}$. Only for annual data ($m = 1$), (11) reduces to (1).

Inequality (11), known as Šidák correction (Šidák 1967), is slightly less conservative than the Bonferroni correction (Bonferroni 1936; Ludescher et al. 2017) $p^{(m)}(x_\nu) \leq \alpha/m$ which holds quite generally when multiple tests are performed on uncorrelated or positively correlated data. However, both (11) and the Bonferroni correction are too conservative for short- or long-term persistent records since the m subrecords are not fully independent of each other.

We find it convenient to describe the effect of persistence in the data by an effective exponent $m_{\mathrm{eff}}(m) \leq m$, such that

$$F(m, \tilde{\alpha}) = 1 - (1 - \tilde{\alpha})^{m_{\mathrm{eff}}}. \tag{12}$$

This choice can be motivated as follows. Assume that in a (synthetic) monthly temperature record the correlations are such that in each year, the temperatures of months 1–3, 4–6, 7–9, and 10–12 are identical but uncorrelated with the

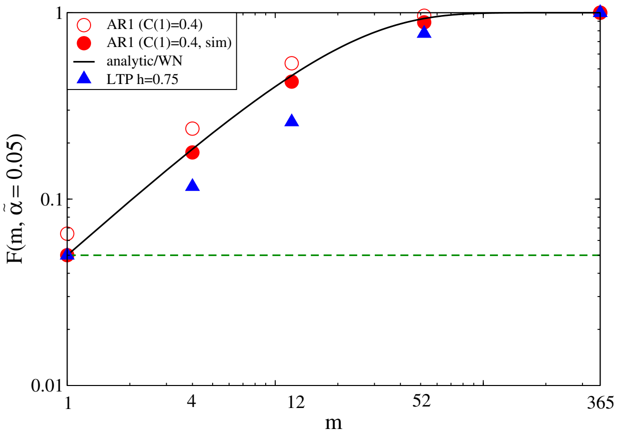

Fig. 1 Family wise error rate $F(m, \tilde{\alpha})$ based on Eq. (1). The figure shows how $F(m, \tilde{\alpha})$ increases with increasing number m of seasons in records of length 50y, for $\tilde{\alpha} = 0.05$. The red circles show F for data generated by an AR(1) process where each record is characterized by the detrended lag-1 autocorrelation $C(1) = 0.4$ between successive months. The open circles refer to the standard approximation of the p values, (2) with (3) and (6), while the full circles refer to a rigorous treatment. The figure also shows $F(m, \tilde{\alpha})$ for long-term persistent records with Hurst exponent $h = 0.75$ (triangles). To obtain F, we generated 20,000 records for each kind of data set. The full line shows the exact curve for Gaussian white noise. The dashed line represents $\alpha = 0.05$, which is the expected value of F in a correct analysis

Table 1 Mean p values and standard deviations σ_p for the 4 lowest p values obtained by the standard approximation (Eq. (1) with (2) and (6)), in short-term correlated records with $C(1) = 0.4$, for annual data ($m = 1$), meteorological seasons ($m = 4$), months ($m = 12$), weeks ($m = 52$), and days ($m = 365$). Shown are averages over 1000 records. The length L of the data is 50 years

m	$\bar{p}(x_m), \sigma_p$	$\bar{p}(x_{m-1}), \sigma_p$	$\bar{p}(x_{m-2}), \sigma_p$	$\bar{p}(x_{m-3}), \sigma_p$
1	0.489, 0.294	–	–	–
4	0.186, 0.163	0.383, 0.207	0.585, 0.210	0.791, 0.172
12	0.069, 0.073	0.142, 0.103	0.215, 0.121	0.290, 0.135
52	0.013, 0.016	0.028, 0.025	0.044, 0.032	0.061, 0.037
365	0.001, 0.001	0.002, 0.002	0.004, 0.003	0.005, 0.004

other temperatures. In this case, $F(m, \tilde{\alpha}) = 1 - (1 - \tilde{\alpha})^{m_{\mathrm{eff}}}$ with $m_{\mathrm{eff}} = 4$.

For the long-term correlated records from Fig. 1, where $\tilde{\alpha} = 0.05$, we have $m_{\mathrm{eff}}(m) \approx 2.4, 5.9$, and 28 for $m = 4, 12$, and 52. For the short-term correlated records where the p values have been obtained numerically, we have $m_{\mathrm{eff}}(m) \approx 3.8, 11.5$, and 47 for $m = 4, 12$, and 52.

However, when using the standard approximation, $m_{\mathrm{eff}}(m)$ even exceeds m for $m = 1$, 4, and 12: $m_{\mathrm{eff}}(1) \approx 1.5$, $m_{\mathrm{eff}}(4) \approx 5.9$, and $m_{\mathrm{eff}}(11) \approx 14.5$. This is a serious inconsistency of the method: Since the seasonal records are considered as independent within this treatment, the family-wise error rate should follow Eq. (10).

The fact that even for annual data ($m = 1$) and for short-term correlations the standard approximation is too liberal, is another serious drawback of this approach [see also (Mitchell et al. 1966; von Storch and Zwiers 1999)], which has been widely used to evaluate climate change (Hartmann et al. 2013). The reason is that even for $m = 1$, (6) is valid only in the limit of $L \to \infty$ and thus is incorrect for the comparatively short records usually considered in climate science.

4 The statistical significance of the largest relative trend

As described above, a record with m seasonal subrecords is called significant, when at least one of its subrecords has a significant trend. Since the largest relative trend in the m subrecords has the smallest p value, the record is significant, when the probability $p^{(m)}_{\max}(x)$ that the largest trend $x \equiv x_{\max}$ is above x, satisfies

$$p^{(m)}_{\max}(x) \le \alpha. \tag{13}$$

Accordingly, when (13) is used as a significance condition, the familiy wise error rate automatically satisfies (9). Inequality (13) has been suggested before, but based on different arguments (Ludescher et al. 2017).

It is easy to see that $p^{(m)}_{\max}(x_{\max})$ is not simply $p^{(m)}(x_{\max})$: For i.i.d. data, the probability that the relative trend in season v is below x, is $1 - p^{(m)}(x)$. Therefore, the probability that in all m seasons the relative trends are below x, is $\left(1 - p^{(m)}(x)\right)^m$. Thus the probability $p^{(m)}_{\max}(x)$ that at least in one season the relative trend is above x, becomes

$$p^{(m)}_{\max}(x) = 1 - \left(1 - p^{(m)}(x)\right)^m. \tag{14}$$

Accordingly, (13) and (14) allow to determine whether the p-value of the largest trend $x = x_{\max}$ in i.i.d. data is significant. By combining (13) with (14) we recover (11).

By definition, $p^{(m)}_{\max}(x_{\max})$ is the minimum p-value one of the m relative trends can have. Therefore, $p^{(m)}_{\max}(x_{\max})$ yields a lower bound for the p values of the $m - 1$ smaller trends: When $p^{(m)}_{\max}(x_{\max})$ is above some threshold α, the p-values of the remaining trends must be also above α.

In Sect. 3 we have argued that the influence of short- and long-term persistence on the family wise error rate can be described by an effective m value. The question is, whether this argument also applies here, i.e., whether the general relation (14) between $p^{(m)}(x)$ and $p^{(m)}_{\max}(x)$ is still valid, at least approximately, for short- and long-term persistent data, when instead of m an effective $m_{\mathrm{eff}}(m)$ is introduced.

To see whether this is the case, we have numerically generated 600,000 short-term correlated monthly records with $C(1) = 0.4$ and 600,000 long-term correlated monthly records with $h = 0.75$, both with length 50y as in Fig. 1. For the long-term correlated records, the mean $C(1)$ value is also close to 0.4. We determined numerically $p^{(m)}(x)$ and $p^{(m)}_{\max}(x)$ for $m = 4$ and 12 (for details, see Section 8). The results are shown in Fig. 2. In each of the 4 panels, the top curve shows $p^{(m)}_{\max}(x)$ and the bottom curve $p^{(m)}(x)$. In black, we show the original curves and in red the approximation

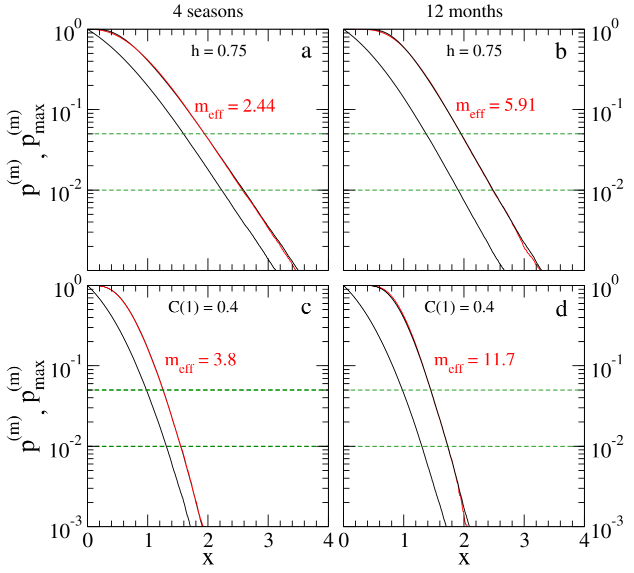

Fig. 2 Adjusted p-value $p^{(m)}_{\max}(x)$ of the largest relative trend (upper curves) compared with $p^{(m)}(x)$ (lower curves) in persistent records. The figure shows, for the monthly long- and short-term persistent data of Fig. 1 with $h = 0.75$ (**a**, **b**) and $C(1) = 0.4$ (**c**, **d**), respectively, that $p^{(m)}_{\max}$ is related to $p^{(m)}(x)$ by the power law (15), with about the same values of $m_{\mathrm{eff}}(m)$ as in Fig. 1. The left panels (**a**, **c**) refer to the 4 meteorological seasons, the right panels (**b**, **d**) to the 12 months. The figure also shows that in the relevant p-regime between 10^{-1} and 10^{-3}, the curves can be well approximated by simple exponentials. The length L of the data is 50 years

$$p_{\max}^{(m)}(x) = 1 - \left(1 - p^{(m)}(x)\right)^{m_{\text{eff}}(m)}, \tag{15}$$

with roughly the same $m_{\text{eff}}(m)$ values as determined in Sect. 3. In all cases, it is very difficult to distinguish between the exact curve $p_{\max}^{(m)}(x)$ and its approximation (15). From (15) we can verify that our guess (12) is correct for arbitrary values of $\tilde{\alpha}$.

Equation (15) generalizes nicely the maximum statistics for i.i.d. numbers to short- and long-term correlated numbers; as in Fig. 1, $m_{\text{eff}}(m)$ describes the effective degrees of freedom in the m seasonal subrecords. When they are all independent, we have m degrees of freedom, and when they are coupled by long-term memory, the degrees of freedom decrease. We will show in Fig. 4 that the stronger the long-term persistence is, the stronger the decrease of m_{eff} is.

Figure 2 also shows how different the p values are for short- and long-term correlations. When the record is short-term persistent, then, for instance, $x_{\max} = 1.6$ is highly significant. However, in the long-term correlated record with roughly the same $C(1)$ value, $x_{\max} = 1.6$ is far from being significant. This shows how crucial it is to use the proper surrogate data when determining the p values of the relative trends. The figure also shows that for $h = 0.75$, the p value of the maximum trend depends only very weakly on m.

Combining (13) with (15) we rediscover (11), where now m is substituted by $m_{\text{eff}}(m)$. While (13–15) describe, in a rigorous way, the p value of the season with the largest relative trend $x = x_{\max}$, they are very conservative for the seasons with the smaller trends. It is possible that by applying (13–15) to the smaller trends, significant trends in these seasons can be overlooked. Our next aim is to obtain better estimations for the adjusted p-values of the nth largest trend $x \equiv x_{\max,\,\text{n}}$.

5 The statistical significance of the smaller relative trends

First, we consider the 2nd largest relative trend $x_{\max,\,2}$. Following the argumentation of Holm (Holm 1979) for correcting for multiple testing, we consider in a Gedankenexperiment a very large set of K records with m seasonal subrecords and eliminate, in each record, randomly one of the m seasonal subrecords. It is clear that in the remaining $K(m-1)$ subrecords, the $K(m-1)$ trends form the same distribution as the Km trends in the Km subrecords and thus have the same p-function $p^{(m)}(x)$.

By definition, the largest trend in the new set of $m-1$ trends cannot be larger than the 2nd largest trend in the original set of m trends. Thus the p value $p_{\max}^{(m-1)}(x)$ of the maximum of these $(m-1)$ trends represents, by construction, an upper bound for the desired p value of the 2nd largest trend.

For i.i.d. data we obtain immediately

$$p_{\max}^{(m-1)}(x) = 1 - \left(1 - p^{(m)}(x)\right)^{m-1}. \tag{16}$$

With the 3rd, 4th, …, $(m-1)$th largest trend we proceed analogously. In general, we obtain an upper bound for the p value of the nth-largest trend by determining the adjusted p value $p_{\max}^{(m+1-n)}(x)$ of the maximum of $(m+1-n)$ trends. By definition, $p_{\max}^{(1)}(x) \equiv p^{(m)}(x)$. For i.i.d. data, we obtain this way

$$p_{\max}^{(m+1-n)}(x) = 1 - \left(1 - p^{(m)}(x)\right)^{m+1-n}, n = 1, \ldots, m. \tag{17}$$

For being significant, $x = x_{\max,\,\text{n}}$ must satisfy the condition

$$p_{\max}^{(m+1-n)}(x) \leq \alpha, \quad n = 1, \ldots, m, \tag{18}$$

which generalizes (13). Relations (10) and (11) yield, for the nth-largest trend $x \equiv x_{\max,\,\text{n}}$, the significance condition

$$p^{(m)}(x) \leq 1 - (1-\alpha)^{\frac{1}{m+1-n}}, \; n = 1, \ldots, m. \tag{19}$$

Note that for the smallest relative trend $x = x_{\max,\text{m}}$, (19) reduces to (1). Inequality (19) is less conservative than the well known Holm–Bonferroni (Holm 1979) correction $p_{\nu}^{(m)} \leq \alpha/(m+1-n)$, which (like the Bonferroni correction) represents upper bounds for the p values in uncorrelated or positively correlated data. For an application of the Holm–Bonferroni correction to the significance of the trends in seasonal records, we refer to (Ludescher et al. 2017).

For short- and long-term persistent data, we found (see Fig. 3) that in analogy to (15),

$$p_{\max}^{(m+1-n)}(x) \simeq 1 - \left(1 - p^{(m)}(x)\right)^{m_{\text{eff}}(m+1-n)}, \tag{20}$$

with an effective exponent $m_{\text{eff}}(m+1-n)$. By definition, $m_{\text{eff}}(1) = 1$. For $m = 12$, the p value functions for the three largest trends nearly coincide.We like to note that in persistent data, in contrast to i.i.d. data, $p_{\max}^{(m+1-n)}$ and $m_{\text{eff}}(m+1-n)$ do not depend only on $(m+1-n)$, but also explicitly on m. For simplicity, and since we are interested only in two values of m, $m = 4$ and 12, we have dropped this dependency in the notation.

By combining (20) and (18) we obtain the significance condition

$$p^{(m)}(x_{\max,\,\text{n}}) \leq 1 - (1-\alpha)^{1/m_{\text{eff}}(m+1-n)}, \tag{21}$$

which is identical to (19) when $m+1-n$ is substituted by $m_{\text{eff}}(m+1-n)$.

The significance condition (21) is one of our central results. It substitutes the condition (1) in uncorrelated, as well as short- and long-term persistent records, and holds for all relative trends, from the largest $(n = 1)$

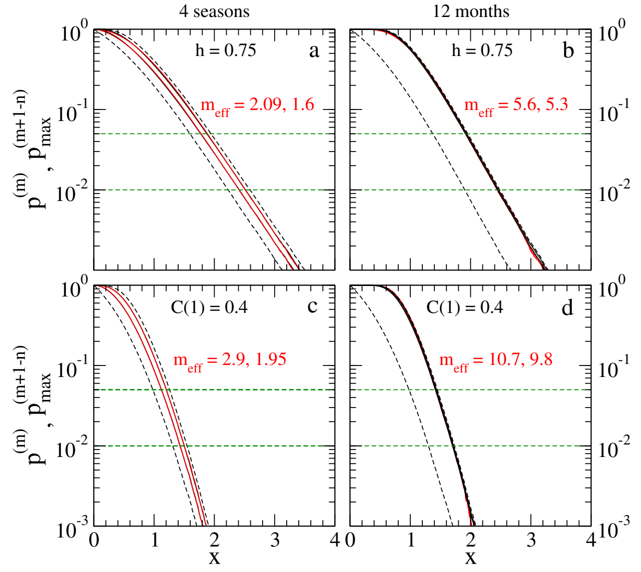

Fig. 3 Adjusted p values $p_{\max}^{(m+1-n)}(x)$ of the 2nd and 3rd largest trends (full lines, red and black) compared with $p^{(m)}(x)$ and $p_{\max}^{(m)}(x)$ from Fig. 2 (dashed lines). The figure shows, for the same data as in Fig. 2, that $p_{\max}^{(m+1-n)}$ is related to $p^{(m)}(x)$ by the power law (21)

to the smallest $(n = m)$ one. For uncorrelated data, $m_{\text{eff}}(m + 1 - n) = m + 1 - n$.

For the short-term persistent monthly data with $C(1) = 0.4$, the values of $m_{\text{eff}}(m + 1 - n)$ are close to $(m + 1 - n)$: $m_{\text{eff}}(4) \approx 3.8$, $m_{\text{eff}}(3) \approx 2.9$, and $m_{\text{eff}}(2) \approx 1.95$ for $m = 4$, and $m_{\text{eff}}(12) \approx 11.7$, $m_{\text{eff}}(11) \approx 10.7$, and $m_{\text{eff}}(10) \approx 9.8$ for $m = 12$.

For the long-term persistent data, the values of the effective exponents $m_{\text{eff}}(m + 1 - n)$ are presented in Fig. 4 and Table 2, for the three largest trends. In addition to $h = 0.75$ chosen in Fig. 2, also the results for other relevant Hurst exponents between 0.5 and 1 are listed. For white noise $(h = 0.5)$, $m_{\text{eff}}(m + 1 - n) = m + 1 - n$. As expected, $m_{\text{eff}}(m + 1 - n)$ decreases with increasing h. For continental temperature data, where the typical h values are between 0.6 and 0.75, $m_{\text{eff}}^{(m)}$ is roughly between 3.6 and 2.4 for the 4 meteorological seasons, and between 11 and 5.9 for the 12 months.

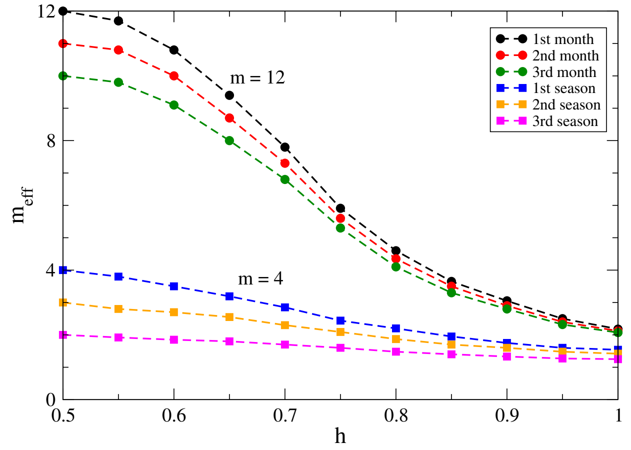

Fig. 4 m_{eff} values of the three largest relative trends $(n = 1, 2, 3)$ as a function of the Hurst exponent h that specifies the strength of the long-term persistence, for $m = 12$ (months) and $m = 4$ (meteorological seasons). For $h = 1/2$, we show the values for white noise: m, $m - 1$, and $m - 2$. For increasing h, all m values tend to unity. The length L of the seasonal records is 50 years

6 The step-down procedure

Adopting the arguments of Holm (Holm 1979) to seasonal climate records, the season with the largest relative trend in a record must be the most significant one. If this trend, for fixed but arbitrary α, turns out not to be statistically significant by (13), then all seasons with lower relative trends also cannot be statistically significant. More generally, when the n-th largest relative trend is not significant, i.e., does not satisfy the condition (18), then none of the lower trends can be significant, i.e., the condition

$$p_{\max}^{(m-n)}(x_{\max,n+1}) \geq p_{\max}^{(m+1-n)}(x_{\max,n}), \quad n = 1, \dots, m, \tag{22}$$

must hold. Accordingly, when the adjusted p value of the $(n + 1)$-th largest trend $x_{\max,n+1}$ is found to be below the adjusted p value of the nth largest trend $x_{\max,n}$ (which may happen when both trends are very close in magnitude), one corrects this by setting $p_{\max}^{(m-n)}(x_{\max,n+1})$ equal to $p_{\max}^{(m+1-n)}(x_{\max,n})$.

Table 2 List of the effective m values m_{eff} shown in Fig. 3 as a function of the Hurst exponent h

(m,m+1–n)\h	0.55	0.6	0.65	0.7	0.75	0.80	0.85	0.90	0.95	1.0
(4,4)	3.8	3.5	3.19	2.85	2.44	2.2	1.95	1.75	1.6	1.54
(4,3)	2.8	2.7	2.55	2.3	2.09	1.87	1.7	1.6	1.48	1.42
(4,2)	1.92	1.85	1.8	1.7	1.6	1.48	1.4	1.33	1.27	1.25
(12,12)	11.7	10.8	9.4	7.8	5.91	4.6	3.65	3.05	2.5	2.18
(12,11)	10.8	10	8.7	7.3	5.6	4.35	3.5	2.9	2.4	2.12
(12,10)	9.8	9.1	8.0	6.8	5.3	4.1	3.3	2.8	2.32	2.07

7 Analytical formula for the adjusted *p* values

In the following, we focus on long-term persistent data since they play a central role in geoscience. Figures 2 and 3 show that in the most relevant regime $10^{-1} < p_{\max}^{(m+1-n)} < 10^{-3}$, the adjusted p values $p_{\max}^{(m+1-n)}(x)$ approximately follow a straight line in the semi-logarithmic plot, i.e.,

$$p_{\max}^{(m+1-n)}(x) \cong 0.05\big(1/5\big)^{\frac{x-x_{0.05}}{x_{0.01}-x_{0.05}}}, n = 1, \dots , m, \qquad (23)$$

where the quantiles $x_{0.05}$ and $x_{0.01}$ are defined by $p_{\max}^{(m+1-n)}(x_{0.05}) = 0.05$ and $p_{\max}^{(m+1-n)}(x_{0.01}) = 0.01$ and depend, for fixed m, on n. We have verified that (23) holds quite generally for long-term persistent data characterized by Hurst exponents between 0.55 and 1, for the same length 50y.

Figure 5 shows $x_{0.05}$ and $x_{0.01}$ as a function of the Hurst exponent h, for $m = 4$ with $n = 1, 2, 3, 4$ and $m = 12$ with $n = 1, 2, 3, 12$. For convenience, we have also listed the values of all quantiles in Table 3. As expected from Fig. 3, for $m = 12$ the quantiles nearly coincide for the 3 largest trends ($n = 1, 2, 3$). While the quantiles increase strongly with h, they increase only comparatively weakly with m. For the maximum trend, the quantiles only depend very weakly on m for $h \geq 0.75$.

Table 3 allows a quick and efficient check for the significance of a trend: The nth largest trend $x_{\max,\,n}$ is significant, when it is between its quantiles $x_{0.05}$ and $x_{0.01}$, and highly significant, when it is above $x_{0.01}$, provided that all larger trends are also significant resp. highly significant (see previous Section). The approximate adjusted p value of $x_{\max,\,n}$ can be obtained from (23).

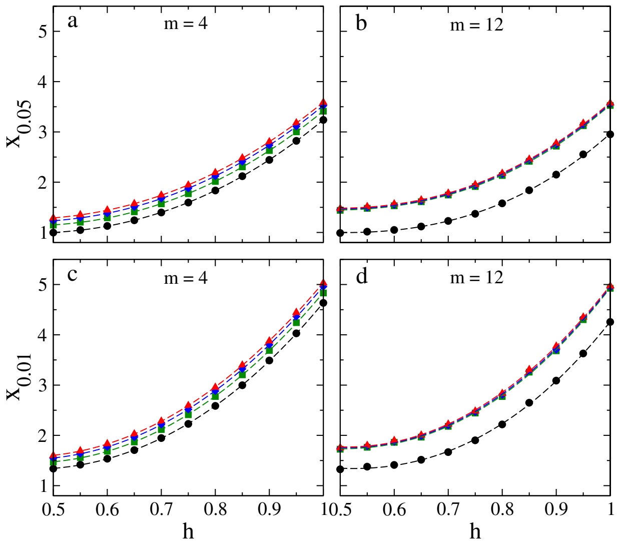

Fig. 5 Dependence of the quantiles $x_{0.05}$ and $x_{0.01}$ on the Hurst exponent h. **a** and **c** show $x_{0.05}$ and $x_{0.01}$ for the 3 meteorological seasons with the largest relative trends (upper 3 curves) and the meteorological season with the smallest relative trend (lowest curve). **b**, **d**: same as (**a**, **c**), but for months. The length L of the data is 50 years

8 Application to climate records

In general, the first step is to analyze the persistence of the considered daily or monthly record $y(i)$ (see Sect. 2). Usually, one of the following two cases occurs:

Table 3 The numerical values of the quantiles $x_{0.05}$ and $x_{0.01}$ shown in Fig. 4 as a function of the Hurst exponent h

(m,m+1–n) \h	0.5	0.55	0.6	0.65	0.7	0.75	0.80	0.85	0.90	0.95	1.0
$x_{0.05}$: (4,4)	1.287	1.352	1.445	1.572	1.742	1.942	2.189	2.482	2.803	3.179	3.588
(4,3)	1.230	1.294	1.381	1.508	1.675	1.875	2.121	2.409	2.735	3.110	3.523
(4,2)	1.147	1.205	1.292	1.411	1.570	1.772	2.015	2.305	2.629	3.002	3.413
(4,1)	0.998	1.049	1.129	1.243	1.399	1.595	1.835	2.120	2.444	2.823	3.238
$x_{0.01}$: (4,4)	1.599	1.693	1.833	2.028	2.281	2.592	2.956	3.398	3.877	4.445	5.031
(4,3)	1.546	1.639	1.772	1.964	2.215	2.516	2.883	3.322	3.802	4.365	4.959
(4,2)	1.469	1.560	1.684	1.870	2.117	2.411	2.773	3.204	3.689	4.242	4.832
(4,1)	1.339	1.418	1.537	1.707	1.945	2.229	2.587	2.997	3.490	4.029	4.634
$x_{0.05}$: (12,12)	1.474	1.513	1.567	1.646	1.781	1.953	2.176	2.456	2.774	3.170	3.573
(12,11)	1.460	1.497	1.550	1.627	1.762	1.935	2.158	2.439	2.747	3.146	3.550
(12,10)	1.440	1.480	1.531	1.612	1.745	1.914	2.133	2.415	2.717	3.124	3.527
(12,1)	0.991	1.015	1.050	1.200	1.229	1.372	1.579	1.841	2.151	2.552	2.950
$x_{0.01}$: (12,12)	1.755	1.799	1.897	2.005	2.222	2.482	2.833	3.309	3.773	4.351	4.970
(12,11)	1.743	1.786	1.874	1.987	2.198	2.465	2.807	3.283	3.706	4.337	4.943
(12,10)	1.724	1.764	1.861	1.965	2.180	2.444	2.773	3.258	3.683	4.301	4.925
(12,1)	1.326	1.377	1.411	1.515	1.668	1.903	2.217	2.651	3.088	3.629	4.255

(a) $y(i)$ shows no persistence. Then we can proceed with (2) and (3) and use (14) and (17) to determine the adjusted p-values.

(b) $y(i)$ is persistent and can be characterized by

$$y(i+1) = by(i) + \zeta_h(i). \tag{24}$$

When $y(i)$ is short-term persistent, then $\zeta_h(i) \equiv \zeta(i)$ represents Gaussian white noise. When $y(i)$ is purely long-term persistent, then $b = 0$ and $\zeta_h(i)$ represents long-term correlated noise with Hurst exponent h. In some cases, as for example, for the Antarctic sea ice extent, the records show both short- and long-term persistence (Yuan et al. 2017; Ludescher et al. 2019).

Equation (24) can be used to generate a large number K of surrogate records $\{y(i)\}$ as described in Sect. 2. In each record, we identify the m seasons and determine their maximum trend $x_{\max}$. This way, we obtain a set of K $x_{\max}$ values.

To determine the adjusted p value $p_{\max}^{(m)}$ in the most efficient way, we follow Sect. 2 and order the K $x_{\max}$ values in descending order, such that $x_{\max}(1) < x_{\max}(2) < \cdots < x_{\max}(K)$. Then, by definition,

$$p_{\max}^{(m)}\left(x = x_{\max}(k)\right) = (K - k + 1)/K \tag{25}$$

for any $k = 1, \ldots, K$, which allows us to determine the desired $p_{\max}^{(m)}(x)$ and the related quantiles.

For obtaining the adjusted p-value of the n-th largest trend $p_{\max}^{(m+1-n)}$, we disregard the first n seasons in all N records and determine the maximum values of the remaining $m + 1 - n$ trends. Then we proceed as above.

In Sects. 5 and 7, we have followed this procedure to obtain the adjusted p values and the related quantiles in both short- and long-term persistent records of length $L = 50y$.

The standard approximation cannot be corrected this way. One way is to use the Bonferroni correction (Bonferroni 1936), where $\tilde{\alpha}$ in (1) is substituted by $\tilde{\alpha}/m$. However, even within this large correction, the results are too liberal for short records, as we have shown in Sect. 3.

9 Trends in Antarctic temperature records

For an important application, we finally apply our results to the temperature trends in Antarctic station data. The warming patterns of Antarctica have recently received a lot of attention which is especially related to the fact that the huge West Antarctic ice sheets belong to the crucial tipping elements in the Earth system (Kriegler et al. 2009; Lenton et al. 2008, 2019).

While ocean warming impacts on ice shelf melting [e.g., (Pritchard et al. 2012)], it is still crucial to carefully monitor and project the temperature trends on and around the southern continent. Recently, Turner et al. (2019) have comprehensively documented, analyzed and interpreted temperature data from 17 stations in Antarctica which all have near-continuous records of more than 30 years in length. Five of the stations (Vernadsky, Esperanza, Marambio, Rothera and Bellingshausen) are at the Antarctic Peninsula or nearby, one station (Orcadas) is on a sub-Antarctic island, and one station (Scott Base) is located South of the Ross Sea. The other 10 stations are located in East Antarctica: Amundsen-Scott is a plateau station located at the South Pole, Vostok is located halfway between the South Pole and the coastline, and 8 stations (Novolazarevskya, Syowa, Mawson, Davis, Mirny, Casey, Dumont d'Urville, and Neumayer) are located along the long coastline of East Antarctica.

In their analysis of the temperature data of these stations, Turner et al. studied annual trends ($m = 1$) and the trends of the 4 seasonal records austral winter, spring, summer, and fall, (i) in a range of 40 years between 1979 and 2018 and (ii) over the full length of the records. Most record lengths vary around 60 years, an exception is Orcadas with 117 years. In both time ranges, they used the standard approximation (2) with (3) and (6) to determine the p values of the trends and used (1) to decide whether the trends were within the bounds of their natural variability or not.

They concluded that 7 stations (Novolazarevskya, Vostok, Scott Base, Rothera, Vernadsky, Esperanza, and Orcadas) showed at least one highly significant warming trend, while 1 station (Dumont d'Urville) showed at least one highly significant cooling trend. Four stations (Casey, Bellingshausen, Amundsen–Scott, and Marambio) showed only a significant warming trend. It is interesting to note that at two stations (Novolazarevskya and Esperanza), the trends were found highly significant only over the full length of the record (57 and 73 years, respectively). Over the shorter period of 40 years none of both trends was found significant.

It is clear from our discussion of the family wise error rate (Fig. 1) that these estimations cannot be correct since the standard approximation considerably overestimates the number of significant records. If the Antarctic temperature data were short-term persistent, we would have to follow the previous Section to obtain the adjusted p values numerically (see Fig. 2). However, since the Antarctic temperature records are long-term persistent, as has been demonstrated in great detail in (Yuan et al. 2015; Ludescher et al. 2016), it is quite easy to obtain the relevant adjusted p values and the relevant bounds of natural variability $x_{0.05}$ and $x_{0.01}$ directly from Fig. 5 and Table 3.

To apply Fig. 5 and Table 3, we have focused on the past 50 years. For Dumont d'Urville and Scott Base we had to consider earlier 50 years sets due to missing data. For Neumayer and Rothera, we took the longest available data set ending in November 2020 and determined the p values numerically. All data are from the Reference Antarctic Data for Environmental Research (READER) dataset (Turner et al. (2004), READER 2021).

First, we determined by DFA2 the Hurst exponents h of all temperature records (3rd column in Table 4.). Then we determined, as described in the Introduction, the trends and the relative trends x (note that the relative trends are positively defined). The results are listed in Table 4. Next, we used Fig. 5 with Table 3 to find the values of $x_{0.05}$ and $x_{0.01}$ and compared them with the relative trends x to see whether a trend was not significant, significant or highly significant.

For example, at Vernadsky station, the Hurst exponent is $h = 0.80$. For the annual data, the trend is 1.457°C. The relative trend x is 1.343, well below $x_{0.05} = 2.41$. Thus the annual trend is not significant. The strongest seasonal trend occurs in austral winter, with $x = 1.379$. Since $x_{0.05}(4, 1) = 2.189$, this trend is not significant and thus, none of the 4 meteorological seasons at the Vernadsky station has a significant warming trend.

In Dumont-d'Urville where $h = 0.65$, the annual temperature decreased in the last 50y considered by – 0.709°C with a relative trend $x = 1.124$. Since $x_{0.05} = 1.58$, this trend is not significant. The largest cooling trend occurs in austral fall, where $x = 1.777$. Since x is between $x_{0.05}(4, 1) = 1.572$ and $x_{0.01}(4, 1) = 2.028$, this trend is significant. The 2nd largest trend where $x = 0.667$ is not significant since $x_{0.05}(4, 2) = 1.508$.

In Vostok ($h = 0.55$), the annual trend ($x = 1.05$) is well below $x_{0.05} = 1.21$ and thus not significant. But the warming in spring ($x = 1.724$) is highly significant, since it is above $x_{0.01} = 1.693$. The 2nd largest trend where $x = 0.414$ is not significant since it is well below $x_{0.05}(4, 2) = 1.294$.

In Esperanza where $h = 0.64$, only the austral summer trend ($x = 1.653 > x_{0.05}(4, 1) = 1.55$) is significant, the second largest trend with $x = 1.439 < x_{0.05}(4, 2) = 1.483$ is not significant.

Finally, in Scott Base only the large spring trend ($x = 1.61 > x_{0.05}(4, 1) = 1.43$) is significant. All other trends are comparatively small and not significant.

Accordingly, in the last 50 years the annual trends of all Antarctic stations were not significant. The island station of Esperanza shows a significant warming trend in austral fall, Dumont-d'Urville showed a significant cooling trend in austral spring, and two stations, Vostok and Scott-Base, show a highly significant resp. significant warming trend in austral spring.

To obtain a more comprehensive picture, we also averaged the 5 Peninsula records (Vernadsky, Esperanza, Marambio, Rothera and Bellingshausen). In the resulting Peninsula record, none of the trends were significant. For the annual data, $x = 1.647$ is well below $x_{0.05} = 1.898$, with $p \approx 0.084$. The season with the largest relative trend is austral fall with $x = 1.71$, which is below

Table 4 Regarded period (December–November), Hurst exponent h, relative trends x and trend magnitudes Δ of 17 Antarctic temperature records during the past 50 years for annual data and the four seasons

Station	Period	Hurst	Annual	DJF	MAM	JJA	SON
Amundsen–Scott	1970–2020	0.53	0.905/0.756	0.436/0.595	0.225/0.328	0.446/0.790	0.825/1.310
Bellingshausen	1970–2020	0.71	1.003/0.704	0.353/– 0.192	0.895/0.811	1.059/1.896	0.349/0.302
Casey	1970–2020	0.66	1.033/– 0.797	0.959/– 0.596	0.602/– 0.932	1.290/– 2.117	0.377/0.457
Vernadsky	1970–2020	0.80	1.343/1.457	0.130/0.076	1.352/1.057	1.379/3.405	0.879/1.290
Davis	1970–2020	0.64	0.192/0.164	0.257/– 0.166	0.138/– 0.225	0.038/– 0.071	0.848/1.117
Syowa	1970–2020	0.71	0.298/0.248	0.057/0.036	0.190/– 0.242	0.579/1.004	0.168/0.192
Mirny	1970–2020	0.64	0.186/– 0.131	0.489/– 0.374	0.508/– 0.683	0.248/– 0.377	0.761/0.910
Mawson	1970–2020	0.61	0.294/0.219	1.183/– 0.797	0.103/0.141	0.394/0.693	0.834/0.841
Orcadas	1970–2020	0.67	1.289/1.024	0.747/0.425	0.466/0.584	1.024/2.037	0.808/1.050
Novolazarevskaya	1970–2020	0.58	0.290/0.170	0.195/– 0.140	0.449/– 0.451	0.457/0.704	0.588/0.568
Esperanza	1970–2020	0.64	1.468/1.523	**1.653***/1.251	1.439/2.783	0.581/1.233	0.538/0.823
Marambio	1970–2020	0.61	1.317/1.557	1.220/1.114	1.337/3.042	0.518/1.291	0.414/0.783
Dumont	1964–2014	0.65	1.124/– 0.709	0.667/– 0.438	**1.777***/– 1.729	0.648/– 1.026	0.456/0.356
Vostok	1970–2020	0.55	1.056/0.914	0.414/0.405	0.093/0.146	0.374/0.732	**1.724****/2.372
Rothera	1977–2020	0.79	1.402/1.694	0.197/0.121	1.730/2.056	1.302/3.402	0.816/1.198
Scott-Base	1966–2016	0.59	0.573/0.554	0.110/– 0.087	0.401/– 0.716	0.152/0.322	**1.611***/2.695
Neumayer	1981–2020	0.56	0.455/– 0.282	0.122/0.104	0.521/– 0.675	0.833/– 1.525	0.737/0.968
Peninsula	1970–2020	0.70	1.647/1.399	1.132/0.539	1.710/1.999	1.116/2.221	0.694/0.839
East Antarctica	1970–2020	0.66	0.102/0.056	0.241/– 0.150	0.373/-0.358	0.144/– 0.178	1.204/0.910

Significant trends (in boldface and marked by one asterisk) occur in Esperanza ($p = 0.034$), Dumont ($p = 0.024$), and Scott-Base ($p = 0.023$). The only highly significant trend (marked by 2 asterisks) occurs in Vostok ($p = 0.0086$). In the averaged data (Peninsula and East Antarctica), none of the trends is significant

$x_{0.05}(4, 1) = 1.742$. The corresponding p value obtained from (23) is $p = 0.055$.

To obtain an East-Antarctica record, we averaged the rest of the stations, apart from Orcadas and Scott-Base. We found that all trends were far from being significant. For the annual data, $x = 0.102$ is well below $x_{0.05} = 1.71$, with a p value close to 1. The season with the largest relative trend is austral spring. Its relative trend $x = 1.204$ is well below $x_{0.05} \approx 1.66$. The exact simulated p value is 0.20.

10 Discussion and conclusions

In this article, we have shown quite generally how to determine whether a trend in a seasonal record is significant or not. Our results are quite general and hold for all kinds of persistence, for Gaussian white noise, as well as for short- and long-term persistent records. We discussed in great detail the standard approximation, which because of its simplicity is very popular among climate scientists and showed that it considerably overestimates the significance, even in short-term persistent records and when only annual data are considered.

The standard approach can also not be applied to long-term persistent records that play an eminent role in climate science. We showed how to determine numerically the relevant (adjusted) p values of the trends. Specifically for records of length 50y, we determined numerically the quantiles $x_{0.05}$ and $x_{0.01}$ and listed them in a table as a function of the Hurst exponent h. Thus, when the Hurst exponent of the record of interest is known, the table allows to find out without much effort, whether a trend is not significant, significant or highly significant.

We applied our analysis to temperature data from Antarctica, but the same procedure can also be applied to other climate records. Examples are river flows which are also long-term persistent [see the pioneering work by Hurst (Hurst 1951)], or the Antarctic sea ice extent (Yuan et al. 2017; Ludescher et al. 2019), which shows both short- and long-term persistence. In principle, the significance of trends in precipitation records (where the memory is not as pronounced as in temperature records) or in drought records (Palmer 1965; Cook et al. 2007; Griffin and Anchukaitis 2014) can be also studied within our approach. However, since precipitation records also show nonlinear (multifractal) correlations (Koscielny-Bunde et al. 2006; Lovejoy and Schertzer 2013), which the models presented here do not account for, results based on the persistence models discussed here can only yield first order approximations for the significance of the trends. We believe that also the question to which extent the recent decrease of the ozone hole (Varotsos and Kirk-Davidoff 2006; Solomon et al. 2016; Kuttippurath and Nair 2017) is significant can be studied by our formalism.

Finally, we like to note that in this article, we discussed purely statistical models for the natural internal climate variability. These models utilize the persistence properties of climate records directly and provide bounds for the natural variability, as well as estimates for the statistical significance of observed trends. Another avenue to estimate the natural internal variability, and thus to detect possible anthropogenic trends, provide general circulation climate models (GCMs) (Bindoff et al. 2013). The GCMs simulate physical processes directly and are employed to estimate natural climate variability. We like to emphasize that the issue of multiple testing in seasonal records, discussed here, applies equally to trend significance estimations by GCMs. Thus to correctly determine the statistical significance of observed seasonal trends and to avoid false discoveries, analogous approaches and corrections as presented here are necessary.

Acknowledgements We like to thank Igor Sokolov for valuable discussions. JL thanks for the support by the East Africa Peru India Climate Capacities (EPICC) project. This project is part of the International Climate Initiative (IKI) of the Federal Ministry for the Environment, Nature Conservation and Nuclear Safety (BMU).

Funding Open Access funding enabled and organized by Projekt DEAL.

Data availability The Antarctic temperature data analyzed in this study are available at the Reference Antarctic Data for Environmental Research (READER) website, www.antarctica.ac.uk/met/READER.

Declarations

Conflict of interest The authors declare that they have no conflict of interest.

References

Bindoff NL et al. (2013) Detection and attribution of climate change: from global to regional. In: Stocker TF, Qin D, Plattner GK, Tignor M, Allen SK, Boschung J, Nauels A, Xia Y, Bex V, Midgley PM (eds) Climate Change 2013: The Physical Science Basis. Contribution of Working Group I to the Fifth Assessment Report of the Intergovernmental Panel on Climate Change. Cambridge University Press, Cambridge, and New York

Blender R, Fraedrich K (2003) Long time memory in global warming simulations. Geophys Res Lett 30(14):1769

Blender R, Raible CC, Lunkeit F (2015) Non-exponential return time distributions for vorticity extremes explained by fractional Poisson processes. Q J Roy Meteorol Soc 141(686):249–257

Blesic S (2020) Applications of statistical physics to study climate phenomena and contribute to overall adaptation efforts. Europhys Lett 132(2):20004

Blesic S, Zanchettin D, Rubino A (2019) Heterogeneity of scaling of the observed global temperature data. J Clim 32(2):349

Bogachev MI, Bunde A (2012) Universality in the precipitation and river runoff. EPL (Europhys Lett) 97(4):48011

Bonferroni CE (1936) Teoria statistica delle classi e calcolo delle probabilita. Pubblicazioni del Istituto Superiore di Scienze Economiche e Commerciali di Firenze 8:3–62

Bromwich DH et al (2014) Central West Antarctica among the most rapidly warming regions on earth. Nat Geosci 6(2):139–145

Bronstein IN, Semendyayev KA, Musiol G, Mühlig H (2004) Handbook of mathematics, 4th edn. Springer, Berlin

Bunde A, Bogachev MI, Lennartz S (2012) Precipitation and river flow: long-term memory and predictability of extreme events. Extrem Events Nat Hazards Complexity Perspect 196:139–152

Bunde A, Ludescher J, Franzke C, Büntgen U (2014) How significant is West Antarctic warming? Nat Geosci 7(4):246–247

Caldwell H et al. (2002) Potential impacts of climate change on freight transport. In: The Potential Impacts of Climate Change on Transportation, Workshop Summary, US Dept. of Transportation, Washington DC

Chapman WL, Walsh JE (2007) A synthesis of Antarctic temperatures. J Clim 20:4096–4117

Cohn TA, Lins HF (2005) Nature's style: naturally trendy. Geophys Res Lett 32:L23402

Cook ER, Seager R, Cane MA, Stahle DW (2007) North American drought: reconstructions, causes, and consequences. Earth Sci Rev 81(1–2):93–134

Dangendorf S et al (2014) Evidence for long-term memory in sea level. Geophys Res Lett 41(15):5530–5537

Eichner JF, Koscielny-Bunde E, Bunde A, Havlin S, Schellnhuber HJ (2003) Power-law persistence and trends in the atmosphere: a detailed study of long temperature records. Phys Rev E 68(4):046133

Franzke C (2010) Long-range dependence and climate noise characteristics of Antarctic temperature data. J Clim 23(22):6074–6081

Griffin D, Anchukaitis KJ (2014) How unusual is the 2012–2014 California drought? Geophys Res Lett 41(24):9017–9023

Hartmann DL et al. (2013) Observations: atmosphere and surface. In: Stocker TF et al. (eds) Climate Change 2013: The Physical Science Basis. Contribution of Working Group I to the Fifth Assessment Report of the Intergovernmental Panel on Climate Change. Cambridge University Press, Cambridge, and New York

Holm S (1979) A simple sequentially rejective multiple test procedure. Scand J Stat 6:65–70

Hurst HE (1951) Long-term storage capacity of reservoirs. Trans Am Soc Civil Eng 116:770–799

IPCC (2014), Climate Change 2014: Impacts, adaptation, and vulnerability. In: Field CB, Barros VR, Dokken DJ, Mach KJ, Mastrandrea MD, Bilir TE, Chatterjee M, Ebi KL, Estrada YO, Genova RC, Girma B, Kissel ES, Levy AN, MacCracken S, Mastrandrea PR, White LL (eds) Contribution of Working Group II to the Fifth Assessment Report of the Intergovernmental Panel on Climate Change. Cambridge University Press, Cambridge and New York

Jones PD, Lister DH (2014) Antarctic near-surface air temperatures compared with ERA-Interim values since 1979. Int J Climatol 35(7):1354–1366

Kantelhardt JW, Koscielny-Bunde E, Kantelhardt JW, Rego HH, Havlin S, Bunde A (2001) Detecting long-range correlations with detrended fluctuation analysis. Phys A 295:441–454

Kiraly A, Bartos I, Janosi I (2006) Correlation properties of daily temperature anomalies over land. Tellus A 58(5):593–600

Koscielny-Bunde E et al (1998) Indication of a universal persistence law governing atmospheric variability. Phys Rev Lett 81(3):729–732

Koscielny-Bunde E, Kantelhardt JW, Braun P, Bunde A, Havlin S (2006) Long-term persistence and multifractality of river runoff and precipitation records. J Hydrol 322(1–4):120–137

Kriegler E, Hall JW, Held H, Dawson R, Schellnhuber HJ (2009) Imprecise probability assessment of tipping points in the climate system. Proc Natl Acad Sci 106(13):5041–5046

Kuttippurath J, Nair PJ (2017) The signs of Antarctic ozone hole recovery. Sci Rep 7:585

Lennartz S, Bunde A (2009a) Trend evaluation in records with long-term memory: application to global warming. Geophys Res Lett 36:L16706

Lennartz S, Bunde A (2009b) Eliminating finite-size effects and detecting the amount of white noise in short records with long-term memory. Phys Rev E 79:066101

Lennartz S, Bunde A, Turcotte DL (2011) Modelling seismic catalogues by cascade models: do we need long-term magnitude correlations? Geophys J Int 184(3):1214–1222

Lenton TM et al (2008) Tipping elements in the earth's climate system. Proc Natl Acad Sci 105(6):1786–1793

Lenton TM, Rockström J, Gaffney O, Rahmstorf S, Richardson K, Steffen W, Schellnhuber HJ (2019) Climate tipping points-too risky to bet against. Nature 575:592–595

Livina VN, Havlin S, Bunde A (2005) Memory in the Occurrence of Earthquakes. Phys Rev Lett 95:208501

Lovejoy S, Schertzer D (2013) The weather and climate: emergent laws and multifractal cascades. Cambridge University Press, Cambridge

Ludescher J, Bunde A, Franzke CL, Schellnhuber HJ (2016) Long-term persistence enhances uncertainty about anthropogenic warming of Antarctica. Clim Dyn 46(1):263–271

Ludescher J, Bunde A, Schellnhuber HJ (2017) Statistical significance of seasonal warming/cooling trends. Proc Natl Acad Sci 114(15):E2998–E3003

Ludescher J, Yuan N, Bunde A (2019) Detecting the statistical significance of the trends in the Antarctic sea ice extent: an indication for a turning point. Clim Dyn 53(1):237–244

Ludescher J, Bunde A, Büntgen U, Schellnhuber HJ (2020) Setting the tree-ring record straight. Clim Dyn 55(11):3017–3024

Malamud BD, Turcotte DL (1999) Self-affine time series I: generation and analyses. Adv Geophys 40:1–90

Mandelbrot BB, Wallis JR (1968) Noah, Joseph, and operational hydrology. Water Resour Res 4(5):909–918

Mitchell JM et al (1966) Climatic change, technical note 79. World Meteorological Organization, Geneva

Monaghan AJ, Bromwich DH, Chapman W, Comiso JC (2008) Recent variability and trends of Antarctic near-surface temperature. Geophys Res 113:D04105

Mudelsee M (2007) Long memory of rivers from spatial aggregation. Wat Resour Res 43:W01202

O'Donnel R, Lewis N, McIntyre S, Condon J (2011) Improved methods for PCA-based reconstructions: case study using the Steig et al. (2019) Antarctic temperature reconstruction. J Clim 24(8):2099–2115

Palmer WC (1965) Meteorological drought. US Department of Commerce, Weather Bureau, Research paper no, p 45

Pritchard HD et al (2012) Antarctic ice sheet loss driven by basal melting of ice shelves. Nature 484(7395):502–505

Rao MS et al (2015) Model and scenario variations in predicted number of generations of *Spodoptera litura* Fab. on peanut during future climate change scenario. PLoS One 10(2):e0116762

Reference Antarctic Data for Environmental Research (READER). www.antarctica.ac.uk/met/READER. Accessed 15 Jan 2021

Rübbelke D, Vögele S (2013) Short-term distributional consequences of climate change impacts on the power sector: who gains and who loses? Clim Change 116(2):191–206

Santer BD et al (2000) Statistical significance of trends and trend differences in layer-average atmospheric temperature time series. J Geophy Res 105:7337–7356

Santhanam MS, Kantz H (2005) Long-range correlations and rare events in boundary layer wind fields. Phys A 345(3–4):713–721

Šidák ZK (1967) Rectangular confidence regions for the means of multivariate normal distributions. J Am Stat Assoc 62(318):626–633

Solomon S, Ivy DJ, Kinnison D, Mills MJ, Neely RR, Schmidt A (2016) Emergence of healing in the Antarctic ozone layer. Science 353(6296):269–274

Steig EJ et al (2009) Warming of the Antarctic ice-sheet surface since the 1957 International Geophysical Year. Nature 457(7228):459–462

Tamazian A, Ludescher J, Bunde A (2015) Significance of trends in long-term correlated records. Phys Rev E 91:032806

Troost C (2015) Agent-based modeling of climate-change adaptation in agriculture: a case study in the Central Swabian Jura. Dissertation, University of Hohenheim

Turner J et al (2004) The SCAR READER project: toward a high-quality database of mean Antarctic meteorological observations. J Clim 17(14):2890–2898

Turner J et al (2016) Absence of 21st century warming on Antarctic peninsula consistent with natural variability. Nature 535:411–415

Turner J et al (2019) Antarctic temperature variability and change from station data. Int J Climatol 40:2986–3007

Varotsos C, Kirk-Davidoff D (2006) Long memory processes in ozone and temperature variations at the region 60°S- 60°N. Atmos Chem Phys 6:4093–4100

von Storch H, Zwiers FW (1999) Statistical analysis in climate research. Cambridge University Press, Cambridge

Vyushin D, Zhidkov I, Havlin S, Bunde A, Brenner S (2004) Volcanic forcing improves atmosphere-ocean coupled general circulation model scaling performance. Geophys Res Lett 31:L10206

Yuan N et al (2015) On the long-term climate memory in the surface air temperature records over Antarctica: a nonnegligible factor for trend evaluation. J Clim 28:5922–5934

Yuan N, Fu Z, Mao J (2010) Different scaling behaviors in daily temperature records over China. Phys A 389:4087–4095

Yuan N, Ding M, Ludescher J, Bunde A (2017) Increase of the Antarctic sea ice extent is highly significant in the Ross Sea. Sci Rep 7:41096