IDŐJÁRÁS

Quarterly Journal of the Hungarian Meteorological Service
Vol. 117, No. 1, January–March 2013, pp. 1-34

On the multiple breakpoint problem and the number of significant breaks in homogenization of climate records

Ralf Lindau* and **Victor Venema**

Meteorological Institute, University of Bonn
Auf dem Hügel 20, D-53121 Bonn, Germany

**Corresponding author E-mail: rlindau@uni-bonn.de*

(Manuscript received in final form November 8, 2012)

Abstract—Changes in instrumentation and relocations of climate stations may insert inhomogeneities into meteorological time series, dividing them into homogeneous subperiods interrupted by sudden breaks. Such inhomogeneities can be distinguished from true variability by considering the differences compared to neighboring stations. The most probable positions for a given number of break points are optimally determined by using a multiple-break point approach. In this study the maximum external variance between the segment averages is used as decision criterion and dynamic programming as optimization method. Even in time series without breaks, the external variance is growing with any additionally assumed break, so that a stop criterion is needed. This is studied by using the characteristics of a random time series. The external variance is shown to be beta-distributed, so that the maximum is found by solving the incomplete beta function. In this way, an analytical function for the maximum external variance is derived. In its differential form our solution shows much formal similarities to the penalty function used in *Caussinus* and *Mestre* (2004), but differs numerically and exhibits more details.

Key words: Climate records, homogenization, multiple break point detection, stop criterion for search algorithms, dynamic programming, penalty term.

1. Introduction

Multiple-century long instrumental datasets of meteorological variables exist for Europe (*Brunetti et al.*, 2006; *Bergström* and *Moberg*, 2002; *Slonosky et al.*, 2001). Such series provide invaluable information on the evolution of the climate. However, between the Dutch Golden Age, the French and the industrial

revolution, the rise and the fall of communism, and the start of the internet age, inevitably many changes have occurred in climate monitoring practices (*Aguilar et al.*, 2003; *Trewin*, 2010). The typical size of temperature jumps due to these changes is similar to the global warming in the 20th century, and the average length of the periods between breaks in the climate records is 15 to 20 years (*Auer et al.*, 2007; *Menne* and *Williams*, 2009). Clearly, such changes interfere with the study of natural variability and secular trends (*Rust et al.*, 2008; *Venema et al.*, 2012).

Technological progress and a better understanding of the measurement process have led to the introduction of new instruments, screens, and measurement procedures (*MeteoSchweiz*, 2000). In the early instrumental period, temperature measurements were often performed under open shelters or in metal window screens on a North facing wall (*Brunetti et al.*, 2006), which were replaced by Montsouri (*Brunet et al.*, 2011), Wild, and various Stevenson screens (*Nordli et al.*, 1997; *Knowles Middleton*, 1966), and nowadays more and more by labor-saving automatic weather stations (*Begert et al.*, 2005). Every screen differs in their protection against radiation, wetting, as well as their quality of ventilation (*Van der Meulen* and *Brandsma*, 2008). Initially many precipitation observations were performed on roofs. As it was realized that many hydrometeors do not enter the gauge due to wind and turbulence, especially in case of snow, the observations were taken nearer the ground, and various types of wind shields were tested leading to deliberate inhomogeneities (*Auer et al.*, 2005). Due to the same effect, any change in the geometry of a rain gauge can lead to unintended inhomogeneities.

Inhomogeneities are frequently caused by relocations, either because the voluntary observer changed, because the observer had to move or because the surrounding was no longer suited for meteorological observations. Changes in the surrounding can lead to gradual or abrupt changes, for example gradual increases in urbanization or growing vegetation or fast changes due to cutting of vegetation, buildings that disrupt the flow or land-use change.

Changes in the observations should be documented in the station history. It is recommended to perform several years of parallel measurements in case of changes (*Aguilar et al.*, 2003). However, it is not guaranteed that metadata is complete, thus statistical homogenization should always be performed additionally. The dominant approach to homogenize climate networks is the relative homogenization method. This principle states that nearby stations are exposed to almost the same climate signal, and thus, the differences between nearby stations can be utilized to detect inhomogeneities (*Conrad* and *Pollack*, 1950). By computing the difference time series, the interannual weather noise, decadal variability, and secular trends are strongly reduced. Consequently, a jump in single station becomes much more salient.

The two fundamental problems of homogenization are that the nearby stations are also inhomogeneous and that typically more than one break is

present. Recent intercomparison studies by *Domonkos* (2011a) and *Venema et al.* (2012) showed that the best performing algorithms are the ones that attack these two problems directly. This study will focus on the multiple-breakpoint problem.

Traditionally the multiple-breakpoint problem is solved by applying single-breakpoint algorithms multiple times. Either a cutting algorithm is applied: the dataset is cut at the most significant break and the subsections are investigated individually until no more breaks are found or the section become too short; see, e.g., *Easterling* and *Peterson* (1995). A variation on this theme is a semi-hierarchical algorithm, in which potential breakpoints are found using the cutting algorithm, but before correcting a potential break its significance is tested anew (*Alexanderson* and *Moberg*, 1997). According to *Domonkos* (2011a), this improvement has a neutral effect on the efficiency of homogenization.

The first algorithms solving the multiple-breakpoint problem directly are MASH (*Szentimrey*, 1996, 1999) and PRODIGE (*Caussinus* and *Mestre*, 1996, 2004). MASH solves the problem with a computationally expensive exhaustive search (*Szentimrey*, 2007). PRODIGE solves the problem in two steps. First, the optimal position of the breaks for a given number of breaks is found using a computationally fast optimization approach called dynamic programming (*Bellman*, 1954; *Hawkins*, 1972). Second, the number of breaks is determined by minimizing the internal variance within the subperiods between two consecutive breaks plus a penalty for every additional break (*Caussinus* and *Lyazrhi*, 1997). The penalty term aims to avoid adding insignificant breaks.

Recently, *Domonkos* (2011b) expanded ACMANT, which is based on the generic PRODIGE algorithm, by searching for common breaks in the annual mean and the size of the annual cycle. *Picard et al.* (2011) developed an alternative version, in which not only pairs, but all data in the network are jointly taken into account for optimization. ACMANT, PRODIGE, and the joint detection method of *Picard et al.* (2011) are implemented in the software package HOMER (*Mestre et al.*, 2012). *Nemec et al.* (2012) used PRODIGE with three different criteria for the assessment of the number of breaks. Beyond dynamic programming, genetic algorithms (e.g., *Li* and *Lund*, 2012; *Davis et al.*, 2012) and simulated annealing (*Lavielle*, 1998) are alternatively used for reducing the computational demand.

Not all inhomogeneities are abrupt changes, some changes are more gradual (*Lindau*, 2006). Such trends are explicitly considered by some homogenization algorithms (*Vincent*, 1998; *Menne* and *Williams*, 2009). Using the HOME benchmark dataset in which 10% of the stations contained a local trend inhomogeneity, a blind experiment with two versions of PRODIGE has been performed (*Venema et al.*, 2012). In the main version only breaks have been used for homogenization, and in the alternative version multiple breaks in one direction are combined into a slope correction. These two versions have a very similar performance. One of the reasons may be that not many local trends

had to be introduced. Still this suggests that trend inhomogeneities can be reasonably well modeled by multiple breaks. Consequently, this paper will only consider break inhomogeneities.

To characterize breaks within a time series, it is helpful to decompose the total variance of the time series into two terms: the internal and the external variance. Consider a time series with k breaks dividing it into $k + 1$ subperiods (*Fig. 1*). In this concept, the variance within the subperiods is referred to as the internal variance, whereas the variance between the means of different subperiods is the external variance. The decomposition with the maximum external variance defines the optimum positions of breaks for a given number of breaks.

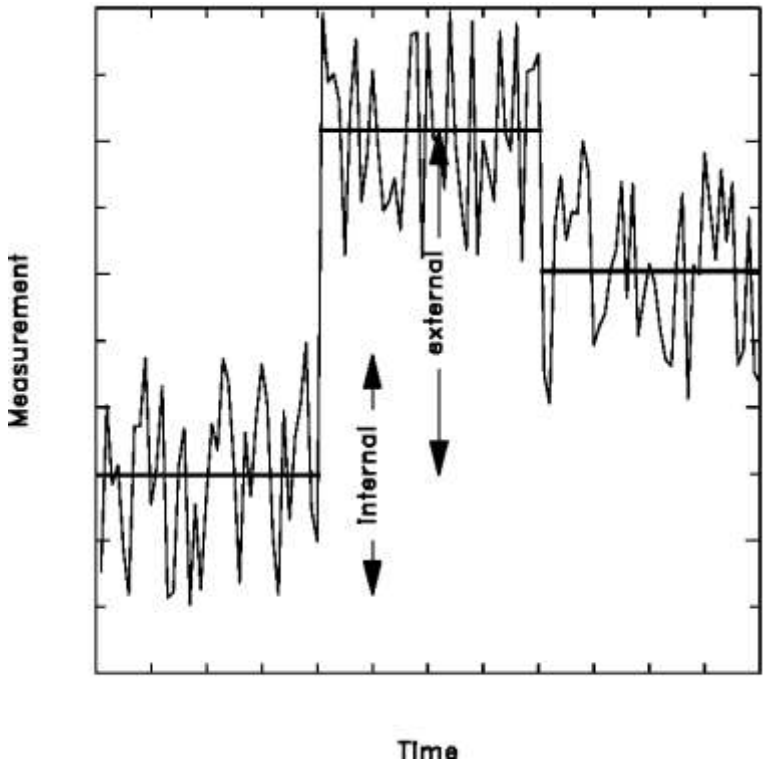

Fig. 1. Sketch to illustrate the occurrence of breaks in climate records and the related expressions, internal and external variance.

As we use internal and external variance as the basic concept to characterize breaks, an exact quantitative formulation is necessary. *Lindau* (2003) discussed the decomposition of variance and showed that the total variance of a time series can be divided into three parts:

$$\frac{1}{n-1}\sum_{i=1}^{N}\sum_{j=1}^{n_j}\left(x_{ij}-\bar{x}\right)^2=\frac{1}{n}\sum_{i=1}^{N}n_i(\bar{x}_i-\bar{x})^2+\frac{1}{n}\sum_{i=1}^{N}\sum_{j=1}^{n_j}\left(x_{ij}-\bar{x}_i\right)^2$$

$$+\frac{1}{n(n-1)}\sum_{i=1}^{N}\sum_{j=1}^{n_j}\left(x_{ij}-\bar{x}\right)^2 \quad . \qquad (1)$$

In Eq. (1), the variance of a time series of length n is considered. It contains N subperiods, each comprising n_i members. Individual members are denoted by x_{ij}, where i specifies the subperiod and j the temporal order within the respective subperiod. The mean of the ith subperiod is denoted by $\bar{x}_i$ and the overall mean of the entire time series by $\bar{x}$, without any index. The total variance on the left hand side is decomposed into the three parts on the right hand side of Eq. (1). These are equal to the external and the internal variance plus, as third term, the error of the total mean. As the last term is constant for a given time series, the sum of internal and external variance is constant, too. Consequently, we can formulate an alternative criterion for the optimum decomposition of a time series into subsegments being a minimum internal variance.

However, two problems arise. The first is of practical nature. The number of possible decompositions is normally too large for a simple test of all permutations. The second is rather fundamental. For a fixed number of breaks, the maximum external variance is actually a reasonable criterion for the optimum decomposition. However, it is obvious that for zero breaks, the entire variance is internal, whereas it is fully external for n–1 breaks. During the transition from 0 to n–1 breaks, more and more variance is converted from internal to external, so that the internal variance is a monotonously falling function of the break number k. Consequently, we need a second criterion for the optimum *number* of breaks. As this is the critical problem for any multiple-breakpoint detection algorithm, the discussion and proposed solution of this problem built the major part of this paper. However, initially also the first minor problem and its solution are shortly described in the following.

There exists a large number of possibilities to decompose a time series of length n into a fixed number of N subsegments: it is equal to $\binom{n-1}{N-1}$. Even for a moderate length of $n = 100$ and ten subsegments, there are already more than 10^{12} combinations, so that the testing of all permutations is mostly not feasible. This problem is already solved by the so-called dynamic programming method, firstly inspired by *Bellman* (1954). Originally designed for economic problems, this method is by now established in many different disciplines, in climate research (*Caussinus* and *Mestre*, 2004) as well as in biogenetics (*Picard et al.*, 2005). As we will also use dynamic programming later on, we describe shortly how we applied this technique.

2. Dynamic programming

We begin with the optimum solution for a single break point. In this case, simple testing of all possibilities is still feasible as only n–1 permutations exist. Afterwards, the best break position together with its respective internal variance is known. The basic idea is now to find an optimum decomposition not only for the entire time series, but also for all truncated variants of any length l. There exist n–1 variants, all beginning with the first time point. The first variant ends at the second time point, the second at the third time point, and the last variant is equal to the entire time series. For each of these variants an optimum position of a single breakpoint is searched and stored together with the criterion on which the decision is made, i.e., its internal variance. In the next step we consider what happens if the truncated variants are filled up to the original length n. In this case the internal variance consists of two contributions: that of the truncated variant, plus that of the added rest. For this step, it is, of course, necessary that the used criterion is additive, which is fulfilled for variances. Consequently, we can test a number of n–1 filled-up variants. That variant, where the combined internal variance is minimal, is then the optimum solution for two break points. The first break is situated within the truncated time series; the second is equal to the length l of the truncated series itself, because here is the break between the two combined time series.

To expand the method from two to three and more breaks, some more work is necessary already at the beginning. So far the truncated variants are always filled up to the entire length n. But the starting point for the proceeding from one to two breaks are, as described above, known previous solutions for all lengths. Consequently, to proceed from two to three breaks, we need not only the best two-break solution for the entire length n, but the solutions for every length. Thus, also all shorter fillings are performed so that we obtain the optimum two-break solution not only for the final time series length n, but also for every shorter length between 2 and n. This set of solutions is then used accordingly as basis to find the three-break solution. Filling up the time series to the full length would be sufficient if we want to stop at three breaks. However, if the method should be continued to higher break numbers, again a full set of three-break solutions is needed.

Thus, the solution for k breaks is found by testing only n–1 truncated and refilled optima, where the truncated part contains already the optimum distribution of k–1 breaks. To perpetuate the method for k+1 breaks, each truncated optimum has to be refilled to all possible length so that a number of cases in the order of n^2 has to be calculated. This reduces the number of cases from the order of $\binom{n}{k}$ to n^2, which facilitates a much faster processing.

3. Outline of the paper

In the above described way, the optimum positions for a given number of breaks can be calculated. Minimum internal variance is serving as criterion, and dynamic programming avoids a time consuming exhaustive search. However, as mentioned above, there is still a problem left. The absolute minimum of internal variance being equal to zero would be attained by inserting n–1 breaks into the time series, which is obviously not the optimum solution. Instead, we need to define which number of breaks is appropriate.

A state-of-the-art method for detecting breaks is PRODIGE (*Caussinus* and *Mestre*, 2004). Although using a log-likelihood method, it is based on the minimization of the internal variance and does not differ essentially from the procedure described here so far. PRODIGE uses a penalty term to ensure that the search stops at a reasonable number of breaks. This penalty term is adopted from *Caussinus* and *Lyazrhi* (1997). Similar to PRODIGE, *Picard et al.* (2005) applied a log-likelihood method to minimize the internal variance, but developed a specific penalty term. Before, they discussed different commonly used penalty terms, such as the Information Criteria AIC and BIC, based on *Akaike* (1973), and found that these penalty terms suffer from different weaknesses.

In the remaining part of this study, we derive an alternative stop criterion based on the idea that the external variance is the key parameter, which defines the optimum solutions. We will use the characteristics of a random standard normal distributed time series as reference. Only if the optimum solution for an additionally inserted break gains significantly more external variance than the expected amount for a random time series, an increased break number is justified. Thus, it is necessary to describe mathematically how the external variance of random data increases with increasing number of breaks, so that it can be used as reference for real data.

In a first step, we derive the statistical distribution that can be expected for the external variance v. In Section 4, we show theoretically that the χ^2 distribution would be a good candidate. In Section 5, we show by empirical tests that the related Beta distribution is even better suited to describe the external variance. To identify the optimum solution for the decomposition, we use, as mentioned, the maximum external variance. Consequently, we have to find the maximum value within a Beta distribution, identical to its exceeding probability, which is performed in Section 6. For that purpose, the definite integral of the Beta distribution, known as the incomplete Beta function, has to be solved. From this formulation, the rate of change of the external variance v for growing break numbers k is derived in Section 7. This derivative dv/dk is then integrated and a formulation for $v(k)$ is presented.

In its differential form our solution shows much formal similarities to the penalty function used in *Caussius* and *Mestre* (2004), but it differs numerically

and exhibits more details. In Section 8 we discuss these differences and propose finally a revision.

4. Theoretical characteristics of random data

Consider a random standard normal distributed time series $N(0,1)$ with k breaks inserted, so that the number of segments N is:

$$N = k + 1 \quad . \tag{2}$$

According to Eq. (1) the external variance v is:

$$v = \frac{1}{n}\sum_{i=1}^{N} n_i(\bar{x}_i - \bar{x})^2 \quad . \tag{3}$$

As standard normal distributed data is considered, $\bar{x} = 0$ and $\sigma_x = 1$. Furthermore, we are interested here in the statistics of external variance for many realizations as produced by $\binom{n-1}{k}$ permutations of break positions for a fixed number of breaks. Averages over these permutations are denoted by brackets, whereas averages over individual data points within a time segment are overlined.

$$[v] = \left[\frac{1}{n}\sum_{i=1}^{N} n_i\,\bar{x}_i^{\,2}\right] \quad . \tag{4}$$

Consider now the segment averages $\bar{x}_i$, which are the critical constituents of $[v]$ in Eq. (4). Their expected mean is equal to zero, since random data with $\bar{x} = 0$ is assumed. Only the finite number of segment members causes the segment means to scatter randomly around zero. As the members of a segment are standard normal distributed, the standard deviation of any segment mean is equal to $1/\sqrt{n_i}$.

$$\bar{x}_i \sim N\left(0, \frac{1}{n_i}\right) \quad . \tag{5}$$

If the segment means are multiplied by the square root of the number of segment members (Eq. (6)), the distribution is broadened in such a way that a standard normal distribution is obtained. These modified means can be defined as to y_i.

$$y_i := \sqrt{n_i}\,\bar{x}_i \sim N(0,1) \quad . \tag{6}$$

Inserting this definition into Eq. (4) leads to:

$$[v] = \left[\frac{1}{n}\sum_{i=1}^{N} {y_i}^2\right] = \frac{N-1}{n} = \frac{k}{n} \quad . \tag{7}$$

The second equal sign in Eq. (7) follows, because the squared sum over standard normal distributed data is N–1, which is directly evident from the definition of standard deviation. Furthermore, the brackets can be omitted as both the total length of the time series n and the number of segments N are constants for all permutations subsumed under the brackets. The last equal sign follows from Eq. (2), which just states that there is always one segment more than breaks.

From Eq. (7), we can conclude the following. The average external variance increases linearly with the number of inserted breaks k. Such a linear increase of v could be expected if one of the $\binom{n-1}{k}$ segmentation possibilities for a given number of breaks is chosen randomly. However, actually we select always the optimum segmentation as given by the above described dynamic programming. Consequently, we are less interested in the expected mean, but in the best of several attempts. In order to conclude such an extreme, the distribution has to be known.

For this purpose, let us go back to Eq. (3) where we insert again Eq. (6). It follows the same relationship as given in Eq. (7), but without averaging brackets, according to:

$$v = \frac{1}{n}\sum_{i=1}^{N} {y_i}^2 \quad . \tag{8}$$

It is striking that Eq. (8) is nearly identical to the definition of a χ^2 distribution, which is as follows: N values are randomly taken out of a standard normal distribution, which are then squared and added up. By repeating this procedure several times, these square sums form a χ^2 distribution with N being the number of degrees of freedom. Remembering that y_i is standard normal distributed, it becomes obvious that Eq. (8) reproduces this definition. The difference is that we divide finally by n. But, hereby, no substantial change is performed, as n is a constant equal to the length of the considered time series.

Consequently, we can conclude that v (actually nv), must be χ^2 distributed with k being the degree of freedom, according to:

$$f(x) = \frac{x^{\frac{k-2}{2}} e^{-\frac{x}{2}}}{2^{\frac{k}{2}} \Gamma\left(\frac{k}{2}\right)} \quad . \tag{9}$$

However, there is an important restriction of this rule. As v is normalized, it is confined between 0 and 1, whereas normal distributed data have no upper limit. The number of breaks k is inversely proportional to the number of segment members n_i. Therefore, the standard deviation of segment averages (Eq. (5)), is small compared to 1 for low break numbers. In this case the normal distribution is a good approximation for $\overline{x_k}$. However, with increasing break number, n_i decreases so that the standard deviation is approaching 1. Assume, e.g., a time series of length 100 with 25 breaks. n_i is then in the order of 4, so that the standard deviation of the $\overline{x_k}$ becomes 0.5 (Eq. (5)). Assuming still a normal distribution is no longer appropriate, as the true frequency for $\overline{x_k}$=1 is zero by definition, whereas the normal distribution at 2 standard deviations is not exactly zero. For the distribution of v it means that we have to expect a kind of confined χ^2 distribution, which is defined exclusively between zero and one. In the next chapter, we will show empirically that this is a Beta distribution. For this purpose, we verify in the following our theoretical considerations by practical tests with random data.

5. Empirical tests with random data

Typical climate time series contain at least 100 data points, which is preventing in general the explicit calculation of the entire distribution as discussed above. However, for n=20, this is still possible and carried out in the following to check our theoretical conclusions. *Fig. 2* shows the development of the external variance v as a function of the number of inserted breaks k.

To obtain statistical quantities, 100 repetitions have been performed. The mean amount of v increases linearly with k, as stated in Eq. (7). Additionally, the minimum and maximum are given for each number of breaks. In realistic cases, i.e., for larger n, the maximum can only be determined by dynamic programming; here the entire distribution could be explicitly calculated. In the following, it is our aim to find a mathematical function determining how the maximum external variance is growing with increasing number of breaks. A first approximation of this solution is already visible in *Fig. 2*. Three estimates are given for the maximum external variance. The central one, where an exponent of 4 is assumed, is in good agreement with the data. Obviously, the external variance v is connected to the break number k by the approximate function:

$$1 - v = \left(1 - \frac{k}{n-1}\right)^4 \quad . \tag{10}$$

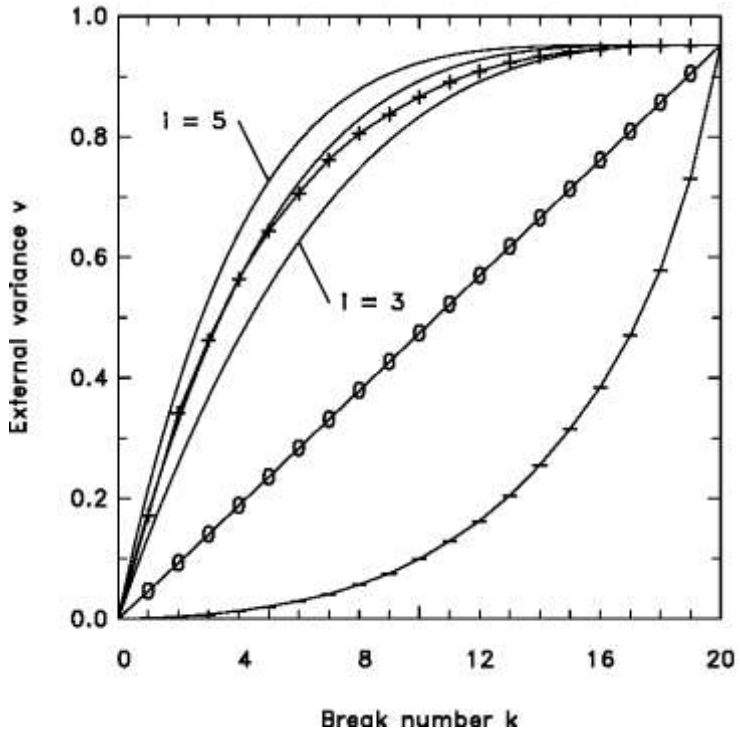

Fig. 2. Mean (0), maximum (+), and minimum (–) external variance as a function of inserted breaks for an $n = 21$ year random time series. For the maximum, three estimates are given: $1–v = (1– k^*)^i$, for $i = 3, 4, 5$, where $k^* = k/(n–1)$ is the normalized break number. For 20 breaks, v reaches not 1, but 0.95, because a fraction of $1/(n–1)$ is covered by the error of the total mean, as given in Eq. (1).

For each break number, *Fig. 2* gives minimum and maximum of the external variance for 100 repetitions. Between these extremes we expect a kind of confined χ^2 distribution. As the shown result is based on numerical calculations, we are able to check our theory. *Fig. 3* shows exemplarily the distribution as obtained from a Monte Carlo experiment for 7 breaks. Differences to the corresponding χ^2 distribution are not large, but noticeable, especially at the tail of the distribution, where the maximum value, we are interested in, occurs. In contrast, the Beta distribution with 7 degrees of freedom is in good agreement with the data. Confirmed by tests with further break numbers, we assume in the following that the external variance is generally Beta distributed. The Beta distribution is formally given by:

$$p(v) = \frac{v^{\frac{k}{2}-1}\,(1-v)^{\frac{n-1-k}{2}-1}}{B\left(\frac{k}{2}\,,\,\frac{n-1-k}{2}\right)} \quad , \tag{11}$$

with p denoting the probability density, v the external variance, and B the Beta function defined as:

$$B(a,b) = \frac{\Gamma(a)\,\Gamma(b)}{\Gamma(a+b)} \quad , \tag{12}$$

with Γ denoting the Gamma function.

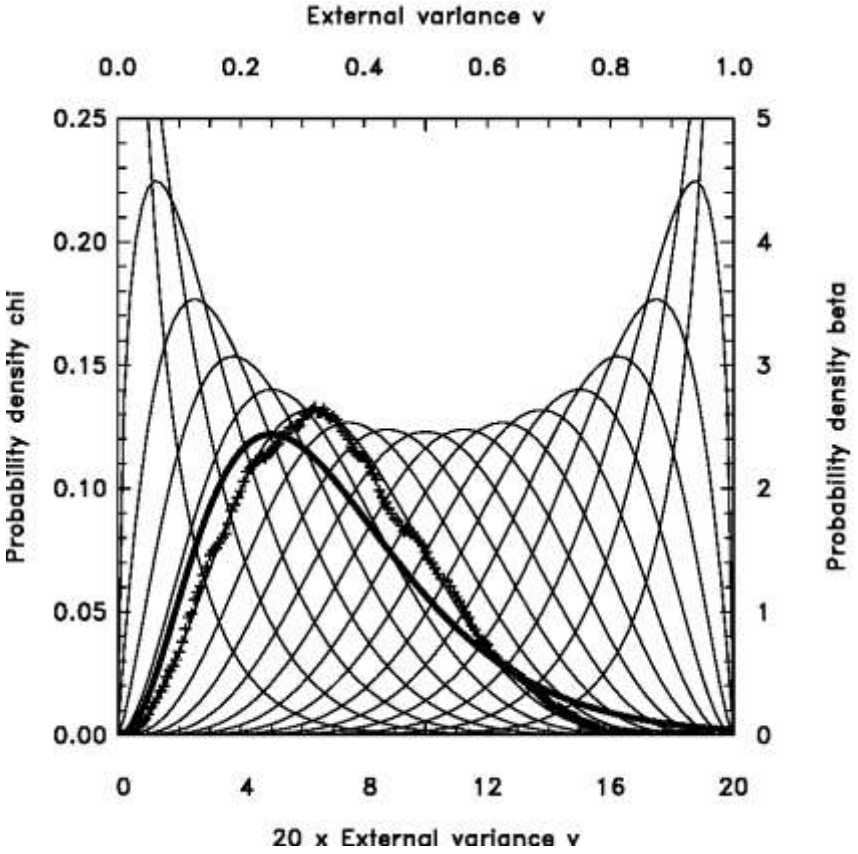

Fig. 3. Probability density for the χ^2 distribution, as given in Eq. (9) for k=7 (thick line). Furthermore, the 20 Beta distributions (thin), and the distribution of random data (crosses) are given. As expected, the data deviates slightly from the χ^2–7 and fits well to the Beta–7 curve. The lower abscissa and the left ordinate is valid for the χ^2 distribution. The upper v-abscissa and the right ordinate are valid for the Beta distribution and the normalized random data.

6. The incomplete Beta function

By Eq. (11) we are so far able to describe the distribution of the external variance *v* depending on length *n* and break number *k*. However, it is the maximum of *v*, which defines the optimum decomposition. Therefore, we need to find the maximum value of Eq. (11), or in other words, the exceeding probability of the Beta distribution, as given by:

$$P(v) = 1 - \int_0^v p\,dv \quad , \tag{13}$$

where the definite integral over a Beta distribution has to be solved, which is referred to as the incomplete Beta function *B*(*a*,*b*,*v*). With this substitution Eq. (13) reads:

$$P(v) = 1 - \frac{B\left(\frac{k}{2}, \frac{n-1-k}{2}, v\right)}{B\left(\frac{k}{2}, \frac{n-1-k}{2}\right)} \quad . \tag{14}$$

For whole numbers the incomplete Beta function is obviously solvable by integration by parts, and the solution is:

$$\frac{B(i, m-i+1, v)}{B(i, m-i+1)} = \sum_{l=i}^{m} \binom{m}{l} v^l (1-v)^{m-l} \quad . \tag{15}$$

By comparing the arguments of the Beta function in Eq. (14) with those in Eq. (15), it follows:

$$i = \frac{k}{2} \quad , \tag{16}$$

and

$$\frac{n-1-k}{2} = m - i + 1 \quad . \tag{17}$$

Inserting Eq. (16) in Eq. (17) we have:

$$m = \frac{n-3}{2} \quad . \tag{18}$$

Since the variables *i* and *m* are defined as integers, Eq. (14) is solvable for even *k* and odd *n*. Replacing *n* and *k* in Eq. (14) by *i* and *m*, it follows:

$$P(v) = 1 - \frac{B(\,i\,,\; m-i+1\,,\; v)}{B(\,i\,,\; m-i+1\,)} \quad . \tag{19}$$

Using Eq. (15), the solution is:

$$P(v) = 1 - \sum_{l=i}^{m} \binom{m}{l} v^l (1-v)^{m-l} \quad . \tag{20}$$

Now we are aiming to replace the 1 in Eq. (20) by using the binomial definition, which is as follows:

$$\sum_{l=0}^{m} \binom{m}{l} a^l b^{m-l} = (a+b)^m \quad . \tag{21}$$

With *a* being *v* and *b* being 1- *v* it follows:

$$\sum_{l=0}^{m} \binom{m}{l} v^l (1-v)^{m-l} = \big(v + (1-v)\big)^m = 1 \quad , \tag{22}$$

so that it is actually possible to replace the 1 in Eq. (20) by a sum from zero to m:

$$P(v) = \sum_{l=0}^{m} \binom{m}{l} v^l (1-v)^{m-l} - \sum_{l=i}^{m} \binom{m}{l} v^l (1-v)^{m-l} \quad . \tag{23}$$

Calculating the sum from zero to *m* minus the sum from *i* to *m*, the sum from zero to *i*–1 is remaining:

$$P(v) = \sum_{l=0}^{i-1} \binom{m}{l} v^l (1-v)^{m-l} \quad . \tag{24}$$

Eq. (24) gives the exceeding probability as a function of external variance for any even break number $k=2i$. Let us again check the obtained equation

numerically by a Monte Carlo computation. For this purpose we create a random time series of the length n=21 and search for the combination of 4 breaks that produces the maximum external variance. Fig. 4 shows the result as obtained by 1000 repetitions. As each individual time series contains $\binom{n-1}{k}=\binom{20}{4}=4845$ possibilities of decomposition, we are dealing with a sample size of 4,845,000. Two conclusions can be drawn. First, the data is in good agreement with Eq. (24). Second, the effective number of combinations is much smaller than the nominal.

To the first conclusion: In *Fig. 4*, vertical lines from ln(0)=1 are drawn down to the exceeding probability that is found in the numerical test data. Thus, the edge of the shaded area gives the probability function for a certain maximum external variance. The according theoretical function as derived from Eq. (24) is given alternatively as a curve. The chosen numbers of n=21 and k=4 can be transformed by Eqs. (16) and (18) to m=9 and i=2. Inserted into Eq. (24) it follows for the depicted example:

$$P(v) = (1-v)^9 + 9v\,(1-v)^8 \quad . \tag{25}$$

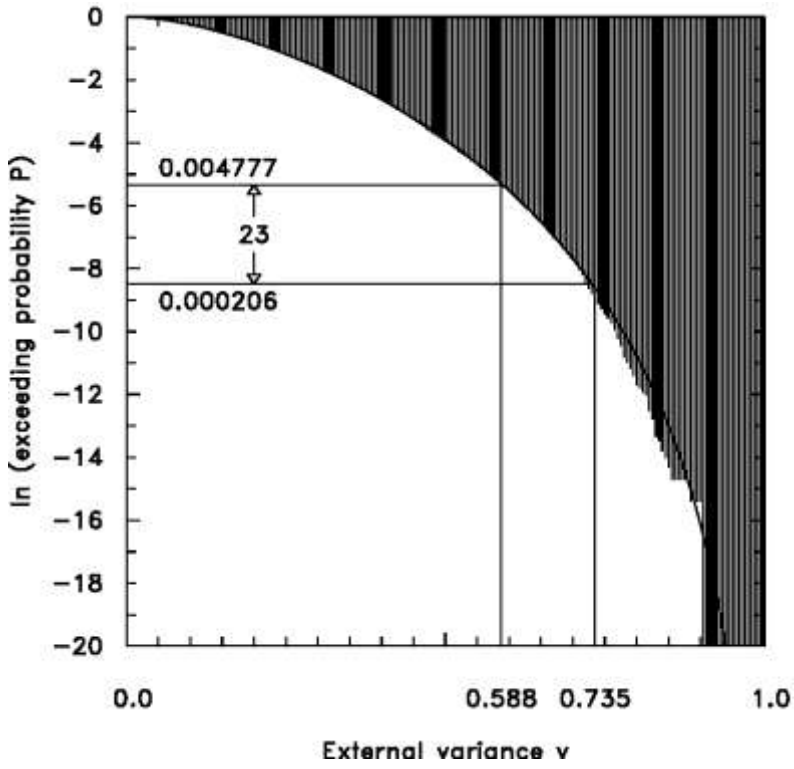

Fig. 4. Logarithmic exceeding probability as a function of external variance for 4 breaks within a 21-year time series. Vertical lines are drawn down from ln(0)=1 to the probability found for random data. The theoretical probability as generally given in Eq. (24) and specified in Eq. (25) is given by a curve. Two special data pairs are indicated, which are discussed in the text.

Fig. 4 shows that the data fits well to Eq. (25) if the probability is not too extreme. For such low probability it is not surprising that the limited Monte Carlo dataset shows more scatter and randomly deviates from the theory.

To the second conclusion: Two reading examples are given in *Fig. 4*. One starts from the exceeding probability of 2.064×10^{-4} (ln (0.0002) = –8.5). This value is equal to $\binom{20}{4}^{-1}$, the reciprocal of the nominal number of combinations for $n=21$ and $k=4$. If all combinations were independent, we could expect a maximum external variance of 0.7350. However, this is not the actually true value, which is already determined as 0.5876 (*Fig. 2*). But we can draw the reverse conclusion: What must be the effective number of combinations for the known external variance? We obtain a value of 4.777×10^{-3}, which is 23 times larger than the starting point. The conclusion is that the effective number of combinations for this special case ($n=21$, $k = 4$) is 23 times smaller than the nominal one, which is equal to $\binom{20}{4}$. The dependency of different solutions is reasonable. Shifting only one break position by one time step creates already a new break combination. However, its external variance will not deviate much from the original.

7. The relative change of variance as a function of increased break number

After confirming Eq. (24) by test data, we can assume its general validity and turn towards more realistic lengths. *Fig. 5* shows the graphs of Eq. (24) for $n=101$ and all even k from 2 to 20. As in *Fig. 4,* the number of independent combinations is estimated by a reversal conclusion from the known results of the maximum external variance. (In this case the results stem from a dynamic programming search as the length of $n = 101$ is too large for an explicit all-permutations-search of the maximum as it was possible for $n = 21$.)

The following question arises: What is the rate of change of the variance, if the number of breaks is increased? Obviously, there are two contributions. First, we skip from the graph in *Fig. 5* valid for k breaks to the next one valid for $k+2$. This causes a certain increase in the external variance, even if the number of combinations would remain constant. Second, there *is* certainly an increased number of permutations, although we showed that the effective number is always smaller than the nominal one.

Fig. 6 gives a sketch of the situation to illustrate how the mathematical formulations for the two components are derived in detail. The exceeding probability P for two arbitrary even break numbers is depicted. To determine the first contribution, we need to know the distance between two neighboring curves in v-direction for a fixed P (*v1– v0* in *Fig. 6*).

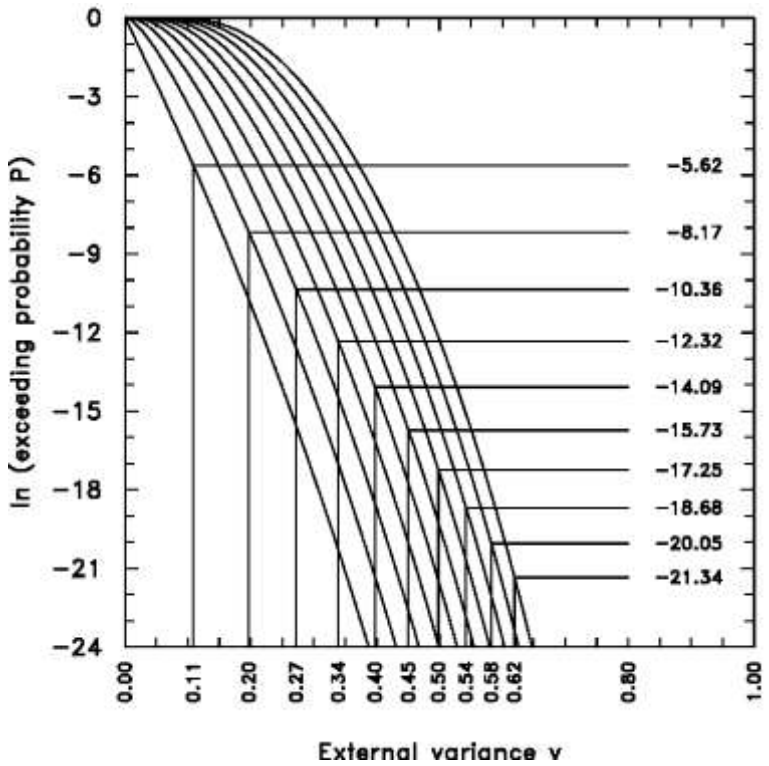

Fig. 5. As *Fig. 4*, but for a 101-year time series and for the ten different break numbers from 2, 4, 6, … , 20. The known external variances for each break number are retranslated into the observed effective exceeding probabilities given as the column at the right edge.

As Eq. (24) is difficult to solve for *v*, we estimate the *v* -distance by the *P*-distance, which is divided by the slope *s*:

$$\left(\frac{dv}{di}\right)_1 = v1 - v0 = \frac{ln(P1) - ln(P2)}{s} \quad . \tag{26}$$

Using the respective *i*-indices for *P1* and *P0* (see *Fig. 6*) we can rewrite:

$$\left(\frac{dv}{di}\right)_1 = \frac{\left(ln\left(P_i(v)\right) - ln\left(P_{i+1}(v)\right)\right)_{v=const}}{s} \quad . \tag{27}$$

This first part of *dv/di* arises because different functions of *P(v)* has to be used. We introduce C_f and refer to it the following as the function contribution:

$$C_f = \left(ln\left(P_{i+1}(v)\right) - ln\left(P_i(v)\right)\right)_{v=const} \quad , \tag{28}$$

so that Eq. (27) can be rewritten:

$$\left(\frac{dv}{di}\right)_1 = -\frac{C_f}{s} \quad . \tag{29}$$

The second contribution is the increase of *v* due to the total decrease of *P* (*v2 – v1* in *Fig. 6*). Geometrically, this can be perceived as a walk down the respective curve.

$$\left(\frac{dv}{di}\right)_2 = v2 - v1 = \frac{ln(P2) - ln(P1)}{s} \quad . \tag{30}$$

Using *i*-indices for *P2* and *P1*, it follows:

$$\left(\frac{dv}{di}\right)_2 = \frac{ln\big(P_{i+1}(v)\big) - ln\big(P_i(v)\big)}{s} \quad . \tag{31}$$

This second part of *dv/di* depends on the increased number of decomposing permutations with growing *i*. Consequently, we refer to the numerator as number contribution C_n, according to:

$$C_n = ln\big(P_{i+1}(v)\big) - ln\big(P_i(v)\big) \quad , \tag{32}$$

and it follows:

$$\left(\frac{dv}{di}\right)_2 = \frac{C_n}{s} \quad . \tag{33}$$

In both cases, changes in *P* are translated into *v* by the slope of the curves. This is appropriate if the curvatures are small and the slopes remain nearly constant. For the relevant parts of the curves this is a good approximation (*Fig. 5*).
Finally, we can summarize Eq. (29) and Eq. (33) to:

$$\frac{dv}{di} = \left(\frac{dv}{di}\right)_2 + \left(\frac{dv}{di}\right)_1 = \frac{C_n - C_f}{s} \quad . \tag{34}$$

To determine *dv/di*, we obviously need three terms, the slope *s*, the function contribution C_f, and the number contribution C_n. These three terms are derived in the following subsections.

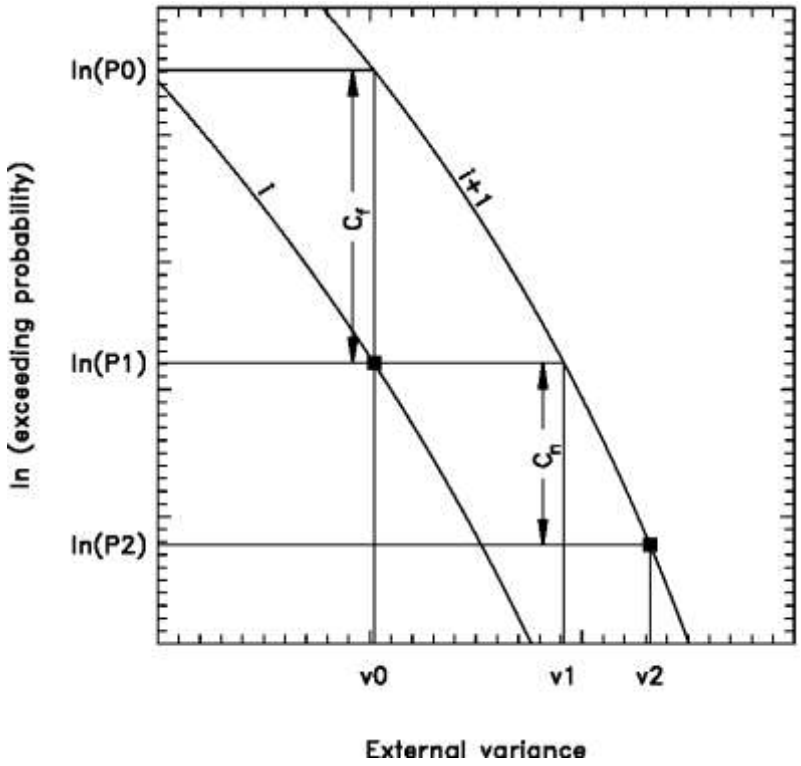

Fig. 6. Sketch to illustrate the total gain of external variance from v0 to v2, when the number of breaks *k* is increased by 2, i.e., from *i* to *i*+1. The first contribution (*v1– v0*) depends on the horizontal distance of the two curves. This contribution is derived in the text by the vertical distance C_f and the slope of the curve. The second contribution (*v2–v1*) occurs due to the increase of possible combinations when the break number is increased. As for the first contribution, it is translated from C_n by using the slope of the depicted curves.

7.1. The slope

The slope *s* of the logarithm of Eq. (24) as it is depicted in *Figs. 5* and *6* is equal to:

$$s = \frac{d}{dv}\Big(ln\big(P(v)\big)\Big) = \frac{1}{P(v)}\frac{dP(v)}{dv} \quad . \tag{35}$$

With Eq. (13) it follows:

$$s = -\frac{p(v)}{P(v)} \quad . \tag{36}$$

Replacing *n* and *k* by *m* and *i* and using the result of Appendix A we can rewrite Eq. (11) to:

$$p(v) \;=\; v^{i-1}\,(1-v)^{m-i}\,(m-i+1)\,\binom{m}{i-1} \quad . \tag{37}$$

Inserting Eq. (24) and Eq. (37) into Eq. (36), it follows:

$$s \;=\; -\,\frac{v^{i-1}\,(1-v)^{m-i}\,(m-i+1)\,\binom{m}{i-1}}{\sum_{l=0}^{i-1}\left(\binom{m}{l}\,v^{l}(1-v)^{m-l}\right)} \quad . \tag{38}$$

In Appendix B we show that the last summand is a good approximation for the sum occurring in the denominator and it follows:

$$s \;=\; -\,\frac{v^{i-1}\,(1-v)^{m-i}\,(m-i+1)\,\binom{m}{i-1}}{\binom{m}{i-1}\,v^{i-1}\,(1-v)^{m-i+1}} \quad , \tag{39}$$

which can be reduced to:

$$s \;=\; -\,\frac{m-i+1}{1-v} \quad . \tag{40}$$

After replacing again *m* and *i* by *n* and *k* it follows:

$$s \;=\; -\,\frac{n-1-k}{2\,(1-v)} \quad . \tag{41}$$

7.2. The function contribution

With Eq. (24) the vertical distance between two neighboring curves as given in *Fig. 6* is:

$$C_f \;=\; ln(P_{i+1}) - ln(P_i) \;=\; ln\left(\frac{\sum_{l=0}^{i}\binom{m}{l}\,v^{l}(1-v)^{m-l}}{\sum_{l=0}^{i-1}\binom{m}{l}\,v^{l}(1-v)^{m-l}}\right). \tag{42}$$

We use again Appendix B and approximate the sums by their last summand:

$$C_f \;=\; ln\left(\frac{\binom{m}{i}\,v^{i}\,(1-v)^{m-i}}{\binom{m}{i-1}\,v^{i-1}\,(1-v)^{m-i+1}}\right) \quad , \tag{43}$$

which can be reduced to:

$$C_f \;=\; ln\left(\frac{\binom{m}{i}\, v}{\binom{m}{i-1}\,(1-v)}\right) \quad . \tag{44}$$

The ratio of consecutive binomial coefficients is equal to (*m*-*i*+1)/*i*:

$$C_f \;=\; ln\left(\frac{(m-i+1)\, v}{i\,(1-v)}\right) \quad . \tag{45}$$

Replacing *m* and *i* again by *n* and *k*, it follows:

$$C_f \;=\; ln\left(\frac{(n-1-k)\, v}{k\,(1-v)}\right) \quad . \tag{46}$$

7.3. The number contribution

The nominal number of combinations grows with growing *k* from $\binom{n-1}{k}$ to $\binom{n-1}{k+1}$. This corresponds to a factor of (*n*–1–*k*)/*k*. However, in *Fig. 4* we show exemplarily for *k* = 4 that the effective number of combinations is lower. In *Fig. 5* the decrease of ln(*P*(*v*)) due to the increase of the effective number of combinations is given in a column at right edge for the even *k* from 2 to 20. From these numbers we derived the actual decreasing factor $C_n = \Delta\ln(P(v))$ and compared it with the nominal (*Table 1*). The nominal decreasing factor for $\Delta k = 1$ is equal to the reciprocal of the growth of combinations $\frac{\binom{n-1}{k}}{\binom{n-1}{k+1}} = \frac{k}{n-1-k}$. Here we need its logarithm; and as the effective decreasing factor is only available for every second *k*, *nom* = –2ln*((n–1–k)/k)* is the proper reference.

From *Table 1* we can extract that the ratio between the effective and nominal factor is rather constant with $r \approx 0.4$, but slightly growing with increasing break number. The growth will be discussed in detail in Appendix C, for the time being we can summarize:

$$C_n \;=\; r\;nom \;=\; -2r\;ln\left(\frac{n-1-k}{k}\right) \quad . \tag{47}$$

Table 1. From *Fig. 5*, C_n, the effective decrease of ln($P(v)$) for the transition from k to k+2 is taken. It is compared to the nominal decrease equal to –2 ln((*n*–1–*k*)/*k*). Finally, the ratio *r* between the effective and nominal factor is given

k_1	k_2	k	eff = Δ ln (P(v))	nom = –2 ln((n–1–k)/k)	r = eff/nom
2	4	3	–2.552	–6.952	0.367
4	6	5	–2.186	–5.889	0.371
6	8	7	–1.963	–5.173	0.379
8	10	9	–1.765	–4.627	0.381
10	12	11	–1.645	–4.181	0.393
12	14	13	–1.514	–3.802	0.398
14	16	15	–1.435	–3.469	0.414
16	18	17	–1.363	–3.171	0.430
18	20	19	–1.292	–2.900	0.446

7.4. The differential equation and its solution

The rate of change of v with regard to k is only half of that with regard to i (compare Eq. (16)):

$$\frac{dv}{dk} = \frac{dv}{di}\frac{di}{dk} = \frac{1}{2}\frac{dv}{di} \quad . \tag{48}$$

Using Eq. (34) it follows:

$$\frac{dv}{dk} = \frac{1}{2}\frac{C_n - C_f}{s} \quad . \tag{49}$$

Inserting our findings for the slope s (Eq. (41)) and for the two contributions C_f and C_n (Eqs. (46) and (47)), the growth of v with growing k is given by:

$$\frac{dv}{dk} = \frac{1-v}{n-1-k}\left(2r\,ln\left(\frac{n-1-k}{k}\right) + ln\left(\frac{(n-1-k)\,v}{k\,(1-v)}\right)\right) . \tag{50}$$

Reducing the fractions under the logarithms by n–1 leads to the normalized break number k^*, defined as:

$$k^* = \frac{k}{n-1} \quad . \tag{51}$$

At the same time, the differential *dk* has to be replaced by:

$$dk = (n-1)\, dk^* \quad , \tag{52}$$

so that Eq. (50) may be rewritten in normalized form:

$$\frac{dv}{dk^*} = \frac{1-v}{1-k^*}\left(2r\, ln\left(\frac{1-k^*}{k^*}\right) + ln\left(\frac{(1-k^*)\, v}{k^*\,(1-v)}\right)\right) \quad . \tag{53}$$

The final main question is now: What is the solution of Eq. (53)? Let us make a first approach to the solution by a very rough estimate for small k^*.

$$\frac{1-k^*}{1-v}\,\frac{dv}{dk^*} = 2r\, ln\left(\frac{1-k^*}{k^*}\right) + ln\left(\frac{(1-k^*)\, v}{k^*\,(1-v)}\right) = \alpha = -C_n + C_f \quad . \tag{54}$$

The first logarithm constituting α, i.e., $-C_n$, is for small k^* in the order of ln(*n*) and it decreases with increasing k^*. The second, C_f, is in the order of ln(v/k^*). Because we know already the approximate solution being $1-v \approx (1-k^*)^4$, we can estimate the second term to about ln(4) (compare Eq. (78) in Appendix B). In contrast to the first term, this term increases with increasing k^* (see Appendix C), because $1-v$ is decreasing faster than $1-k^*$. Assuming n = 101, an estimate for α is:

$$\alpha \approx 2r\, ln(n) + ln(4) \approx 2 \cdot 0.4\ \ln(100) + ln(4) = 5.07 \quad . \tag{55}$$

If α were actually constant, the integration of Eq. (54) would be easy:

$$\frac{1}{1-v}\, dv = \frac{\alpha}{1-k^*}\, dk^* \quad , \tag{56}$$

$$-\, ln(1-v) = -\,\alpha\, ln(1-k^*) \quad , \tag{57}$$

$$1-v = (1-k^*)^{\alpha} \quad , \tag{58}$$

which is rather similar to the already known approximate solution (Eq. (10)), except that the exponent found in Eq. (55) is higher. This already shows that the assumptions made to estimate *s*, C_f, and C_n were reasonable.

For a more accurate solution let us go back to the performance of the random data that we already used above to verify our theory. By these data we can check how well the rough estimate of a constant α is fulfilled in reality. *Fig. 7* shows that such an estimate is actually not too bad, which is the reason

for Eq. (58) being rather close to the true solution. For a more precise solution, we fit a function to $\alpha(k^*)$ and obtain:

$$\frac{1-k^*}{1-v}\,\frac{dv}{dk^*} \;=\; \frac{1-k^*}{2}\;ln\left(\frac{1-k^*}{k^*}\right) + 2\;ln(5) \quad . \tag{59}$$

Eq. (59) may be rewritten as:

$$\frac{1}{1-v}\;dv \;=\; \left(\frac{1}{2}\;ln\left(\frac{1-k^*}{k^*}\right) + \frac{2\;ln(5)}{1-k^*}\right)\;dk^* \quad , \tag{60}$$

which is easy to integrate. Its solution is:

$$1-v \;=\; (1-k^*)^a\left(\frac{1-k^*}{k^*}\right)^{bk^*} \quad , \tag{61}$$

with $a = 2\ln(5) + 1/2$ and $b = -1/2$.

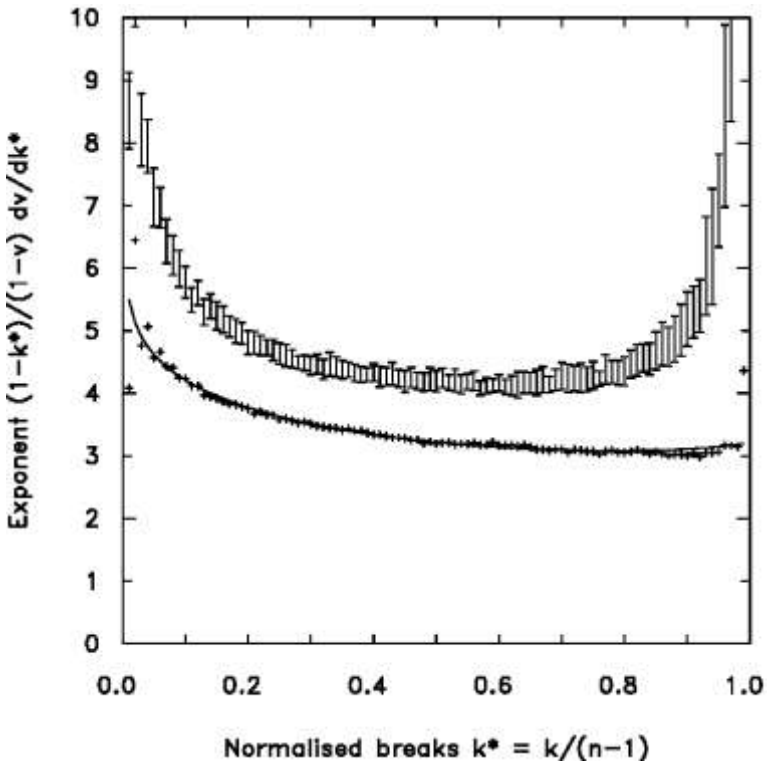

Fig. 7. Exponent α as given in Eq. (54) as a function of the normalized break number k^* for random data (crosses). These data consists of 1000 random 101-year time series. The vertical bars connect the 90 and 95 percentiles. The thin line is giving the function according to Eq. (59).

In *Fig. 7*, the function for the exponent α as given in Eq. (59) fits well to the results derived from random data. However, the relative gain of external variance is larger for even values compared to their uneven neighbors, especially for low values. This feature is reasonable, as it needs always a pair of breaks to isolate a subsegment. To produce the data, we performed 1000 repetitions, mainly to reduce the scatter. However, the repetitions can also be exploited to derive the variability of the solution. Consequently, not only the mean, but also the 90 and 95 percentiles are given. The average exponent starts for low normalized break numbers at about 5. This means that the external variance grows at the beginning 5 times faster than the normalized break number. This behavior is found for random data. When such a variance growth will occur in real data, we can be rather sure that no true break is present as it is normal for random data which has by definition no real break. The 95 percentile is for the first breaks as large as nearly 10. Thus, in only 5% of the cases, the external variance grows by a factor of more than 10 times faster than k^*. Hence, this value can be used as limit to distinguish true from spurious breaks. For the first break numbers it reaches nearly 10, decreasing rapidly to about 5 for $k^* = 0.1$.

8. Discussion of the penalty term

Within the homogenization algorithm PRODIGE (*Caussinus* and *Mestre*, 2004), the following expression is minimized to estimate the number of predicted breaks.

$$C_k(Y) \;=\; \ln\left(1 - \frac{\sum_{j=1}^{k+1} n_j\left(\overline{Y}_j - \bar{Y}\right)^2}{\sum_{i=1}^{n}(Y_i - \bar{Y})^2}\right) + \frac{2\,(k+l)}{n-1}\; ln(n) \;= \min \quad . \quad (62)$$

The numeric value of $C_k(Y)$ depends on the data Y and the number of breaks k and consists of two opposite contributions. Firstly, the logarithm of the normalized internal variance, and secondly, a penalty term, originally proposed by *Caussinus* and *Lyazrhi* (1997). Whereas the first is decreasing with larger k, the second is increasing. Using our notations for the same terms, Eq. (62) can be rewritten as:

$$\ln(1 - v) \;+\; \frac{2k}{n-1}\; ln(n) \;=\; \min \quad . \qquad (63)$$

In Eq. (63), we combined k and l, the number of breaks and the number of outliers to a single number. Splitting off an outlier is identical to the separation of a subperiod of length 1. Consequently, it is not necessary to treat outliers

separately. If we further use the normalized break number k^* according to Eq. (51) instead of k, we can rewrite:

$$\ln(1-v) \; + \; 2k^* \, ln(n) \; = \; \min \quad . \tag{64}$$

To find the break number k^* for which the expression is minimal, the first derivative with respect to k^* is set to zero:

$$-\frac{1}{1-v}\frac{dv}{dk^*} + 2\,ln(n) \; = \; 0 \quad , \tag{65}$$

which can be rewritten to:

$$\frac{1}{1-v}\frac{dv}{dk^*} \; = \; 2\,ln(n) \quad . \tag{66}$$

For a given time series, the length n is constant. Consequently, we can conclude from Eq. (66), that PRODIGE uses a fixed number, equal to 2 ln(n), as stop criterion. If the relative gain of the external variance falls below that constant, no further breaks are added and the final break number is reached. However, from Eq. (59) we know the function for the relative gain of external variance in detail; it just has to be divided by $1-k^*$.

$$\frac{1}{1-v}\frac{dv}{dk^*} \; = \; \frac{1}{2}\,ln\left(\frac{1-k^*}{k^*}\right) + \frac{2\,ln(5)}{1-k^*} \quad . \tag{67}$$

Fig. 8 shows this function for a time series of length 101. Additionally, six exceeding values for probabilities from 1/4 to 1/128 are given, based on 5000 repetitions. These curves are approximately equidistant. For comparison, the constant as proposed by *Caussinus* and *Mestre* (2004) and rewritten in Eq. (66) is given, which is about 9 (exactly 2 ln(101)) for n = 101.

As the exceeding values are computed for random data, they can be interpreted as error probability. The 1% error line (exactly 1/128) at the upper end of the family of curves in *Fig. 8* starts at a variance gain of about 15, and reaches, for k^* = 0.1, a value of about 8.

In the climatologically interesting range of small k^*, the numeric value of the mean variance gain (lowest line in *Fig. 8*) is equal to about 5 and can be interpreted as following. For random data, which contains no break by definition, the relative external variance grows on average with each additionally inserted break 5 times faster than expected by a simple linear approach. Such a linear approach just supposes that each break adds the same amount of external variance. For n = 101 this would be one percent per break. In reality, the data contains larger jumps just by chance, comprising not only 1%,

but 5% of the remaining variance. In seldom cases, these highest jumps contain even 15% of the remaining variance, but the probability for that is only about 1% (uppermost line in *Fig. 8*). As the number of tentatively inserted breaks is growing, the highest jumps are already used before, so that the amount of the remaining decreases. Increasing the break number from 9 to 10 ($k^* = 0.1$) gains in average still 5%, as for the lowest break numbers, but the maximum value, exceeded in 1% of the cases, drops from 15% to 8%.

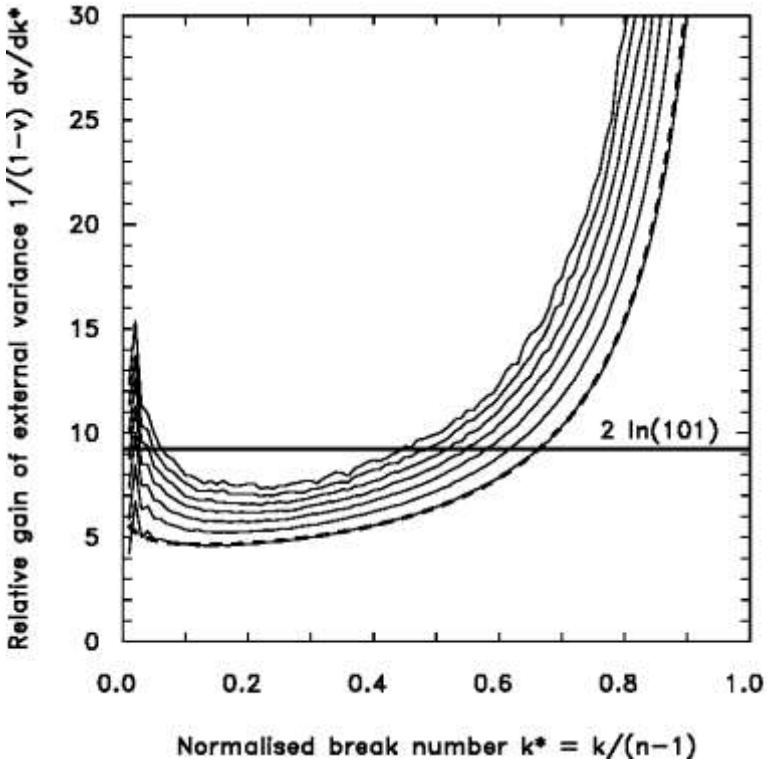

Fig. 8. Relative gain of external variance as a function of normalized breaks k^* for a time series length of $n = 101$. The dashed fat curve denotes the theoretical value as given by Eq. (67). The solid thin curves are showing the data results as obtained by 5000 repetitions. The lowest indicates the mean, which is largely congruent with the theory. The upper ones give the exceeding value for probabilities from 2^{-2}, 2^{-3}, … , 2^{-7}. For comparison, the constant 2 ln(n) proposed as stop criterion by *Caussinus* and *Lyazrhi* (1997) is given by the horizontal line.

The Lyazrhi constant of 2 ln(n) as proposed by *Caussinus* and *Mestre* (2004) is equal to about 9 for $n = 101$. At the beginning, i.e. for one break, this value lies in the middle of the family of error curves in *Fig. 8*. Thus, it corresponds here to an error of about 5%. At $k^* = 0.08$, i.e. for 8 breaks, the horizontal line is leaving the area covered by error curves. Thus, the error level decreases below 1%. Assuming continued equidistance, the horizontal line will reach areas with errors of less than 0.1% at $k^* = 0.15$. Thus, for low break numbers, PRODIGE accepts breaks, even if the error is relatively high (about

5%). In contrast, higher break numbers are effectively suppressed. Only breaks are accepted that add an amount of variance, which would occur randomly with a probability of less than 1%.

The choice of the Lyazrhi constant appears to be rather artful. For the first breaks, it allows errors of about 5%, which is a widely accepted error margin. However, for more than 8 breaks (within a time series of 101 data points), the method is much more rigid. Obviously, the preexisting knowledge is used that such high numbers of breaks are per se unlikely, so that a suppression is reasonable.

In *Fig. 9*, the corresponding features for shorter time series (n = 21) are given. Compared to n = 101, the average variance gain remains unchanged, showing that Eq. (67) is universally valid. However, the exceeding values increase, and the distances between the error curves grow by a factor of 5. This indicates that the growing factor is inversely proportional to the time series length n. In contrast, the Lyazrhi constant even decrease, although only slightly due to its logarithmic form. The direction of change of the Lyazrhi constant for different time series length is contradicting our findings for random data and should be studied further. However, instrumental climate records comprise often about 100 data points, and for such lengths the constant is chosen rather well.

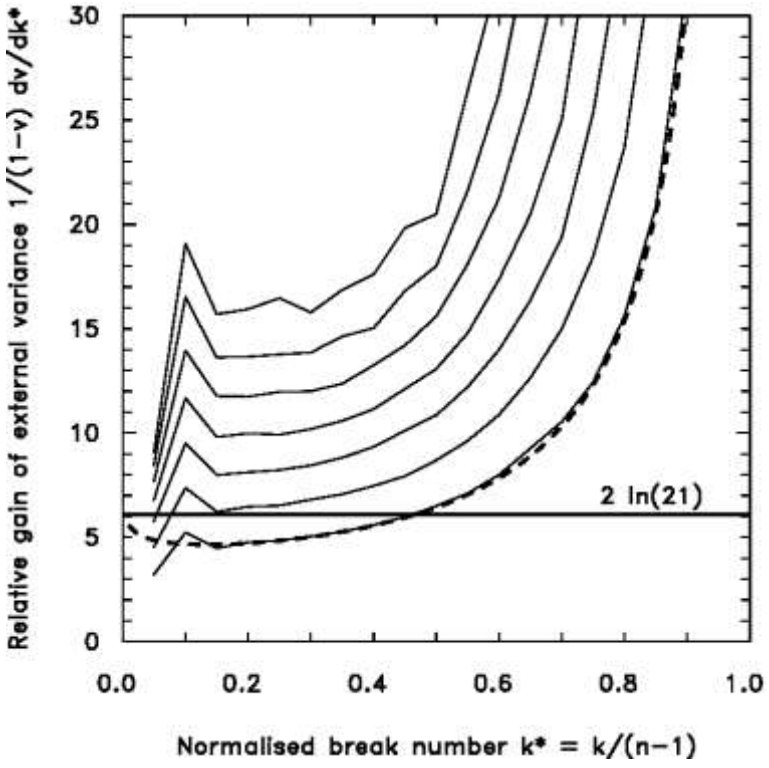

Fig. 9. As *Fig. 8*, but for n = 21. The average variance gain remains unchanged compared to *Fig. 8*, because Eq. (67) is universally valid. However, the exceeding values increase inversely proportional to n. In contrast, the constant of *Caussinus* and *Lyzrhi* (1997) decreases with decreasing n.

9. Conclusions

The external variance, defined as the variance of the subperiods' means, is shown to be the key parameter to detect breaks in climate records. Maximum external variance indicates the most probable combination of break positions. We analyzed the characteristics of the external variance occurring in random data and derived a mathematical formulation (Eq. (61)) for the growth of its maximum with increasing number of assumed breaks. As random data includes by definition no break, this knowledge can be used as null hypothesis to separate true breaks in real climate records more accurately from noise. In this way, it helps to enhance the valuable information from historical data.

***Acknowledgement*—**We are grateful for advice from *Peter Domonkos* and *Tamas Szentimrey*. This work has been performed within the project Daily Stew supported by the Deutsche Forschungsgemeinschaft DFG (VE 366/5).

Appendix A

Consider the Beta function in Eq. (11):

$$B\left(\frac{k}{2}, \frac{n-1-k}{2}\right) = B(i, m-i+1) = \frac{\Gamma(i)\,\Gamma(m-i+1)}{\Gamma(i+m-i+1)} = \frac{(i-1)!\;(m-i)!}{m!} \quad . \tag{68}$$

Multiplication of both the numerator and denominator with *m–i+1* leads to:

$$B\left(\frac{k}{2}, \frac{n-1-k}{2}\right) = \frac{(i-1)!\;(m-i+1)!}{(m-i+1)\,m!} \quad . \tag{69}$$

Remembering the definition of binomial coefficients being $\binom{n}{k} = \frac{n!}{k!\,(n-k)!}$, we can write:

$$B\left(\frac{k}{2}, \frac{n-1-k}{2}\right) = \left((m-i+1)\binom{m}{i-1}\right)^{-1} \quad . \tag{70}$$

Appendix B

Consider the individual summands of the sum as defined in Eq. (24). The factor of change *f* between a certain summand and its successor is:

$$f = \frac{\binom{m}{l_i} v}{\binom{m}{l_i - 1}(1-v)} \quad , \tag{71}$$

where l_i runs from zero to i. The ratio of consecutive binomial coefficients can be replaced, and it follows:

$$f = \frac{(m - l_i + 1)\, v}{l_i(1-v)} \quad . \tag{72}$$

m and *i* can be replaced by *n* and *k*:

$$f = \frac{(n - 1 - l_k)\, v}{l_k(1-v)} \quad . \tag{73}$$

Inserting *k* instead of l_k is a lower limit for *f* because $(n–1–l_k)/l_k$, the rate of change of the binomial coefficients, is decreasing monotonously with *k*:

$$f > \frac{(n - 1 - k)\, v}{k(1-v)} \quad . \tag{74}$$

Normalize *k* by 1/(*n*–1):

$$f > \frac{(1 - k^*)\, v}{k^*(1-v)} \quad . \tag{75}$$

The approximate solution is known with $1–v = (1–k^*)^4$, see Eq. (10).

$$f > \frac{(1 - k^*)\ (1 - (1 - k^*)^4)}{k^*(1 - k^*)^4} \quad , \tag{76}$$

$$f > \frac{1 - (1 - k^*)^4}{k^*(1 - k^*)^3} \quad , \tag{77}$$

for $k \to 0$:

$$f > \frac{1 - (1 - 4k^*)}{k^*(1 - 3k^*)} = \frac{4k^*}{k^*(1 - 3k^*)} = \frac{4}{1 - 3k^*} = 4 \quad , \tag{78}$$

for $k \to 1$:

$$f > \frac{(1-k^*)^{-3} - (1-k^*)^4}{k^*} = \frac{\infty - 0}{1} = \infty \quad . \qquad (79)$$

We can conclude that each element of the sum given in Eq. (24) is by a factor f larger than the prior element. For small k^* the factor f is greater than about 4 and grows to infinity for large k^*. Consequently, we can approximate the sum by its last summand according to:

$$P(v) = \sum_{l=0}^{i-1} \binom{m}{l} v^l (1-v)^{m-l} \approx \binom{m}{i-1} v^{i-1} (1-v)^{m-i+1} \quad . \qquad (80)$$

Appendix C

Once the solution for $v(k^*)$ is available (Eq. (61)), a more accurate estimation of the function contribution C_f is possible. So far, we approximated the sum given in Eq. (24) by its last summand, as discussed in Appendix B. Now we are able to check the impact of this approximation. Using the known solution, we calculated two versions of C_f. First, by taking into account only the last summand as in Eq. (43) and alternatively the complete term, as given in Eq. (42). *Fig. 10* shows these two estimates of C_f as dashed lines. The upper one denotes the full solution, the lower the approximation. Their difference remains limited, which confirms our findings in Appendix B. As discussed in Eq. (55), C_f starts for low k^* at about ln(4) and rises to infinity for high k^*.

Concerning the number contribution C_n, we applied so far only a rough estimate as given in Eq. (47), assuming a constant ratio between effective and nominal combination growths. Actual values for C_n are listed in *Table 1* for low break numbers. However, they are numerically computable up to about $k^* = 0.75$. In *Fig. 10*, these values for C_n are given as crosses. They are multiplied by –1, as $-C_n$ contributes to the exponent α. We fitted a function of the form:

$$(ak^* + b)\, ln\left(\frac{1-k^*}{k^*}\right) + c \quad , \qquad (81)$$

to the data, which is depicted by the lower full curve in *Fig. 10*, and obtained for the coefficients:

$$a_1 = 0.5, \quad b_1 = 0.55, \quad c_1 = 0.4 \quad .$$

A similar fit to C_f is given by the upper full curve in *Fig. 10*. Here the coefficients are:

$$a_2 = -1.0, \quad b_2 = -0.15, \quad c_2 = 2.7 \quad .$$

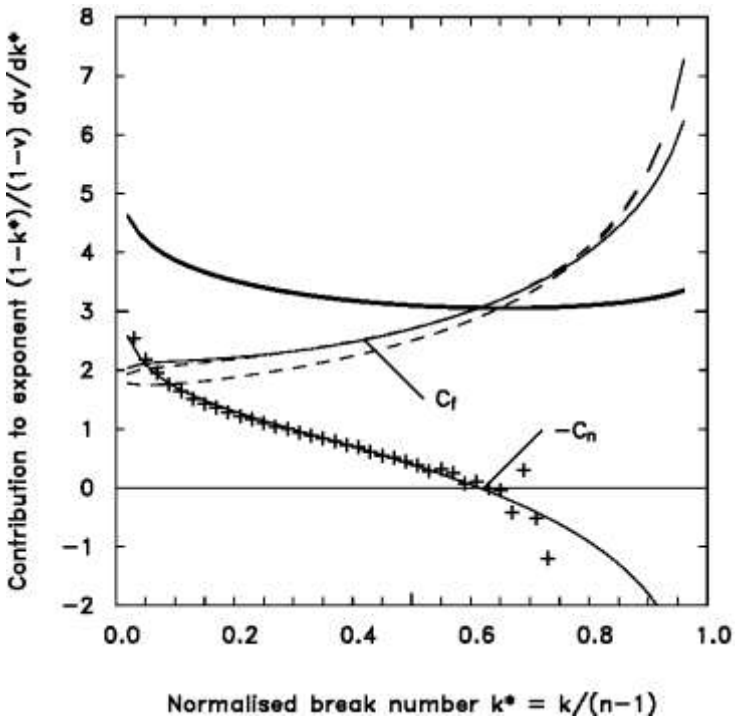

Fig. 10. Contributions of C_f and $-C_n$ to the exponent $= \frac{1-k^*}{1-v}\,\frac{dv}{dk^*}$. The two dashed lines are reconstructions of C_f from the known solution of $v(k^*)$, as given in Eq. (61). The solid line gives a fitted function for C_f. Crosses denote data for C_n connected likewise by a fitted curve. The sum of the two contributions is given by the fat line.

The sum of two curves yields then an alternative estimation for the exponent α. It is depicted as a fat line in *Fig. 10* and characterized by the sum of the coefficients:

$$a_3 = -0.5, \quad b_3 = 0.4, \quad c_3 = 3.1 \quad .$$

This alternative estimate is in good agreement (please compare *Fig. 7* lowest line with *Fig. 10* uppermost fat line) with the solution derived directly from the data as given in Eq. (59), where the coefficients are:

$$a_4 = -0.5, \quad b_4 = 0.5, \quad c_4 = 2\ln(5) = 3.2 \quad .$$

We see that Eq. (59), so far directly based on a fit to the data, is as well understandable from the theory as the sum of the two contributions C_f and $-C_n$.

References

Aguilar, E., Auer, I., Brunet, M., Peterson, T.C., and *Wieringa, J.,* 2003: Guidelines on climate metadata and homogenization. World Meteorological Organization, *WMO-TD No. 1186*, WCDMP No. 53, Geneva, Switzerland, 55 pp.

Akaike, H., 1973: Information theory and an extension of the maximum likelihood principle, In *2nd International Symposium on Information Theory*, Akademiai Kiado, Budapest, 267–281.

Alexandersson, H. and *Moberg, A.*, 1997: Homogenization of Swedish temperature data. 1. Homogeneity test for linear trends. *Int. J. Climatol. 17*, 25–34.

Auer, I., Böhm, R., Jurkovic, A., Orlik, A., Potzmann, R., Schöner, W., Ungersböck, M., Brunetti, M., Nanni, T., Maugeri, M., Briffa, K., Jones, P., Efthymiadis, D., Mestre, O., Moisselin, J.M., Begert, M., Brazdil, R., Bochnicek, O., Cegnar, T., Gajic-Capkaj, M., Zaninovic, K., Majstorovic, Z., Szalai, S., Szentimrey, T., and *Mercalli, L.*,2005: A new instrumental precipitation dataset for the Greater Alpine Region for the period 1800–2002, *Int. J. Climatol. 25*, 139–166.

Auer, I., Böhm, R., Jurkovic, A., Lipa, W., Orlik, A., Potzmann, R., Schöner, W., Ungersböck, M., Matulla, C., Briffa, K., Jones, P., Efthymiadis, D., Brunetti, M., Nanni, T., Maugeri, M., Mercalli, L., Mestre, O., Moisselin, J.M., Begert, M., Müller-Westermeier, G., Kveton, V., Bochnicek, O., Stastny, P., Lapin, M., Szalai, S., Szentimrey, T., Cegnar, T., Dolinar, M., Gajic-Capka, M., Zaninovic, K., Majstorovic, Z., and *Nieplova, E.*, 2007: HISTALP – historical instrumental climatological surface time series of the Greater Alpine Region, *Int. J. Climatol*. 27, 17–46.

Begert, M., Schlegel, T. and *Kirchhofer, W.*, 2005: Homogeneous temperature and precipitation series of Switzerland from 1864 to 2000. *Int. J. Climatol. 25*, 65–80.

Bellman, R., 1954: The Theory of Dynamic Programming, *Bull. Am. Math. Soc. 60*, 503–516. doi: 10.1090/S0002-9904-1954-09848-8, MR 0067457.

Bergström, H. and *Moberg, A.*, 2002: Daily air temperature and pressure series for Uppsala (1722–1998). *Climatic Change, 53*, 213–252.

Brunet, M., Asin, J., Sigro, J., Banon, M., Garcıa, F., Aguilar, E., Esteban Palenzuela, J., Peterson, T.C., and *Jones, P.*, 2011: The minimization of the screen bias from ancient Western Mediterranean air temperature records: an exploratory statistical analysis. *Int. J. Climatol. 31*, 1879–1895.

Brunetti, M., Maugeri, M., Monti, F., and *Nannia, T.*,2006: Temperature and precipitation variability in Italy in the last two centuries from homogenised instrumental time series. *Int. J. Climatol. 26*, 345–381.

Caussinus H. and *Lyazrhi, F.*,1997: Choosing a linear model with a random number of change-points and outliers. *Ann. Inst. Stat. Math. 49*, 761–775.

Caussinus, H. and *Mestre, O.*, 1996: New mathematical tools and methodologies for relative homogeneity testing. *Proc. First Seminar for Homogenization of Surface Climatological Data*, Budapest, Hungary, Hungarian Meteorology Service, 63–72.

Caussinus, H. and *Mestre, O.*, 2004: Detection and correction of artificial shifts in climate series. *Appl. Statist. 53*, part 3, 405–425.

Conrad, V. and *Pollak. C.*, 1950: *Methods in climatology*, Harvard University Press, Cambridge, MA, 459 pp.

Davis R.A.,. Lee, T.C.M., and *Rodriguez-Yam, G.A.*, 2012: Structural break estimation for nonstationary time series models. *J. Am.. Stat. Assoc*. *101*, 223–239.

Domonkos, P., 2011a: Efficiency evaluation for detecting inhomogeneities by objective homogenization methods. *Theor. Appl. Climatol. 105*, 455–467.

Domonkos, P., 2011b: Adapted Caussinus-Mestre Algorithm for Networks of Temperature Series (ACMANT). *Int. J. Geosci. 2*, 293–309.

Easterling, D.R. and *Peterson, T.C.*, 1995: A new method for detecting undocumented discontinuities in climatological time series. *Int. J. Climatol., 15*, 369–377.

Hawkins, D.M., 1972: On the choice of segments in piecewise approximation. *J. Inst. Maths. Applics., 9*, 250–256.

Knowles Middleton, W.E., 1966: A history of the thermometer and its use in meteorology. The John Hopkin Press, Baltimore, Maryland. 249 pp.

Lavielle, M., 1998: Optimal segmentation of random processes. *IEEE Trans. Signal Processing, 46*, 1365–1373.

Li, S. and *Lund, R.*, 2012: Multiple Changepoint Detection via Genetic Algorithms. *J. Climate 25*, 674–686.

Lindau, R., 2003: Errors of Atlantic Air-Sea Fluxes Derived from Ship Observations., *J. Climate, 16*, 783–788.

Lindau, R., 2006: The elimination of spurious trends in marine wind data using pressure observations. *Int. J. Climatol*. 26, 797–817.

Menne M.J. and *Williams Jr., C.N.*, 2009: Homogenization of temperature series via pairwise comparisons. *J. Climate, 22*, 1700–1717.

Mestre O., Domonkos, P., Picard, F., Auer, I., Robin, S., Lebarbier, E., Böhm, R., Aguilar, E., Guijarro, J., Vertachnik, G., Klancar, M., Dubuisson, B., and *Stepanek, P.*, 2012: HOMER: homogenisation software in R –methods and applications. *Időjárás 117*, 47–67.

MeteoSchweiz, 2000: Alte meteorologische Instrumente (Old meteorological instruments). Bundesamt für Meteorology und Klimatologie (MeteoSchweiz), Zürich, 190 p.

Nemec J., Gruber, G., Chimani, B., and *Auer, I.*, 2012: Trends in extreme temperature indices in Austria based on a new homogenized dataset. *Int. J. Climatol.*, DOI: 10.1002/joc.3532.

Nordli, P.O., Alexandersson, H., Frich, P., Förland, E.J., Heino, R., Jonsson, T., Tuomenvirta, H., and Tveito, O.E., 1997: The effect of radiation screens on Nordic time series of mean temperature. Int. J. Climatol. 17, 1667–1681.

Picard, F., Robin, S., Lavielle, M., Vaisse, C., and Daudin, J.-J., 2005: A statistical approach for array CGH data analysis, BMC Bioinformatics 6, 27.

Picard F., Lebarbier, E., Hoebeke, M., Rigaill, G., Thiam, B., and *Robin, S.*, 2011: Joint segmentation, calling, and normalization of multiple CGH profiles. *Biostatistics 12*, 413–428.

Rust, H.W., Mestre, O., and *Venema, V.K.C.*, 2008: Less jumps, less memory: homogenized temperature records and long memory. *J. Geophys. Res. Atmos. 113*, D19110.

Slonosky, V.C., Jones, P.D. and *Davies, T.D.*, 2001: Instrumental pressure observations and atmospheric circulation from the 17th and 18th centuries: London and Paris, *Int. J. Climatol. 21*, 285–298.

Szentimrey, T., 1996: Statistical procedure for joint homogenisation of climatic time series. *Proceedings of the First seminar of homogenisation of surface climatological data*, Budapest, Hungary, 6–12 October 1996, 47–62.

Szentimrey, T., 1999: Multiple Analysis of Series for Homogenization (MASH). *Proceedings of the second seminar for homogenization of surface climatological data*, Budapest, Hungary; WMO, WCDMP-No. 41, 27–46.

Szentimrey, T., 2007: Manual of homogenization software MASHv3.02. Hungarian Meteorological Service, 65 p.

Trewin, B., 2010: Exposure, instrumentation, and observing practice effects on land temperature measurements. *WIREs Clim. Change, 1*, 490–506.

Van der Meulen, J.P. and *Brandsma, T.*, 2008: Thermometer screen intercomparison in De Bilt (The Netherlands), part I: Understanding the weather-dependent temperature differences. *Int. J. Climatol. 28*, 371–387.

Venema, V., Mestre, O., Aguilar, E., Auer, I., Guijarro, J.A., Domonkos, P., Vertacnik, G., Szentimrey, T., Stepanek, P., Zahradnicek, P., Viarre, J., Müller-Westermeier, G., Lakatos, M., Williams, C.N., Menne M.J., Lindau, R., Rasol, D., Rustemeier, E., Kolokythas, K., Marinova, T., Andresen, L., Acquaotta, F., Fratianni, S., Cheval, S., Klancar, M., Brunetti, M., Gruber, Ch., Prohom Duran, M., Likso, T., Esteban, P., and *Brandsma, Th.*, 2012: Benchmarking homogenization algorithms for monthly data. *Clim. Past 8*, 89–115.

Vincent, L.A., 1998: A technique for the identification of inhomogeneities in Canadian temperature series. *J. Climate 11*, 1094–1104.