The Uncertainty of Break Positions Detected by Homogenization Algorithms in Climate Records

Short title: The Uncertainty of Break Positions

by

Ralf Lindau and Victor Venema

Meteorological Institute University of Bonn

Auf dem Hügel 20

D-53121 Bonn

Germany

Email: rlindau@uni-bonn.de

Phone: +49 228 735185

Fax: +49 228 735188

Abstract

Long instrumental climate records suffer from inhomogeneities due to, e.g., relocations of the stations or changes in instrumentation, which may introduce sudden jumps into the time series. These inhomogeneities may mask or strengthen true trends. Relative homogenization algorithms use the difference time series of a candidate station with neighboring stations to identify such breaks (changepoints). Modern multiple breakpoint methods search for the optimum segmentation, which is characterized by minimum internal variance within the segments and maximum external variance between the segment means.

We analyze the accuracy of these homogenization methods and concentrate on the uncertainty in the position of the break. Due to unavoidable random noise in the difference time series, the segmentation method may find a shifted break position, which attains a higher external variance than the true one. Different lengths of potentially exchanged subsegments are considered; that one providing the largest external variance will be chosen as possibly erroneous optimum. We will show that the variances of shifted segmentations can be described as Brownian motion with drift, where the signal-to-noise ratio (SNR) defines the drift size.

Available formulae for one-sided and continuous Brownian motion with drift are expanded to two-sided and discrete processes as they occur in praxis. The error probability increases strongly for SNRs lower than 1. Thus when the internal variance is larger than the variance introduced by the breaks, the probability of finding the right break position is small.

1 Introduction

Operational meteorological observations are available for the past centuries and provide valuable information to monitor the climate and changes therein (Rennie et al., 2014; Rhode et al., 2013; Morice et al., 2012; Lawrimore et al., 2011). However, relocations of stations as well as changes in the measuring techniques and surroundings introduce non-climatic changes (inhomogeneities) into climate time series, which makes the analysis of raw data problematic (Aguilar et al., 2003). Homogenization algorithms are able to detect and correct inhomogeneities by considering the difference time series of the candidate station with neighboring stations. Searching the segmentation that explains most of the variance by a minimum number of breaks is an objective and accurate way to determine the positions of all breaks in a time series (Caussinus and Mestre, 2004; Lu et al., 2010; Lindau and Venema, 2013). Therefore, this multiple breakpoint method is often applied by modern homogenization algorithms and will be used in this study.

While we have a reasonable qualitative idea of the quality of homogenized data from benchmarking (Williams et al., 2012; Venema et al., 2012) and validation studies (e.g., Buishand, 1982; Easterling and Peterson, 1995; Beaulieu et al., 2008), an important shortcoming in homogenization is the lack of quantitative uncertainty estimates. The uncertainties due to remaining inhomogeneities in the data are determined by several factors. Important ones are that not all breaks in the candidate station can be detected, the uncertainty in the estimation of correction parameters due to insufficient well-correlated neighbouring stations, the uncertainty in the corrections due to remaining inhomogeneities in the references and finally, the topic of this paper, the uncertainty in the date of the break.

Whereas previous benchmarking studies were mainly interested in the performance of homogenization methods, the upcoming benchmarking study by the International Surface Temperature Initiative (ISTI) aims at being able to estimate uncertainties (Stott and Thorne, 2010; Thorne et al., 2011; Chandler et al., 2012; Willett et al., 2014). By producing a benchmark dataset that mimics the station network of the real global data holding and by inserting highly realistic inhomogeneities this benchmarking initiative promises to deliver the most accurate numerical estimates to date of the uncertainties due to remaining inhomogeneities.

Inhomogeneities are a significant part of the uncertainty budget of climate change studies. The Global Historical Climate Network version 3 (GHCNv3), for example, has been homogenized with the Pairwise Homogenization Algorithm (Menne and Williams, 2009). Lawrimore et al. (2011) reported that the homogenized temperature trend is 0.8°C per century between 1880 and 2012. Whereas, the trend in the raw temperature record is only 0.6°C; the 0.2°C difference is due to homogenization adjustments. Given that homogenization can reduce biases, but not fully remove them (Menne and Williams, 2009), it is expected that the data still contains remaining inhomogeneities and that this 0.2°C is an underestimate. Such biases and uncertainties need to be quantified.

On the global scale many biases average away. Conversely, the biases and uncertainties on the regional scale are much larger (Auer et al., 2005; Brunet et al., 2011; Brunetti et al., 2006). Uncertainties will vary from station to station, mainly determined by the signal-to-noise ratio (Domonkos, 2013), which depends on the correlation of a candidate station to its neighbours.

The best effort to date to compute and communicate uncertainties due to remaining inhomogeneities has been made for the HadCRUT dataset (Brohan et al., 2006; Morice et al., 2012). In this study, the influence of these complicated temporally and spatially correlated errors is communicated by generating an ensemble with stochastically simulated inhomogeneities. In all stations errors are introduced to model remaining urbanization errors (0.055°C/century) and biases due to improvements in exposure of the sensors before 1930. Remaining unbiased random breaks are modelled as 40-year long periods perturbed by a normal random number with a standard deviation of 0.4°C. Because this dataset contains data homogenized by the suppliers, these errors do not depend on the detected inhomogeneities and are the same for all stations. They also do not depend on the station density or the urbanity of the stations.

The uncertainty estimates for the date of the break this article proposes may also help validation of homogenization methods with detection scores. For the HOME benchmark it was found that there was only a modest correlation between the detection scores for the breaks and the performance of the homogenization methods with respect to climatologically important error measures, such as the uncertainty in the stations trends (Venema et al., 2012). This could suggest that detecting a large number of breaks is less important than previously thought; reliably detecting the large breaks and other parts of the algorithms may be more important. However, it is also possible that the better methods were able to detect smaller inhomogeneities, but were punished because these have an inherent larger uncertainty in their date. With the results of this paper, it may be possible to develop detection scores that take this uncertainty into account and that would provide a fairer comparison of the breaks detected by methods with different sensitivities.

In section 2.1 we begin our investigation with the variance a given segmentation is able to explain. This parameter serves in most multiple-breakpoint algorithms as criterion to identify breakpoints. If a false segmentation with shifted break positions explains by chance more variance than the true one, it is preferred by mistake and an error of break position occurs. This critical variance difference is shown to be a quadratic function of a stochastic variable (2.2). By analyzing null points and slope of this function (2.3 to 2.6), we show that the probability to gain more variance by a false segmentation can be described by a Brownian motion with negative drift, where the drift strength is determined by the signal-to-noise ratio (2.7). Because the maximum variance is utilized in homogenization algorithms, we need to describe where the maximum of a Brownian motion with drift occurs.

In section 3 we discuss the available formula for continuous processes (Buffet, 2003). However, there are two shortcomings. First, break detection is obviously a discrete process and second, it is a two-sided process, because deviations are possible to both sides of the true break position. Therefore, section 4 is dedicated to discrete Brownian motion with drift. Initially, we concentrate on the description of one-sided processes. We decompose the probability of a total maximum into two factors: the backward (past) probability and the forward (future) probability, where the latter is shown to be equal to the hit rate (4.1). In 4.2 we give an estimate for the probability of a backward maximum and an exact solution for the two first steps of the process. These are shown to be sufficient to estimate the hit rate with high accuracy (4.3). In section 4.4, the alterations necessary for the two-sided process are derived. Section 5 concludes.

2 From maximum external variance to Brownian motion with drift

In this section we will show that the probability to obtain a false, i.e. shifted, break position can be described statistically by a Brownian motion with drift. More specific: the probability that a break position of a climate time series is erroneously shifted by k temporal units is equal to the probability that the maximum of a Brownian motion with drift is located at the k^{th} step. We start with the explained variance of a segmentation, in the following referred to as the external variance.

2.1 The maximum external variance as break criterion

Technically, the total variance of a time series can be decomposed into two parts: the external variance between the different segment means and the internal variance within the segments (Lindau, 2003). The maximum external variance is then used as decision criterion for the optimal segmentation, which defines the proposed break positions. Obviously, position errors do occur, whenever the external variance of a false segmentation with shifted break positions attains more external variance than the true one. Thus, the difference of external variance between the right and false segmentation is the key parameter to describe position errors.

2.2 The difference of external variance between false and correct segmentation

A scheme of a position error is given in Fig. 1, showing a time series with an obvious break in the middle that should be detected. The two segments of lengths n_1 and n_2 have initially the means x_1 and x_2 as indicated by the fat horizontal lines. If a subsegment of length k with mean x_0 is erroneously exchanged from segment 2 to segment 1, the means change to x$_1$' and x$_2$' (as denoted by the thin horizontal lines). In this example, x$_1$' is strongly reduced, because x$_2$ was originally much lower than x$_1$; x$_2$' differs only slightly and randomly, because only a short piece is missing. Together it means that the two segment averages converge, which reduces the external variance. Consequently, this wrong segmentation would be rejected. In this paper we determine the probability of the seldom cases that it is not rejected.

The criterion to define the optimum segmentation is maximum external variance. The difference Δv between a wrong and the correct segmentation is derived in detail in Appendix A and given by:

$$n\,\Delta v \;=\; -f_1\,{x_1}^2 \;+\; f_2{x_2}^2 + 2x_0(f_1x_1 - f_2x_2) + (f_2 - f_1)\,{x_0}^2 \tag{1}$$

where n denotes the total length of the time series; the above factors f$_1$and f$_2$, and the factor f$_0$ used in Eq. (8) are defined as:

$$f_1 \coloneqq \frac{kn_1}{n_1 + k} \tag{2}$$

$$f_2 \coloneqq \frac{kn_2}{n_2 - k} \tag{3}$$

$$f_0 \coloneqq \frac{k^2(n_1 + n_2)}{(n_1 + k)(n_2 - k)} \quad = \quad f_2 - f_1 \tag{4}$$

Thus, the change of external variance is only a function of the three means and the lengths of the two segments and of the exchanged subsegment.

In the next step we replace x$_0$ in Eq. (1) based on the following consideration: The mean of the exchanged subsegment x$_0$ is equal to x$_2$, the segment where it stem from, plus a random scatter variable δ:

$$x_0 = x_2 + \delta \tag{5}$$

where δ depends on the internal variance σ^2 and the length *k* of the exchanged subperiod.

$$\delta = \frac{\sigma}{k}\sum_{i=1}^{k}\delta_i \qquad , \qquad \delta_i \sim \mathcal{N}(0{,}1) \tag{6}$$

so that δ is a normal distributed variable with variance σ^2/k. We insert Eq. (5) into Eq. (1) and divide by the square of the jump height D, which is:

$$D^2 = (x_1 - x_2)^2 \tag{7}$$

and obtain (as described in detail in Appendix B) the normalized variance gain v*:

$$v^* \coloneqq \frac{n\,\Delta v}{D^2} = -f_1 + 2f_1\varepsilon + f_0\varepsilon^2 \tag{8}$$

with

$$\varepsilon = \frac{\delta}{|D|} = \frac{\sigma}{k|D|}\sum_{i=1}^{m} \delta_i \quad , \quad \delta_i \sim \mathcal{N}(0,1) \tag{9}$$

Eq. (8) shows that v*, which is the decision criterion for break detection, is a quadratic function of a random variable ε. ε is normal distributed with zero mean and standard deviation $\sigma/(k|D|)$ (Eq. 9). Using the definition of the signal-to-noise ratio:

$$SNR := \frac{\left|\frac{D}{2}\right|}{\sigma} \tag{10}$$

we see that the standard deviation of ε depends on the length of the exchanged segment k and on the signal-to-noise ratio.

2.3 The null points and first derivative of the variance gain

If the right-hand side of Eq. (8) becomes positive, the shift of the break position by *k* items leads to increased external variance so that this solution is preferred by mistake. The zero points for Eq. (8) are given by:

$$-f_1 + 2f_1\varepsilon_0 + f_0{\varepsilon_0}^2 = 0 \tag{11}$$

Leading to:

$$\varepsilon_0 = -a \pm \sqrt{a^2 + a}\,, \qquad with\; a = \frac{f_1}{f_0} \tag{12}$$

Factoring out *a* and using the first two terms of the series expansion of the square root yield:

$$\varepsilon_0 = -a \pm a\sqrt{1 + \frac{1}{a}} \cong -a \pm a\left(1 + \frac{1}{2a}\right) \tag{13}$$

As illustrated in Fig. 2, there are two solutions for ε_0:

$$\varepsilon_0 = -2a - \frac{1}{2} \quad \vee \quad \varepsilon_0 = \frac{1}{2} \tag{14}$$

To obtain an estimate for the factor *a*, we consider f_1 and f_0 (Eqs. 2 and 4). If both segments have comparable lengths ($n_1 \cong n_2$) and the exchanged subsegment *k* is much shorter than the two original segments, we can approximate:

$$f_0 = \frac{k^2(n_1 + n_2)}{(n_1 + k)(n_2 - k)} \cong \frac{2n_1 k}{{n_1}^2} = \frac{2k^2}{n_1} \tag{15}$$

$$f_1 = \frac{kn_1}{n_1 + k} \cong k \tag{16}$$

An approximation for *a* is then:

$$a = \frac{f_1}{f_0} \cong \frac{n_1}{2k} \gg 1 \tag{17}$$

Thus, there are two zero points:

$$\varepsilon_0 \cong -\frac{n_1}{k} \quad \vee \quad \varepsilon_0 = \frac{1}{2} \tag{18}$$

Let us consider the rate of change at the zero points. The derivative of Eq. (8) is:

$$\frac{dv^*}{d\varepsilon}(\varepsilon) = 2f_1 + 2f_0\varepsilon \tag{19}$$

We insert the positive zero point of Eq. (18) and use Eq. (4) to replace f_0:

$$\frac{dv^*}{d\varepsilon}(\varepsilon_0) = f_1 + f_2 \tag{20}$$

From Eq. (16) f_1 can be estimated by k; an analogous consideration provides the same result for f_2. Thus, an estimate for the slope at the positive zero point is:

$$\frac{dv^*}{d\varepsilon}(\varepsilon_0) \cong 2k \tag{21}$$

As the parabola is symmetric, the slope is identical in size, but opposite in sign at the negative zero point.

2.4 Data simulation

So far our considerations were purely theoretical. To confirm them empirically, we generated 10,000 random time series of length 100 with internal standard deviation of 1 and inserted a jump of the double height in the middle so that the signal-to-ratio is equal to 1. In this way, we tested the validity of Eq. (8). Subperiods of different lengths (from *k* = 1 to 9) are exchanged between the segments and the variance gain v* and ε are computed and displayed in a scatterplot (Fig. 3). Each data pair (ε ; v*) is plotted as a number denoting the respective length *k*, which are close together for large k. The expected theoretical parabolas (Eq. 8) for each *k* are given as thin lines for comparison. Theory and simulations are in good agreement. The variance gain becomes positive for ε larger than ½ and is well describable by a linear function with slope 2*k*. However, only the right side of the parabola (compare Fig. 2) around the positive zero point is covered by data. The theoretically predicted negative one is obviously not present in our data.

2.5 The negative solution

The random variable ε is normalized by the jump height (Eq. 9). Consequently, the negative solution, which requires strong negative ε, occurs only if the random scatter is much larger than the jump height. Such a case is illustrated in Fig.4. The first member of the second segment is drastically disturbed. If it is erroneously included into the first segment the mean falls here even below that originally found for segment 2, whereas vice versa the mean for segment 2 rises higher than the original level of segment 1. This produces additional internal variance in both segments for all but the exchanged extreme value. However, its own contribution to the internal variance decreases, because the new segment average moved closer after the exchange. If the latter effect dominates, the internal variance decreases so that the exchange of an extreme value is estimated to be the better segmentation. This is obviously an abnormal situation. Therefore, we reject the negative solution in Eq. (14) and concentrate exclusively on the positive solution $\varepsilon_0 = 0.5$ for the rest of this study.

2.6 The positive solution

The positive solution is easy to interpret (Fig. 5). If a subsegment adjacent to the true break is randomly lifted by more than half of the jump height, its average is nearer to the neighboring segment mean than to its own. Thus, including it to the neighboring segment will reduce the internal variance so that an erroneous break position is obtained.

The appropriate criterion to assess the optimal length of such an exchanged subsegment is obviously that the hatched area in Fig. 5 is maximal. The values of the times series itself are given by the internal variance σ times a random standard normal variable δ_i. The blank area under the zero line has to be subtracted, which is equal to half of the jump height. Finally we have to sum over an increasing number of columns and search for the maximum. Thus, the optimum length *k* is obtained by the following maximization:

$$\sum_{i=1}^{k} (\sigma\delta_i - |D/2|) = \max \qquad (22)$$

Dividing by σ and factoring out the constant SNR from the summation yields:

$$\sum_{i=1}^{k} (\delta_i) \; - \; k\,SNR = \max \qquad (23)$$

Thus, the sum over *k* values of a random standard normal variable minus a linear term in *k* has to be maximized.

This descriptive solution can be alternatively obtained by theoretical considerations. We start with the parabola equation (8) and perform a linear approximation around the zero point:

$$v^* = \frac{dv^*}{d\varepsilon}(\varepsilon_0)(\varepsilon - \varepsilon_0) \tag{24}$$

Inserting the positive zero point (Eq. (18)) and the estimated slope here (Eq. (21)) yields:

$$v^* = 2k\left(\varepsilon - \frac{1}{2}\right) \tag{25}$$

With the definition of ε in Eq. (9), it follows:

$$v^* \;=\; 2\,\frac{\sigma}{|D|}\sum_{i=1}^{k}\delta_i \;-\; k \tag{26}$$

We search for the maximum of v*, which is not altered when we multiply by the constant SNR. Consequently, we obtain again Eq. (23) and its descriptive derivation is confirmed theoretically.

2.7 Brownian motion with drift

Eq. (23) and alternatively Eq. (26) are the key findings of section 2. They can be interpreted as following. For the true break position, *k* is zero and Eq. (26) yields zero variance as reference. If a segmentation, which is shifted by *k* items, attains a positive variance v*, it is erroneously assessed to be superior. If even more than one shifted segmentations turns out to be positive the largest one is chosen as optimum. The probability of such events is simply describable by a sum over a successively expanded sequence of a normal distributed random variable minus a linear term in *k*, the length of the sequence. Such a process is known as Brownian motion with drift. The drift is always negative and given by the SNR. Normally, Brownian motion is considered as temporal process; in our case the deviation *k* from the true breakpoint plays an analogous role. As we are searching for the probability that the maximum of variance is reached at a certain deviation *k*, we need to know the distribution of the time of the maximum of a Brownian motion with drift.

3 Brownian motion with drift as continuous process

For continuous time processes the distribution of the location of the maximum of a Brownian motion with drift is given by Buffet (2003).

$$f(s) = 2\left[\frac{1}{\sqrt{s}}\,\varphi\big(\mu\sqrt{s}\big) + \mu\Phi\big(\mu\sqrt{s}\big)\right] \times \left[\frac{1}{\sqrt{t-s}}\,\varphi\big(\mu\sqrt{t-s}\big) - \mu\Phi\big(-\mu\sqrt{t-s}\big)\right], 0 < s < t \tag{27}$$

where φ and Φ denote the standard normal density and distribution function, respectively, and μ is the drift, which in our case equals the negative SNR. Thus, for equal jump height and scatter μ is equal to -1. The time axis of a Brownian motion corresponds in our case to temporal distance from the true break. Thus, *t* in Eq. (27) corresponds to the length of the entire segment and *s* to the break shift *k*, which can be assumed to be much smaller than *t*. In this case ($k \ll t$) the second factor in Eq. (27) can be neglected, as φ becomes 0 and Φ becomes 1 for large arguments.

For three different values of the drift, the full Eq. (27) will next be compared to (i) numerical simulations of a discrete Brownian motion with drift and (ii) a complete break search simulation using the same data as described in section 2.4. In this way, we tested (i) the similarity of continuous and discrete time processes and (ii) the proper description of deviations in break positions by a Brownian motion with drift as it has been postulated in section 2.

First, Eq. (27) is computed for drifts of -0.5, -1.0, and -2.0. The three resulting curves are given as dashed lines in Fig.6. As mentioned above, the size of the drift corresponds to the signal-to-noise ratio. For d = -0.5, e.g., the scatter is two times larger than the standard deviation introduced by the jump height. Such large scatter leads to a slow decrease of the curve, showing relatively high probabilities also for large deviations of the estimated break positions. For equal signal and noise, a deviation of two time steps occurs, e.g., with a probability of exp(-3.1) = 0.045. When the scatter is halved (SNR=2), it is drastically reduced to exp(-6.7) = 0.001.

However, Eq. (27) is valid for continuous time processes only. Such solutions are not readily transferrable to discrete time steps, we are interested in here. Therefore, we simulated 100,000 Brownian motion processes that are discrete in time and compared their statistics to the results of Eq. (27). The discrete-case results are given as circles. A third type of results is given as crosses. Here, a complete break search is performed using the maximum external variance as decision criterion.

All three types of processes are generally in good agreement. However, the hit rate, i.e. deviations of zero, which is an important detection score in homogenization, is not well reproduced by Eq. (27). The reason is that the hit rate is not well defined for continuous processes.

4 Brownian motion with drift as discrete process

Break search is a discrete process and the hit rate will play a key role when we come to two-sided processes in section 4.4. Therefore, we derive in the following an alternative way to estimate the distribution of the time of the maximum of a Brownian motion with drift, now for discrete processes.

4.1 Backward and Forward Maxima

Consider a discrete Brownian motion *B* with drift *d*:

$$B_k = kd + \sum_{i=1}^{k} \delta_i \tag{28}$$

where the initial state is $B_0 = 0$ and δ_i denotes a random standard normal variable. Please note that Eq. (28) is in accordance to Eq. (23), but that the negative SNR now is referred as to drift *d*. The aim is to determine the probability that the absolute maximum of *B* occurs at time step *k*. The element B_k is the maximum of a Brownian motion, if two conditions are fulfilled: It is both a forward maximum and a backward maximum (Fig. 7). A backward maximum is defined by being larger than all its predecessors; for a forward maximum this is true for all successors. The probability to be the absolute maximum is then obtainable by multiplying the back- and forward probabilities.

B_0 being the absolute maximum means that no error in the determination of the break position occurred; we refer to this probability as the hit rate *h*. Hit rate and forward probability are identical, as each arbitrary time step can be regarded as new starting point of an alternative Brownian motion. Consequently, we can write for the probability of the absolute maximum:

$$P\left(B_k = \max_{0 \leq i < \infty} B_i\right) = h\, p_{kk} \tag{29}$$

where p_{ij} denotes the probability that the preliminary maximum after i time steps is found at time step j (so that p_{kk} denotes the backward maximum). Thus, to compute the probability that a Brownian motion reaches its maximum at a certain time step, we need the hit rate h and the probability of backward maximum p_{kk}.

4.2 The Probability of Backward Maximum

A necessary condition for an element B_k to become a backward maximum is that it is positive; otherwise B_0 would be larger. To be positive, the sum of k random standard normal variables must exceed $k|d|$ (see Eq.(28)), which is the amount subtracted by the drift term (please note that we are restricted to negative drifts). The standard deviation of the sum of k random standard normal values is equal to $\sqrt{k}$ so that the probability to exceed $k|d|$ is equal to $1 - \Phi(|d|\sqrt{k})$, where Φ denotes the standard normal cumulative distribution function. Since only negative drifts are considered, this expression is equivalent to $\Phi(d\sqrt{k})$. To avoid confusion due to the negative sign of d, we define the following abbreviation for the exceeding probabilities:

$$\Phi_k := 1 - \Phi(|d|\sqrt{k}) = \Phi(d\sqrt{k}) \tag{30}$$

Thus, an upper limit for the probability of B_k to be a backward maximum is Φ_k, which describes that it has to be at least positive. The reason for the actually reduced probability is the possible existence of previous B_i, which are also positive and thus competing candidates for the maximum. Normally, Φ_k itself is already rather small; e.g. for d = -1 and k = 3 as small as 0.042. However, this does not mean that the existence of other competitors is even smaller. On the contrary, B_k is based on all antecessors and if it reaches positive values, the entire path to B_k can be expected to scatter around zero. Therefore, a reasonable estimate for the number of further competitors is $(k-1)/2$, each of them having the same chance to be the maximum as B_k itself. Altogether, there are $(k+1)/2$ competitors, so that the possibility to be the maximum is reduced by the reciprocal:

$$p_{kk} \approx \frac{2}{k+1}\,\Phi_k \tag{31}$$

Although Eq. (31) is only a rough estimate, it fits well for d = -1, as shown in Fig. 8.

For the first two steps, an exact derivation of the backward probability is possible. Moreover, these two values are needed in section 4.3 to determine the hit rate.

Consider the first step of a Brownian motion B_1 with negative drift d. B_1 is the preliminary maximum, if the first increment δ_1 is larger than $|d|$. As the variable δ is standard normal distributed, the probability to exceed $|d|$ is equal to Φ_1 as given by the distribution function Eq. (30), and we can write:

$$p_{11} = P(\delta_1 > |d|) = \Phi_1 \tag{32}$$

Fig. 9 illustrates the situations after the first and after the second step of the Brownian motion. After the first step, the y-axis in Fig. 9a is not yet relevant. The blank area and the sparsely slanted hatched area denote the region where B_0 and B_1 are the maxima, corresponding to the probabilities p_{10} and p_{11}, respectively. Thus, the area defined by Eq. (32) is initially marked by sparsely slanted hatched area.

After the second step of the Brownian motion, the narrow vertical hachure is added, denoting the area where B_2 is the new maximum. This situation is considered in Fig. 9b. B_2 has become the new maximum, if $B_2 > 0$ and $B_2 > B_1$. This is the case, if the two inequalities $\delta_1 + \delta_2 > 2|d|$ and $\delta_2 > |d|$ hold true (narrow vertically hachure). In the double hatched area the maximum is transferred during the second step from B_1 to B_2 and the area keeping B_1 as maximum is reduced. But also the blank area, denoting that B_0 is the maximum, is reduced as it contributed the hatched upper triangle to B_2. What is the probability represented by the different areas? The probability p_{21} (sparsely hatched solely) is easy to determine:

$$p_{21} \;=\; P\,(\delta_1 > |d| \;\wedge\; \delta_2 < |d|) \;=\; \Phi_1(1 - \Phi_1) \tag{33}$$

Deriving the probability p_{22} (all narrow hatches) is more complicated. Let us begin with the total area above the 45° line in Fig. 9b. It contains the probability that the sum of two independent normal distributed variables is exceeding 2, which is equal to:

$$P(\delta_1 + \delta_2 > 2|d|) = \Phi_2 \tag{34}$$

The area given by Eq. (34) is larger than that covered by p_{22}, we are aiming at. So let us cut this area into two equal pieces along the $\delta_1 = \delta_2$ line. Normally, the probability contained in an area is not simply halved, when the area is bisected. However, here it does hold true, because the intersecting line runs towards the zero point, to which the two-dimensional normal density function is point-symmetrical. Thus:

$$P\,(\delta_1 + \delta_2 > 2|d| \;\wedge\; \delta_2 > \delta_1) \;=\; \frac{\Phi_2}{2} \tag{35}$$

Now, half of the double-hatched square passed over from B_1 to B_2 is missing. However, we know the probability of the full square. It is that of two independent variables both exceeding d:

$$P\,(\delta_1 > |d| \;\wedge\; \delta_2 > |d|) \;=\; {\Phi_1}^2 \tag{36}$$

This area has to be halved along the $\delta_1 = \delta_2$ line:

$$P\,(\delta_2 > |d| \;\wedge\; \delta_1 > \delta_2) \;=\; \frac{{\Phi_1}^2}{2} \tag{37}$$

The probability p$_{22}$ is then obtained by the summation of Eqs. (35) and (37).

$$p_{22} \;=\; P\,(\delta_1 + \delta_2 > 2|d| \;\wedge\; \delta_2 > |d|) \;=\; \frac{\Phi_2 + {\Phi_1}^2}{2} \qquad (38)$$

Finally, the probability p$_{20}$ can be concluded as the remaining part of p_{21} (Eq. (33)) and p_{22} (Eq. (38)):

$$p_{20} \;=\; 1 - \Phi_1(1 - \Phi_1) \;-\; \frac{\Phi_2 + {\Phi_1}^2}{2} \qquad (39)$$

4.3 The Hit Rate

In Section 4.1, we defined the hit rate h as the probability that the starting point of a Brownian motion B_0 = 0 remains as maximum throughout the entire process, i.e. that all following B$_i$ are negative. We generated 100,000 random Brownian motion processes to simulate the hit rate for drifts of -2, -1, -0.5, and -0.1 as a function of the length of the process (Fig. 10). Only for d = -0.1, the process has not entirely converged after 50 steps. However, Brownian motion with such a small drift size describes processes, where the scatter is 10 times larger than the standard deviation introduced by the jump height. In such situations, it is clear that break positions are hardly detectable. The errors are very large and the distribution of deviation becomes nearly uniform. These case are of minor interest for practical purposes. However, as an asymptotical solution for large scatter the zero drift case is discussed in Appendix C.

For larger negative drift sizes, i.e. lower scatter, the hit rate converges rather rapidly. For breaks that are half as large as the scatter (d = -0.5), the hit rate h reaches its constant level of 0.529 after about 10 steps of the Brownian motion. If it did not become positive so far, it is very unlikely that the process is creating a positive value later on. Consequently, the initial value remains as maximum, which means that the break position is determined correctly. For the critical case (d = -1), when scatter and breaks are of equal size, the hit rate increases to 0.8 and is reached faster. Already for d= -2, break positions are obviously rather easy to detect, which is documented by a hit rate of 0.976.

The hit rate h is an important measure of error statistics and interesting by itself. However, according to Eq. (29), the hit rate is also necessary to conclude the aimed total probability of a certain deviation k from its preliminary backward probability p_{kk}. On the other hand, hit rate and backward probability are connected via:

$$h + h\sum_{i=1}^{\infty} p_{ii} = 1 \qquad (40)$$

so that h might be concluded from p_{ii}. Eq. (40) describes that the backward probabilities p_{ii} has to be multiplied by h to obtain the final probabilities for the deviations. Their sum plus the hit rate itself adds up to 1. In section 4.2, we derived the first two backward probabilities p_{11} and p_{22}. In this section, we show that they are sufficient to compute h with high accuracy. We transform Eq. (40) and approximate the right-hand side. The idea is the following: instead of the multiplying the right-hand side of Eq. (41) by h < 1, we can alternatively stop the summation earlier at a finite k.

$$1 - h = h \sum_{i=1}^{\infty} p_{ii} \approx \sum_{i=1}^{k} p_{ii} \qquad (41)$$

In Appendix D, we show that *k* = 2 is the optimum choice. It follows:

$$h = 1 - p_{11} - p_{22} \qquad (42)$$

Inserting Eq. (32) and Eq. (38):

$$h = 1 - \Phi_1 - \frac{\Phi_2 + {\Phi_1}^2}{2} \qquad (43)$$

For drift *d* = -1, the two first exceeding probabilities are $\Phi_1 = 0.158$ and $\Phi_2 = 0.079$. Using Eq. (43), *h* is estimated to 0.79, which is in good agreement with the simulated value of 0.80 (Fig.10). Fig. 11 shows such comparisons of estimated and true hit rate for all drifts between -0.05 and -4. It confirms that Eq. (43) is a good estimate for the hit rate.

4.4 Two-sided deviations

So far, we restricted our analyses to one-sided processes. However, deviations are possible to both sides of the true break position. In this section, we determine how the deviation probabilities are distributed for such two-sided processes.

It is random scatter, which is responsible for possible deviations on both sides. Due to its randomness, it produces maxima independently, but in the same way on both sides. To simulate the two-sided situation, we may simply repeat the procedure of the first on the second side. As for the first side, any maximum appearing at a non-zero time step *k* indicates that an error in the break detection occurred. Consequences for the hit rate *h* are directly clear: it is reduced from *h* to h^2, because initially correct cases may become incorrect by the independent second trial. However, also two competing maxima, each on one side may occur. In this case, only one of them, the larger one, is chosen.

For one-sided processes the probability $P1_i$ for a certain deviation *i* is equal to the hit rate times the backward probability as given in Eq. (29). The probability that a competitor occurs on the other side is equal to 1- *h*, whereas in *h* cases no competitor appears.

$$P1_k = h\, p_{kk} = (1 - h)\, h\, p_{kk} + h^2 p_{kk} \qquad (44)$$

If a competitor appears (first summand), the probability to gain is 0.5, because no side is preferred concerning the height of produced maxima. Consequently, in these cases the probability is halved. Otherwise (second summand), the original maximum is confirmed. The two-sided probability for a certain deviation is then:

$$P2_k \;=\; \frac{1-h}{2}\, h\, p_{kk} + h^2 p_{kk} \;=\; \frac{1+h}{2}\, h\, p_{kk} \qquad (45)$$

Thus, for two-sided processes, the one-sided probabilities are reduced by the factor $(1+h)/2$ (compare Eq. 29).

We perform the reverse proof: The sum over all deviations of Eq. (45) is:

$$2\sum_{i=1}^{\infty} P2_i \;=\; (1+h)\; h \sum_{i=1}^{\infty} p_{ii} \qquad (46)$$

By Eq. (40), we can substitute the right-side sum:

$$2\sum_{i=1}^{\infty} P2_i \;=\; (1+h)\; h\; \frac{1-h}{h} \;=\; 1-h^2 \qquad (47)$$

Together with the hit rate, which is reduced to h^2, the overall sum of probabilities equals again 1, which is a necessary condition for the correctness of Eq. (45).

Fig. 13 shows the consequences for practical applications for three SNRs. Error probabilities for SNR = 1 are given by the thick line attaining a hit rate of 64%. The hit rate drops from 95% for SNR = 2 to 29% for SNR = ½. We can conclude that for SNR > 1 the break search becomes quickly very exact, whereas it becomes at the same time very inexact for SNR < 1.

5 Conclusions

Non-climatic breakpoints in climate time series cause abrupt jumps with constant time segments in between multiple breakpoints. Multiple breakpoint homogenization algorithms use as detection criterion the maximum external variance between the means of such segments. The difference of external variance between the correct and a wrong segmentation with shifted break positions is shown to be a quadratic function of a random normal variable ε (Eq. 8) that depends on the length of the erroneously exchanged subsegment and the signal to noise ratio (SNR), which is defined as the ratio of the external and internal standard deviation between and within the segments, respectively (Eq. 10).

The potentially gained variance by selecting the wrong segmentation can be described by a Brownian motion with drift, where the strength of the (negative) drift is given by the SNR. Since the maximum variance is relevant here, we need to know the time, when the maximum occurs. Furthermore, to determine probabilities for these events, the distribution of the time of the maximum of a Brownian motion with drift had to be determined. We expanded the available equation for continuous and one-sided processes (Buffet, 2003) to discrete and two-sided processes, as they occur in homogenization.

The probability for a certain deviation is decomposed into the backward and the forward probability, where the latter is identical to the hit rate h. An estimate for the backward probability is given by Eq. (31); exact solutions are derived for deviations of one (Eq. 32) and two time steps (Eq. 38), which are sufficient to determine the hit rate with high accuracy (Eq. 43).

For two-sided processes the hit rate is squared leading to a reduction; the probabilities of non-zero deviations (summed over both sides) are increased in return by the factor 1+h (Eq. 45). We showed that the error probability for SNRs less than 1 is high (Fig. 13). As consequence, break search algorithms will generally show large uncertainties in the date of the break when the internal variance is larger than the variance introduced by the breaks.

Appendix A

Consider the contribution of two neighboring segments of length n_1 and n_2 with the means x_1 and x_2 to the external variance v_a of a time series with length n:

$$nv_a = n_1 {x_1}^2 + n_2 {x_2}^2 + rest \qquad (A1)$$

where *rest* stands for the external variance of the other segments. If the first k members of the second segment, which have the average x_0, are erroneously shifted from segment 2 to segment 1, the two segment means are changed to:

$$x_{1n} = \frac{n_1 x_1 + k x_0}{n_1 + k} \qquad (A2)$$

and

$$x_{2n} = \frac{n_2 x_2 - k x_0}{n_2 - k} \qquad (A3)$$

After the exchange of the boundary subsegment, the altered external variance v_b is given by:

$$nv_b = (n_1 + k)\, {x_{1n}}^2 + (n_2 - k)\, {x_{2n}}^2 + rest \qquad (A4)$$

Inserting Eq. (A2) and Eq. (A3) into Eq. (A4) leads to:

$$nv_b = \frac{n_1{}^2 x_1{}^2 + 2kn_1x_0x_1 + k^2x_0{}^2}{n_1 + k} + \frac{n_2{}^2x_2{}^2 - 2kn_2x_0x_2 + k^2x_0{}^2}{n_2 - k} + rest \qquad (A5)$$

The change of external variance is given by subtracting Eq. (A1) from Eq. (A5):

$$n(v_b - v_a) = \frac{n_1{}^2 - n_1(n_1 + k)}{n_1 + k}\, x_1{}^2 + \frac{n_2{}^2 - n_2(n_2 - k)}{n_2 - k}\, x_2{}^2 + 2\frac{kn_1}{n_1 + k}x_1x_0 - 2\frac{kn_2}{n_2 - k}x_2x_0 + \left(\frac{m^2}{n_1 + k} + \frac{m^2}{n_2 - k}\right)x_0{}^2 \qquad (A6)$$

which can be reduced to:

$$n(v_b - v_a) = \frac{kn_1}{n_1 + k}\, x_1{}^2 + \frac{kn_2}{n_2 - k}\, x_2{}^2 + 2\frac{kn_1}{n_1 + k}x_1x_0 - 2\frac{mk}{n_2 - k}x_2x_0 + \frac{k^2(n_1 + n_2)}{(n_1 + k)(n_2 - k)}x_0{}^2 \qquad (A7)$$

We introduce the following abbreviations:

$$f_0 := \frac{k^2(n_1 + n_2)}{(n_1 + k)(n_2 - k)} \qquad (A8)$$

$$f_1 := \frac{kn_1}{n_1 + k} \qquad (A9)$$

$$f_2 := \frac{kn_2}{n_2 - k} \qquad (A10)$$

Additionally, the factor f_0 can be replaced by $f_2 - f_1$, since:

$$f_2 - f_1 = \frac{kn_1}{n_1 + k} + \frac{kn_2}{n_2 - k} = \frac{kn_2(n_1 + k) - kn_1(n_2 - k)}{(n_1 + k)(n_2 - k)} = \frac{k^2(n_1 + n_2)}{(n_1 + k)(n_2 - k)} = f_0 \qquad (A11)$$

so that Eq.(A7) can be reduced to:

$$n(v_b - v_a) = -f_1\, x_1{}^2 + f_2\, x_2{}^2 + 2x_0(f_1x_1 - f_2x_2) + (f_2 - f_1)\, x_0{}^2 \qquad (A12)$$

Appendix B

The mean of the exchanged subsegment x_0 is equal to x_2 plus a random scatter variable δ:

$$x_0 = x_2 + \delta \tag{B1}$$

Inserting Eq. (B1) into Eq. (A12), the change of external variance becomes:

$$n(v_b - v_a) = -f_1\, x_1{}^2 + f_2\, x_2{}^2 + 2(x_2 + \delta)(f_1 x_1 - f_2 x_2) + (f_2 - f_1)\,(x_2 + \delta)^2 \tag{B2}$$

Eq. (B2) can be expanded to:

$$\begin{aligned} n(v_b - v_a) = {} & -f_1\, x_1{}^2 + f_2 x_2{}^2 + 2f_1 x_1 x_2 - 2f_2 x_2{}^2 + 2\delta f_1 x_1 - 2\delta x_2 + f_2\,(x_2{}^2 + 2\delta x_2 + \delta^2) \\ & - f_1\,(x_2{}^2 + 2\delta x_2 + \delta^2) \end{aligned} \tag{B3}$$

and reduced to:

$$n(v_b - v_a) = -f_1\, x_1{}^2 + 2f_1 x_1\, x_2 - f_1 x_2{}^2 + 2\delta f_1 x_1 + f_2 \delta^2 - 2\delta f_1 x_2 - f_1 \delta^2 \tag{B4}$$

which can be further reduced to:

$$n(v_b - v_a) = -f_1\,(x_1 - x_2)^2 + 2f_1(x_1 - x_2)\delta + (f_2 - f_1)\delta^2 \tag{B5}$$

Dividing Eq. (B5) by $D^2 = (x_1 - x_2)^2$ and replacing $f_2 - f_1$by f_0 (Eq. A11) yields a quadratic equation in ε:

$$v^* := \frac{n(v_b - v_a)}{D^2} = -f_1 + 2f_1\varepsilon + f_0\varepsilon^2 \tag{B6}$$

Appendix C: Zero Drift

Brownian motion without drift gives the asymptotical solution for infinite noise relative to the jump height. We denote the probability that the running maximum after *i* steps is located at time step *j* as *p*(i, j). The Brownian motion starts at *i* = 0 with B_0 = 0 so that p(0,0) is 1. After the first step p(1,0) and p(1,1) are both 0.5, because x_1 is equally likely positive or negative. In the second step, there are two possibilities that B_2 becomes the new maximum. Either B_0 or B_1 is the preliminary maximum. The chance to overtake it from B_1 is 0.5, that to gain it from B_0 is 0.25.

By each step a certain amount of probability is passed over from the previous time steps B_0 to B_{i-1} towards B_i. The gain factors are ½ for B_{i-1}, ¼ for B_{i-2} and 1/6 for B_{i-3} etc. so that a recursive formula for the probability of the preliminary last step to be the maximum is:

$$p(i,i) = \sum_{j=0}^{i-1} \frac{p(i-1,j)}{2(i-j)} \qquad (C1)$$

The probabilities of the previous time steps have to be reduced accordingly:

$$p(i,j) = p(i-1,j)\frac{2(i-j)-1}{2(i-j)} \qquad (C2)$$

In Fig. 12, we used the recursive Eqs. (C1) and (C2) and depicted the results for a discrete-time Brownian motion of 50 steps. The probability for each step to be the maximum of the process is given by vertical columns. For continuous-time processes, Buffet (2003) gives an according function:

$$f(s) = \frac{1}{\pi\sqrt{s\,(t-s)}}, \quad 0 < s < t \qquad (C3)$$

this curve is also shown for comparison.

The probability of the initial time step to persist as maximum p(n,0) is of special interest since it gives the hit rate of finding the correct break position, where n is the length of the considered subperiod. According to Eq. (C2), p(n,0) is:

$$\prod_{n=1}^{n} \frac{2i-1}{2i} = \frac{\binom{2n-1}{n-1}}{2^{2n-1}} \qquad (C4)$$

which converges for large n to:

$$\lim_{n\to\infty} \frac{\binom{2n-1}{n-1}}{2^{2n-1}} = \frac{1}{\sqrt{\pi n}} \qquad (C5)$$

For n = 50, Eq. (C5) yields 0.08, which is in good agreement with the height of the first column in Fig. 12. However, from the curve for a continuous-time process (Eq. (C3)), this value is hardly to extract. Similar considerations for the Brownian motion *with* drift made the full analysis performed in section 4 necessary.

Appendix D

In Section 4.3, Eq. (41), the question arises, which number of summands *k* on the right-hand side is optimal to estimate the left-hand side.

$$h \sum_{i=1}^{\infty} p_{ii} \approx \sum_{i=1}^{k} p_{ii} \qquad (D1)$$

As Figure 8 shows, the decrease of backward probabilities p_{ii} is nearly linear in logarithmic scale so that the sum of p_{ii} can be approximated by an infinite geometric series with constant shrinking factor q. Geometric series $\sum_{k=0}^{\infty} a_0 q^k$ have the limit $a_0/(1-q)$, so that Eq. (D1) may be rewritten to:

$$h \frac{p_{11}}{1-q} \approx \sum_{i=0}^{k-1} p_{11} q^k \qquad (D2)$$

It follows:

$$h \approx (1-q)\left(1 + q + q^2 + \cdots + q^{k-1}\right) = 1 - q^k \qquad (D3)$$

For drift d = -1, the slope in Fig. 8 is about -0.7. The factor q can be estimated by:

$$q \cong e^{-0.7} \cong {}^{1}\!/_{2} \qquad (D4)$$

From Fig. 10 we know that h = 0.8. Obviously, k = 2 is the best exponent to bring both sides of Approximation (D3) in agreement. This holds true also for other drift sizes, as Fig. 11 shows.

References

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

Auer I, Böhm R, Jurković 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, Gajić-Čapka M, Zaninović K, Majstorović Ž, Szalai S, Szentimrey T, Mercalli L. 2005. A new instrumental precipitation dataset for the greater Alpine region for the period 1800-2002. *Int. J. Climatol.* **25:** 139-166.

Beaulieu C, Seidou O, Ouarda TBMJ, Zhang X, Boulet G, Yagouti A. 2008. Intercomparison of homogenization techniques for precipitation data. *Water Resour. Res.* **44**: W02425, DOI: 10.1029/ 2006WR005615.

Brohan P, Kennedy J, Harris I, Tett SFB, Jones PD. 2006. Uncertainty estimates in regional and global observed temperature changes: a new dataset from 1850. *J. Geophys. Res.* **111**:D12106.

Brunet M, Asin J, Sigró J, Bañon M, García F, Aguilar E, Palenzuela JE, Peterson TC, 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, Nanni T. 2006. Temperature and precipitation variability in Italy in the last two centuries from homogenized instrumental time series. *International Journal of Climatology* **26:** 345–381.

Buffet E. 2003. On the time of the maximum of a Brownian motion with drift, *J. Appl. Math. Stoch. Anal.* **16**: 3, 201-207.

Buishand TA. 1982. Some methods for testing the homogeneity of rainfall records. *J. hydrol.* **58:** 11-27.

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

Chandler RE, Thorne P, Lawrimore J, Willett K. 2012. Building trust in climate science: data products for the 21st century. *Environmetrics* **23:** 373–381. DOI: 10.1002/env.2141.

Domonkos P. 2013. Efficiencies of inhomogeneity-detection algorithms: comparison of different detection methods and efficiency measures. *Journal of Climatology*, Art. ID 390945, DOI: 10.1155/2013/390945.

Easterling DR, Peterson TC. 1995. A new method for detecting undocumented discontinuities in climatological time series. *Int. J. Climatol*. **15:** 369-377.

Lawrimore JH, Menne MJ, Gleason BE, Williams CN, Wuertz DB, Vose RS, Rennie J. 2011. An overview of the Global Historical Climatology Network monthly mean temperature data set, version 3. *J. of Geophysical Research-Atmospheres* **116:** (D19121). DOI: 10.1029/2011jd016187.

Lindau R. 2003. Errors of Atlantic air-sea fluxes derived from ship observations. *J. Clim.* **16:**, No.4, 783-788.

Lindau, R, Venema VKC. 2013. On the multiple breakpoint problem and the number of significant breaks in homogenisation of climate records. *Idojaras, Quarterly journal of the Hungarian Meteorological Service* **117:** no. 1, 1-34.

Lu, QQ, Lund R, Lee TCM. 2010. An MDL approach to the climate segmentation problem. *The Annals of Applied Statistics* **4**: 299--319. DOI: 10.1214/09-AOAS289.

Menne MJ, Williams Jr. CN. 2009. Homogenization of temperature series via pairwise comparisons. *J. Climate* **22:** 1700–1717.

Morice, CP, Kennedy JJ, Rayner NA, Jones PD. 2012. Quantifying uncertainties in global and regional temperature change using an ensemble of observational estimates: The HadCRUT4 dataset. *J. Geophys. Res.* **117:** D08101. DOI: 10.1029/2011JD017187.

Rennie JJ, Lawrimore JH, Gleason BE, Thorne PW, Morice CP, Menne MJ, Williams CN, Gambi de Almeida W, Christy JR, Flannery M, Ishihara M, Kamiguchi K, Klein-Tank AMG, Mhanda A, Lister DH, Razuvaev V, Renom M, Rusticucci M, Tandy J, Worley SJ, Venema V, Angel W, Brunet M, Dattore B, Diamond H, Lazzara MA, Le Blancq F, Luterbacher J, Mächel H, Revadekar J, Vose RS, Yin X. 2014. The International Surface Temperature Initiative global land surface databank: Monthly temperature data version 1 release description and methods. Accepted, *Geoscience Data Journal*.

Rohde R, Muller RA, Jacobsen R, Muller E, Perlmutter S, Rosenfeld A, Wurtele J, Groom D, Wickham C. 2013. A new estimate of the average Earth surface land temperature spanning 1753 to 2011. *Geoinfor Geostat: An Overview* 1:1. DOI: 10.4172/2327-4581.1000101.

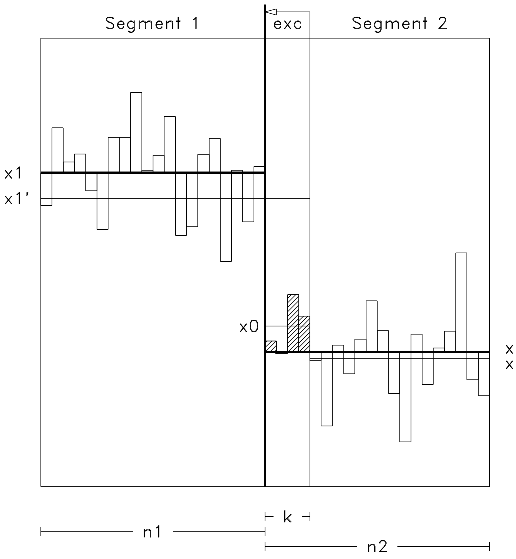
Stott PA, Thorne PW. 2010. How best to log local temperatures? *Nature* **465:** 158–159, DOI: 10.1038/465158a.

Thorne PW, Willett, KM, Allan RJ, Bojinski S, Christy JR, Fox N et al. 2011. Guiding the creation of a comprehensive surface temperature resource for twenty-first-century climate science. *Bull. Amer. Meteor. Soc.* **92:** ES40–ES47.

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

Willett K, Williams C, Jolliffe I, Lund R, Alexander L, Brönniman S, Vincent LA, Easterbrook S, Venema V, Berry D, Warren R, Lopardo G, Auchmann R, Aguilar E, Menne M, Gallagher C, Hausfather Z, Thorarinsdottir T, Thorne PW. 2014. Concepts for benchmarking of homogenisation algorithm performance on the global scale. *Geosci. Instrum. Method. Data Syst. Discuss.*, **4**: 235-270, DOI: 10.5194/gid-4-235-2014.

Williams CN, Menne MJ, Thorne PW. 2012. Benchmarking the performance of pairwise homogenization of surface temperatures in the United States. *J. Geophys. Res.* **117:** D05116. DOI: 10.1029/2011JD016761.



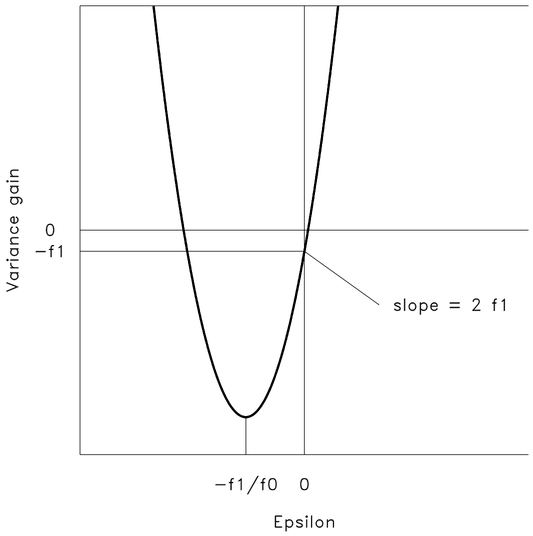
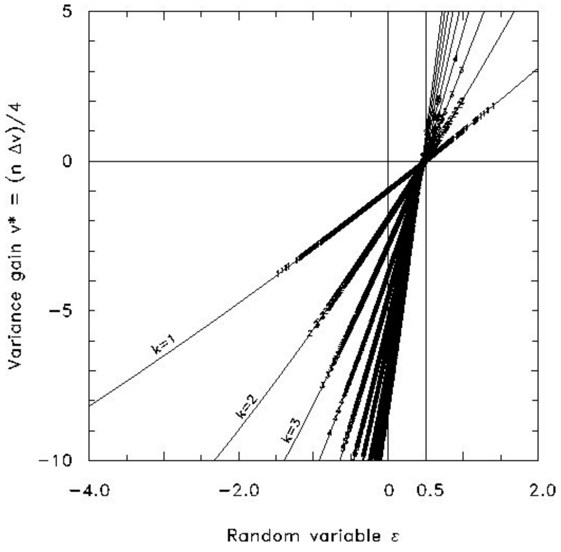


Fig. 1: Sketch to illustrate position errors occurring if a subsegment is erroneously exchanged from segment 2 to segment 1.

Fig.2: Illustration of the variance gain as a function of the random variable ε. The parabola is zero for ε equal to 0.5 and -2f1/f0 -0.5 ≈ -n1/k, respectively.



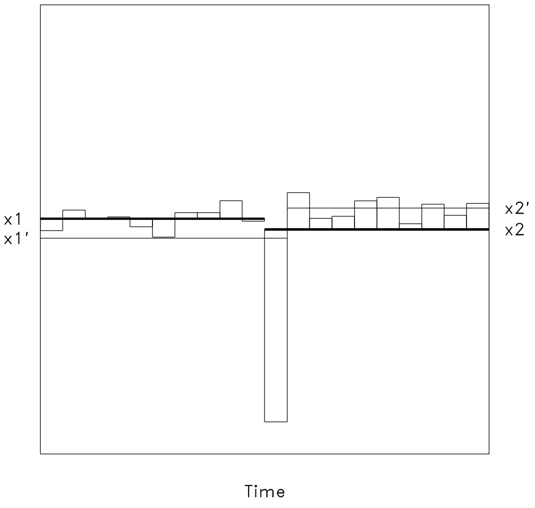
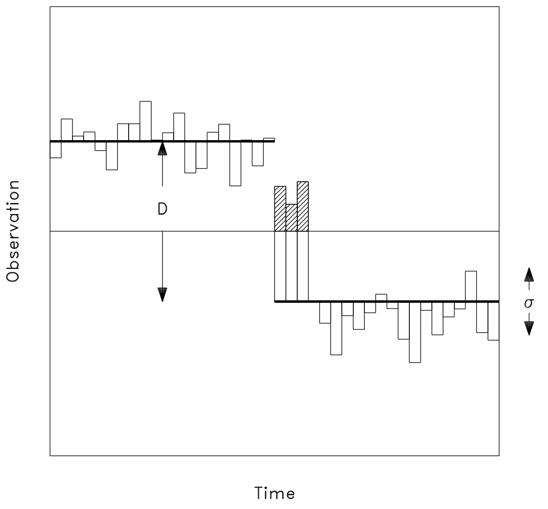


Fig.3: Simulation result obtained from 10,000 random time series of length 100 with SNR = 1. The variance gain is confirmed to be a function of the random variable ε and length of exchanged subsegment k.

Fig.4: Condition to obtain the negative solution as given in Fig. 2 and Eq. (18). If the extreme value adjacent to the break is exchanged, the segment averages x1 and x2 (fat horizontal lines) change to x1' and x2' (thin), which thereby diverge.



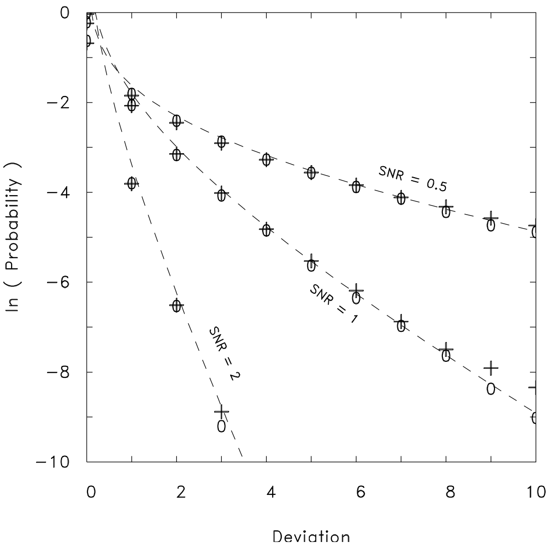
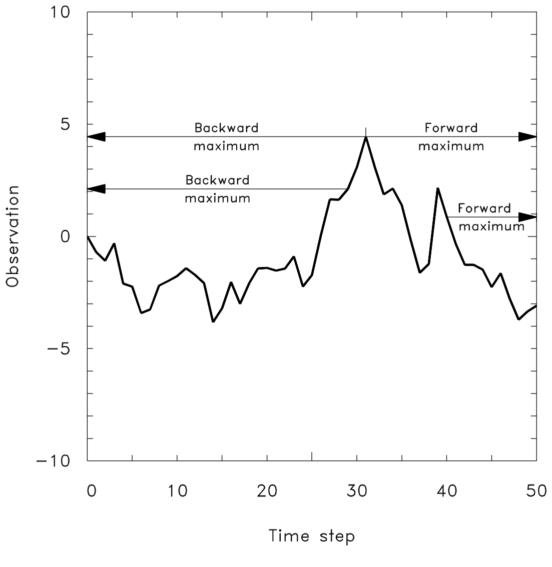


Fig.5: Condition for a positive solution with jump height D and internal variance σ. If an adjacent subsegment rises above the D/2 line, it is erroneously incorporated to the neighbor segment. The maximum hatched area is the critical parameter.

Fig.6: Distribution of the time of the maximum of a Brownian motion with three different drifts (-0.5, -1.0, and -2.0), corresponding to SNR of equal (but positive) size. Results for continuous time processes are obtained by Eq. (27) and denoted by dashed lines. Alternative results for discrete time processes obtained by simulations with 100,000 repetitions are given by circles and crosses. The crosses denote full break search simulations, the circles just Brownian motion.



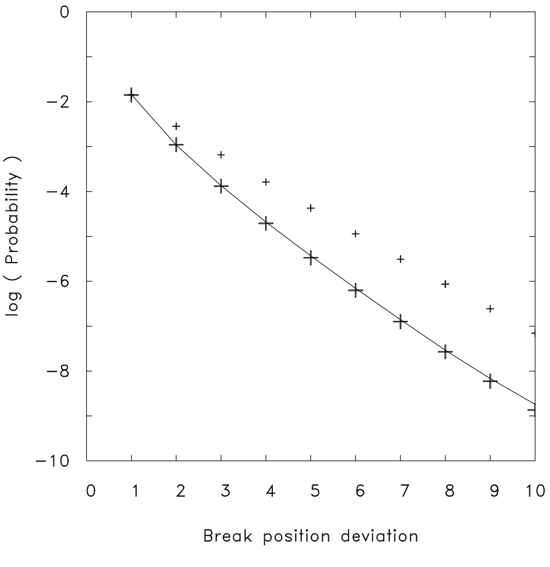


Fig. 7: Sketch to illustrate the expressions backward and forward maximum.

Fig. 8: Backward probability for SNR = 1. Large crosses denote the mean results obtained from simulations with 1,000,000 times series of lengths 100 (truth). The line shows Eq. (31), which simply uses the standard normal distribution function (small crosses) and a reduction factor.

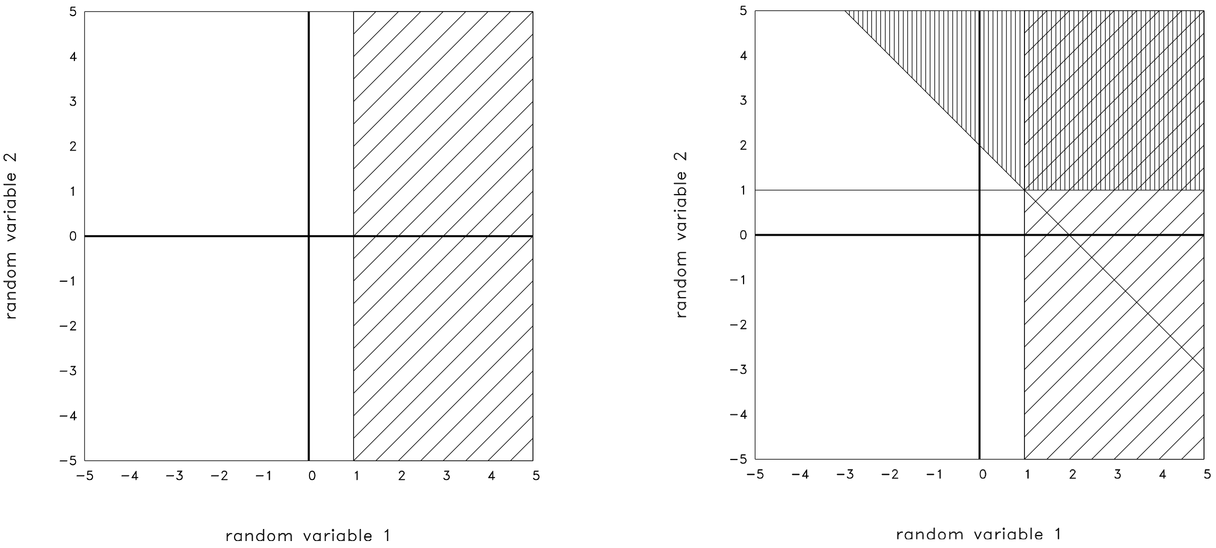
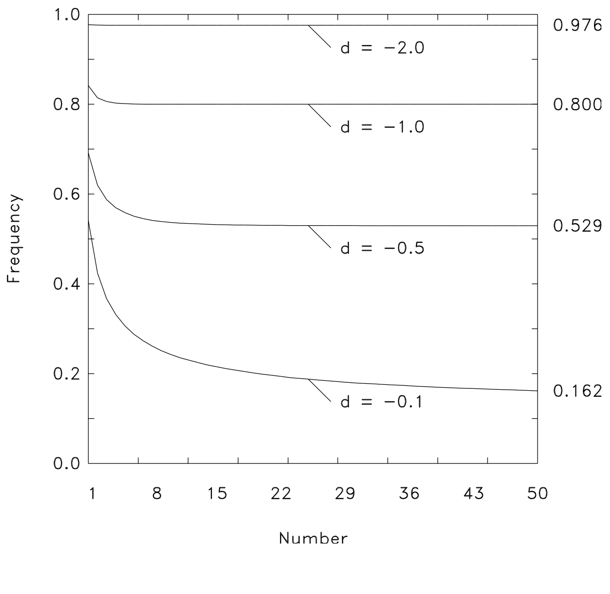


Fig.9: Sketch of the probabilities for the time of the maximum of a Brownian motion with drift -1 after one (left panel: only blank and sparsely slanted hatched area for B_0 and B_1, respectively), and after two steps (right panel: additional narrow vertical hatched area for B_2). This area should be multiplied with the two-dimensional normal density function to compute the probability. For other drift sizes the axes has to be scaled accordingly.



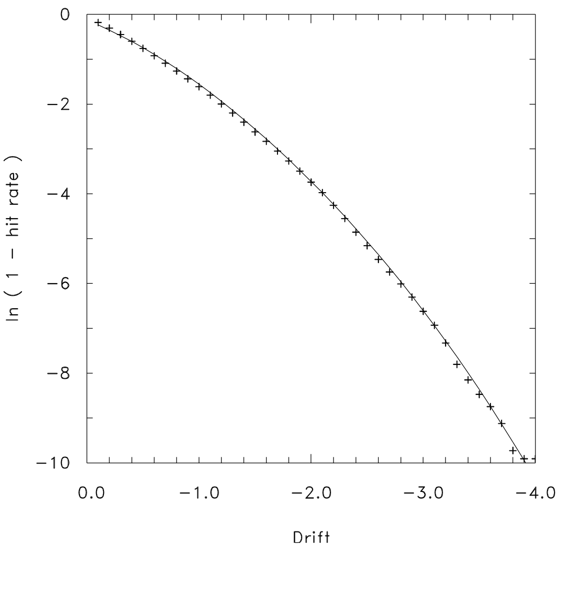
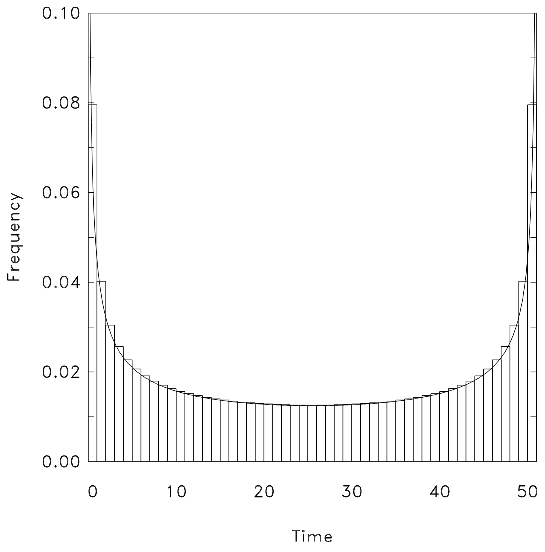


Fig.10: Probability to remain permanently below zero for four Brownian motions with different drifts as a function of the length of the process. This probability is discussed as hit rate h in the text.

Fig.11: Hit rate as a function of drift. The true values obtained by simulation are given by the solid line; crosses denote estimations by Eq. (43).



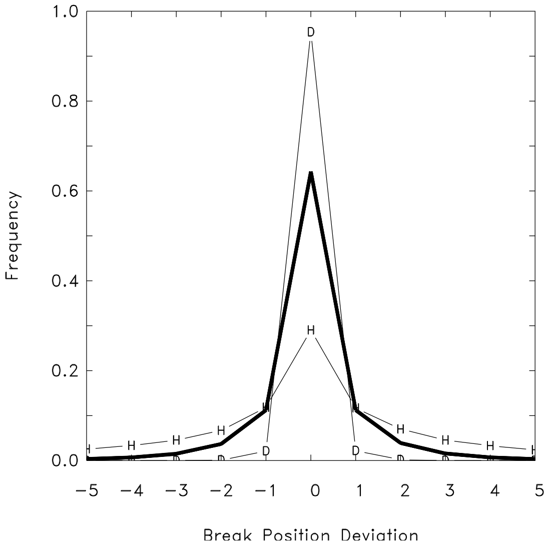


Fig.12: Distribution of the time of the maximum for a 50-step Brownian motion without drift.

Fig.13: Probability to miss the correct breakpoint for a two-sided process. The distributions for three different drift sizes are given (-0.5, -1.0, -2.0). The thick line denotes SNR of 1 (d = -1), half and doubled signals are marked by H and D, respectively.