Earth-Science Reviews 190 (2019) 310–322

Contents lists available at ScienceDirect

Earth-Science Reviews

journal homepage: www.elsevier.com/locate/earscirev

Trend analysis of climate time series: A review of methods

Manfred Mudelsee

Alfred Wegener Institute Helmholtz Centre for Polar and Marine Research, Bussestrasse 24, 27570 Bremerhaven, Germany
Climate Risk Analysis, Kreuzstrasse 27, Heckenbeck, 37581 Bad Gandersheim, Germany

ARTICLE INFO

Keywords:
Bootstrap resampling
Global surface temperature
Instrumental period
Linear regression
Nonparametric regression
Statistical change-point model

ABSTRACT

The increasing trend curve of global surface temperature against time since the 19th century is the icon for the considerable influence humans have on the climate since the industrialization. The discourse about the curve has spread from climate science to the public and political arenas in the 1990s and may be characterized by terms such as “hockey stick” or “global warming hiatus”. Despite its discussion in the public and the searches for the impact of the warming in climate science, it is statistical science that puts numbers to the warming. Statistics has developed methods to quantify the warming trend and detect change points. Statistics serves to place error bars and other measures of uncertainty to the estimated trend parameters. Uncertainties are ubiquitous in all natural and life sciences, and error bars are an indispensable guide for the interpretation of any estimated curve—to assess, for example, whether global temperature really made a pause after the year 1998.

Statistical trend estimation methods are well developed and include not only linear curves, but also change-points, accelerated increases, other nonlinear behavior, and nonparametric descriptions. State-of-the-art, computing-intensive simulation algorithms take into account the peculiar aspects of climate data, namely non-Gaussian distributional shape and autocorrelation. The reliability of such computer age statistical methods has been testified by Monte Carlo simulation methods using artificial data.

The application of the state-of-the-art statistical methods to the GISTEMP time series of global surface temperature reveals an accelerated warming since the year 1974. It shows that a relative peak in warming for the years around World War II may not be a real feature but a product of inferior data quality for that time interval. Statistics also reveals that there is no basis to infer a global warming hiatus after the year 1998. The post-1998 hiatus only seems to exist, hidden behind large error bars, when considering data up to the year 2013. If the fit interval is extended to the year 2017, there is no significant hiatus. The researcher has the power to select the fit interval, which allows her or him to suppress certain fit solutions and favor other solutions. Power necessitates responsibility. The recommendation therefore is that interval selection should be objective and oriented on general principles. The application of statistical methods to data has also a moral aspect.

1. Introduction

A univariate time series is a sample of data values in dependence on time, where for each element of a set of time points, $t(i)$, there exists one corresponding data point, $x(i)$. The time values are assumed to increase strictly monotonically with the counter, i, that means, $t(1) < t(2) < t(3)$, and so forth. The data size or sample size, n, is the number of time–data pairs. The compact notation for a time series sample is $\{t(i),x(i)\}_{i=1}^{n}$. The spacing of a time series is defined as $d(i) = t(i) - t(i-1)$. An evenly spaced time series has a constant spacing. The methods presented in this review can be applied to evenly and unevenly spaced time series. The average spacing of a time series, which is given by $\bar{d} = [t(n) - t(1)]/(n-1)$, is also called time resolution. Although the methods explained in this review are presented for univariate series, it is principally possible to extend them to multivariate time series, where several data points, $x(i)$, $y(i)$, $z(i)$, and so forth, are available at time $t(i)$.

The GISTEMP time series (Fig. 1) is a reconstruction of global surface temperature based on land and ocean data. The x-values are the temperature anomalies relative to the 1951–1980 mean in units of degrees Celsius. The t-values are the years from 1880 to 2017. This is an evenly spaced series of size $n = 138$, and the time resolution is 1 year. The method of calculation of GISTEMP from a number of records from globally distributed measurement stations is described by Hansen et al. (2010).

The data generating system considered in this article is the climate. It is the climate system that is in the center of interest of climate re-

E-mail address: mudelsee@climate-risk-analysis.com.

https://doi.org/10.1016/j.earscirev.2018.12.005
Received 19 June 2018; Received in revised form 17 October 2018; Accepted 6 December 2018
Available online 11 December 2018

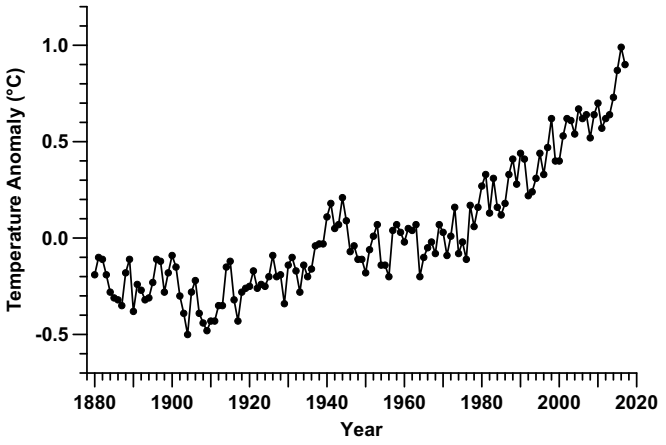

Fig. 1. Global surface temperature time series. Shown against time is the temperature anomaly relative to the 1951–1980 mean. The series is called GISTEMP after its producer, the NASA Goddard Institute for Space Studies, New York, NY, United States of America. (The data from 1880 to 2016 were downloaded from https://data.giss.nasa.gov/gistemp/graphs/graph_data/Global_Mean_Estimates_based_on_Land_and_Ocean_Data/graph.txt on 3 January 2018; the value for 2017 was taken from https://www.giss.nasa.gov/research/news/20180118/ on 19 January 2018.)

searchers, not a particular time series. The time series serves only to learn about the climate by means of statistical analysis of the time series data. The target of the learning procedure considered in this article is the trend, which is, loosely speaking, the long-term systematic change of the mean value over time. This concept is specified in Section 2. The act of learning is called estimation. It is part of applied statistical science. Since the x-values are affected by uncertainties from the measurements, and since the data size is less than infinity, an estimation is never exact. Therefore, an uncertainty measure has to accompany an estimation result to allow the assessment of the accuracy of the estimation result. Estimates without error bars are useless. Statistical science has developed methods for trend estimation and uncertainty determination, which support climate science. This review explains those statistical methods in detail and at a level that is accessible to non-experts. It is based on the book by the author (Mudelsee, 2014). It gives a brief account of the historical development of the methods. As an illustrative example, we study temperature trends for the instrumental period (Fig. 1). The explained methods can evidently be applied to any type of climate time series, also from the deep past.

2. Climate

Climate is a paradigm of a complex system, which comprises the atmosphere, the hydrosphere, the cryosphere, the lithosphere, the biosphere, and the interactions among these compartments (Stocker et al., 2013). For the analysis of climate data, it is useful to employ conceptualizations—mathematical equations—which reduce the complexity. This goes via the introduction of statistical random variables, X, which are representing a climate variable (e.g., temperature) with not exactly known value. The random variables are concatenated over time to form a stochastic process, $X(i)$. Note the distinction between the process, written capitalized, and the non-capitalized numerical value, $x(i)$. This distinction stems from statistical science Brockwell and Davis, 1991). One simple climate equation is

$$X(i) = X_{\mathrm{trend}}(i) + S(i) \times X_{\mathrm{noise}}(i). \quad (1)$$

Eq. (1) decomposes climate into a trend and a noise component. The noise component has mean zero; $E[X_{\mathrm{noise}}(i)] = 0$ for $i = 1, \ldots, n$, where E is the expectation operator. In other words, the center of location of the distribution of $X_{\mathrm{noise}}(i)$ is zero. The center of location for the climate variable, $X(i)$, is, hence, described by the time-dependent component $X_{\mathrm{trend}}(i)$, the trend.

The noise component has standard deviation unity; $STD[X_{\mathrm{noise}}(i)] = 1$ for $i = 1, \ldots, n$, where STD is the standard deviation operator. In other words, the spread of the distribution of $X_{\mathrm{noise}}(i)$ is unity. The spread for the climate variable, $X(i)$, around the trend, $X_{\mathrm{trend}}(i)$, is, hence, described by the time-dependent scaling function $S(i)$, the variability. Whereas $X(i)$ and $X_{\mathrm{trend}}(i)$ have physical units (e.g., degrees Celsius), $X_{\mathrm{noise}}(i)$ has not. The units are being brought into the noise component via the function $S(i)$.

Eq. (1) is a mathematical representation of the definition of climate in terms of mean and variability. This definition was developed around the end of the 19th century by Austrian, German and Russian–German researchers (Brückner, 1890; Hann, 1901; Köppen, 1923). Today, at the beginning of the 21st century, it seems appropriate to extend the definition to a full statistical description of the climate system (Stocker et al., 2013)—to one that includes not only the first statistical moment (expectation) or the second (standard deviation), but also higher orders and extremes. A climate equation that includes also an extreme component was presented by Mudelsee (2010: Chapter 1).

The noise component, $X_{\mathrm{noise}}(i)$, has a distribution with mean zero and standard deviation unity, but the full shape of the distribution is not prescribed. Climate variables often show noise distributions that differ in shape from a normal or Gaussian, bell-shaped form. The noise component for climate variables often exhibits autocorrelation: if $X_{\mathrm{noise}}(i)$ is positive, then $X_{\mathrm{noise}}(i+1)$ is likely also positive. This "memorizing ability" of climate may act on many timescales. It is also called persistence or serial dependence. These two climate noise ingredients, non-Gaussian shape and autocorrelation, have to be taken into account in the estimation and uncertainty determination. This can be done by suitable statistical techniques (Section 3). If they are are ignored, then there is a risk of bias and overstatements stemming from too small uncertainties.

Eq. (1) has the noise term added. It may be an interesting option to describe climate by means of multiplicative noise, but this seems not yet to have been seriously tried.

3. Regression

Regression is a statistical method to estimate the trend component, $X_{\mathrm{trend}}(i)$, from the climate Eq. (1). It can also be applied to a climate equation with an extreme component (Mudelsee, 2014). Required is a statistical regression model of the trend. The input to the regression method is the time series, $\{t(i), x(i)\}_{i=1}^{n}$. If the statistical model employs parameters, then the output is in the form of estimated parameter values with uncertainty measure. This review article considers the simple linear model (Section 3.1) and more complex nonlinear models (Section 3.2). It is also possible to not specify a parametric model and instead estimate $X_{\mathrm{trend}}(i)$ by means of smoothing techniques (Section 3.3).

3.1. Linear regression

The linear regression describes $X_{\mathrm{trend}}(i)$ by means of two parameters, namely the intercept, β_0, and the slope, β_1. The model is "on the process level" given by

$$X(i) = \beta_0 + \beta_1 \times T(i) + S(i) \times X_{\mathrm{noise}}(i). \quad (2)$$

$T(i)$ is the time variable assigned to $X(i)$.

The ordinary least-squares (OLS) estimation minimizes the sum of squares of differences between data and the linear fit. It is "on the sample level" given by

$$SSQ(\beta_0, \beta_1) = \sum_{i=1}^{n} [x(i) - \beta_0 - \beta_1 \times t(i)]^2 \quad (3)$$

The estimates, denoted with a "hat" as $\widehat{\beta}_0$ and $\widehat{\beta}_1$, can be calculated

by setting the first derivatives of $SSQ(\beta_0,\beta_1)$ with respect to β_0 and β_1 equal to zero. This yields two linear equations—called normal equations—for two unkwnows, which can be analytically solved. "Analytical" means that the solution consists of two simple formulas,

$$\widehat{\beta}_0 = \left[\sum_{i=1}^{n} x(i) - \widehat{\beta}_1 \times \sum_{i=1}^{n} t(i)\right]/n, \tag{4}$$

$$\widehat{\beta}_1 = \left\{\left[\sum_{i=1}^{n} t(i)\right] \times \left[\sum_{i=1}^{n} x(i)\right]/n - \sum_{i=1}^{n} t(i) \times x(i)\right\} \times \left\{\left[\sum_{i=1}^{n} t(i)\right]^2/n - \sum_{i=1}^{n} [t(i)]^2\right\}^{-1}. \tag{5}$$

It can be shown that the second derivatives of $SSQ(\beta_0,\beta_1)$ at the solution point $(\widehat{\beta}_0,\widehat{\beta}_1)$ are positive, which means that the extremum is a minimum.

There exist analytical, "classical" formulas for the uncertainty measures for $\widehat{\beta}_0$ and $\widehat{\beta}_1$. However, these formulas are based on the assumptions of Gaussian distributional shape and absent autocorrelation—a quite unrealistic situation for climate data (Section 2). A superior method is moving block bootstrap (MBB) resampling Künsch, 1989). A description of the MBB for the application to linear OLS regression is given in Appendix A.

Fig. 2a shows the linear regression model fitted by means of OLS to the GISTEMP time series. The intercept estimate with MBB standard error is

$$\widehat{\beta}_0 \pm \mathrm{se}_{\widehat{\beta}_0} = -14.0^{\circ}\mathrm{C} \pm 1.6^{\circ}\mathrm{C}. \tag{6}$$

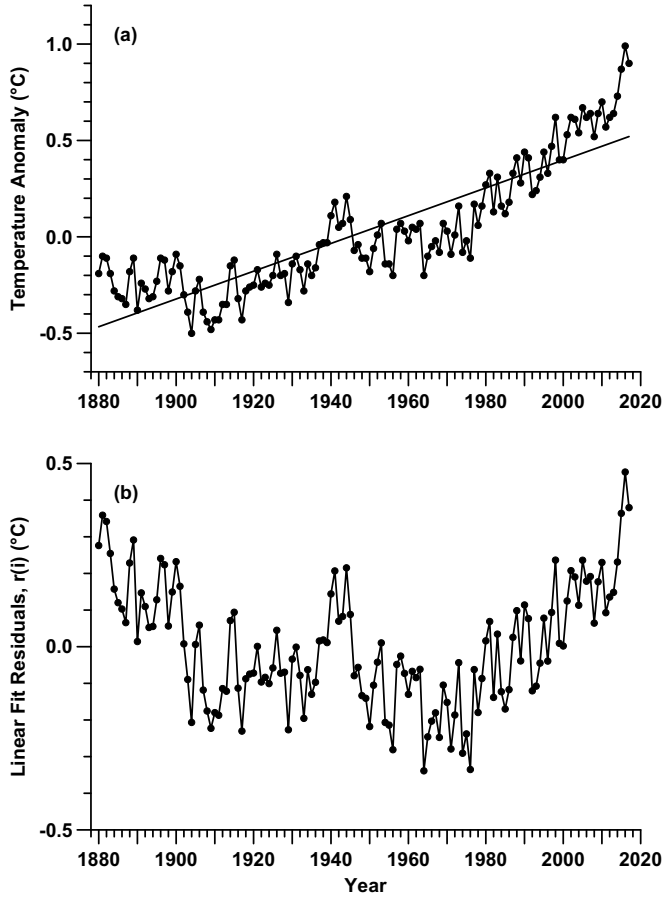

Fig. 2. GISTEMP, linear OLS regression. (a) Time series and linear fit (solid line); (b) residuals against time.

The slope estimate with MBB standard error is

$$\widehat{\beta}_1 \pm \mathrm{se}_{\widehat{\beta}_1} = 0.0072^{\circ}\mathrm{C/a} \pm 0.0008^{\circ}\mathrm{C/a}. \tag{7}$$

Fig. 2b shows the linear regression residuals (Eq. A.1) against time. It appears obvious to the eye that there is still structure in the residuals; this structure consists of (1) a long-term trend and (2) systematic shorter-term deviations, such as the positive residuals at around the year 1940.

In other words, the linear model is not well suited to describe the trend for the GISTEMP time series. The intercept and slope estimates should therefore be interpreted with caution (see Section 4).

Fig. 3a shows a histogram of the linear regression residuals for the GISTEMP time series. It appears that there exists a mild right-skewness, a deviation from the unskewed Gaussian shape.

Fig. 3b shows a lag-1 scatterplot of the linear regression residuals. It appears obvious that there is autocorrelation because the cloud of points is oriented along the 1:1 line. The estimated persistence time (Mudelsee, 2002) on the residuals is

$$\widehat{\tau} = 4.0 \text{ a} \pm 1.0 \text{ a}. \tag{8}$$

Fig. 3 demonstrates on the GISTEMP time series that the assumptions of Gaussian shape and absent autocorrelation are violated. Therefore, it is important to employ the MBB (or other bootstrap variants) for the calculation of uncertainty measures (see Section 4).

Efron (1979, 1982) collected and synthesized earlier works by others and presented ordinary (i.e., pointwise or block length unity) bootstrap resampling. He showed theoretically on simple estimation problems—including Gaussianity and absent autocorrelation—the correctness of the bootstrap. In other words, he showed that resampling yields the same values for the uncertainty measures as the classical approach, which is based on assumptions such as the Gaussian. The advantage of the ordinary bootstrap is that it yields reliable uncertainty measures also for situations where the distributional shape is unknown and may deviate from a Gaussian. The bootstrap "lets the data speak for themselves." The reason of the bootstrap's success is that resampling from the residuals preserves the distributional shape of the noise component. Singh (1981) soon pointed out that the ordinary bootstrap fails for autocorrelated noise components because it does not preserve the autocorrelation. Künsch (1989) presented the MBB (Fig. 4) as a method to preserve autocorrelation (over the length of a block). Efron and Tibshirani (1986) and Hall (1988) pursued the construction of confidence intervals from bootstrap replications. Efron and Tibshirani (1993) wrote a textbook on the bootstrap. Efron and Hastie (2016) showed how the bootstrap is embedded into the modern computer age statistical methodology.

The validity of the MBB can be tested by means of Monte Carlo simulation experiments, which generate artificial data from stochastic processes with known properties. One such test is about the coverage of confidence intervals. For example, what coverage does a 95% confidence interval (for a parameter estimate, say, $\widehat{\beta}_1$) achieve. In other words, what is the fraction of simulations for which the confidence interval does contain the prescribed, and hence known, value (for β_1). There exist other autocorrelation-preserving bootstrap methods than mentioned, and there exists a variety of confidence interval construction methods. Mudelsee (2014) gives an overview of those methodical aspects, presents own Monte Carlo tests, and shows applications to climate time series.

The previous paragraphs have addressed the computational aspect of linear regression. Another, key aspect is the suitability of the linear model (Eq. 2). For climatological applications, this means the question whether a linear increase (or decrease) is not too simple for describing the trend component. Model suitability can be evaluated graphically via various types of plots of the regression residuals (Eq. A.1). These realizations of the noise process should nominally not exhibit more structure than the assumed AR(1) autocorrelation model (Montgomery

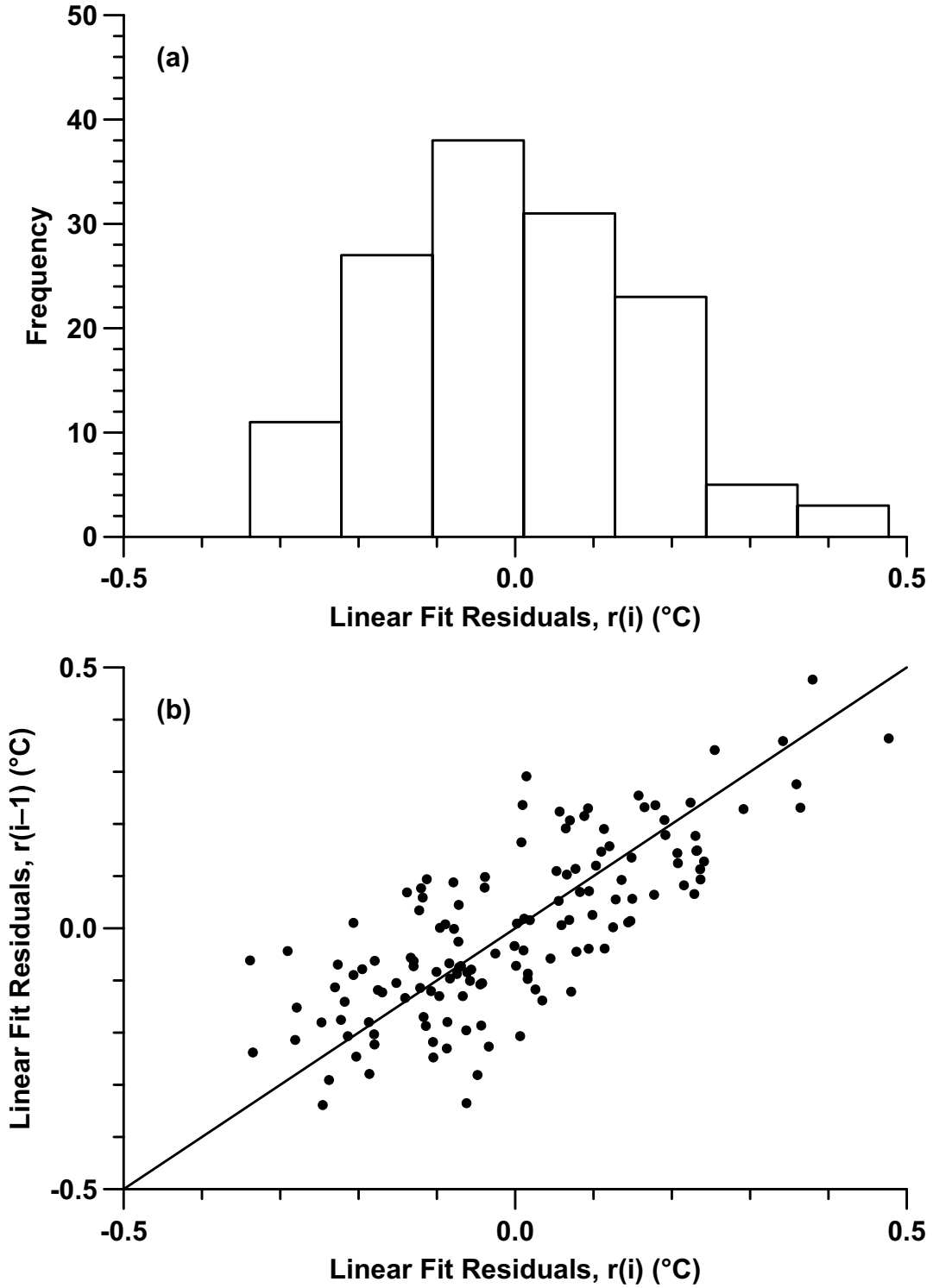

Fig. 3. GISTEMP, linear OLS regression, residuals. (a) Histogram; the number of classes follows the rule by Scott (1979). (b) Lag-1 scatterplot (filled circles) with 1:1 line (solid line).

and Peck, 1992) (see Section 4). If the linear model is found too simple, then nonlinear models may lead to further insights (Section 3.2).

There are more estimation procedures for the linear model than minimizing the OLS sum (Eq. 3). Weighted least-squares (WLS) employs a weighting and minimizes the sum

$$SSQW(\beta_0, \beta_1) = \sum_{i=1}^{n} [x(i) - \beta_0 - \beta_1 \times t(i)]^2 / S(i)^2. \quad (9)$$

The idea is that points, i, with smaller variability, $S(i)$, contribute heavier to the calculation of the regression parameter estimates. This may lead to smaller classical standard errors than from using OLS (Sen and Srivastava, 1990). The challenge with WLS is that $S(i)$ is usually unknown and has to be estimated. For example, Mudelsee and Raymo (2005) fitted ramp models to $S(i)$ per eye for data documenting the glaciation of the Northern Hemisphere in the Pliocene. Generalized least-squares (GLS) employs another sum of squares, which includes

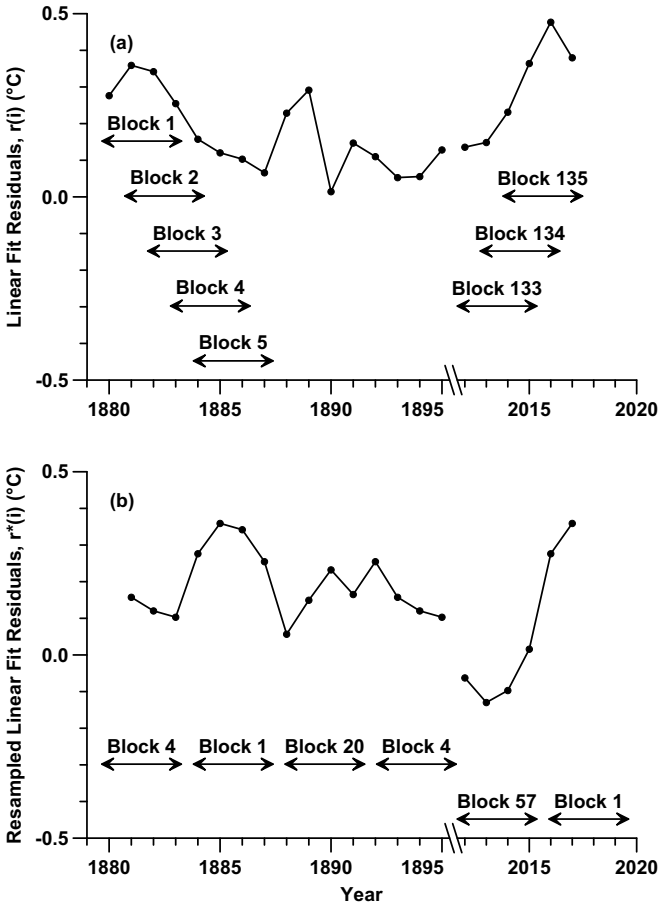

Fig. 4. GISTEMP, MBB resampling. (a) Residual time series; note broken time axis. The series is segmented into blocks of length $l = 4$. (b) MBB resample; note broken time axis. The resample is obtained by randomly concatenating blocks. Only two points (i.e., half a block) are required for the last two points. It is possible to resample a block several times. The shown selection (4,1,20,4, …,57,1) is random; other calls of the MBB procedure may generate different selections.

also the autocorrelation structure (Montgomery and Peck, 1992). Also GLS may lead to smaller classical standard errors than OLS or WLS (Sen and Srivastava, 1990). In his book, Mudelsee (2014: Fig. 4.2) calculated standard-error ratios (GLS over OLS) to quantify this effect in dependence on the strength of the (equivalent) autocorrelation coefficient, finding modest improvements of GLS over OLS. The challenge with GLS is that not only $S(i)$, but also the autocorrelation parameter is usually unknown and has to be estimated. The WLS and GLS challenges can be met by means of iterative procedures (Prais and Winsten, 1954). These are based on making initial guesses of variability (and autocorrelation), performing a first estimation, updating the variability (and autocorrelation) estimates via usage of the regression residuals, and repeating the estimation. This procedure is iterated until parameter estimates and least-squares sum do not change significantly (measured by means of the accuracy of the representation of real numbers in the computer).

There are more autocorrelation models than AR(1), although it is a matter of debate how useful they are in climatological practice. An argument for usage of the simple AR(1) model is that the complexity can be put into the other terms of the climate equations, such as nonlinear trend components (Section 3.2) or an extreme component. Another argument is that for usage in MBB resampling, only rough knowledge about the AR(1) autocorrelation parameter is required for selecting the block length (Eq. A.5). That means, the AR(1) parameter should in many situations capture a good portion of the autocorrelation structure and deliver reliable uncertainty measures from MBB resampling. The AR(1) model has an exponentially (i.e., fast) decaying autocorrelation function (Brockwell and Davis, 1991; Priestley, 1981). It is hence called a short-memory model. In case of long-memory autocorrelation (Beran, 1994; Doukhan et al., 2003; Robinson, 2003), where the autocorrelation function decays hyperbolically (i.e., slowly), the MBB may have to be replaced by a subsampling procedure to yield reliable uncertainty measures (Lahiri, 2003). Subsampling means that the resample has fewer points than the original sample and that only one random block is drawn (Politis et al., 1999). The difficulty here is to adequately set the subsampling length (Lahiri, 2003). The book by Mudelsee (2014; Table 4.5) presents a Monte Carlo experiment on subsampling block length selection in linear OLS regression. To summarize, the simple AR(1) autocorrelation model has a good empirical and theoretical justification in climatology, and uncertainty measures from MBB resampling based on the AR(1) model (Eq. A.5) should be reliable. There may be situations, however, where long-memory models should be considered as alternatives, but usage of those has then to be justified on basis of a climatological theory. One example is river runoff, for which a hydrological explanation of long memory (i.e., the Hurst phenomenon) via spatial aggregation of AR(1) components in a river network was given by Mudelsee (2007).

Paleoclimate time series, $\{t(i), x(i)\}_{i=1}^{n}$, are mostly obtained via the measurement of proxy variables for the climate on natural archives, such as marine sediment cores. (Another archive of paleoclimate is formed by historical documents.) The timescale construction is based on absolute dating of fixpoints in the archive and a statistical regression model for the accumulation of the archive. A certainly not exhaustive list of early works on timescale construction is the following: Bennett (1994); Agrinier et al. (1999); Bennett and Fuller (2002); Buck and Millard (2004); Drysdale et al. (2004); Blaauw and Christen (2005); Heegaard et al. (2005); Spötl et al. (2006); Haslett and Parnell (2008); Blaauw (2010); Klauenberg et al. (2011); Scholz and Hoffmann (2011); Blaauw and Heegaard (2012); Fohlmeister (2012); Hendy et al. (2012); Hercman and Pawlak (2012). Absolute dating inevitably shows measurement errors, and the accumulation model as well is subject to systematic and random errors. These error influences can be captured by means of a statistical accumulation model, which is able to generate simulated timescales, $t^*(i)$. The simulated timescales, in turn, are fused into the resample, $\{t^*(i), x^*(i)\}_{i=1}^{n}$, which is used for generating the bootstrap replications (Step 8). The topic of uncertain timescales has not been a focus of statisticians in the past, and the methodology of quantitatively assessing the influence of timescale uncertainties (Mudelsee (2010) is, at the time of writing (mid 2018), not very far advanced.

3.2. Nonlinear regression

Climate is a complex system, and usually the trend component is better described by a nonlinear than a linear regression function. Fig. 5 shows a selection of useful nonlinear functions.

The parabolic model (Fig. 5e),

$$\boldsymbol{X}(i) = \beta_0 + \beta_1 \times T(i) + \beta_2 \times T(i)^2 + S(i) \times \boldsymbol{X}_{\text{noise}}(i), \tag{10}$$

is, strictly speaking, not a nonlinear model because $T(i)^2$ can be seen as a second variable. This allows usage of multivariate linear regression estimation (von Storch and Zwiers, 1999), which is straightforward. "Real" nonlinear regression models, such as the saturation function (Fig. 5f),

$$\boldsymbol{X}(i) = \beta_0 + \beta_1 \times \{1 - \exp[-\beta_2 \times (T(i) - \beta_3)]\} + S(i) \times \boldsymbol{X}_{\text{noise}}(i), \tag{11}$$

are therefore nonlinear in the parameters.

Estimation of nonlinear trend models is usually more complex than of linear models because numerical procedures have to be employed. Although there are several types of procedures, the general estimation

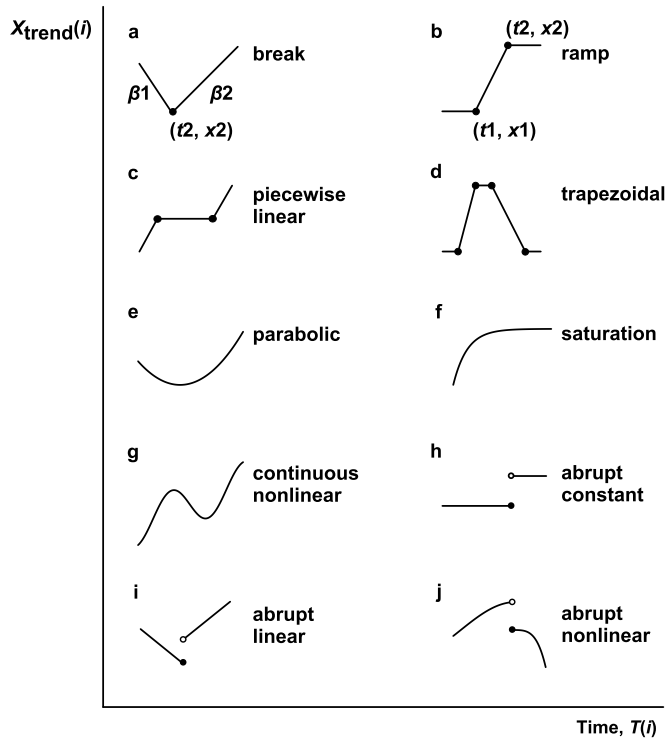

Fig. 5. Nonlinear regression models. The change points in the continuous change-point models (a, b, c, and d) are indicated by filled circles. The change points in discontinuous, abrupt models (h, i, and j) are indicated by open/filled circles. The parameterization of the models is explained for two cases. First, the break model (a) is defined by four parameters: $t2$, change-point time; $x2$, change-point level; $\beta1$, left slope; and $\beta2$, right slope. Second, the ramp model (b) is also defined by four parameters: $t1$, left change-point time; $x1$, left change-point level; $t2$, right change-point time; and $x2$, right change-point level.

concept may be outlined as follows. The parameters of the model span a hyperspace. The estimation consists in minimizing a cost function (e.g., least-squares). The best estimate is a point in the hyperspace for which the cost function is minimal. The problem consists in finding that point. This is also called an optimization problem. The general concept and mathematical foundations of nonlinear regression are treated in the books by Gallant (1987) and Seber and Wild (1989). Methods to solve optimization problems are treated by Michalewicz and Fogel (2000).

Particularly useful for climatological trend estimations are nonlinear change-point functions (Fig. 5) because they serve to quantify climate transitions. In particular, the break and the ramp models have been used by climate researchers.

The break regression model (Fig. 5a) is described by four parameters, $t2$, $x2$, $\beta1$, and $\beta2$. The break model can be fitted to a time series, $\{t(i),x(i)\}_{i=1}{}^{n}$, that means, the parameters can be estimated, by means of WLS. Since the break function is not differentiable with respect to time at $t2$, the WLS estimate for $t2$ cannot be obtained by taking the derivative. Mudelsee (2009) presented a brute-force estimation—for all trial $t2$ points from the set $\{t(i)\}_{i=i1}{}^{i2}$, the optimal $x2$, $\beta1$, and $\beta2$ values are calculated (via the derivatives)—which allows to calculate the WLS sum. Finally, that value for $t2$ is taken, for which the WLS sum is minimal. This is a global optimization technique for estimating $t2$, which is associated with high computing costs. However, for typical data sizes in climatology (say, up to a few tens of thousands of points), it is a feasible procedure on modern computers. By tailoring the search range (i.e., selecting $i1$ and $i2$), it is possible to reduce the computing costs. Since the $t2$ estimate is taken from the original time values, there may be an estimate superior to the brute-force estimate somewhere in between the original time values. However, this issue of a

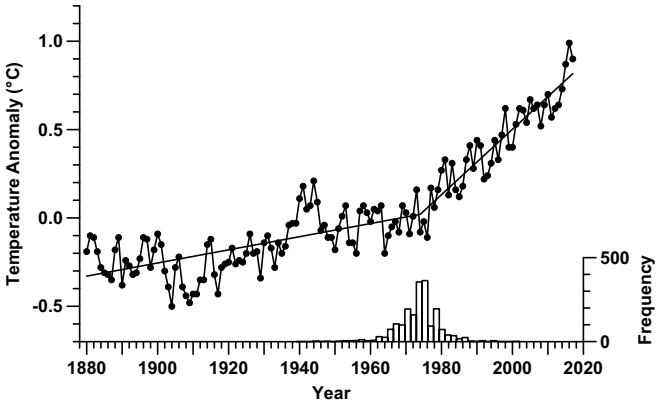

Fig. 6. GISTEMP time series and break regression curve (solid line) fitted by means of OLS. Also shown is a histogram of the 2000 bootstrap replications of the change-point time.

superior fine search would only be relevant if the standard error for the brute-force determined value of $\widehat{t2}$ is smaller than the spacing at around $\widehat{t2}$. On the contrary, the practical observation (Mudelsee, 2014) is that usually the spacing is smaller and is, hence, not the accuracy-limiting factor.

There are no classical standard errors for the break model parameters. However, the MBB resampling, presented in Section 3.1, can also be applied to the residuals of the break model fit. Mudelsee (2014) presents Monte Carlo tests of confidence intervals for the break model parameter estimates based on MBB resampling. The software for fitting a break trend model to data, BREAKFIT (Mudelsee, 2009), includes also graphical residual analysis.

Fig. 6 shows the break regression model fitted by means of OLS to the GISTEMP time series. The estimate of the change-point time with MBB standard error is

$$\widehat{t2} \pm \mathrm{se}_{\widehat{t2}} = 1974.0 \pm 5.9. \quad (12)$$

The estimate of the change-point level is

$$\widehat{x2} \pm \mathrm{se}_{\widehat{x2}} = 0.02^{\circ}\mathrm{C} \pm 0.06^{\circ}\mathrm{C}. \quad (13)$$

The estimate of the left slope is

$$\widehat{\beta1} \pm \mathrm{se}_{\widehat{\beta1}} = 0.0037^{\circ}\mathrm{C/a} \pm 0.0007^{\circ}\mathrm{C/a}. \quad (14)$$

The estimate of the right slope is

$$\widehat{\beta2} \pm \mathrm{se}_{\widehat{\beta2}} = 0.0185^{\circ}\mathrm{C/a} \pm 0.0026^{\circ}\mathrm{C/a}. \quad (15)$$

It appears obvious to the eye that the break fit is superior to the linear fit (Fig. 2a) to describe the trend in the GISTEMP time series. There is still some structure left in the form of shorter-term deviations, such as the positive excursions at around the year 1940. Owing to the better fit quality, the estimates may be helpful for a climatological interpretation (Section 4).

The estimated change-point time (Eq. 12) is rather stable with respect to the selection of the fit interval, as a sensitivity study reveals. Replacing the lower interval bound of 1880 by 1890, 1900, 1910, 1920, or 1930, had no effect (within error bars) on $\widehat{t2}$. Also replacing the upper interval bound of 2017 by 2010 had no effect (within error bars) on $\widehat{t2}$. Only further shrinking the fit interval has some effect. For example, for the fit interval 1940–2000, the result of the break fit is $\widehat{t2} = 1966.0 \pm 3.8$.

The procedure for the ramp regression model (Fig. 5b) is rather similar to the procedure for the break model. The ramp is described by four parameters, $t1$, $x1$, $t2$, and $x2$. The ramp model can be fitted to a time series by means of WLS. Since the ramp function is not differentiable with respect to time at $t1$ and $t2$, the WLS estimates for $t1$ and $t2$ cannot be obtained by taking the derivatives. Mudelsee (2000)

presented a brute-force estimation—for all trial $t1$–$t2$ combinations from the set $\{t(i)\}_{i=i1}^{i2}$, with the constraint $t1 < t2$, the optimal $x1$ and $x2$ values are calculated (via the derivatives)—which allows to calculate the WLS sum. Finally, that pair of values for $t1$–$t2$ is taken, for which the WLS sum is minimal. This is a global optimization technique for estimating the $t1$–$t2$ combination for the ramp fitting, which is associated with even higher computing costs (which are roughly proportional to $n^2/2$) than fitting the break (roughly proportional to n for $i1 = 1$ and $i2 = n$). Even here, it is a feasible procedure on modern computers. By tailoring the search ranges for $t1$ and $t2$, it is possible to reduce the computing costs. Also in case of the ramp, the issue of a superior fine search would only be relevant if the standard errors for the brute-force determined values of $\widehat{t1}$ and $\widehat{t2}$ are smaller than the spacing at around the respective estimates. The practical observation (Mudelsee, 2014), that usually the spacing is not the accuracy-limiting factor, applies also to the ramp.

There are no classical standard errors for the ramp model. However, the MBB resampling can also be applied to the residuals of the ramp model fit. The original ramp paper (Mudelsee, 2000) employed the stationary bootstrap (SB) instead of MBB. The SB is a bootstrap variant with geometrically distributed (i.e., non-constant) block lengths (Politis and Romano, 1994) aimed at ensuring the stationarity of the resample generating process. Olatayo (2014) adapted the SB by invoking a truncated geometric distribution of block lengths. Lahiri (2003) presents theoretical comparisons between the MBB, the SB, the autoregressive boostrap (ARB), and other bootstrap variants. The ARB (Freedman and Peters, 1984; Peters and Freedman, 1984; Efron and Tibshirani, 1986; Findley, 1986; Bose, 1988) is a semi-parametric bootstrap variant, which employs an AR(1) model for the autocorrelation but keeps the bootstrap's idea to resample from the data to preserve the distributional shape. The practical conclusion (Mudelsee, 2014) is that the deviations among the methods appear minor and that exclusive usage of the MBB for estimation problems cannot be condemned. Mudelsee (2014) presents Monte Carlo tests of confidence intervals for the ramp model parameter estimates based on ARB resampling. The software for fitting a ramp trend model to data, RAMPFIT (Mudelsee, 2000), includes graphical residual analysis. Also given is a version with MBB resampling (RAMPFITc). RAMPFITc can also be used for linear WLS and OLS regression (Section 3.1) via fixing $t1 = t(1)$ and $t2 = t(n)$.

Fig. 7 shows the ramp regression model fitted by means of OLS to a selected time interval (1880–1974) of the GISTEMP time series. The estimate of the left change-point time is

$$\widehat{t1} \pm \mathrm{se}_{\widehat{t1}} = 1923.0 \pm 8.3. \tag{16}$$

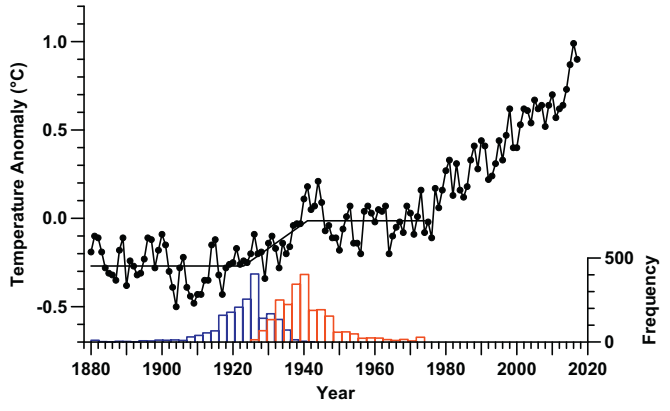

Fig. 7. GISTEMP time series and ramp regression curve (solid line) fitted by means of OLS. The fit interval is [1880;1974] (i.e., $n = 95$). Also shown are histograms of the 2000 bootstrap replications of the change-point times $\widehat{t1}$ (blue) and $\widehat{t2}$ (red), respectively. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

The estimate of the left change-point level is

$$\widehat{x1} \pm \mathrm{se}_{\widehat{x1}} = -0.270^{\circ}\mathrm{C} \pm 0.024^{\circ}\mathrm{C}. \tag{17}$$

The estimate of the right change-point time is

$$\widehat{t2} \pm \mathrm{se}_{\widehat{t2}} = 1941.0 \pm 8.6. \tag{18}$$

The estimate of the right change-point level is

$$\widehat{x2} \pm \mathrm{se}_{\widehat{x2}} = -0.014^{\circ}\mathrm{C} \pm 0.028^{\circ}\mathrm{C}. \tag{19}$$

It appears obvious to the eye—and it can be tested by means of the supplied links to data and software—that the ramp fit over the full interval would not be superior to the break fit (Fig. 6). Therefore only the earlier interval, before the change-point time for the break model (Eq. 12), is investigated by means of the ramp model. The ramp fit finds a transition in the form of a warming from the 1920s to the 1940s (see Section 4).

3.3. Nonparametric regression

Instead of identifying the trend component, $X_{\mathrm{trend}}(i)$, with a linear or certain nonlinear function with parameters to be estimated, the smoothing method estimates the trend at a time point, T', by, loosely speaking, averaging the data points, $X(i)$, within a neighbourhood around T'. A simple example is the running mean, where the points inside a window are averaged and the window runs along the time axis. Statistical science recommends to replace the non-smooth weighting window (points inside receive constant, positive weight and points outside zero weight) by a smooth kernel function, K. The kernel trend estimator after Gasser and Müller (1979, 1984) is given by

$$\widehat{X}_{\mathrm{trend}}^{\mathrm{GM}}(T) = h^{-1} \sum_{i=1}^{n} \left[\int_{s(i-1)}^{s(i)} K\left(\frac{T-y}{h}\right) dy \right] X(i). \tag{20}$$

The kernel function used by the software (KERNEL) for making the illustration of the method (Fig. 8), is the Epanechnikov kernel, $K(y) = 0.75 \times (1 - y^2)$. The bandwidth parameter, h, is crucial because

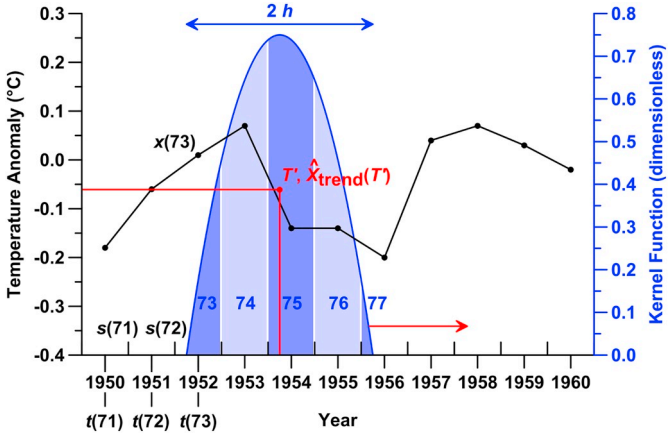

Fig. 8. Nonparametric regression using the smoothing after Gasser and Müller (1979, 1984) with Epanechnikov kernel. The GISTEMP time series (black curve) for the interval from 1950 ($i = 71$) to 1960 ($i = 81$) is used for illustration. The sequence $\{s(i)\}$ is given as the time midpoints (time axis, upper tick marks). The time for the shown trend estimation is $T' = 1953.75$. The kernel function (blue curve) has a bandwidth of $h = 2$ years. The area under the kernel is divided into several subareas (dark and light blue shading). The subareas correspond (blue numbers) to the integrals in Eq. (20). (For example, $x(73)$ is multiplied by the size of subarea 73.) The trend estimate, $\widehat{X}_{\mathrm{trend}}(T')$, is obtained by subarea-weighting the affected x-values (Eq. 20). The values $\widehat{X}_{\mathrm{trend}}(T')$ and T' are marked red. The kernel is moved along the time axis (red arrow) to estimate the trend at other time values, T'. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

it determines the uncertainty measures for the trend estimate. On the other hand, the choice of K (Epanechnikov, Gaussian, and so forth), is more of "cosmetic" (Diggle, 1985) interest. KERNEL places the integration bound, $s(i)$, in the middle between two time points (Fig. 8). KERNEL further sets

$$s(0) = 1.5 \times t(1) - 0.5 \times t(2) \quad (21)$$

and

$$s(n) = 1.5 \times t(n) - 0.5 \times t(n-1). \quad (22)$$

Note that the selection of the sequence $s(i)$ in particular, and the Gasser–Müller smoothing procedure in general, can be performed on unevenly spaced time series. The trend can be estimated for all points, T', within the observation interval, $[t(1); t(n)]$. The kernel functions are modified near the interval boundaries (Gasser and Müller, 1979, 1984), so that the trend can be estimated also there.

An uncertainty measure for the estimated trend curve, $\widehat{X}_{\mathrm{trend}}(T')$, is essential for assessing the significance of the ups and downs in the estimate, whether these variations constitute real features or are generated by the noise. A pointwise standard-error band can be constructed from the standard-error intervals for the trend estimate as follows. Let $\mathrm{se}(T')$ denote the standard error for $\widehat{X}_{\mathrm{trend}}(T')$. The standard-error interval for the time value T' is given by $[\widehat{X}_{\mathrm{trend}}(T') - \mathrm{se}(T'); \widehat{X}_{\mathrm{trend}}(T') + \mathrm{se}(T')]$. The band is obtained by concatenating the upper bounds, $\widehat{X}_{\mathrm{trend}}(T') + \mathrm{se}(T')$, for the full time interval, $t(1) \le T' \le t(n)$, and by concatenating the lower bounds, $\widehat{X}_{\mathrm{trend}}(T') - \mathrm{se}(T')$. A stricter test of the significance of the ups and downs than from using $1 \times \mathrm{se}(T')$ is obtained by using $2 \times \mathrm{se}(T')$.

The KERNEL software has implemented MBB resampling to calculate the bootstrap standard errors. Resampling is performed on the residuals of the nonparametric regression (i.e., data minus fit). Also the other methodological steps (autocorrelation estimation, block length selection, and so forth) are analogous to the MBB procedure for the linear model (Section 3.1). This brings the twofold advantage of the MBB to the nonparametric regression: (1) distributional robustness and (2) consideration of autocorrelation. There exist bootstrap adaptations Davison and Hinkley, 1997; Härdle, 1990) aimed at further increasing the accuracy of resampling-derived uncertainty measures. Autocorrelation may constitute a major hurdle for applying guidelines for bandwidth selection and for setting the confidence level in confidence interval construction; see Mudelsee (2014: Sections 4.3.1 and 4.3.2) and references cited therein for more details. The general advice is to consider theoretical knowledge about the data generating system (i.e., the climate), to try many settings, to "play" with the bandwidth, h, and study the sensitivity of the resulting trend estimations.

Fig. 9 shows the trend estimated by means of nonparametric regression for the GISTEMP time series. The bandwidth of 5 years was predefined to inspect mid- and shorter-term (decadal-scale) variations, such as the warming in the years around 1940, and to smooth away faster variations. The nonparametric trend estimate has a relative maximum at the year 1942 with a peak value of 0.083 °C. This warming was statistically significant in the sense that on the flanks of the peak, the temperature trend values have upper uncertainty bounds, $\widehat{X}_{\mathrm{trend}}(T') + 2 \times \mathrm{se}(T')$, that are below the peak value (Fig. 9). "Relative maximum" means that the 1942-peak is not the largest value; from the year 1979 onward, the trend estimate is always greater than the upper uncertainty bound of 0.142 °C for the 1942-peak. For an interpretation, see Section 4.2.

4. Discussion

The previous Section 3 presented the statistical concepts and technical details of trend estimation with uncertainty-measure determination. The GISTEMP time series of global surface temperature for the interval from 1880 to 2017 (Fig. 1) served to illustrate the methods. This Section 4 pursues the interpretation of the statistical results. It highlights topics that are relevant not only with regard to the GISTEMP record, but to climate time series in general. The first topic is model suitability (Section 4.1), to which related are (1) fit-interval selection and (2) the interplay between data availability and the complexity of the questions (i.e., statistical models) that can be addressed by means of analysis of the data.

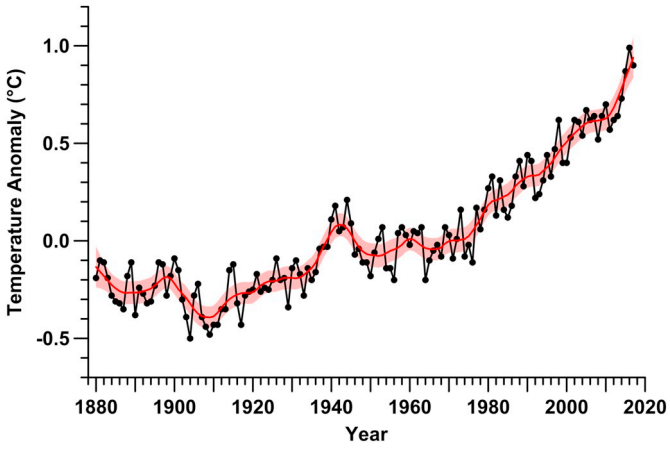

Fig. 9. GISTEMP, trend estimation by means of nonparametric regression. Gasser–Müller smoothing (Eq. 20) with an Epanechnikov kernel and a bandwidth of $h = 5$ a is applied to the GISTEMP time series (black) to obtain the trend (red solid line) with a 2-se(T') band (red shaded). (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

The second topic of discussion (Section 4.2) puts a caveat on the 1942-peak in temperature due to data-quality problems caused by changes in the measurement praxis of sea-surface temperature (SST) during World War II.

Finally, Section 4.3 is devoted to the analysis of global temperature for the years after and before the warm year 1998, which led some people interested in climate—mainly amateurs from the blogosphere, but also a few professionals—to suspect that a hiatus occurred.

4.1. Model suitability

The philosophical principle of parsimony—Ockham's razor—applied to the context of trend estimation, posits that simple regression models are preferable to complex ones. But not too simple. The linear model, with just two parameters, was assessed by eye as too simple for describing the trend for the GISTEMP time series (Fig. 2). On the other hand, the break model appeared as clearly better suited (Fig. 6).

The reduced sum of squares, SSQ_ν, penalizes for usage of a high number, p, of parameters by means of dividing SSQ at the estimation point by the degrees of freedom,

$$\nu = n - p. \quad (23)$$

See Bevington and Robinson (1992) for a practical view of the reduced sum of squares and the degrees of freedom. The concept of degrees of freedom was introduced by Fisher (1922).

The linear fit to the GISTEMP time series has a reduced sum of squares of

$$SSQ_\nu = SSQ(\widehat{\beta}_0, \widehat{\beta}_1)/(n-2) = 0.027\ (^\circ\mathrm{C})^2. \quad (24)$$

The break fit to the same data set has a reduced sum of squares of

$$SSQ_\nu = SSQ(\widehat{t2}, \widehat{x2}, \widehat{\beta 1}, \widehat{\beta 2})/(n-4) = 0.013\ (^\circ\mathrm{C})^2. \quad (25)$$

This result confirms quantitatively the superiority of the break over the linear model.

The ramp fit to the GISTEMP time series over the full interval from 1880 to 2017 with $n = 138$ (not shown) would yield a rather large

reduced sum of squares,

$$SSQ_\nu = SSQ(\widehat{t1}, \widehat{x1}, \widehat{t2}, \widehat{x2})/(n-4) = 0.018\ (°C)^2. \quad (26)$$

This result confirms quantitatively the superiority of the break over the ramp model for this data set. If the full interval is to be analyzed, then the four parameters are better invested into the break than the ramp.

However, the ramp may be useful to quantify the onset of the 1942-warmth (Fig. 7). For the selected time interval from 1880 to 1974, with $n = 95$, the fit measure is

$$SSQ_\nu = SSQ(\widehat{t1}, \widehat{x1}, \widehat{t2}, \widehat{x2})/(n-4) = 0.011\ (°C)^2, \quad (27)$$

which is comparable to the quality of the break fit.

There are other quantitative measures of the fit than the reduced sum of squares, such as Akaike's information criterion (AIC) or a corrected version of it (AICC), which takes into account the number of fit parameters (Mudelsee, 2014). However, climate researchers should not slavishly follow the values of the fit measures. It is important to keep in mind (1) background knowledge about the data generating system and (2) what the concrete research questions are. For example, if the interest is in quantifying a climate transition, then the questions may consist in: When did the transition start, when did it end, what was the amplitude? These questions directly invoke the parametric ramp regression model, and the estimated parameters provide the answers (with error bars).

The example of fitting a ramp to quantify the onset of the 1942-warmth illustrates the importance of the selection of the fit interval. A kind of an analogue to Ockham's razor would suggest that having many data points (a large fit interval) is preferable to having only few. But not too many points. This allows to extract the fit interval that is still compatible with a relatively simple model. In case of the 1942-warmth (Fig. 7), the upper bound of the fit interval of 1974 was determined via the estimated change-point time in the break fit (Fig. 6). In other words, the interval from 1974 to the present was excluded from the ramp fit to keep the compatibility.

If the interest is less in quantifying a climate transition or determining change points but more in describing the trend over the full time interval, then the nonparametric regression model may be a good choice (Fig. 9).

To summarize, there is no golden rule what particular model to employ. The advice is rather to “play” with the models and parameter setting, to observe the sensitivity of the results, and to acquire an intuition for what is real and what is noise. The final task, however, is to honestly communicate those findings.

4.2. World War II bias

The statistical finding of a warming in the years around 1940 and a relative maximum at the year 1942 with a kernel-determined peak value of 0.083 °C (Fig. 9) assumes a rather homogeneous quality of the data (Fig. 1) over time. However, the consultation of the metadata (Thompson et al., 2008) shows that the method to measure the temperature of the upper meters of the sea changed during World War II. On British ships, SST was measured differently compared to American ships. The relative contributions to the combined average SST was not constant over time during that period, and they are not exactly known. Furthermore, measuring temperature was not the prime objective on the ships traveling through the seas at that time, which means a reduced data quality. Since the oceans cover about two thirds of the surface of the Earth, the derived global surface temperature curve may be considerably distorted during the interval around World War II. Statistical science calls such a systematic distortion bias. It is not straightforward to quantify the bias (Hansen et al., 2017), and climate researchers are trying to improve the metadata situation (Kent et al., 2017). This problem is not unique to GISTEMP (Fig. 1). The two other “competing” series, published by the Climatic Research Unit of the University of East Anglia in the United Kingdom and the National Oceanic and Atmospheric Administration of the United States of America, respectively, also rely on the SST measurements.

Due to the World War II bias, the results of the OLS fit of a ramp to the interval from 1880 to 1974 (Fig. 7), that means, the quantification of the onset of the warming, could be questionable. One remedy may be to employ WLS, with lower weights put to the data for the World War II period (1939–1945). Another reason for using WLS is that the number of measurement stations, and hence the reliability of the temperature values, grew gradually from the 19th to the 21st century (Hansen et al., 2010). The repetition of the fit of the ramp with WLS uses the following function, $S(i)$, for weighting. For the year 1880, $S(i)$ is set equal to 0.2 °C and for the year 2017 equal to 0.05 °C (Hansen 2018, personal communication). For the years in between those end points, $S(i)$ is set as the linear connection between those points—with the exception of the interval 1939–1945. For this period, $S(i)$ is set equal to 0.3 °C, which corresponds to the rough bias estimate of SST made by Folland et al. (1984).

The WLS results are as follows. The estimate of the left change-point time is

$$\widehat{t1} \pm \mathrm{se}_{\widehat{t1}} = 1912.0 \pm 13.3. \quad (28)$$

The estimate of the left change-point level is

$$\widehat{x1} \pm \mathrm{se}_{\widehat{x1}} = -0.281°C \pm 0.034°C. \quad (29)$$

The estimate of the right change-point time is

$$\widehat{t2} \pm \mathrm{se}_{\widehat{t2}} = 1959.0 \pm 10.5. \quad (30)$$

The estimate of the right change-point level is

$$\widehat{x2} \pm \mathrm{se}_{\widehat{x2}} = -0.011°C \pm 0.024°C. \quad (31)$$

To summarize the results, the warming in the years around 1940 becomes less significant when taking into account the World War II bias by means of WLS. However, bias is a systematic distortion, and there may be ways to improve the metadata situation (Kent et al., 2017). This would allow to quantify the bias better with the help of a statistical model with a structure that is based on the physics of the SST recording. This bias model may then lead to more accurate data and reduce the uncertainties that affect the interpretation of global temperatures in the 1940s.

4.3. Suspected global warming hiatus

The story of the suspected hiatus came up in the months before the 15th session of the Conference of the Parties to the United Nations Framework Convention on Climate Change (COP 15), which was held in Copenhagen in December 2009. It was kept alive for a couple of years after the conference. The story is as follows. Global temperature, on the rise since several decades (Figs. 2a, 6, and 9), reached a peak in 1998 (the GISTEMP temperature anomaly for that year is 0.62 °C) and then stopped in its rise.

The story can be translated into two quantitative hypotheses. First, there was a break point at the year 1998. Second, the slope of the temperature trend after 1998 was zero. The hypotheses can be tested by means of parametric change-point regressions with the break model. To simulate what was known in the immediate years after COP 15, the analysis “plays” and varies the upper bound of the fit interval.

The results of this exercise are as follows (Fig. 10, Table 1). As regards the first hypothesis, change-point time estimates of 1998 are found for two selected fit intervals (1992–2011 and 1992–2013). However, the standard errors for that value are considerable (3 years). If the fit interval is extended to 1992–2017, then the 1998 estimate disappears and a new, “cheaper” (in terms of SSQ) estimate of 2013 emerges. This fit solution shows then an opposite behavior—an

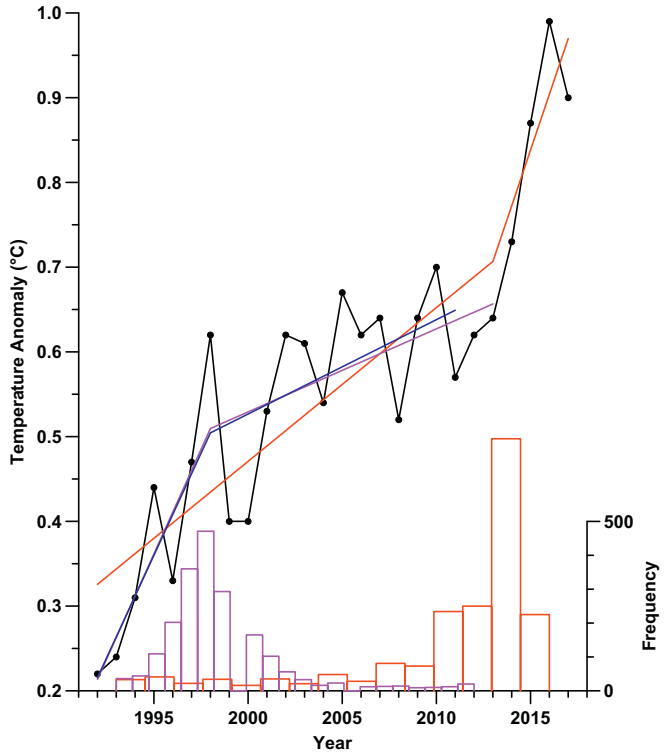

Fig. 10. GISTEMP and the suspected global warming hiatus. Break trend models are fitted by means of OLS to various intervals (1992–2017, $n = 26$, red line; 1992–2013, $n = 22$, purple line; 1992–2011, $n = 20$, blue line). Also shown are histograms of the 2000 bootstrap replications of the change-point time for the intervals 1992–2017 (red) and 1992–2013 (purple), respectively; the histogram for 1992–2011 (not shown for legibility) looks very similar to that for 1992–2013. See Table 1 for parameter estimates. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

Table 1
GISTEMP and the suspected global warming hiatus, results of break fits (Fig. 10).

Fit Interval	$\widehat{t2}$ (year)	$\widehat{\beta 1}$ (°C/a)	$\widehat{\beta 2}$ (°C/a)
1992–2011	1998 ± 3	0.048 ± 0.040	0.011 ± 0.025
1992–2013	1998 ± 3	0.049 ± 0.035	0.010 ± 0.019
1992–2017	2013 ± 6	0.018 ± 0.031	0.066 ± 0.043

accelerated warming in the years after 2013. The standard error for the year-2013 estimate is even larger (6 years) than for the year-1998 estimates (fit intervals 1992–2011 and 1992–2013).

As regards the second hypothesis, the right slopes indicate a warming in the years after 1998 (Table 1). However, the standard errors for the slopes render them statistically insignificant. If the fit interval is extended to 1992–2017, then the right slope becomes significant (strong warming since the year 2013)—and the left slope becomes insignificant.

The second hypothesis, of zero slope of the temperature trend after 1998, can also be tested by means of nonparametric regression. Instead of the trend, its first derivative with respect to time, becomes the target of the kernel estimation after Gasser and Müller (1979, 1984). The analysis (Fig. 11) employs the same kernel function (Epanechnikov) and bandwidth ($h = 5$ a) as used for trend estimation (Fig. 9).

The zoom on the recent time interval (Fig. 11b) reveals that for the years from, roughly, 2005 to 2010, the slope is within standard-error band indistinguishable from zero. This result agrees with the findings

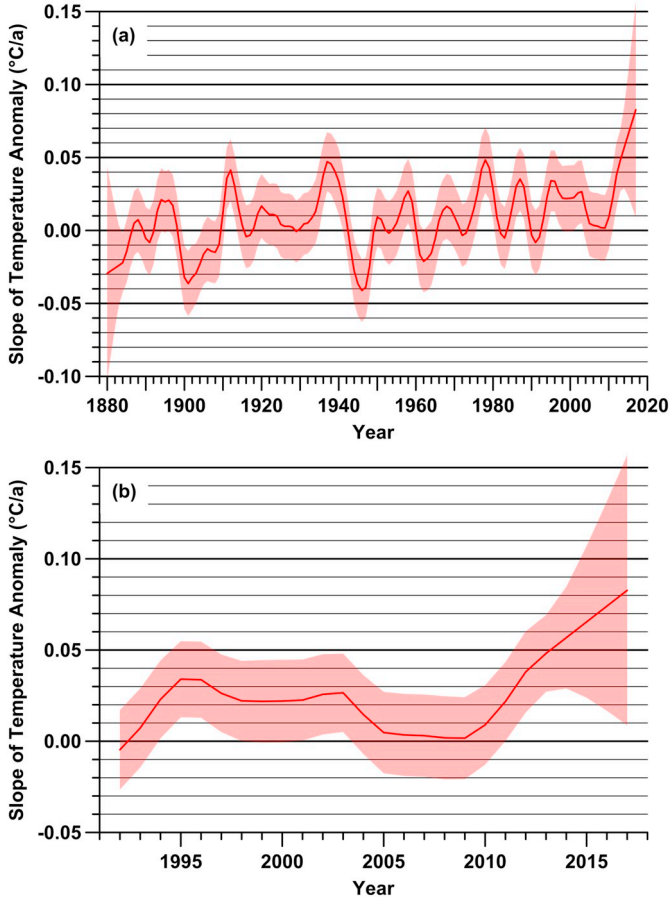

Fig. 11. GISTEMP and the suspected global warming hiatus. Gasser–Müller smoothing (Eq. 20) with an Epanechnikov kernel and a bandwidth of $h = 5$ a is applied to the GISTEMP time series (Fig. 1) to obtain the first derivative of the trend (red solid line) with a 2-se(T') band (red shaded). (a) Full time interval, 1880–2017; (b) zoomed time interval, 1992–2017. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

obtained from fitting the parametric break model (Fig. 10).

After 2010, the slope increased and has been since then significantly positive. In earlier periods since 1880, the slope is mostly positive as well. There have been only occasionally cooling periods, such as in the years after 1900 (Fig. 11a).

To summarize, it seems impossible to show with a decent level of statistical significance that a hiatus in surface temperature occurred in the years after 1998. The number of data points is too small, and the statistical uncertainties are too large. This effect is exacerbated by the presence of serial dependence (Fig. 3b), which reduces the number of independent data points Mudelsee (2014: Chapter 2).

The resulting estimates for the change-point time and the slopes before and after depend on the selected fit interval (Table 1). For example, if the fit interval is extended from 1992–2013 to 1992–2017, then the estimated change-point time of 1998 disappears in the break fit (Fig. 10).

This observation points toward a general issue in trend estimation on time series. The researcher has the power to select the fit interval, which allows her or him to suppress certain fit solutions and favor other solutions. The interval selection should therefore be objective and oriented on general principles, such as to have a long interval that still yields results compatible with a simple trend model.

The power (to select a fit interval) has therefore to be accompanied by responsibility: to employ objective criteria and also reveal the sensitivity of results to the fit-interval selection. Ultimately, however, the

misuse of statistics is a social, cultural and ethical problem, for which technical fixes are doomed (Saltelli and Stark, 2018).

5. Conclusions

At the time of writing (mid 2018), the statistical methodology of trend estimation is well elaborated. Some development may come in the form of GLS estimation techniques for nonlinear regression functions (Fig. 5), such as the break or the ramp models. Another direction is the development of estimation routines for the piecewise linear model (Fig. 5c). Furthermore, the fitting of multiple change-point models (i.e., more than two change points) is of genuine interest. This is technically challenging and likely necessitates the implementation of advanced optimization techniques, such as genetic algorithms (Michalewicz and Fogel, 2000). The reward of such a technology may consist in a reduction of the problem of fit-interval selection. An interesting example in that regard is the analysis of regional temperatures in the Alpine region (Battaglia and Protopapas, 2012), which demonstrated an accelerated warming in several phases. As regards nonparametric regression, it appears that the potential of that method (standard-error band and derivative estimation) for climatology has only occasionally been appreciated. One example is the search for ^{14}C plateaus (i.e., zero slope) in marine sedimentary records from the Holocene (Sarnthein et al., 2015).

The statistical methodology of uncertainty determination for climate time series is well elaborated, as far as uncertainties stemming from measurement or proxy errors in the climate variable, X, are concerned. Bootstrap methods take into account deviations from Gaussian shape. Blocking variants of the bootstrap, such as the MBB, take into account autocorrelation. This means that the two major peculiarities of climate time series—non-Gaussian shape and autocorrelation—can be successfully dealt with in the statistical analysis. As a result, it is possible to avoid unrealistically small error bars from ignored autocorrelation, which could lead to overstatements. On the other hand, as regards timescale errors in the variable T, this is a topic where further research will be quite relevant for paleoclimatology. The book by Mudelsee (2014) contains some algorithms, simulation tests, and references on that emerging field. Also the Bayesian view of probability can be adopted for research (Parnell et al., 2015; Blaauw et al., 2018).

As time proceeds, new data sources become available and existing time series are updated, which leads to new insights. This is a part of normal science (Kuhn, 1970), and the climate community should not expect too many big surprises. As an example, the suspected global warming hiatus in the years after 1998 disappears when considering not only the interval up to 2013 for the fit, but rather the interval up to 2017. This example illustrates that the selection of the fit interval for trend estimation—has also a moral aspect. Climate researchers should be aware of this.

Acknowledgments

I thank the referee for a helpful and constructive review. I thank James Hansen (The Earth Institute, Columbia University, New York, NY, United States of America), Philip Jones (Climatic Research Unit, University of East Anglia, Norwich, United Kingdom), Gerrit Lohmann (Alfred Wegener Institute Helmholtz Centre for Polar and Marine Research, Bremerhaven, Germany), Andrew Parnell (School of Mathematics and Statistics, University College Dublin, Dublin, Ireland), and Gavin Schmidt (NASA Goddard Institute for Space Studies and Center for Climate Systems Research, Columbia University, New York, NY, United States of America) for comments on a previous manuscript version. The used software packages (BREAKFIT, KERNEL, RAMPFIT, and RAMPFITc) are freely available at http://www.climate-risk-analysis.com (19 June 2018).

Appendix A. Appendix

We describe the MBB for the application to linear regression in the following steps. It is detailed on an algorithmic level, ready to be implemented in various forms of software.

Step 1: OLS estimation

Use the time series, $\{t(i), x(i)\}_{i=1}^{n}$, and calculate the OLS estimates of intercept and slope (Eqs. 4 and 5). See, for example, the GISTEMP time series (Fig. 2a).

Step 2: Residuals

Calculate the residuals as

$$r(i) = x(i) - \widehat{\beta}_0 - \widehat{\beta}_1 \times t(i). \tag{A.1}$$

See, for example, the GISTEMP time series (Figs. 2b and 3).

Step 3: Autocorrelation estimation

Use the residuals to calculate the autocorrelation coefficient estimate. For even spacing, this is given by

$$\widehat{a} = \left[\sum_{i=2}^{n} r(i) \times r(i-1)\right] \Big/ \left\{\sum_{i=2}^{n} [r(i)]^2\right\}. \tag{A.2}$$

The underlying mathematical autocorrelation model is the first-order autoregressive (AR(1)) process, where a noise value "remembers" only its own immediate past (Priestley, 1981). For climate variables, the AR(1) model has been found useful empirically (Gilman et al., 1963) and theoretically (Hasselmann, 1976) for describing the autocorrelation structure. The AR(1) model for even spacing has one parameter, $-1 < a < 1$, which measures the strength of the "memory". The noise component for climate data usually has $a \geq 0$ (Mudelsee, 2014).

For uneven spacing, the procedure is more complex. First, the persistence time estimate, $\hat{\tau}$, has to be numerically calculated (i.e., there is no analytical formula) as the minimizer of the least-squares sum,

$$S(\tilde{\tau}) = \sum_{i=2}^{n} [r(i) - \exp\{-[t(i) - t(i-1)]/\tilde{\tau}\} \times r(i-1)]^2. \tag{A.3}$$

The idea behind that formula (Mudelsee, 2002) is that the memory is the larger, the smaller the time distance from a time point, $t(i)$, to the preceding time point, $t(i-1)$, is. The persistence time, $\tau \geq 0$, is the parameter of the AR(1) model for uneven spacing (Robinson, 1977; Mudelsee, 2002).

The estimated equivalent autocorrelation coefficient is the uneven-spacing analogue to the estimated autocorrelation coefficient,

$$\widehat{a}_{\mathrm{equ}} = \exp(-\overline{d}/\hat{\tau}). \tag{A.4}$$

Step 4: Block length selection

The MBB resamples blocks of a certain length of residuals (see Step 6). An optimal block length selector (Carlstein, 1986; Sherman et al., 1998) is

$$l_{\mathrm{opt}} = NINT\{[6^{1/2} \times \hat{a}/(1-\hat{a}^2)]^{2/3} \times n^{1/3}\}, \tag{A.5}$$

where *NINT* is the nearest integer function. (If $\hat{a}$ approaches zero, then set $l_{\mathrm{opt}} = 1$; if $\hat{a}$ approaches unity, then set $l_{\mathrm{opt}} = n - 1$.) Eq. (A.5) is for even spacing; in case of uneven spacing, use Eq. (5) with $\hat{a}_{\mathrm{equ}}$ instead of $\hat{a}$. "Optimality" should not be seen too strictly. The key point is that the block length should be long enough to capture a sufficient amount of the autocorrelation. There exist other block length selectors (Mudelsee, 2014; Section 3.3). There exist also bias-corrected versions of $\hat{a}$ and $\hat{a}_{\mathrm{equ}}$ (Mudelsee, 2014: Chapter 2), but this is less relevant here.

Step 5: Counter

Set the resampling counter, b, equal to 1.

Step 6: Start (MBB resampling)

Use l_{opt} for the MBB (Fig. 4) to generate a random set of residuals, $\{r^*(i)\}_{i=1}{}^n$.

Step 7: Resample

Calculate the resample, $\{t(i), x^*(i)\}_{i=1}{}^n$, via

$$x^*(i) = \hat{\beta}_0 + \hat{\beta}_1 \times t(i) + r^*(i). \tag{A.6}$$

The residuals are employed as realizations of the noise component, $S(i) \times X_{\mathrm{noise}}(i)$.

Step 8: Replications

Use the resample, $\{t(i), x^*(i)\}_{i=1}{}^n$, and calculate the OLS estimates of intercept and slope (Eqs. 4 and 5). These quantities, $\hat{\beta}_0^{*b}$ and $\hat{\beta}_1^{*b}$, are called replications; they capture the noise influence on the estimation uncertainty.

Step 9: End (MBB resampling)

Increase b by 1 and go back to Step 6 until $b = B = 2000$ replications exist for both intercept and slope.

Step 10: Uncertainty measure

A simple uncertainty measure is the bootstrap standard error, which is the sample standard deviation over the replications:

$$\mathrm{se}_{\hat{\beta}_0} = \sum_{b=1}^{B} [\hat{\beta}_0^{*b} - \langle \hat{\beta}_0^{*} \rangle]^2/(n-1), \tag{A.7}$$

$$\mathrm{se}_{\hat{\beta}_1} = \sum_{b=1}^{B} [\hat{\beta}_1^{*b} - \langle \hat{\beta}_1^{*} \rangle]^2/(n-1), \tag{A.8}$$

where the sample means are given by

$$\langle \hat{\beta}_0^{*} \rangle = \sum_{b=1}^{B} \hat{\beta}_0^{*b}/B, \tag{A.9}$$

$$\langle \hat{\beta}_1^{*} \rangle = \sum_{b=1}^{B} \hat{\beta}_1^{*b}/B. \tag{A.10}$$

Other uncertainty measures, such as confidence intervals, can also be calculated from the replications (Efron and Tibshirani, 1993).

References

Agrinier, P., Gallet, Y., Lewin, E., 1999. On the age calibration of the geomagnetic polarity timescale. Geophys. J. Int. 137 (1), 81–90.

Battaglia, F., Protopapas, M.K., 2012. An analysis of global warming in the Alpine region based on nonlinear nonstationary time series models. Stat. Methods Appl. 21 (3), 315–334.

Bennett, K.D., 1994. Confidence intervals for age estimates and deposition times in late-Quaternary sediment sequences. Holocene 4 (4), 337–348.

Bennett, K.D., Fuller, J.L., 2002. Determining the age of the mid-Holocene *Tsuga canadensis* (hemlock) decline, eastern North America. Holocene 12 (4), 421–429.

Beran, J., 1994. Statistics for Long-Memory Processes. Chapman and Hall, Boca Raton, FL 315 pp.

Bevington, P.R., Robinson, D.K., 1992. Data Reduction and Error Analysis for the Physical Sciences, 2nd ed. McGraw-Hill, New York 328 pp.

Blaauw, M., 2010. Methods and code for 'classical' age-modelling of radiocarbon sequences. Quat. Geochronol. 5 (5), 512–518.

Blaauw, M., Christen, J.A., 2005. Radiocarbon peat chronologies and environmental change. Appl. Stat. 54 (4), 805–816.

Blaauw, M., Heegaard, E., 2012. Estimation of age–depth relationships. In: Birks, H.J.B., Lotter, A.F., Juggins, S., Smol, J.P. (Eds.), Tracking Environmental Change Using Lake Sediments: Data Handling and Numerical Techniques. Springer, Dordrecht, pp. 379–413.

Blaauw, M., Christen, J.A., Bennett, K.D., Reimer, P.J., 2018. Double the dates and go for Bayes—impacts of model choice, dating density and quality on chronologies. Quat. Sci. Rev. 188, 58–66.

Bose, A., 1988. Edgeworth correction by bootstrap in autoregressions. Ann. Stat. 16 (4), 1709–1722.

Brockwell, P.J., Davis, R.A., 1991. Time Series: Theory and Methods, 2nd ed. Springer, New York 577 pp.

Brückner, E., 1890. Klimaschwankungen seit 1700 nebst Bemerkungen über die Klimaschwankungen der Diluvialzeit. Geographische Abhandlungen 4 (2), 153–484.

Buck, C.E., Millard, A.R. (Eds.), 2004. Tools for Constructing Chronologies: Crossing Disciplinary Boundaries. Springer, London 257 pp.

Carlstein, E., 1986. The use of subseries values for estimating the variance of a general statistic from a stationary sequence. Ann. Stat. 14 (3), 1171–1179.

Davison, A.C., Hinkley, D.V., 1997. Bootstrap Methods and Their Application. Cambridge University Press, Cambridge 582 pp.

Diggle, P., 1985. A kernel method for smoothing point process data. Appl. Stat. 34 (2), 138–147.

Doukhan, P., Oppenheim, G., Taqqu, M.S. (Eds.), 2003. Theory and Applications of Long-Range Dependence. Birkhäuser, Boston 719 pp.

Drysdale, R.N., Zanchetta, G., Hellstrom, J.C., Fallick, A.E., Zhao, J., Isola, I., Bruschi, G., 2004. Palaeoclimatic implications of the growth history and stable isotope (δ^{18}O and δ^{13}C) geochemistry of a middle to late Pleistocene stalagmite from central-western Italy. Earth Planet. Sci. Lett. 227 (3–4), 215–229.

Efron, B., 1979. Bootstrap methods: another look at the jackknife. Ann. Stat. 7 (1), 1–26.

Efron, B., 1982. The Jackknife, the Bootstrap and Other Resampling Plans. SIAM, Philadelphia, PA 92 pp.

Efron, B., Hastie, T., 2016. Computer Age Statistical Inference: Algorithms, Evidence, and Data Science. Cambridge University Press, New York 475 pp.

Efron, B., Tibshirani, R., 1986. Bootstrap methods for standard errors, confidence intervals, and other measures of statistical accuracy (with discussion). Stat. Sci. 1 (1), 54–77.

Efron, B., Tibshirani, R.J., 1993. An Introduction to the Bootstrap. Chapman and Hall, London 436 pp.

Findley, D.F., 1986. On bootstrap estimates of forecast mean square errors for autoregressive processes. In: Allen, D.M. (Ed.), Computer Science and Statistics. North-Holland, Amsterdam, pp. 11–17.

Fisher, R.A., 1922. On the interpretation of χ^2 from contingency tables, and the calculation of *P*. J. R. Stat. Soc. 85 (1), 87–94.

Fohlmeister, J., 2012. A statistical approach to construct composite climate records of dated archives. Quat. Geochronol. 14, 48–56.

Folland, C.K., Parker, D.E., Kates, F.E., 1984. Worldwide marine temperature fluctuations 1856–1981. Nature 310 (5979), 670–673.
Freedman, D.A., Peters, S.C., 1984. Bootstrapping an econometric model: some empirical results. J. Bus. Econ. Stat. 2 (2), 150–158.
Gallant, A.R., 1987. Nonlinear Statistical Models. Wiley, New York 610 pp.
Gasser, T., Müller, H.-G., 1979. Kernel estimation of regression functions. In: Gasser, T., Rosenblatt, M. (Eds.), Smoothing Techniques for Curve Estimation. Springer, Berlin, pp. 23–68.
Gasser, T., Müller, H.-G., 1984. Estimating regression functions and their derivatives by the kernel method. Scand. J. Stat. 11 (3), 171–185.
Gilman, D.L., Fuglister, F.J., Mitchell Jr., J.M., 1963. On the power spectrum of "red noise". J. Atmos. Sci. 20 (2), 182–184.
Hall, P., 1988. Theoretical comparison of bootstrap confidence intervals (with discussion). Ann. Stat. 16 (3), 927–985.
Hann, J., 1901. Lehrbuch der Meteorologie. Tauchnitz, Leipzig 805 pp.
Hansen, J., Ruedy, R., Sato, M., Lo, K., 2010. Global surface temperature change. Rev. Geophys. 48 (4), RG4004. https://doi.org/10.1029/2010RG000345.
Hansen, J., Sato, M., Kharecha, P., von Schuckmann, K., Beerling, D.J., Cao, J., Marcott, S., Masson-Delmotte, V., Prather, M.J., Rohling, E.J., Shakun, J., Smith, P., Lacis, A., Russell, G., Ruedy, R., 2017. Young people's burden: requirement of negative CO_2 emissions. Earth Syst. Dyn. 8 (3), 577–616.
Härdle, W., 1990. Applied Nonparametric Regression. Cambridge University Press, Cambridge 333 pp.
Haslett, J., Parnell, A., 2008. A simple monotone process with application to radiocarbon-dated depth chronologies. Appl. Stat. 57 (4), 399–418.
Hasselmann, K., 1976. Stochastic climate models: part I. Theory. Tellus 28 (6), 473–485.
Heegaard, E., Birks, H.J.B., Telford, R.J., 2005. Relationships between calibrated ages and depth in stratigraphical sequences: an estimation procedure by mixed-effect regression. Holocene 15 (4), 612–618.
Hendy, E.J., Tomiak, P.J., Collins, M.J., Hellstrom, J., Tudhope, A.W., Lough, J.M., Penkman, K.E.H., 2012. Assessing amino acid racemization variability in coral intracrystalline protein for geochronological applications. Geochim. Cosmochim. Acta 86, 338–353.
Hercman, H., Pawlak, J., 2012. MOD-AGE: an age–depth model construction algorithm. Quat. Geochronol. 12, 1–10.
Kent, E.C., Kennedy, J.J., Smith, T.M., Hirahara, S., Huang, B., Kaplan, A., Parker, D.E., Atkinson, C.P., Berry, D.I., Carella, G., Fukuda, Y., Ishii, M., Jones, P.D., Lindgren, F., Merchant, C.J., Morak-Bozzo, S., Rayner, N.A., Venema, V., Yasui, S., Zhang, H.-M., 2017. A call for new approaches to quantifying biases in observations of sea surface temperature. Bull. Am. Meteorol. Soc. 98 (8), 1601–1616.
Klauenberg K, Blackwell PG, Buck CE, Mulvaney R, Röthlisberger R, Wolff EW (2011) Bayesian glaciological modelling to quantify uncertainties in ice core chronologies. Quat. Sci. Rev. 30(21 − 22): 2961–2975.
Köppen, W., 1923. Die Klimate der Erde: Grundriss der Klimakunde. de Gruyter, Berlin 369 pp.
Kuhn, T.S., 1970. The Structure of Scientific Revolutions, 2nd ed. University of Chicago Press, Chicago 210 pp.
Künsch, H.R., 1989. The jackknife and the bootstrap for general stationary observations. Ann. Stat. 17 (3), 1217–1241.
Lahiri, S.N., 2003. Resampling Methods for Dependent Data. Springer, New York 374 pp.
Michalewicz, Z., Fogel, D.B., 2000. How to Solve It: Modern Heuristics. Springer, Berlin 467 pp.
Montgomery, D.C., Peck, E.A., 1992. Introduction to Linear Regression Analysis, 2nd ed. Wiley, New York 527 pp.
Mudelsee, M., 2000. Ramp function regression: a tool for quantifying climate transitions. Comput. Geosci. 26 (3), 293–307.
Mudelsee, M., 2002. TAUEST: a computer program for estimating persistence in unevenly spaced weather/climate time series. Comput. Geosci. 28 (1), 69–72.
Mudelsee, M., 2007. Long memory of rivers from spatial aggregation. Water Resour. Res. 43 (1). https://doi.org/10.1029/2006WR005721. W01,202.
Mudelsee, M., 2009. Break function regression: a tool for quantifying trend changes in climate time series. Eur. Phys. J. Spec. Topics 174 (1), 49–63.
Mudelsee, M., 2010. Climate Time Series Analysis: Classical Statistical and Bootstrap Methods, 1st ed. Springer, Dordrecht 474 pp.
Mudelsee, M., 2014. Climate Time Series Analysis: Classical Statistical and Bootstrap Methods, 2nd ed. Springer, Cham, Switzerland 454 pp.
Mudelsee, M., Raymo, M.E., 2005. Slow dynamics of the Northern Hemisphere Glaciation. Paleoceanography 20 (4), PA4022. https://doi.org/10.1029/2005PA001153.
Olatayo, T.O., 2014. Truncated geometric bootstrap method for time series stationary process. Appl. Math. 5 (13), 2057–2061.
Parnell, A.C., Sweeney, J., Doan, T.K., Salter-Townshend, M., Allen, J.R.M., Huntley, B., Haslett, J., 2015. Bayesian inference for palaeoclimate with time uncertainty and stochastic volatility. Appl. Stat. 64 (1), 115–138.
Peters, S.C., Freedman, D.A., 1984. Some notes on the bootstrap in regression problems. J. Bus. Econ. Stat. 2 (4), 406–409.
Politis, D.N., Romano, J.P., 1994. The stationary bootstrap. J. Am. Stat. Assoc. 89 (428), 1303–1313.
Politis, D.N., Romano, J.P., Wolf, M., 1999. Subsampling. Springer, New York 347 pp.
Prais, S.J., Winsten, C.B., 1954. Trend Estimators and Serial Correlation. Cowles Commission, Yale University, New Haven, CT 26 pp. (Discussion Paper No. 383).
Priestley, M.B., 1981. Spectral Analysis and Time Series. Academic Press, London 890 pp.
Robinson, P.M., 1977. Estimation of a time series model from unequally spaced data. Stoch. Process. Appl. 6 (1), 9–24.
Robinson, P.M. (Ed.), 2003. Time Series with Long Memory. Oxford University Press, Oxford 382 pp.
Saltelli, A., Stark, P., 2018. Statistics: a social and cultural issue. Nature 553 (7688), 281.
Sarnthein, M., Balmer, S., Grootes, P.M., Mudelsee, M., 2015. Planktic and benthic ^{14}C reservoir ages for three ocean basins, calibrated by a suite of ^{14}C plateaus in the glacial-to-deglacial Suigetsu atmospheric ^{14}C record. Radiocarbon 57 (1), 129–151.
Scholz, D., Hoffmann, D.L., 2011. StalAge – an algorithm designed for construction of speleothem age models. Quat. Geochronol. 6 (3–4), 369–382.
Scott, D.W., 1979. On optimal and data-based histograms. Biometrika 66 (3), 605–610.
Seber, G.A.F., Wild, C.J., 1989. Nonlinear Regression. Wiley, New York 768 pp.
Sen, A., Srivastava, M., 1990. Regression Analysis: Theory, Methods, and Applications. Springer, New York 347 pp.
Sherman, M., Speed Jr., F.M., Speed, F.M., 1998. Analysis of tidal data via the blockwise bootstrap. J. Appl. Stat. 25 (3), 333–340.
Singh, K., 1981. On the asymptotic accuracy of Efron's bootstrap. Ann. Stat. 9 (6), 1187–1195.
Spötl, C., Mangini, A., Richards, D.A., 2006. Chronology and paleoenvironment of Marine Isotope Stage 3 from two high-elevation speleothems, Austrian Alps. Quat. Sci. Rev. 25 (9–10), 1127–1136.
Climate change 2013: the physical science basis. In: Stocker, T.F., Qin, D., Plattner, G.-K., Tignor, M.M.B., Allen, S.K., Boschung, J., Nauels, A., Xia, Y., Bex, V., Midgley, P.M. (Eds.), Working Group I Contribution to the Fifth Assessment Report of the Intergovernmental Panel on Climate Change. Cambridge University Press, Cambridge 1535 pp.
Thompson, D.W.J., Kennedy, J.J., Wallace, J.M., Jones, P.D., 2008. A large discontinuity in the mid-twentieth century in observed global-mean surface temperature. Nature 453 (7195), 646–649.
von Storch, H., Zwiers, F.W., 1999. Statistical Analysis in Climate Research. Cambridge University Press, Cambridge 484 pp.