Record-values, non-stationarity tests and extreme value distributions

By R.E. Benestad

The Norwegian Meteorological Institute, PO Box 43, 0313, Oslo, Norway [*]

Submitted: August 18, 2003

Revised: December 14, 2004

Accepted.

Published: Global and Planetary Change **vol 44**, issue 1-4, p.11-26

[*] *Corresponding author: R.E. Benestad, rasmus.benestad@met.no, The Norwegian Meteorological Institute, PO Box 43, 0313 Oslo, Norway, phone +47-22 96 31 70, fax +47-22 96 30 50*

ABSTRACT

The chance of seeing new record-values in a stationary series is described by a simple mathematical expression. The expected probability for new record-values is then used to estimate the expectation number for new parallel records in N independent stations at a given time. This probability is then compared with the observed number of new records. A confidence interval about the theoretical expectation number is estimated using Monte-Carlo integrations, and a χ^2-type test can be used to assess whether the rate of observed new records is higher than expected for a number of stationary series. The results of this record-statistics test suggest that the observed rate of record-warm monthly mean temperatures at 17 stations around the world may be unexpectedly high.

In addition to testing N independent series as a group, it is also possible to examine the number of records set in a single series of a given length. An expression is derived for how the number of records varies with the length of the series, and a confidence interval is estimated using Monte-Carlo integrations. Taking the mean number of records from 17 climate stations spread around the globe, it is shown that by the end of the 20th century, it is higher than expected if the series had been stationary.

The record-statistics tests can be used to identify non-stationarities which are problematic for extrapolations of return-periods and return-values from fitting the tails of extreme value distributions. The results for the monthly mean temperature from 17 stations world-wide point to the presence of non-stationaries, implying that a projection will under-estimate the future return-values while over-estimating the return-periods for the monthly mean temperature if the warming trend continues.

Key words: Record-value statistics extremes temperature

Introduction

The estimation of probability distribution functions (p.d.f.) for time series usually assume that the series comprises of *independent and identically distributed* (iid) random variables (Raqab, 2001; von Storch and Zwiers, 1999; Balakrishnan and Chan, 1998) (identical distribution implies stationarity and homogeneity). Extreme value statistics that aim at estimating return values and return periods (which is in effect an extrapolation) normally involve fitting the tails of distribution functions to the data. The presence of non-stationarities (here referring to non-constant p.d.f.) in the form of long-term trends in the location (e.g. mean value) or range (e.g. variance) can result in serious biases and spurious results when used for extrapolation beyond the examined interval. For example, climatic events with an estimated return period of 40 years in the present-day climate may in the future occur about every 8 year on average (Palmer and Räisänen, 2002). In the area of climate change, one is often interested in non-stationary time series, such as the global mean temperature (Jones et al., 1998) which indicates a warming trend (Houghton et al., 2001). Another aspect of climate change is the possibility of changing amplitudes or variability (Schär et al., 2004).

Record-event type statistics exists in the literature (Bunge and Goldie, 2001; Feller, 1971; Glick, 1978; Nevzorov, 1987; Nagaraja, 1988; Ahsanullah, 1989, 1995; Arnold et al., 1998; Balakrishnan and Chan, 1998; Bairamov and Eryilmaz, 2000; Raqab, 2001), but has not been widely applied within the climate research community. One reason for this may be that practicable record-event statistics is not well developed for climate studies. A simple and practicable method for applying record-event statistics to climatological data has been proposed by Vogel et al. (2001) and Benestad (2003). An important question regarding extreme values is how often we can expect a new record-event, assuming a series of iid random variables. Balakrishnan and Chan (1998) considered the distribution of record-values and the timing of each event. They applied Monte-Carlo simulations to derive confidence limits for testing whether a current record-event was consistent with an iid process. Benestad (2003) used a simpler method to examine the number of new records in monthly mean global temperature (Jones et al., 1998), monthly maximum 24-hourly precipitation, and absolute monthly maximum & minimum temperature in the Nordklim data set (Tuomenvirta et al., 2001).

The record-event statistics discussed in these papers assume a null-hypothesis of iid random values, and a falsification of the tests of iid may indicate the presence of non-stationarities (i.e. the variables are not identically distributed) or dependence between the observations. In this case, parallel series are examined, and in order to conclude whether the series are non-stationary, it is necessary to rule out the possibility of dependence between sequences and serial dependence within sequences. If such tests can identify non-stationarities, then they are useful in conjunction with fitting extreme value distributions.

In this paper, the analysis of Benestad (2003) will be extended to long series of monthly mean temperature from different parts of the world (Hansen et al., 2001). The results from the record-event analysis will be discussed in conjunction with the more traditional extreme value distributions as well as situations when the iid tests may fail. Two types of iid tests will also be compared in terms of their sensitivity to dependencies.

Data & Method

Data

A subset of NASA Goddard Institute for Space Studies (GISS) temperature series from Hansen et al. (2001)[*], listed in Table 1, was used for the record statistics analysis. The

[*] URL `http://www.giss.nasa.gov/data/update/gistemp/`

analysis requires long time series since new records are infrequent and shorter series give less precise results. Hence, one criterion was that the observations at least should extend into the 19th century and series with large gaps of missing data were excluded. The stations were selected according to whether they were up-to-date (last year is '2003'), their length (first year before '1900'), and gaps of missing values (only stations with less than 100 missing values were used). Other criteria for selecting stations was that they were spread out in order to avoid spatial dependencies and that there were no clear visual indications of inhomogeneities. The presence of inter-stations dependencies can influence the analysis (Vogel et al., 2001), and in order to avoid dependent data series, only one station was used in places where several stations might be affected by the same synoptic temperature anomalies, and therefore a final set of 17 stations was used in the iid tests described here (Figure 1: named stations). Each station provided 12 different series, one for each calendar month, and the total number of series were therefore $N = 12 \times 17 = 204$. The N series of length n were combined to form an $n \times N$ data matrix $\mathbf{X}$, and the mathematical symbol x_i will be used here to denote one single series whereas $\vec{x}_n$ will refer to N simultaneous (parallel) observations at time n.

Long series are also required for good fits to the tails of distributions and to model extremes. The Central England Temperature (CET) record (Jones, 1999) was used independently to illustrate how generalised extreme value distributions (GEVs) may change over time, as this series is the longest available instrumental record and spans over 1659–2003. The CET data were obtained from the Climate Research Unit (University of East Anglia)[*] and the Hadley Centre[†] Internet pages.

[*] http://www.cru.uea.ac.uk/ mikeh/datasets/uk/cet.htm

[†] http://www.met-office.gov.uk/research/hadleycentre/CR_data/Monthly/HadCET_act.txt

Method

The probability of the last value in a series of n iid observations being the highest (i.e. setting a new record) with no ties for the maximum value can be estimated according to:

$$p_n(1) = \frac{1}{n}. \tag{1}$$

The notation adopted here is $p_n(1)$ denoting *one* new record for the nth observation, and $p_n(0)$ would be the probability of seeing no new records.

For short time-series, the chance of seeing a new record value is higher than for long time records, and there is little change in probability for $n \gg 1$. Because of the reciprocal power of n, equation (1) can be expressed as linear equation in the logarithmic values of p and n:

$$\ln[p_n(1)] = -\ln(n). \tag{2}$$

It is possible to relate the theoretical probability to the empirical data by utilising an expectation value for N stations defined as $E_n = Np_n(1)$. Then the theoretical probability $p_n(1)$ can be compared with $\hat{E}_n/N$ (this quantity is henceforth referred to as the 'record-density'), where $\hat{E}_n$ is the empirical estimate of E_n. For each sequence, let $\Upsilon(x_n)$ be 1 if the value x_n is a new record, and otherwise zero:

$$\Upsilon(x_n) = \begin{cases} 0 & \text{for } x_n \le \max[x_1, x_2, ..., x_{n-1}] \\ 1 & \text{for } x_n > \max[x_1, x_2, ..., x_{n-1}] \end{cases} \tag{3}$$

Then, for N parallel observations $\hat{E}_n = \sum \Upsilon(\vec{x}_n)$, summing the simultaneous record-events.

The expected number of records for a *single* stationary time series of length n can be estimated according to $\mathcal{E}(n) = \sum_{i=1}^{n} p_i(1)$. A similar aggregated statistic can be estimated for a network of stations according to

$$\hat{\mathcal{E}}(n) = \sum_{i=1}^{n} \hat{E}_i/N. \tag{4}$$

A small number of missing values does not have a significant effect on the aggregated analysis, since the record-densities $\hat{E}_n/N$ are taken as the mean number of records over valid parallel observations only. Missing values reduces the number of independent variables, hence making the estimates less precise. In this study, the focus has been on the aggregated statistics as opposed to a single series, and the term 'estimated number of record' will henceforth refer to equation (4).

Note that the expression for the probability in equation (1) does *not* describe a p.d.f. and $\int p_n(x)dx \neq 1^*$. The analysis of Benestad (2003) demonstrated through sets of Monte-Carlo simulations that the mathematical framework given by equations (1) and (2) provide a good description of the record-event incidence, and that $\mathcal{E}(n) = \sum_{i=1}^{n} 1/i$ gives a good estimate of the number of record-events in a series with iid random values simulated through Monte-Carlo integrations.

It can furthermore be shown that the probability of seeing at least one new record-event in many parallel independent series of length n increases with the number of series N:

[*] Therefore, the expected number records for the entire record $\mathcal{E}(n)$ cannot be estimated using the expression $\int np_n(1)dn$, but must involve an elaborate chain of combined conditional probabilities for series of lengths $i = 1...n$.

$$P_n^N(1) = 1 - \left(1 - \frac{1}{n}\right)^N . \tag{5}$$

The dashed curve in Figure 2 shows the probability of seeing at least one new record for measurement $i = 1, 2, ..., n$ in a set of N independent series, according to equation (5). For $N = 17 \times 12 = 204$ there is a good chance of seeing new parallel record-values, even after a 100 measurements (years), in contrast to a single series and to the expectation number divided by the number of series ($E_n/N = p_n(1)$; solid black). Because the low values of E_n/N and the persistence of high $P_n^N(1)$ for high n, the deviation from the record-statistics from iid becomes more clear for higher values of n (i.e. long series).

It is important to test whether the occurrence of record-events deviates significantly from the expected rate. The iid tests described here provide a simple and practicable means to examine sequences for whether they are non-stationary, with an emphasis on extremes. A set of Monte-Carlo integrations was used to test the actual observations with the theory, involving $1000 \times N$ stochastic independent and stationary (all values are drawn from the same distribution) Gaussian series of same length as the observations, produced with a random number generator*. The Monte-Carlo simulations through out this study involved replacing the actual station observations with stochastic numbers ($\mathbf{X} \rightarrow \mathbf{X_{MC}}$), preserving the dimensions ($n \times N$) of the data matrix. The record-statistics for each of these replacements was obtained through *identical* subsequent processing to that of the real observations. Hence the record-time statistics from the Monte-Carlo simulations are ensured to belong to the same universe as the observation-based analysis, as $\hat{E}_n/N$ can only take values i/N for $i = 0, 1, ..., N$ in the Monte-Carlo simulations as well as in the analysis of the station series. There were two types of Monte-Carlo

* The `rnorm()` function in `R` which returns random values pertaining to a Gaussian distribution of zero mean and unit standard deviation.

results obtained in this study: i) confidence intervals for χ^2-type tests, and ii)confidence intervals for the estimated number of record-events ($\hat{\mathcal{E}}$). For the first kind, the χ^2-statistic null-distribution was derived from 1000 corresponding values of $\hat{E}_n/N$ calculated using stochastic numbers ($\mathbf{X_{MC}}$) instead of actual observations ($\mathbf{X}$). The confidence interval for $\hat{\mathcal{E}}$ was based on a null-distribution derived from 1000 estimates of $\hat{\mathcal{E}}$, each derived from N stochastic series of length n instead of actual observations. The same procedure was used as for the observations (equation (4)).

Results

The timing of the record-events can be seen in Figure 3, which shows the timing of record events found when time runs forwards (lower part of the figure, henceforth referred to as the 'forward' analysis) as well as backward in time ('backward' analysis, upper part). It is apparent from Figure 3 that the number of records becomes much more rare with time (number of observations) in case of the 'backward' analysis compared to the 'forward' analysis. This difference is consistent with a long-term positive trend and the series being non-stationary, suggesting that the data are inconsistent with the null-hypothesis of being iid. There are only a few indications of the record-events taking place simultaneously (clusters of points) after the first few years (clustering in the beginning is expected). In the 'forward' analysis, there is a slight tendency of clustering of records, whereas such clusters of record-events are virtually absent in the 'backward' results. The few clusters seen in Figure 3 may be an indication of some dependency among the sequences as well as within these in terms of times when new records are set. The effect of dependencies will be examined more closely later on.

Empirical estimates are obtained for the expectation value E_n of number of new parallel records seen at the nth observation for a set of 17×12 independent series, and these estimates can be compared with the expected number of record-events. Figure 4a shows such a comparison, where the probability is plotted with the record-densities, which is defined as the estimated expectation value $\hat{E}_n$ divided by the number of records N. In general, the empirical estimates at first sight appear to follow the expected values, albeit with some deviations from the confidence region. The relationship between the theory and empirical data can be scrutinised in more detail if the log-relations in equation (2) are used, and Figure 4b shows that the empirical data do not lie on a straight line as a good fit would do. Rather, the points representing the 'forward' analysis suggest a higher than expected record-densities toward the end of the series. A type of χ^2-test* applied to $\hat{E}_i/N$ and $p_i(1); i = 1, 2, ..., n$ suggests that these distributions are statistically different at the 5% level (using the first type of Monte-Carlo simulation described above). The results from the 'backward' analysis also suggest a deviation from the null-hypothesis of iid, but not with a statistical significance at the 5% level. This kind of asymmetry can arise from slight tendency of clustering of record-events in the 'forward' analysis (Figure 3).

A different line of tests can be performed on the expected number of records $\mathcal{E}(n)$ (Figure 5) for a single series or a combination of series. The latter gives more precise results and obtains higher statistical power. Here, the value of $\mathcal{E}(n)$ is taken as the mean over all the stations and months: $\hat{\mathcal{E}} = (1/N) \sum_{i=1}^{n} \hat{E}_i$. The expression $\mathcal{E}(n) = \sum_{i=1}^{n} 1/i$ can be approximated as $\ln(n)$ for large values of n (i.e. the summation is replaced by an integral), and $\mathcal{E}(n) \approx \ln(n)$, so that an exponential scale can be used for the y-axis to reproduce the approximately linear relation between $\exp[\hat{\mathcal{E}}(n)]$ and n. The 95% confi-

* a standard χ^2-test (Wilks, 1995, p. 133, eq. 5.18) applies to p.d.f.s, whereas in this case a similar methodology is applied to $p_n(1)$ which is *not* a p.d.f. The test statistic is $\sum_{i=1} \frac{(\hat{E}_i/N - p_n(1))^2}{p_n(1)}$.

dence region is again estimated using Monte-Carlo simulations with $1000 \times N$ stochastic series. The comparison between the null-hypothesis of iid values and the empirical data in Figure 5 indicates that the incidence of record-events is initially low in the early part of the record for the 'forward' analysis, but increases and is high toward the end. Conversely, the 'backward' analysis yields a high rate of new records in the beginning and low frequency towards the end.

The two types of iid-tests (χ^2-based and $\mathcal{E}(n)$) yield slightly different results, as the latter gives a stronger indication of a deviation from iid. This difference leads on to the question of whether this difference can be attributed to dependencies.

Benestad (2003) examined the effect of serial correlation on the expected number of records through a set of Monte-Carlo simulations ($\hat{\mathcal{E}}_{MC}(n)$), and found the central location of the $\hat{\mathcal{E}}_{MC}(n)$ distribution to be insensitive to serial correlation (dependencies). In this context, dependencies imply a smaller effective number of parallel sequences N. Two identical series would not alter the number of records $\hat{\mathcal{E}}(n)$ since its derivation involves the mean record-event of the series. The sensitivity of $\hat{\mathcal{E}}_{MC}(n)$ to N was explored further through a set of Monte-Carlo simulation (Figure 6; open circles), and corresponding confidence limits were estimated (shown as lines). A linear least-squares regression was applied to the 2.5% and 97.5% quantiles of $\hat{\mathcal{E}}_{MC}(n)$ versus $\log(n)$ since these approximately formed a linear relationship, and Table 2 lists the results of the regression for different values of N. In accordance with Benestad (2003), the values for $\hat{\mathcal{E}}(n)$ were insensitive to N, whereas the confidence interval was smaller for larger N. Hence, dependencies affect the statistical significance but not the number of records $\hat{\mathcal{E}}(n)$. A similar exercise was carried out to examine the effect of dependencies on the χ^2-type test (Figure 6; solid grey circles). The mean value of the $\chi^2_{MC}(n)$ distributions is more sensitive to N than the $\hat{\mathcal{E}}(n)$ statistic. The 95% confidence interval of $\chi^2_{MC}(n)$ was also sensitive to N. In other words, Monte-Carlo simulations with different values for N show that the values of $\hat{\mathcal{E}}(n)$ are insensitive to the inter-sequential dependencies whereas χ^2 is more sensitive, and the confidence interval for both are clearly affected by the number of independent series.

In order to study the effect of potential serial and inter-station dependencies even further, the data matrix was sub-sampled by selecting only 4 instead of 12 months per year, each of which separated by two intervening months. Different months were selected from the most adjacent stations in order to reduce the effects of spatial correlations, as lagged inter-station cross-correlations are smaller than simultaneous inter-station correlations. Table 3 shows which months were selected from each station series, Figure 7 shows a plot of the correlation matrix, and Figure 8 shows that there are few coinciding record-events, suggesting generally low and insignificant cross-correlations. The analysis shown in Figure 5 was repeated for the subset ($N = 68$) and the results are presented in Figure 9: the results from the sub-set suggest that the series are non-stationary and not merely deviant from iid.

To test of the sensitivity of the iid testing to the type of distribution, a new set of Monte-Carlo simulations was conducted, replacing the normal distribution with gamma, generalised extreme value distribution (GEV), and binomial distributions (`rgamma()`, `rgev()`, `rbinom()` respectively). The gamma and GEV distributions used three different shape parametres ([1, 10, 0.1] and [1,10,-1] respectively) and the Binomial distribution was constructed by taking the number of trials to be 1000 and a probability $p = 0.1$. The value of $\hat{\mathcal{E}}(n)$ (n=107) were 5.246532, 5.243208, 5.244004, 5.244242, 5.246963, 5.244904, 5.248688, and 4.918285 for the normally distributed, the 3 gamma distributions, the 3 GEV distributions, and the Binomial distribution respectively. With the exception of the Binomial distribution, all the distributions essentially gave the same relationship between

$\hat{\mathcal{E}}(n)$ and n (Figure 10).

Discussion & Conclusions

The reason why the Binomial case deviates from the others at $n \geq 10$ is that this distribution typically produces only $\sim$ 50–60 different descrete values for each set of simultaneous realisation ($\vec{x}_n$) of random numbers, whereas the other distributions yield $\sim N$ different values (rational numbers). In theory, the Binomial distributions with 1000 trials produces descrete numbers $x_i \in [0, 1, 2, ..., 1000]$. The effect of constraining the values to a small finite set can be illustrated by considering a simplified ideal case where x_i consist of random descrete numbers but with $x_i \in [x_1, x_2, x_3]$ (analogous to states in quantum mechanics), $i = 1, 2, ..., n$. Then $\hat{\mathcal{E}}(n)$ can never exceed the value of 3, regardless of how long the series is. This is true for both 'forward' and 'backward' analysis. Hence, for discrete numbers following the Binomial distribution, the probability for setting a new record-value $p_n(1) \neq 1/n$ because the values of $\mathbf{X}$ are confined to a finite set of descrete values and many ties are present. The Binomial distribution results illustrate one limitation of the iid tests, that they only truely work for unconstrained rational numbers and not for series holding descrete levels with a small and finite number of levels ($< n$). This 'finite set' aspect of the method may have relevance for cases where instruments have a fixed range, the readings are truncated to a few decimal places, and the instrument's upper range is close to the highest readings. If the highest values are cut-off, then both 'forward' and 'backward' analysis will yield too low values for $\hat{\mathcal{E}}(n)$ for large values of n. In this respect, low $\hat{\mathcal{E}}(n)$ in both 'forward' as well as 'backward' analysis may suggest poor instrument performance.

The record-statistics described here only utilises part of the information contained in the data, and one may argue that it has a low statistical power. While this argument may be valid for a single series, the iid testing is well-suited for aggregating many different parallel series, regardless of the distribution of the individual sequence, and such aggregations improve the statistical power of the analysis as long as the criterion of independence is valid. Therefore the iid tests are well-suited for studying global change. Furthermore, the advantages of the iid tests performed here are their simplicity and that they make no assumptions about the distribution of the data. Also, these iid tests are fundamentally different to other types of extreme value analysis (e.g. general extreme value distributions) and hence serve as a valuable complement.

The null-hypothesis of iid random variables was rejected at the 5% confidence level in all the tests, except for χ^2-test of the 'backward' analysis of the record-density. One reason why this 'backward' test did not detect any unexpected record-events may the clustering in time of record-events may have caused an asymmetry in the χ^2 statistic. A clustering in either 'forward' or 'backward' analysis may be sufficient to indicate dependencies (dependencies do not vanish by changing order). In this case, the test of number of record-events $\hat{\mathcal{E}}(n)$ appears to be superior to the χ^2-test because it is less sensitive to dependencies, and both the 'forward' and 'backward' analyses indicate that the number of record-events is outside the 95% confidence interval after 1985. The $\hat{\mathcal{E}}(n)$ statistic may be more sensitive with larger n since $d\mathcal{E}(n)/dn \rightarrow 0$ as $n \rightarrow \infty$ for iid processes whereas non-stationarities often lead to persistently high frequency of new record-events.

The results from the record-event analysis have indicated that the empirical data are inconsistent with the null-hypothesis of iid since new record-warm monthly mean temperatures are observed at a higher rate than a similar set of stationary series of independent values would imply. If dependencies can be ruled out, then this is a test of stationarity with respect to extreme values and has implications for studies on extremes.

For instance, extreme value distributions (Coles, 2001) used for inferring return-values and return-periods may provide spurious results since the tails of the distributions may change in the future as a result of such non-stationarities. Figure 11 illustrates this by presenting an analysis of return-values and return-periods for the Central England August Temperature based on a GEV*. Two GEV fits have been found for the first (grey) and the second half (black) of the series, and a general increase in the return-values between these two periods is evident. This increase is also consistent with the notion of non-stationarity (since 'backward' $\hat{\mathcal{E}}(n)$ is below whereas 'forward' is within the upper range of the 95% confidence interval) and a shift in the tail of the distribution. However, in spite of non-stationarities, return-values and periods obtained through GEV modelling may still be useful concepts if they are considered as quantiles summarising the marginal distribution of the data.

A simple test against non-stationarity based on the fact that the 95% confidence intervals are close to linear in $\log(n)$ can be utilised to indicate whether the GEV results are representative for the future or not. A regression analysis (Table 2) gives the linear relationship between the lower and upper confidence limits for a number of N. In this case, the high number of observed record-events for the GISS series suggests that the GEV may over-estimate the return-period (or under-estimate return-value) for future events, assuming the trend continues.

The analysis revealed that the monthly mean GISS temperatures are not stationary, but exhibit a positive long-term trend and inter-decadal fluctuations. A long-term trend may constitute a climate change, possibly caused by anthropogenic greenhouse gases (Houghton et al., 2001). However, other factors may also affect the climate, such

* Applying a general extreme value (GEV) distribution from the `evd` package (Stephenson, 2003) in `R` and using all values greater than the 75% percentile to fit the GEV.

as landscape changes or solar activity. The presence of so-called “urban heat islands” and other inhomogeneities have not been accounted for in this analysis (Peterson, 2003), and the cause for the change in the data has not been identified in this study. It is interesting to note the recent record-warm summer months in Europe for 2003 (Schär et al., 2004), which may be consistent with the conclusion of this study that new record-events are occurring more frequently than expected.

Acknowledgement

This work was done under the Norwegian Regional Climate Development under Global Warming (RegClim) programme, and was supported by the Norwegian Research Council (Contract NRC-No. 120656/720) and the Norwegian Meteorological Institute. The analysis was carried out using the `R` (Ellner, 2001; Gentleman and Ihaka, 2000) data processing and analysis language, which is freely available over the Internet (URL http://www.R-project.org/). Valuable comments from an anonymous referee have also improved the paper.

References

Ahsanullah, M. 1989. Introduction to Record Statistics. Prentice Hall: Hemel Hempstead, U.K.

Ahsanullah, M. 1995. Record Statistics. Nova Science Publishers Inc.: Commack, N.Y., USA.

Arnold, B.C., Balakrishnan, N., and Nagaraja, H.N. 1998. Records. John Wiley: New York.

Bairamov, I.G., and Eryilmaz, S.N. 2000. Distributional properties of statistics based on minimal spacing and record exceedance statistics. Journal of Statistical Planning and Inference, **90**, 21–33.

Balakrishnan, N., and Chan, P. S. 1998. On the normal record values and associated inference. Statistics & Probability Letters, **39**, 73–80.

Benestad, R.E. 2003. How often can we expect a record-event? Climate Research, **25**, 3–13.

Bunge, J., and Goldie, C.M. 2001. Handbook of Statistics. Stochastic Processes: Theory and Methods, vol. 19. Chap. Record sequences and their applications.

Coles, S.G. 2001. An Introduction to Statistical Modeling of Extreme Values. Springer: London.

Ellner, S.P. 2001. Review of R, Version 1.1.1. Bulletin of the Ecological Society of America, **82**(April), 127–128.

Feller, W. 1971. An Introduction to Probability Theory & Its Applications. 2 edn. John Wiley & Sons: New York.

Gentleman, R., and Ihaka, R. 2000. Lexical Scope and Statistical Computing. Journal of Computational and Graphical Statistics, **9**, 491–508.

Glick, N. 1978. Breaking Records and Breaking Boards. Amer. Math. Monthly, **85**, 12–26.

Hansen, J., Ruedy, R., Sato, M., Imhoff, M., Lawrence, W., Easterling, D., Peterson, T., and Karl, T. 2001. A closer look at United States and global surface temperature change. Journal of Geophysical Research, **106**, 23,947–23,963.

Houghton, J.T., Ding, Y., Griggs, D.J., Noguer, M., van der Linden, P.J., Dai, X., Maskell, K., and Johnson, C.A. 2001. Climate Change 2001: The Scientific Basis. Contribution of Working Group I to the Third Assessment Report of IPCC. International Panel on Climate Change, (Available from www.ipcc.ch).

Jones, P. D., Raper, S. C. B., Bradley, R. S., Diaz, H. F., Kelly, P. M., and Wigley, T. M. L. 1998. Northern Hemisphere surface air temperature variations, 1851–1984. J. Clim. Appl. Met., **25**, 161–179.

Jones, P.D. 1999. Classics in physical geography revisited. Progress in Physical Geography, **23**(3), 425–428. The Central England Temperature (CET) and Manley.

Nagaraja, H.N. 1988. Record Values and Related Statistics - A Review. Comun. Statist. Theo. Meth., **17**, 2223–2238.

Nevzorov, V.B. 1987. Records. Theo. Prob. Appl., **32**, 201–228.

Palmer, T.N., and Räisänen, J. 2002. Quantifying the risk of extreme seasonal precipitation events in a changing climate. Nature, **415**, 512–514.

Peterson, T.C. 2003. Assessment of Urban Versus Rural In Situ Surface Temperatures in the Contiguous United States: No Difference Found. Journal of Climate, **16**, 2941–3071.

Raqab, M.Z. 2001. Some results on the moments of record values from linear exponential distribution. Mathematical and Computer Modelling, **34**, 1–8.

Schär, C., Vidale, P.L., Lüthi, D., Frei, C., Häberli, C., Linger, M.A., and Appenzeller, C. 2004. The role of increasing temperature variability in European summer heatwaves.

Nature, **427**, 332–336.

Stephenson, A. 2003. Functions for extreme value distributions. http://cran.r-project.org/.

Tuomenvirta, H., Drebs, A., Førland, E., Tveito, O.E., Alexandersson, H., Laursen, E.V., and Jónsson, T. 2001. Nordklim data set 1.0. KLIMA 08/01. met.no, P.O.Box 43 Blindern, N-0313 Oslo, Norway (www.met.no).

Vogel, R.M., Zafirakou-Koulouris, A., and Matalas, N.C. 2001. The Frequency of Record Breaking Floods in the United States. Water Resources Research, **37**, 1723–1731.

von Storch, H., and Zwiers, F.W. 1999. Statistical Analysis in Climate Research. Cambridge University Press: Cambridge, UK.

Wilks, D.S. 1995. Statistical Methods in the Atmospheric Sciences. Academic Press: Orlando, Florida, USA.

TABLE 1. A list of long station seriess from the GISS climate station temperature series with details about population, coordinates and record length. A subset of these stations, marked in column 1, was selected to reduce spatial dependencies: only 17 out of these 42 stations were used for the subsequent analysis.

	Location	population	Latitude	longitude	start	stop
			°N	°E	year	year
1	Aberdeen/Dyce	210000	57.2	-2.2	1871	2003
	Akola	168000	20.7	77.1	1875	2003
	Bangalore	1654000	13	77.6	1875	2003
	Belfast/Alder	552000	54.6	-6.2	1834	2003
	Berlin-Tempel	3021000	52.5	13.4	1701	2003
2	Bismarck/Mun.	50000	46.8	-100.8	1875	2003
	Bombay/Cola	5971000	18.9	72.8	1842	2003
3	Buenos Aires	9927000	-34.6	-58.5	1856	2003
	Christchurch	165000	-43.5	172.5	1864	2003
4	Concord Usa	36000	43.2	-71.5	1871	2003
	Curitiba	844000	-25.4	-49.3	1885	2003
	Dar-El-Beida	1365000	36.7	3.2	1856	2003
	Dijon	150000	47.3	5.1	1845	2003
	Dublin Airpor	680000	53.4	-6.2	1831	2003
	Enisejsk	20000	58.5	92.2	1871	2003
5	Funchal	38000	32.6	16.9	1864	2003
	Geneve-Cointr	320000	46.2	6.1	1753	2003
6	Honolulu, Oah	836000	21.4	-157.9	1883	2003
7	Ishigakijima	35000	24.3	124.2	1897	2003
	Jacksonville U/A To Waycro	898000	30.4	-81.7	1872	2003
	Kharkiv	1444000	50	36.1	1892	2003
	Larissa	72000	39.6	22.4	1899	2003
8	Lisboa/Geof	1100000	38.7	-9.2	1854	2003
9	Moskva	8011000	55.8	37.6	1779	2003
10	Nassau Airpor	134000	25.1	-77.5	1855	2003
	Omsk	1014000	55	73.4	1887	2003
	Poona	1135000	18.5	73.8	1876	2003
11	Portland/Int.	1414000	45.6	-122.6	1873	2003
12	Saentis	0	47.2	9.3	1883	2003
13	Sao Paulo	7034000	-23.5	-46.6	1887	2003
	Saratov	856000	51.6	46	1836	2003
14	Seychelles In	0	-4.7	55.5	1894	2003
	Strasbourg	252000	48.5	7.6	1801	2003
	Tampa/Int.,Fl	1995000	28	-82.5	1825	2003
15	Thessaloniki	482000	40.5	23	1892	2003
16	Thiruvanantha	410000	8.5	77	1837	2003
	Trier-Petrisb	100000	49.8	6.7	1788	2003
17	Turuhansk	0	65.8	87.9	1881	2003
	Valladolid	228000	41.6	-4.8	1866	2003
	Washington/Na	3734000	38.9	-77	1820	2003
	Wroclaw Ii	523000	51.1	16.9	1792	2003
	Zurich (Town)	718000	47.4	8.6	1836	2003

TABLE 2. a) Upper and lower levels of 95% confidence interval for number of record-events in one series as estimated by applying a linear regression to the Monte-Carlo results in Figure 5 assuming the relationship $y = \exp(\mathcal{E}(n))$, $x = n$, and using a linear relationship between y and x.

	Estimate	Std. Error	t value	Pr(>|t|)
Upper				
(Intercept)	-2.4022	0.5785	-4.15	0.0005
x	2.3658	0.0093	254.34	0.0000
Lower				
(Intercept)	2.9810	0.3619	8.24	0.0000
x	1.3730	0.0058	235.98	0.0000

b) Best-fit coefficients for linear relationship between confidence limits and length of sequence (n) from a series of Monte-Carlo simulations. The coefficients given are for the equation $q_i(n) = \log[a + b \times n]$. The mean value of $\hat{\mathcal{E}}(n)$ was insensitive to N.

N	$q_{0.025}(n)$		$q_{0.975}(n)$	
	a	b	a	b
10	4.29	0.55	-32.65	6.17
20	4.65	0.73	-17.21	4.29
30	4.31	0.92	-11.83	3.70
50	3.83	1.06	-6.56	3.03
68	3.44	1.13	-4.78	2.82
75	3.47	1.15	-5.06	2.75
100	3.59	1.22	-5.27	2.68
204	2.98	1.37	-2.40	2.37
500	2.28	1.50	-0.85	2.10
600	2.24	1.52	-0.91	2.08
1000	1.92	1.58	-0.17	2.01

TABLE 3. Details of the sub-sampling of the station data in order to reduce effects of dependencies. For each station, only 4 months were used: 'FMAN' denotes February, May, August and November, 'MJSD' means March, June, September and December, wheras 'JAJO' refers to January, April, July, and October. $N = 68$.

Months used	Station
FMAN	Aberdeen
FMAN	Bismarck
FMAN	Buenos-Aires
MJSD	Concord
MJSD	Funchal
FMAN	Honolulu
FMAN	ishigakijima
JAJO	Lisboa
JAJO	Moskva
FMAN	Nassau
MJSD	Portland
MJSD	Saentis
MJSD	Sao Paulo
FMAN	Seychelles
FMAN	Thessaloniki
MJSD	Thiruvanantha
FMAN	Turuhansk

Figure 1. Map showing the climate stations used in the record-event statistics. The named ball symbols show the locations included in the record-event analysis, whereas the unnamed filled grey circles show sites excluded for reasons such as spatial correlations or large missing value gaps.

Figure 2. A comparison between the expectation number divided by the number of series (E_n/N; solid black) and the theoretical probability of seeing at least one new record ($P_n^N(1)$; dashed grey).

Figure 3. The timing of record-events incidents. The lower half of the figure shows 'forward' analysis whereas the upper part gives the timing of record-events when running backward in times ('backward' analysis). The vertical axis represent different sequences (204 in each analysis).

Figure 4. a) The probability of seeing a new record for $N = 17 \times 12 = 204$ (17 stations each with 12 monthly series) series against length of series. b) a Log-log plot showing the relationship between the theoretical and empirical values of E_n/N. The grey-filled circles show the results from the 'forward' and the diamonds from the 'backward' analysis. The grey line indicates the best linear fit between theory and empirical data for the 'forward' and black for the 'backward' analysis (merely shown to guide the eye). The dashed lines indicate the diagonal and the boundaries of the 95% confidence region. The time axis is the number of observations, which in this case is one per year. The 'χ^2-results' shown in panel b are from a 'standard χ^2-test', but derived from the test described in the text, applied to $p_n(1)$ and *not* a p.d.f.

Figure 5. A comparison between the theoretical expected number of record-events and the observed number as a function of the length of series. The observed numbers are taken as the aggregated value of all $N = 17 \times 12 = 204$ series. The vertical axis represent the number of record events $\hat{\mathcal{E}}(n)$ and the horizontal axis gives the length of sequence n.

Figure 6. Monte-Carlo simulations testing the sensitivity of the χ^2-type and the $\hat{\mathcal{E}}(n)$ tests to the value of N. Both the mean value (points) and the confidence intervals (lines) of the distributions are shown.

Figure 7. The cross-correlation matrix for the station series $N = 68$ sub-set described in Table 3.

Figure 8. Same as Figure 3 but for the N=68 sub-sample (Table 3).

Figure 9. The iid test shown in Figure 5 repeated for the smaller subset (Table 3) in order to reduce the effect of potential inter-dependencies and the cross-correlation between the various series was small (Figure 7).

Figure 10. A comparison between $\hat{\mathcal{E}}(n)$ estimated through a set of Monte-Carlo simulations (N=1000) and n for a number of processes with different p.d.f.

Figure 11. Analysis of return-values and return-periods based on a best-fit GEV. The grey line shows the fit to the first half of the data whereas the black line shows for the second half. The figure shows confidence intervals only for the second half of the series (the points and CI for the first half show similar spread around the grey curve). All values greater than the 75% percentile have been used to fit the GEV. Test of iid: 'forward' gives $\hat{\mathcal{E}}(n) = 7.08$ while 'backward' yields 4.58 (95% confidence interval: 5.25 – 7.67).

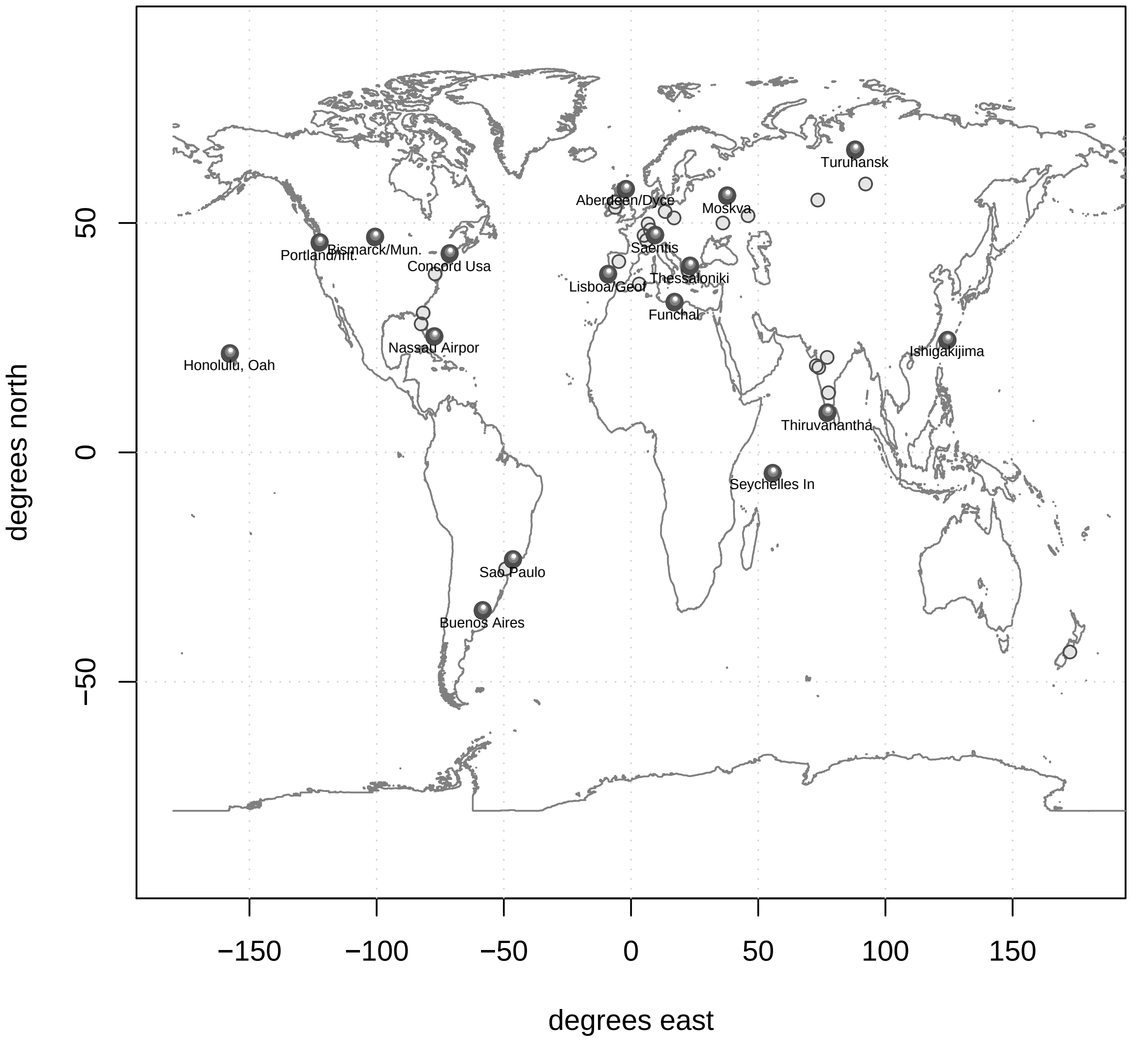
Turuhansk
Aberdeen/Dyce
Moskva
Portland/Int.
Bismarck/Mun.
Concord Usa
Saentis
Lisboa/Geof
Thessaloniki
Funchal
Nassau Airpor
Honolulu, Oah
Ishigakijima
Thiruvanantha
Seychelles In
Sao Paulo
Buenos Aires
degrees north
50
0
−50
−150
−100
−50
0
50
100
150
degrees east

Figure 1.

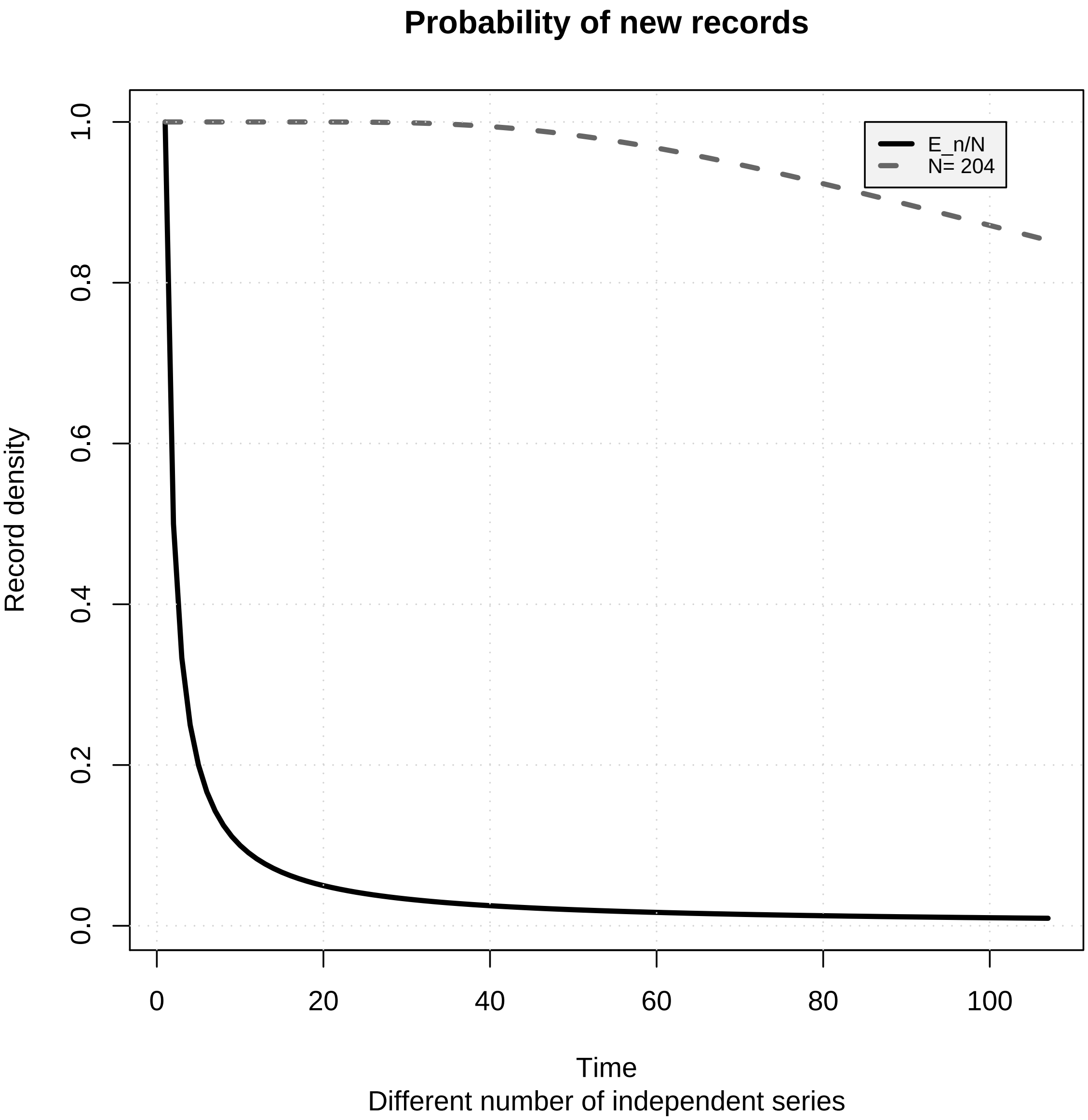
Probability of new records
Record density
0.0
0.2
0.4
0.6
0.8
1.0
0
20
40
60
80
100
Time
Different number of independent series
E_n/N
N= 204

Figure 2.

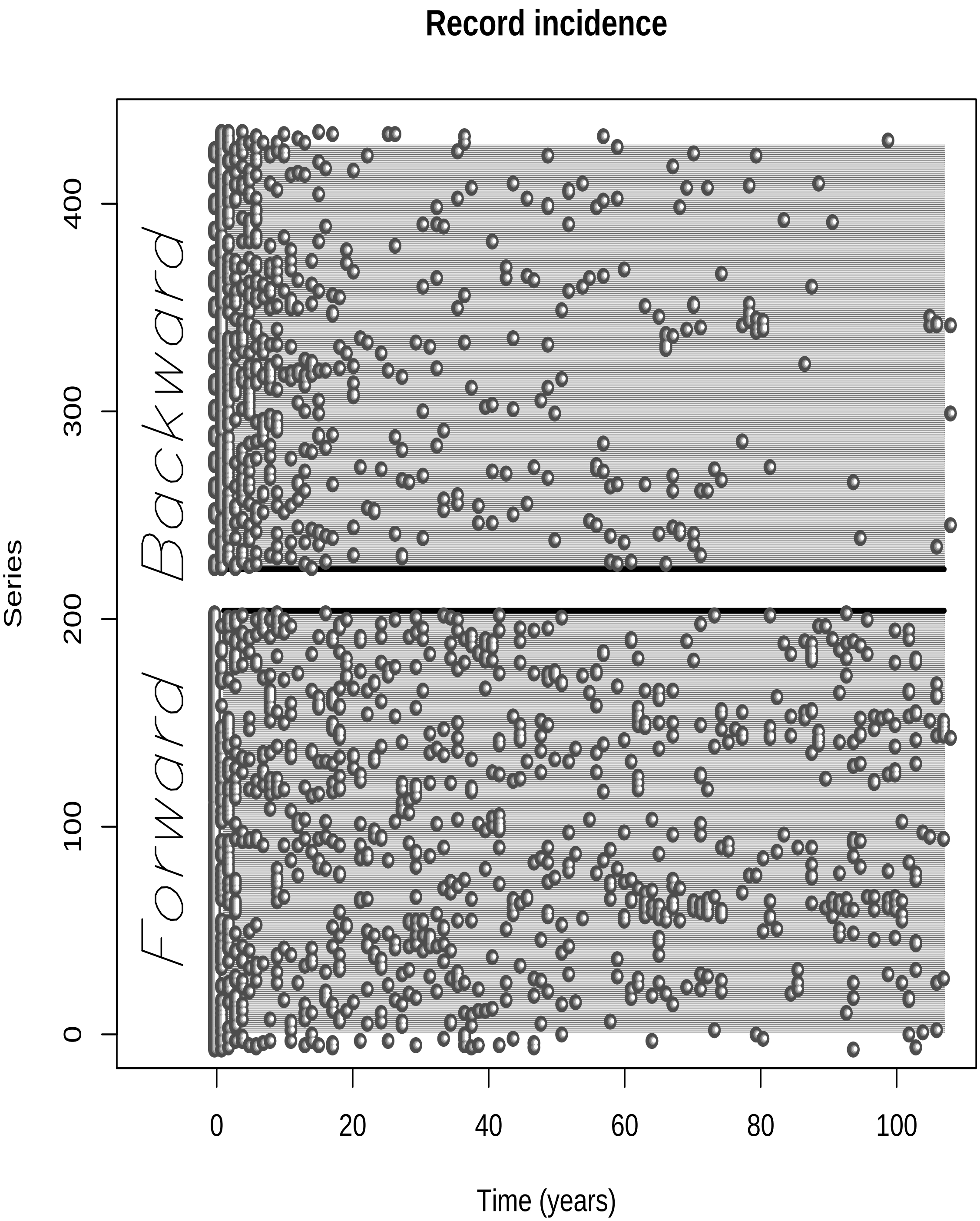
Record incidence
Series
Time (years)
Backward
Forward
0
100
200
300
400
0
20
40
60
80
100

Figure 3.

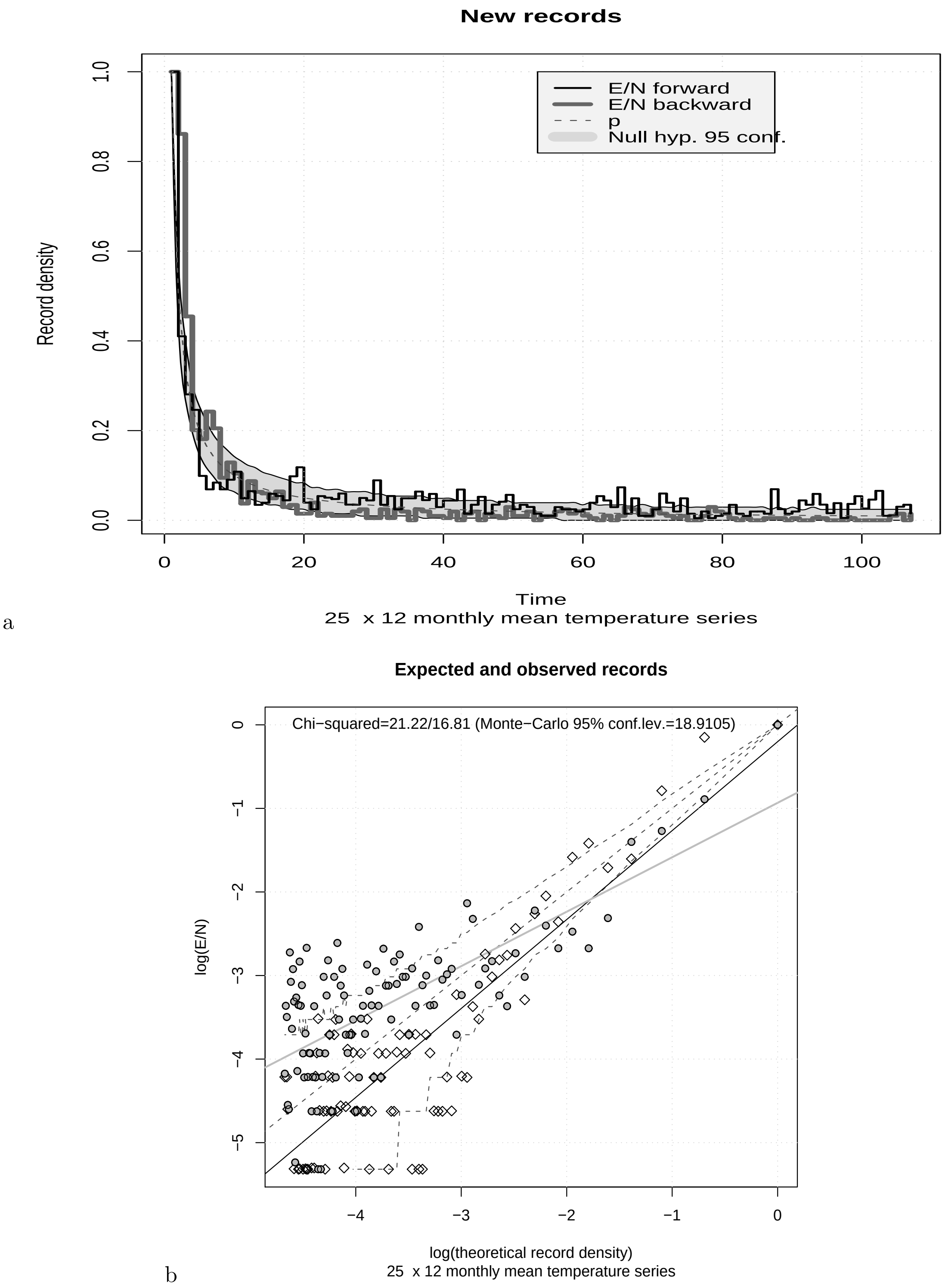
New records
Record density
1.0
0.8
0.6
0.4
0.2
0.0
E/N forward
E/N backward
p
Null hyp. 95 conf.
0
20
40
60
80
100
Time
25 x 12 monthly mean temperature series
a
Expected and observed records
Chi−squared=21.22/16.81 (Monte−Carlo 95% conf.lev.=18.9105)
log(E/N)
0
−1
−2
−3
−4
−5
−4
−3
−2
−1
0
log(theoretical record density)
25 x 12 monthly mean temperature series
b

Figure 4.

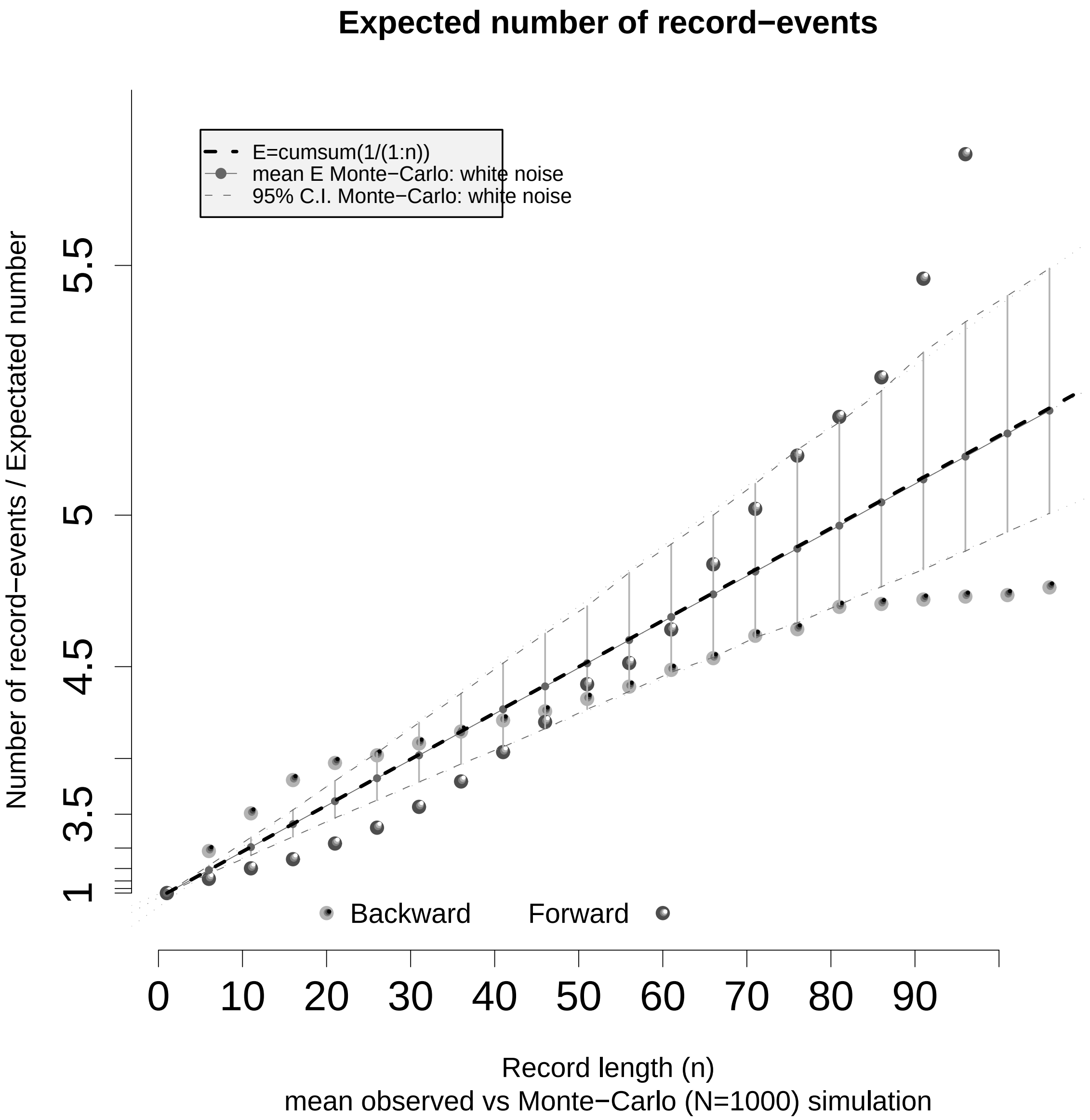
Expected number of record−events
E=cumsum(1/(1:n))
mean E Monte−Carlo: white noise
95% C.I. Monte−Carlo: white noise
Number of record−events / Expectated number
1
3.5
4.5
5
5.5
Backward
Forward
0
10
20
30
40
50
60
70
80
90
Record length (n)
mean observed vs Monte−Carlo (N=1000) simulation

Figure 5.

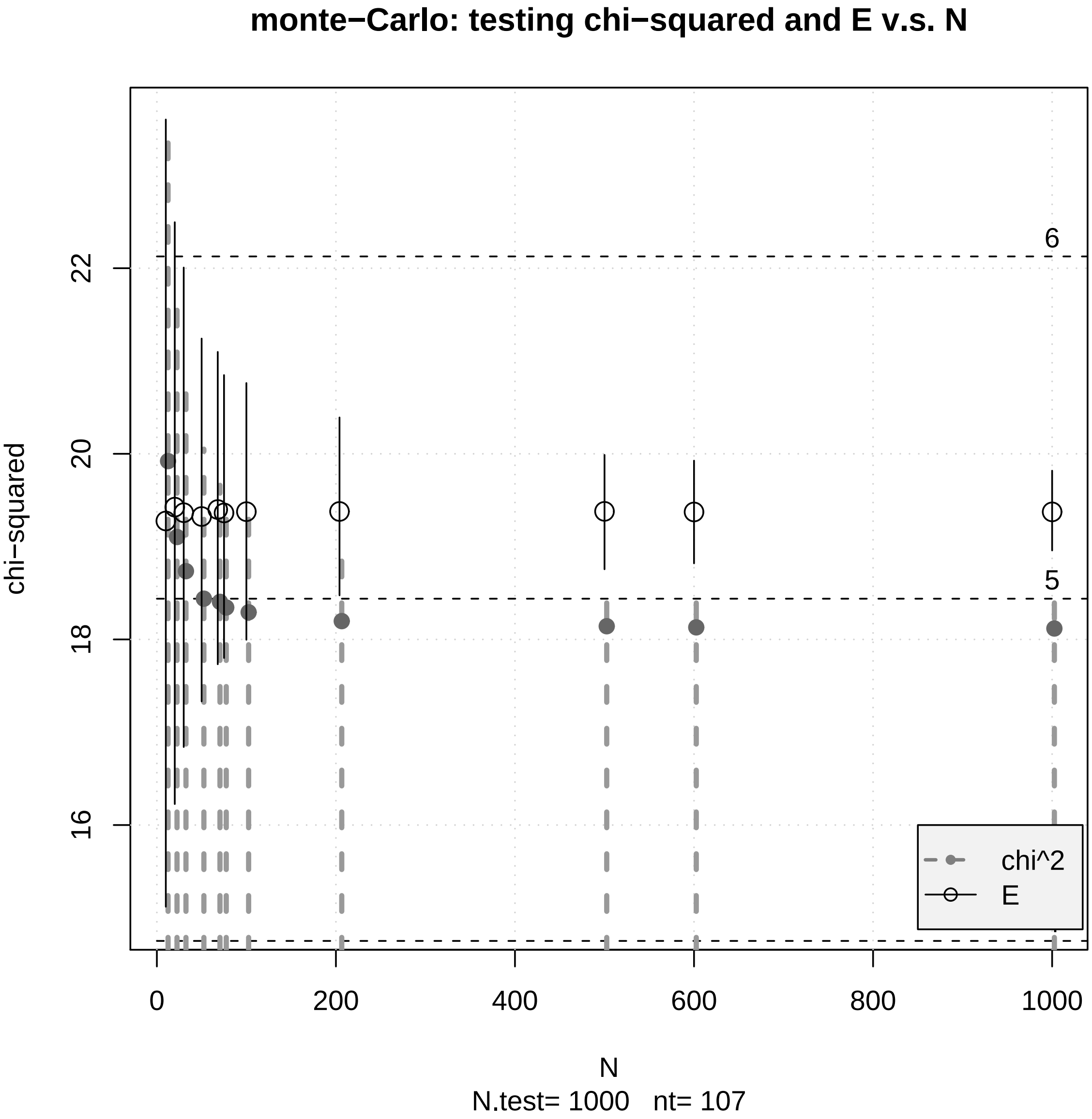
monte−Carlo: testing chi−squared and E v.s. N
chi−squared
16
18
20
22
0
200
400
600
800
1000
N
N.test= 1000 nt= 107
6
5
chi^2
E

Figure 6.

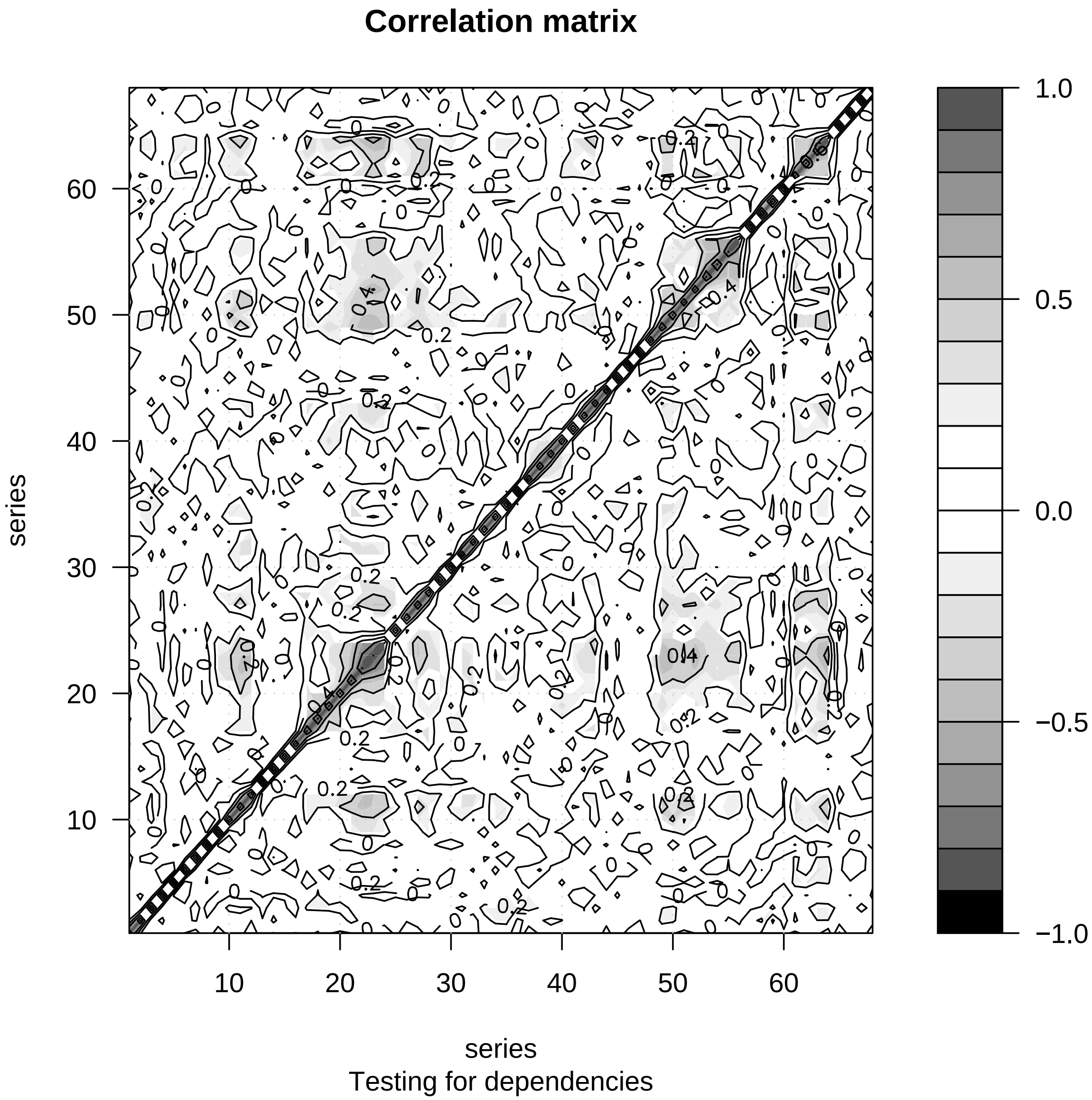
Correlation matrix
series
series
Testing for dependencies
10
20
30
40
50
60
10
20
30
40
50
60
1.0
0.5
0.0
−0.5
−1.0
0
0.2
0.4

Figure 7.

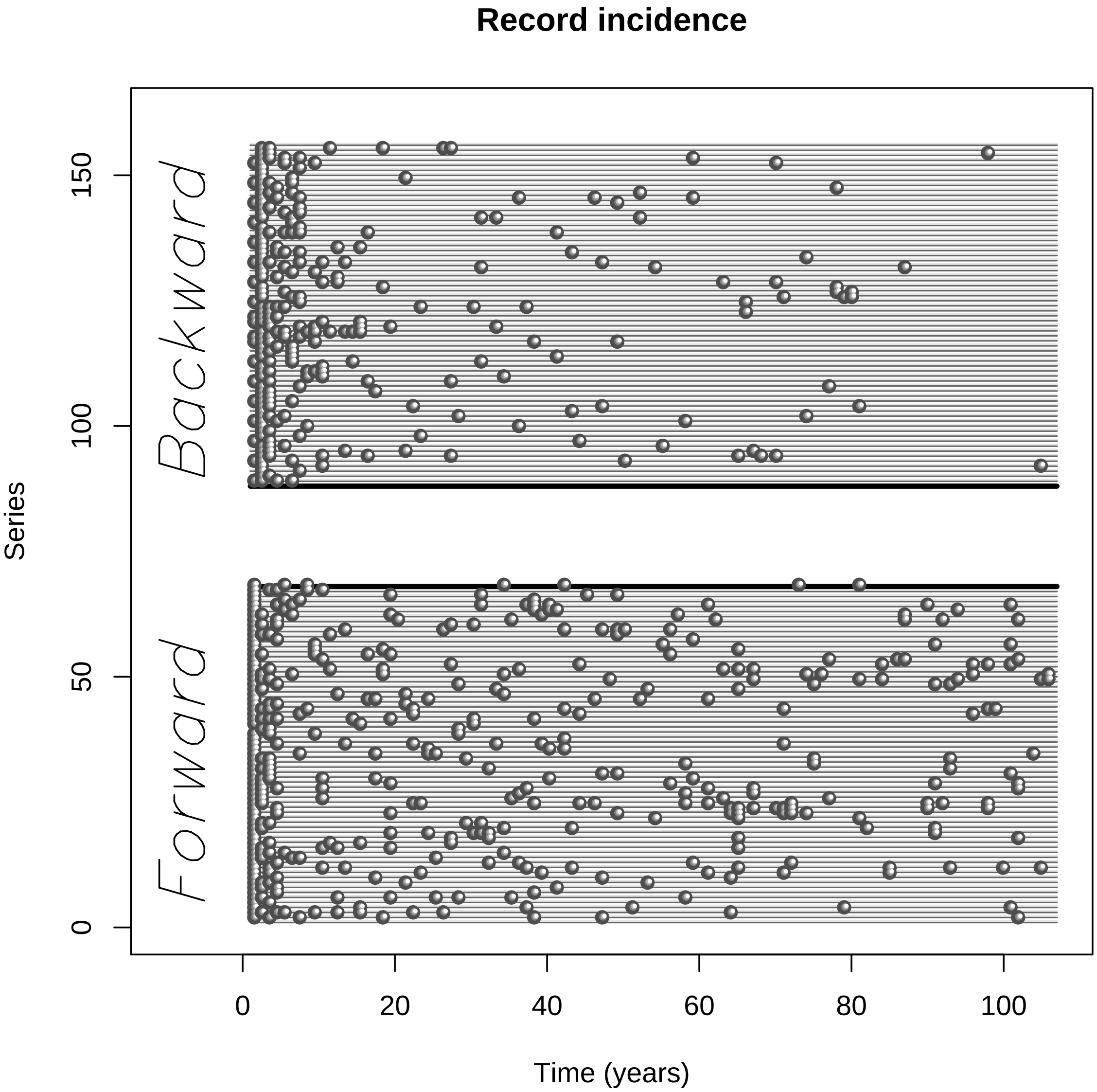
Record incidence
Series
Time (years)
0
50
100
150
0
20
40
60
80
100
Backward
Forward

Figure 8.

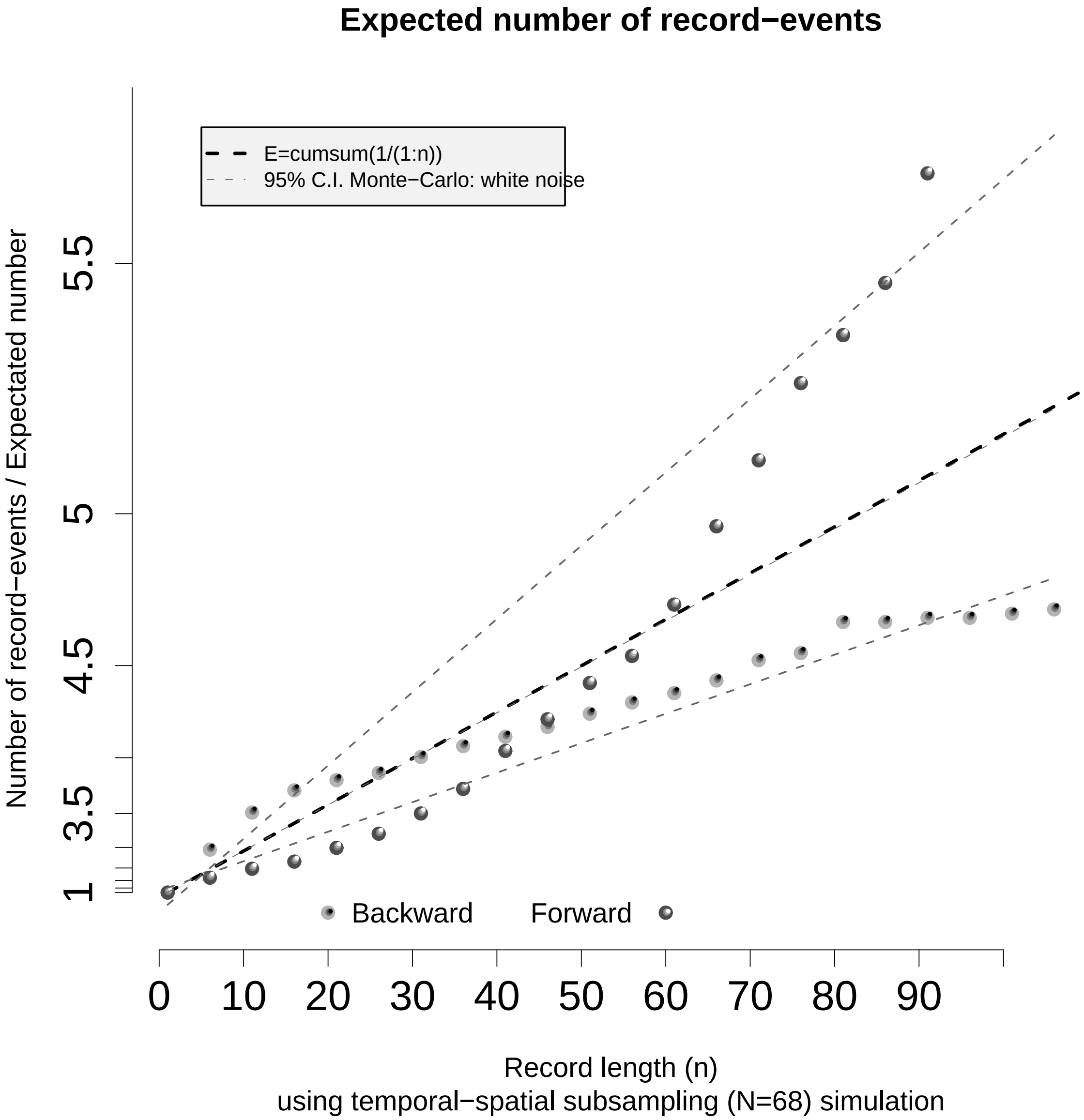
Expected number of record−events
E=cumsum(1/(1:n))
95% C.I. Monte−Carlo: white noise
Number of record−events / Expectated number
1
3.5
4.5
5
5.5
Backward
Forward
0
10
20
30
40
50
60
70
80
90
Record length (n)
using temporal−spatial subsampling (N=68) simulation

Figure 9.

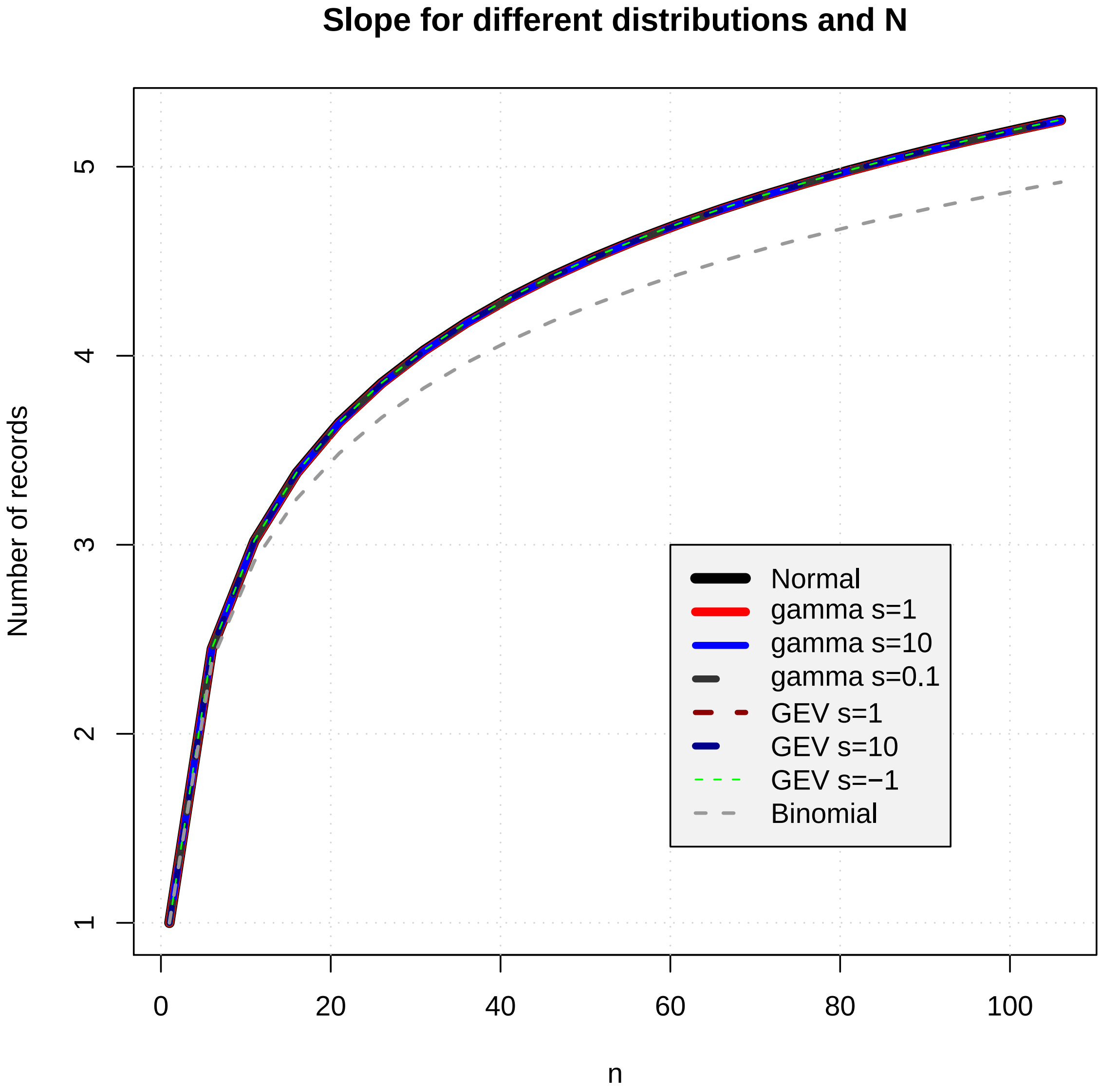
Slope for different distributions and N
Number of records
n
1
2
3
4
5
0
20
40
60
80
100
Normal
gamma s=1
gamma s=10
gamma s=0.1
GEV s=1
GEV s=10
GEV s=−1
Binomial

Figure 10.

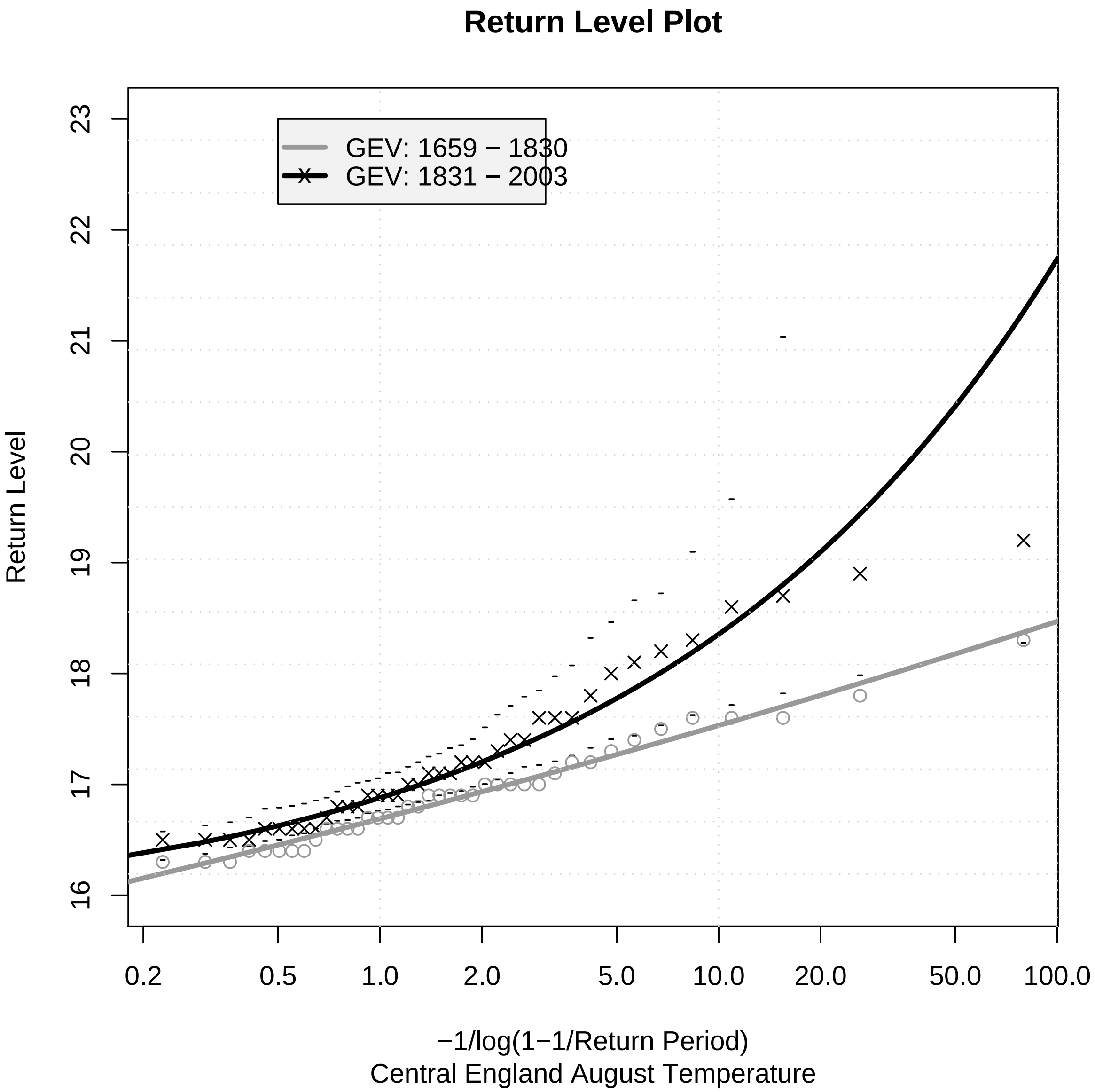
Return Level Plot
GEV: 1659 − 1830
GEV: 1831 − 2003
Return Level
16
17
18
19
20
21
22
23
0.2
0.5
1.0
2.0
5.0
10.0
20.0
50.0
100.0
−1/log(1−1/Return Period)
Central England August Temperature

Figure 11.