Journal of Statistical Software

March 2023, Volume 106, Issue 5. *doi: 10.18637/jss.v106.i05*

RecordTest: An R Package to Analyze Non-Stationarity in the Extremes Based on Record-Breaking Events

Jorge Castillo-Mateo
University of Zaragoza

Ana C. Cebrián
University of Zaragoza

Jesús Asín
University of Zaragoza

Abstract

The study of non-stationary behavior in the extremes is important to analyze data in environmental sciences, climate, finance, or sports. As an alternative to the classical extreme value theory, this analysis can be based on the study of record-breaking events. The R package **RecordTest** provides a useful framework for non-parametric analysis of non-stationary behavior in the extremes, based on the analysis of records. The underlying idea of all the non-parametric tools implemented in the package is to use the distribution of the record occurrence under series of independent and identically distributed continuous random variables, to analyze if the observed records are compatible with that behavior. Two families of tests are implemented. The first only requires the record times of the series, while the second includes more powerful tests that join the information from different types of records: upper and lower records in the forward and backward series. The package also offers functions that cover all the steps in this type of analysis such as data preparation, identification of the records, exploratory analysis, and complementary graphical tools. The applicability of the package is illustrated with the analysis of the effect of global warming on the extremes of the daily maximum temperature series in Zaragoza, Spain.

Keywords: extreme value analysis, hypothesis of stationarity, non-parametric tests, records, R.

1. Introduction

Time series data in many fields need to be examined for evidence of structural trends or shifts over time. In general, these analyses focus on the study of changes in the mean behavior, however changes in the extremes, i.e., in the tails of the distribution, are also of great interest. Extreme events typically cause large impacts because society and ecosystems are not adapted

to them, so that their study is essential. Examples of the importance of the analysis of the extremes appear in environmental sciences (large wildfires), climate (heat waves), hydrology (floods), finance (market risk), sports (limits of human capabilities), and many others.

The numerous studies about non-stationarity in the mean have been favored by the availability of easy-to-use software to compute non-parametric tests, e.g., the Mann-Kendall test (MK; Mann 1945; Kendall and Gibbons 1990). However, there does not exist similarly simple software to analyze trends, change-points, or non-stationary behavior in the extremes. Detection of this type of behaviors is complicated because extremes are rare by definition. Specific tools are required since it is difficult to link its evolution to the mean; i.e., if the magnitude of a trend in the mean is small in terms of the variability of the series, or if there are changes in the variability, the effect on the extremes might not be evident. Tools to analyze non-stationary behavior in the tails of the distribution are also required as validation tools in statistical modeling. Specific analysis of the capability to reproduce the extremes is essential in a validation analysis, since models that represent the entire distribution of a series tend to badly fit the tails and to yield important biases in extreme value statistics.

In this situation, the need for statistical tools to analyze the non-stationary behavior in the extremes and records of a series is clear. This is the aim of the R (R Core Team 2022) package **RecordTest** (Castillo-Mateo 2022b) described in this paper; the package is available from the Comprehensive R Archive Network (CRAN) at `https://CRAN.R-project.org/package=RecordTest` and on GitHub at `https://github.com/JorgeCastilloMateo/RecordTest`. The package includes non-parametric tests and graphical tools based on records to detect non-stationary behavior, such as trends and change-points, in the extremes of a series. These tools can be applied to serially correlated series with seasonal behavior, if they are previously prepared applying an approach based on splitting the series. All the tools for the data preparation are also implemented. In addition, the inferential tools in the package are able to jointly analyze $M \geq 1$ series with possibly different distributions. This property is useful to analyze split series and also in spatial analysis, to study series from different locations and obtain global conclusions over the area of interest.

Classical methods for the analysis of extreme events are the block maxima and the excesses over threshold. Both of them require to fit the tails of the distribution using parametric models such as generalized extreme value (GEV) and Pareto (GP) distributions, and Poisson processes (Coles 2001). The tools in **RecordTest** are based on a different approach, the analysis of the occurrence of record events and its comparison with the behavior of records in the classical record model (CRM; Arnold, Balakrishnan, and Nagaraja 1998). The CRM describes the distribution of the records of a series $(X_1, \ldots, X_T)^\top$ of independent and identically distributed (IID) continuous random variables (RVs). An important advantage of this approach is that it yields non-parametric tools due to the probabilistic properties of records. In particular, the fact that the distribution of the record occurrence under the CRM does not depend on the underlying distribution of the X_t's allows the use of distribution-free statistics and Monte Carlo methods. **RecordTest** includes the tests of this type proposed by Foster and Stuart (1954), Diersen and Trenkler (1996), Benestad (2003, 2004), Cebrián, Castillo-Mateo, and Asín (2022), Castillo-Mateo (2022a), and some extensions thereof proposed in this paper. The package also includes useful graphical tools based on the behavior of the record occurrence under the CRM. We found in Cebrián *et al.* (2022) that the power of the records tests is high, e.g., it is between 0.80 and almost 1.00 for a sample size of $M = 12$ series of length $T = 50$ and an alternative with a linear trend in the mean which has a magnitude of about

2.5% of the standard deviation. The MK test for the mean is more powerful when the series follow a normal distribution with a linear trend in the mean. However, records tests, which focus on the tails of the distribution, tend to be more powerful than the MK test in series with one or two-side bounded distributions or distributions with one or two light tails, such as GEV and GP, often used in extreme value analysis.

Many questions of interest in the analysis of non-stationary behavior in the tails are directly related to records, and the tools in **RecordTest** are specifically useful for this type of analysis. The study of records is common in sports (Gembris, Taylor, and Suter 2002, 2007), but it is also essential in environmental sciences, for the study of floods (Vogel, Zafirakou-Koulouris, and Matalas 2001), earthquakes (Van Aalsburg, Newman, Turcotte, and Rundle 2010; Yoder, Turcotte, and Rundle 2010), or avalanches (Shcherbakov, Davidsen, and Tiampo 2013). In the context of climate change, an important question is the effect of global warming on the number of record-breaking temperatures and precipitation events (Benestad 2003, 2004; Coumou, Robinson, and Rahmstorf 2013; Wergen and Krug 2010; Lehmann, Coumou, and Frieler 2015; Lehmann, Mempel, and Coumou 2018). The study of records is also of interest in physics, in the theory of spin-glasses or high temperature superconductors, in evolutionary biology, or in finances for the study of stock prices (see Wergen 2013, and references therein).

There are many packages for analyzing the existence of trends and non-stationary behavior using non-parametric tests, but most of them focus on the analysis of the mean of the distribution. For example, in R, **Kendall** (McLeod 2022) computes the MK test, and **modifiedmk** (Patakamuri and O'Brien 2021) implements modified versions of trend tests for serially correlated data. **trend** (Pohlert 2020) includes a great variety of tests such as Cox-Stuart, (seasonal) MK and Hirsch-Slack tests for trend detection, and Lanzante, Pettitt and Buishand tests for change-point detection. **pyMannKendall** (Hussain and Mahmud 2019) is a Python (Van Rossum *et al.* 2011) implementation of non-parametric MK trend analysis, which brings together eleven types of tests. The R package **npcp** (Kojadinovic 2023) provides non-parametric CUSUM change-point detection tests sensitive to changes in the mean, the variance, the covariance, or the autocovariance in univariate or multivariate observations, as well as a test for detecting changes in the distribution of independent block maxima. Focusing on the analysis of extremes, there are quite a few R packages, such as **evir** (Pfaff and McNeil 2018), which even includes the function `evir::records()` for extracting records. Some of them include relevant tools for testing and modeling non-stationarity. For example, **extRemes** (Gilleland and Katz 2016) and **ismev** (Heffernan and Stephenson 2018) include non-stationary models for univariate extremes, **evd** (Stephenson 2002) has some functionality for non-stationary estimation but the main emphasis is on bivariate extremes, **SpatialExtremes** (Ribatet 2022) and **texmex** (Southworth, Heffernan, and Metcalfe 2020) analyze a multivariate framework, and **NHPoisson** (Cebrián, Abaurrea, and Asín 2015) fits non-homogeneous Poisson processes for peak over threshold analysis. All these packages offer a parametric approach based on the fit of GEV, GP, and Poisson processes. A similar approach is used in the Python package **pyextremes** (Bocharov 2022), which includes methods such as block maxima, peaks over threshold and fitting of GEV and GP distributions; and in the MATLAB (The MathWorks Inc. 2022) package **NEVA** (Cheng, AghaKouchak, Gilleland, and Katz 2014), which allows the fitting of both stationary and non-stationary GEV and GP distributions in a Bayesian framework. However, as far as we know, there does not exist any statistical software package for testing a non-stationary behavior in records. The aim of **RecordTest** is to fill this gap and provide a comprehensive toolkit to assess significant deviations from

a stationary behavior in the tails of the distribution and to characterize when it occurs and which features are affected. The study of records provides a new approach from a different point of view than the block maxima and excesses over threshold approaches. Beyond the intrinsic interest of the records, there are some differences that make this approach useful for other types of analysis. An advantage of the inference tools in **RecordTest** is that no previous modeling is needed. In addition, the study of records allows the simultaneous analysis of the lower and upper tails of the distribution by including both upper and lower records in the analysis.

The outline of the paper is as follows. Section 2 introduces some basic definitions and properties of the main variables related to the record occurrence. Section 3 describes the functions and capabilities of the package, including data preparation, exploratory analysis, statistical tests and graphical tools. Section 4 illustrates the use of **RecordTest** to analyze the non-stationary behavior in the tails of the daily maximum temperature series in Zaragoza, Spain. A summary of the paper and some future work are given in Section 5.

2. Classical record model and deviations from stationarity

The statistical tools for detecting non-stationarity in the extremes implemented in **RecordTest** are based on the properties of the record occurrence in series of IID continuous RVs, i.e., the CRM. This section reviews some basic concepts and the probabilistic results that are the basis of those tools.

2.1. Variables to characterize the record occurrence

Let $(X_1, \ldots, X_T)^\top$ be a series of RVs. An observation X_t is called an upper record (or simply a record) if its value exceeds that of all previous observations, i.e., if $X_t > \max\{X_1, \ldots, X_{t-1}\}$. By virtue of this definition, X_1 is always a trivial record. Analogously, X_t is a lower record if $X_t < \min\{X_1, \ldots, X_{t-1}\}$. Let $(I_1, \ldots, I_T)^\top$ be the series of record indicator RVs defined by

$$I_t = \begin{cases} 1 & \text{if } X_t \text{ is a record,} \\ 0 & \text{otherwise.} \end{cases} \tag{1}$$

Then, the number of records is defined by the record counting process, $(N_1, \ldots, N_T)^\top$, where

$$N_t = I_1 + I_2 + \cdots + I_t, \tag{2}$$

and subsequently, the series of record times, $(L_1, \ldots, L_{N_T})^\top$, is defined by

$$L_i = \min\{t \mid N_t = i\}. \tag{3}$$

Finally, although they are not directly related to the occurrence, the series $(R_1, \ldots, R_{N_T})^\top$ of record values is defined by $R_i = X_{L_i}$.

All the tools implemented in this package assume that there are M independent series of length T available, i.e., there is a sequence $\mathbf{X} = (\mathbf{X}_1, \ldots, \mathbf{X}_M)$ of independent series where $\mathbf{X}_m = (X_{1m}, \ldots, X_{Tm})^\top$ for $m = 1, \ldots, M$. However, most of the tools can be applied even with $M = 1$. The M series can be the result of splitting the original series, or series measured at different spatial points, for example. Given $\mathbf{X}$, we define the sequences of record

indicators, $\mathbf{I} = (\mathbf{I}_1, \ldots, \mathbf{I}_M)$, and the sequences of the number of records, $\mathbf{N} = (\mathbf{N}_1, \ldots, \mathbf{N}_M)$, where $\mathbf{I}_m = (I_{1m}, \ldots, I_{Tm})^\top$ and $\mathbf{N}_m = (N_{1m}, \ldots, N_{Tm})^\top$ for $m = 1, \ldots, M$. In a similar way, obvious definitions deal with record times and record values, $\mathbf{L}$ and $\mathbf{R}$, respectively.

2.2. Probabilistic properties of the record occurrence

Arnold *et al.* (1998) present the essential topics related to the theory of records. An important result states that, in the CRM, the series $(I_1, \ldots, I_T)^\top$ consists of mutually independent RVs with *Bernoulli*(p_t) distribution where p_t, the probability of observing a new record at time t, is

$$p_t = \mathsf{P}(I_t = 1) = \frac{1}{t}, \qquad t = 1, \ldots, T. \tag{4}$$

As a consequence, the distribution of I_t, N_t, and L_i does not depend on the common continuous distribution of the X_t's. This property allows the definition of the non-parametric statistical tests and graphical tools available in **RecordTest**.

Concerning the number of records, under the CRM, N_T converges in distribution as $T \to \infty$ to a normal distribution with mean and variance

$$\mathsf{E}(N_T) = \sum_{t=1}^{T} \frac{1}{t} \qquad \text{and} \qquad \mathsf{VAR}(N_T) = \sum_{t=2}^{T} \frac{1}{t}\left(1 - \frac{1}{t}\right). \tag{5}$$

These expressions are obtained from the fact that N_T is a sum of independent Bernoulli RVs. To give some intuition about the model, note that $\mathsf{E}(N_T) \approx \log T + \gamma$ and $\mathsf{VAR}(N_T) \approx \log T + \gamma - \pi^2/6$ where γ is the Euler constant, and both expressions tend to infinity. However, under the CRM, records are not common and their occurrence becomes scarcer for larger values of T.

Turning to the notation with M independent series, if we assume that the M series have the same probabilities of record, i.e., $p_{tm} \equiv p_t$, the maximum likelihood estimates (MLEs) of these probabilities are

$$\hat{p}_t = \frac{I_{t1} + \cdots + I_{tM}}{M}, \qquad t = 1, \ldots, T, \tag{6}$$

where the sum of record indicators above follows an exact binomial distribution with M trials and probabilities of success $p_t = 1/t$. If $M = 1$, the variability associated with the estimates is large. As the number of series M increases, the estimates become more accurate and precise.

2.3. Analysis of non-stationarity in the tails of the distribution

The aim of all the inference tools in **RecordTest** is to detect a non-stationary behavior in the occurrence of records in a time series, and more generally in the upper (and lower) tail of the distribution of the series. When we refer to a non-stationary behavior we mean any deviation from the CRM in the generating system of records. The underlying idea in all the tools is to study if the occurrence of observed records is compatible with the expected behavior of the occurrence of records under the CRM, i.e., in a series of IID continuous RVs. Under the assumption that the RVs in the series are independent, any deviation from the expected behavior of records will give evidence of a change over time in the distribution, i.e., non-stationarity. In many real problems, a non-stationary behavior in a non-seasonal series is due to the existence of any type of trend.

The tests in the package are based on record probabilities. Since these probabilities are known under the CRM, the null hypothesis of all the tests is

$$H_0 : p_{tm} = 1/t, \qquad \text{for all } t = 1, \ldots, T, \text{ and } m = 1, \ldots, M, \tag{7}$$

with $p_{tm} = \mathsf{P}(I_{tm} = 1)$. Different alternative hypotheses, one-sided, two-sided or the existence of a change-point, can be of interest, and tests for each of them are proposed in **RecordTest**. The one-sided alternative claims that the probabilities of record are either greater or less than the values given by the null hypothesis. This increase or decrease may be originated by the existence of a positive or a negative trend in location, or by an increase or decrease of variability.

Two different families of tools are implemented in the package, the first only requires to know the record times of the series, while the second requires to have the entire series available. The idea of the second family, first suggested by Foster and Stuart (1954), is that more powerful tests are obtained by joining the information from different types of records instead of using only one type. More precisely, from one series $(X_1, \ldots, X_T)^\top$, four different types of records can be computed: the upper and lower records in the forward and in the backward series (or directions). The backward series $(X_T, \ldots, X_1)^\top$ is obtained by reversing the order of the terms. For example, the upper record indicators in the backward series are

$$I_t^{(BU)} = \begin{cases} 1 & \text{if } X_{T-t+1} > \max\{X_T, \ldots, X_{T-t+2}\}, \\ 0 & \text{otherwise.} \end{cases} \tag{8}$$

To distinguish what type of records a statistic or RV refers to, the corresponding superscripts F (forward), B (backward), U (upper), or L (lower) are added between brackets. Given the symmetry of the problem, under the CRM, the probability of record p_t is $1/t$ for the four types of records.

3. Functions and capabilities

RecordTest provides a framework for the analysis of non-stationary behavior in the extremes of a series using records. It covers all the steps in the analysis: data preparation, identification of the records, exploratory analysis, a wide range of statistical tests, and complementary graphical tools. This section describes the functions grouped according to their objective.

3.1. Data preparation and record variables

The main argument of the inference functions in the package is a vector $(X_1, \ldots, X_T)^\top$ or a matrix $\mathbf{X}$. If only the record times are available, they have to be transformed into a series $(X_1, \ldots, X_T)^\top$ with those record times. This transformation is implemented by the function `series_record()`, whose arguments are the record times, `L_upper` or `L_lower`, and optionally the record values, `R_upper` or `R_lower`. Note that inference based on this new series only makes sense for the tools that use the types of records that are introduced as an argument: upper, lower, or both.

All the functions allow missing values represented by `NA`. The way to deal with this is to replace them by `-Inf` for upper records and `Inf` for lower records, so they are records only if they appear at $t = 1$.

Split series

In many real problems, the original series has to be split into M subseries, for instance, to remove the seasonal behavior (Hirsch, Slack, and Smith 1982). As an example, if a daily series with annual seasonality, $(X_{1,1}, X_{1,2}, \ldots, X_{1,365}, X_{2,1}, X_{2,2}, \ldots, X_{T,365})^\top$ where X_{tm} is the variable on day m within year t, is split into 365 series, one for each day within year,

$$\begin{pmatrix} X_{1,1} & X_{1,2} & \cdots & X_{1,365} \\ X_{2,1} & X_{2,2} & \cdots & X_{2,365} \\ \vdots & \vdots & & \vdots \\ X_{T,1} & X_{T,2} & \cdots & X_{T,365} \end{pmatrix}_{T \times 365}, \tag{9}$$

the resulting subseries (columns) do not show seasonal behavior. In addition, since the consecutive observations in each subseries are now separated by one year, the serial correlation can be assumed to be zero. The distribution of the 365 series, which correspond to series at different calendar days, will not be the same due to the seasonal effect. However, the T RVs in each series, which correspond to variables measured at the same calendar day across years, may be identically distributed. Note that the null hypothesis H_0 of the inference tools in **RecordTest** is that each of the M series available are sequences of IID continuous RVs, but no assumption about the equal distribution of the M series is required. This functionality is implemented by `series_split()` that splits the series in argument `X` into `Mcols` subseries and arrange them as the matrix in (9).

Uncorrelated series

All the statistical methods implemented in the package assume that the M series under study are independent. If we have a set of dependent series, we should extract a subset of independent series from them before applying the inference tools. The function `series_uncor()` extracts a subset of uncorrelated series from the set available. This function has the arguments `test.fun`, a function to implement the desired correlation or dependence test, and `alpha`, that establishes the significance level. The default function is the standard `stats::cor.test()` with a significance level $\alpha = 0.05$, i.e., two series are considered uncorrelated if the Pearson correlation between them is not significantly different from zero at that significance level. Although zero correlation does not imply independence, this is a usual approach in most real data problems because dependence manifests itself as some level of linear correlation and testing dependence is not possible in general. However, more sophisticated functions could be used for testing dependence in other particular situations. For example, `extRemes::taildep.test()` or `evd::evind.test()` could be considered to test dependence at the extremes of the series.

We explain the algorithm with `test.fun = stats::cor.test` as an illustration. The iterative procedure to be used for selection is specified by the argument `type`. If the series have a sequential order, for example they are measured in consecutive days, the argument `type = "adjacent"` should be used and the following approach is applied: given that the kth series is in the subset, correlation between series k and $k+1$ is tested; if the correlation is not significant, series $k+1$ is included in the subset, otherwise correlation between series k and $k+2$ is tested. This step is repeated until a series $k+j$ that is not significantly correlated with series k is found. This approach does not test the pairwise correlation of all the series in the final set, but it is adequate in situations where dependence is expected between consecutive series. If we want to test the pairwise correlation of all the series we use `type = "all"`.

This procedure only keeps series that are not significantly correlated with any other selected series, which gives more evidence in favor of the correlation matrix of the chosen series being diagonal.

Breaking record ties

The CRM assumes that the X_t's are continuous RVs so that the probability of ties is zero. However, ties and *weak* records (observations equal to the current record value) can occur in a series even if the variable is continuous because the measured values are rounded. The function `series_ties()` gives the percentage of weak records in a series. It is important to know this percentage since, if it is high, the number of records will be lower than expected under the CRM, even if the series is IID (Wergen, Volovik, Redner, and Krug 2012).

If the number of weak records is high, the function `series_untie()`, which applies a simple procedure to break ties, should be used. It adds to each element in the series a uniform random value in the range $(-u/2, u/2)$, where u is the precision unit of the observations, so that weak records disappear. This procedure reproduces the fact that some of the weak records would have been records if they had not been rounded.

Backward series

The function `series_rev()` gives the backward series of the argument `X`. `X` can be a vector or a matrix, and in the last case the output is the backward series of each column.

Record variables

RecordTest includes functions to compute all the record RVs introduced in Section 2.1, given a series $(X_1, \ldots, X_T)^\top$ or a matrix $\mathbf{X}$. `I.record()` computes the record indicators $\mathbf{I}$, using a S3 method, and `N.record()` computes the observed cumulative number of records up to time t, $\mathbf{N}$. Additionally, the record times $\mathbf{L}$ and record values $\mathbf{R}$ are computed by `L.record()` and `R.record()`, respectively. The arguments of all these functions are: `X`, the vector or matrix to analyze; `record`, a character string, `"upper"` or `"lower"`, indicating the type of records to be calculated; and `weak`, a logical argument to indicate whether weak records are considered. The function `p.record()` computes the MLEs $\hat{p}_t$'s in (6).

Under the CRM, N_t follows a Poisson binomial distribution. The package includes functions, `dpoisbinom()`, `ppoisbinom()`, `qpoisbinom()`, and `rpoisbinom()`, to compute the density, distribution, and quantile function, and a random generation for the Poisson binomial distribution using the algorithm by Hong (2013).

3.2. Exploratory data analysis

Records plot. The function `records()` plots the time series $(X_1, \ldots, X_T)^\top$ and identifies the upper and lower records observed in the series; one or both directions can be specified in argument `direction = c("forward", "backward", "both")`. This plot helps to detect asymmetries between the four types of records. If we have to analyze the extreme behavior of M series, an alternative is to summarize them into a single series, calculating the mean or the maximum in each time t and apply this function. Another alternative is the following plot.

Plot of the times of record. The function `L.plot()` plots the record times in M series, (L_{im}, m), for $i = 1, \ldots, N_{Tm}$, and $m = 1, \ldots, M$. The M series are represented in the vertical axis, and the record times in each series in the horizontal axis. The display includes four panels, one for each type of records, FU, FL, BU, and BL. This plot helps to study the hypothesis of the CRM since clear differences, especially in the number of points in the second half of the plots, suggest non-stationarity in the tails.

3.3. Tests and plots for non-stationarity based on one type of records

The tools described in this section can be applied even when the only information available is the record times. Three families of tests, based on the number of records, the probabilities of record and the likelihood of the record indicators, are implemented.

Number of records

The function `N.test()` includes several tests based on statistics related to the number of records. The general expression of the statistic is

$$N^{\omega} = \sum_{m=1}^{M} \sum_{t=1}^{T} \omega_t I_{tm}, \tag{10}$$

where the ω_t's are weights given to the records according to their position in the series. The reason of using weights is that records at high values of t are less likely to occur so that, if they occur, they give more evidence against the null hypothesis H_0. Thus, the use of weights makes records at high t to increase more the value of the statistic. The weights are controlled by the argument `weights` that must be a function. Diersen and Trenkler (1996) recommend linear weights $\omega_t = t - 1$, i.e., `weights = function(t) t - 1`. Cebrián *et al.* (2022) propose a score-sum test that is a particular case of this statistic with weights $\omega_1 = 0$ and $\omega_t = t^2/(t-1)$ $(t = 2, \ldots, T)$, i.e., `weights = function(t) ifelse(t == 1, 0, t^2 / (t - 1))`. Both types of weights are asymptotically equivalent and increase the power of the test (Cebrián *et al.* 2022).

Under the null hypothesis H_0 in (7), N^{ω} is asymptotically normal as $M \to \infty$ with mean and variance

$$\mathsf{E}_0(N^{\omega}) = M \sum_{t=1}^{T} \omega_t \frac{1}{t} \qquad \text{and} \qquad \mathsf{VAR}_0(N^{\omega}) = M \sum_{t=2}^{T} \omega_t^2 \frac{1}{t}\left(1 - \frac{1}{t}\right). \tag{11}$$

When $\omega_t = 1$, N^{ω} is the raw number of records, it follows an exact Poisson binomial distribution, and it is asymptotically normal also in T. The argument `distribution` indicates the distribution to compute the p value, `"normal"` or `"poisson-binomial"`. With w_t's which are not equal to 0 or 1, only `distribution = "normal"` can be used. Alternatively, in any situation, the p value can be estimated using Monte Carlo simulations with `simulate.p.value = TRUE`. This is not often necessary since, even when N^{ω} is not asymptotically normal in T, the size based on the normal distribution is reasonably satisfactory even with $M = 1$.

Another test is based on an estimation of the variance instead of the variance under the null hypothesis H_0,

$$\tilde{N}_S^{\omega} = \frac{N^{\omega} - \mathsf{E}_0(N^{\omega})}{\sqrt{\widehat{\mathsf{VAR}}(N^{\omega})}}, \tag{12}$$

where $\widehat{\mathsf{VAR}}(N^\omega)$ is the unbiased sample variance. The resulting test can be applied when $M > 1$, and it is more robust against serial correlation. Under the null hypothesis H_0, $\tilde{N}_S^\omega$ follows an asymptotic t_{M-1} distribution, and it is implemented using `distribution = "t"`.

All the tests in `N.test()` can be applied to any of the four types of records (FU, FL, BU, or BL). The argument `record` indicates the type of records, `"upper"` or `"lower"`, to be analyzed. If backward records were desired, `series_rev(X)` has to be used as argument. Other arguments of the function are `alternative`, that indicates the alternative hypothesis, `"greater"` or `"less"`. The argument `correct` indicates whether a continuity correction should be made in the normal or t distribution approximations, which is recommended. The last arguments can be used in most of the tests in the package.

The output of `N.test()` and most tests in the package is a list of class '`htest`', which contains the components `statistic`, `parameter`, `p.value`, `alternative`, `estimate`, `method`, and `data.name`.

Plot of the number of records. The asymptotic results described above can be used to analyze graphically the null hypothesis H_0. The function `N.plot()` plots the observed values $(t, \bar{N}_t^\omega)$, where $\bar{N}_t^\omega = \sum_{m=1}^{M} \sum_{j=1}^{t} \omega_j I_{jm}/M$, together with the expected values under the null hypothesis H_0, $\mathsf{E}_0(\bar{N}_t^\omega)$. As an alternative to confidence intervals of $\mathsf{E}(\bar{N}_t^\omega)$, reference intervals (RIs) defined by the lower and upper $\alpha/2$th percentiles of the distribution of $\bar{N}_t^\omega$ under the null hypothesis H_0 are plotted, i.e.,

$$\mathsf{E}_0\left(\bar{N}_t^\omega\right) \pm z_{\alpha/2} \sqrt{\mathsf{VAR}_0\left(\bar{N}_t^\omega\right)}, \tag{13}$$

with $\mathsf{E}_0(\bar{N}_t^\omega)$ and $\mathsf{VAR}_0(\bar{N}_t^\omega)$ in (11) taking account of the average instead of the sum, and $z_{\alpha/2}$ the upper $\alpha/2$th percentile of the standard normal distribution. If the observed data follow the null distribution, they are expected to lie inside a particular RI $100(1-\alpha)\%$ of the time. It is noteworthy that the resulting RIs are not independent and the resulting bands are not reference bands at a $1-\alpha$ confidence level. However, they are useful to observe deviations from stationarity in the evolution of the number of records, and to identify the time point from which this deviation is significant.

Different weights can be specified with the argument `weights`. Several types of records can be plotted in the same graph using the argument `record`, which is a logical vector of length four (FU, FL, BU, and BL) that specifies the records to be plotted. There are two options to calculate the backward records, `backward = "T"` indicates that the backward number of records up to time t are calculated in the series observed up to time T, $(X_T, \ldots, X_1)^\top$, and `backward = "t"` in the series observed up to time t, $(X_t, \ldots, X_1)^\top$.

Probabilities of record

F test for linear regression. The function `p.regression.test()` implements a test based on the fit of a regression model to the record probabilities p_t as a function of time. Under the null hypothesis H_0, the MLEs $\hat{p}_t$'s in (6) satisfy $\mathsf{E}_0(\hat{p}_t) = 1/t$. `p.regression.test()` implements an F test to compare the null model $\mathcal{M}_0 : \mathsf{E}(t\hat{p}_t) = 1$ against a model $\mathcal{M}_1$ where the expectation is a function of t, specified by the argument `formula`. The default is $\mathcal{M}_1 : \mathsf{E}(t\hat{p}_t) = \beta_0 + \beta_1 t$, $t = 2, \ldots, T$, that is `y ~ x`. More complicated time trends can be used, e.g., a quadratic trend with `formula = y ~ poly(x, degree = 2)`.

Given that the response $\hat{p}_t$ has a non-constant variance even under the null hypothesis H_0, the regression model has to be fitted using weighted least squares with weights $1/\mathsf{VAR}_0(t\hat{p}_t)$. A simulation study showed that the size of this test is satisfactory for $M > 10$. As in previous test functions, the p value can be estimated using Monte Carlo simulations with `simulate.p.value = TRUE`.

Plot of the probabilities of record. The function `p.plot()` represents the points $(t, t\hat{p}_t)$ for $t = 1, \ldots, T$, and the fitted linear model described in `p.regression.test()`, to evaluate the goodness of fit of the model. The fitted regression line can be replaced, e.g., by a locally estimated scatterplot smoothing (LOESS), using `smooth.method = stats::loess`. RIs based on the binomial distribution of $M\hat{p}_t$ when the null hypothesis H_0 is true are added to the plot if `conf.int = TRUE`. These RIs are independent and they are helpful to detect any substantial departure from the CRM at particular times t. The plot can be displayed on different scales: using `plot = "2"`, $\hat{p}_t$ is represented against t; and using `plot = "3"`, a logarithmic scale is used in both axes.

χ^2 goodness-of-fit test. The function `p.chisq.test()` implements a Pearson's χ^2 test,

$$M \sum_{t=2}^{T} \left(\hat{p}_t - \mathsf{E}_0\left(\hat{p}_t\right)\right)^2 \left(\frac{1}{\mathsf{E}_0\left(\hat{p}_t\right)} + \frac{1}{1 - \mathsf{E}_0\left(\hat{p}_t\right)}\right), \tag{14}$$

comparing the observed and expected probabilities of record and no-record (see Benestad 2003, 2004, for more details). Under the null hypothesis H_0, the distribution of the statistic is asymptotically χ^2_{T-1}.

The size of the test is not appropriate for small M. In those cases, the function gives a warning message and it is convenient to estimate the p value using Monte Carlo simulations with `simulate.p.value = TRUE`.

Likelihood ratio and score tests

The functions `lr.test()` and `score.test()` compute the family of tests by Cebrián *et al.* (2022) to study the null hypothesis H_0 based on the likelihood and the score function of the record indicators $\mathbf{I}$. The main difference with the previous tests is that they can be used to test both one-sided and two-sided alternatives. Although a different statistic has to be used in each case, we only have to indicate the adequate alternative in the argument `alternative` which can take values `"two.sided"`, `"greater"`, or `"less"`.

The default alternative in the two functions is that all the $T \times M$ probabilities p_{tm} may be different, with any restriction. Using `probabilities = "equal"`, both statistics are modified to study a particular case, that the probabilities in the M series are equal, although possibly different to $1/t$. According to Cebrián *et al.* (2022), those tests are less powerful than the tests for the general alternative, even if the probabilities in the M series are equal. In general, the tests in `score.test()` are more powerful and are recommended.

3.4. Tests and plots for non-stationarity based on different types of records

The power of the tests based on one type of records is improved by joining the information from the four types of records. Two families of this type of test are implemented in **RecordTest**. In both cases, the first step is to obtain the statistic described in the previous section for each type of records, and then build a joint statistic, or combine the p values of the resulting tests.

Tests based on joint statistics

The function `foster.test()` implements the tests based on joint statistics developed by Foster and Stuart (1954), with the possibility of adding weights, as suggested by Diersen and Trenkler (1996). Seven different statistics can be selected with `statistic = c("D", "d", "S", "s", "U", "L", "W")`. All the tests, apart from `"S"` and `"s"`, study the stationarity against the alternative of a trend in the mean. The statistics `"d"`, `"D"` and `"W"` analyze non-stationary behavior in both tails using two or four types of records,

$$d^{\omega} = N^{\omega,(U)} - N^{\omega,(L)} = \sum_{m=1}^{M}\sum_{t=1}^{T}\omega_t \left(I_{tm}^{(U)} - I_{tm}^{(L)}\right), \tag{15}$$

$$D^{\omega} = d^{\omega,(F)} - d^{\omega,(B)} = \sum_{m=1}^{M}\sum_{t=1}^{T}\omega_t \left(I_{tm}^{(FU)} - I_{tm}^{(FL)} - I_{tm}^{(BU)} + I_{tm}^{(BL)}\right), \tag{16}$$

$$W^{\omega} = \sum_{t=1}^{T}\omega_t \left(I_{tm}^{(FU)} + I_{tm}^{(BL)}\right). \tag{17}$$

The statistics in `"U"`, $U^{\omega} = N^{\omega,(FU)} - N^{\omega,(BU)}$, and `"L"`, $L^{\omega} = N^{\omega,(BL)} - N^{\omega,(FL)}$, only use the two types of upper or lower records and they are useful to detect trends only in the right or the left tail, respectively. The statistics in `"S"` and `"s"` study the existence of a trend in variation, and they are defined as $s^{\omega} = N^{\omega,(U)} + N^{\omega,(L)}$ and $S^{\omega} = s^{\omega,(F)} - s^{\omega,(B)}$. The statistics without weights are asymptotically normal in both T and M, and the weighted statistics only in M; although their size based on the normal distribution is satisfactory even with $M = 1$.

As explained in the definition of $\tilde{N}_S^{\omega}$ in (12), more robust statistics against serial correlation are obtained when the above statistics are standardized in mean and sample variance. The new statistics have an asymptotic t_{M-1} distribution and they are computed with `distribution = "t"`.

Plot of the Foster-Stuart statistics. The function `foster.plot()` plots the observed values of one of the statistics defined above, selected with `statistic`, obtained with the series observed up to time t, for every $t = 1, \ldots, T$. The plot also includes the expected values and RIs based on the normal approximation of the distribution of the statistic under the null hypothesis H_0. It is useful to detect the time t for which the stationarity hypothesis fails.

Global test. The function `global.test()` also computes a joint statistic, but it combines the statistic selected in `FUN`, say $\mathcal{X}$, which must be one of the two-sided tests in `p.regression.test()`, `p.chisq.test()`, `lr.test()`, or `score.test()`. By default, the global statistic $\mathcal{X}^G = \mathcal{X}^{(FU)} + \mathcal{X}^{(FL)} + \mathcal{X}^{(BU)} + \mathcal{X}^{(BL)}$ is used, but some terms can be omitted using argument `record`. The distribution of $\mathcal{X}^G$ is unknown, but the p value is estimated by Monte Carlo simulations.

Tests based on combined p values

The functions `brown.method()` and `fisher.method()` compute tests that combine the p values resulting from applying the tests with asymptotic normal distribution to the different types of records.

The function `fisher.method()` implements the general Fisher's method to combine the p values from any set of independent tests with the same null hypothesis; the vector of p values is the only argument. It can be applied to any test but, in the context of records, it is used to join the p values of the records that are asymptotically independent, i.e., FU and FL, BU and BL, FU and BL, or FL and BU.

The function `brown.method()` implements an algorithm to combine the p values of the tests in `N.test()` to any subset of the four types of records, selected with `record`. Since the p values are dependent, the algorithm is based on the Brown's method: the combined p values,

$$-2\left(\log(pv^{(FU)}) + \log(pv^{(FL)}) + \log(pv^{(BU)}) + \log(pv^{(BL)})\right), \tag{18}$$

follow a $c\chi^2_f$ distribution with scale parameter c and degrees of freedom f that depend on the covariances of the p values. In general, this test is more powerful than the previous ones and the (seasonal) MK test when the series follow a GP or some types of GEV distributions with a linear drift in location (see Cebrián *et al.* 2022, for the details).

3.5. Tests for change-point detection

The function `change.point()` implements a family of tests to study the null hypothesis H_0 against the alternative hypothesis of an unknown change-point t_0, i.e.,

$$H_1 : p_{tm} = 1/t, \quad t = 1, \ldots, t_0 \quad \text{and} \quad p_{tm} \neq 1/t, \quad t = t_0 + 1, \ldots, T, \tag{19}$$

for $m = 1, \ldots, M$. Note that these tests aim to detect the beginning of the non-stationary behavior in the tails, not in the mean. The test statistic given by Castillo-Mateo (2022a) is

$$K^{\omega} = \max_{1 \leq t \leq T} \left| \frac{N_t^{\omega} - \mathsf{E}_0(N_t^{\omega})}{\sqrt{\mathsf{VAR}_0(N_T^{\omega})}} - \frac{\mathsf{VAR}_0(N_t^{\omega})}{\mathsf{VAR}_0(N_T^{\omega})} \frac{N_T^{\omega} - \mathsf{E}_0(N_T^{\omega})}{\sqrt{\mathsf{VAR}_0(N_T^{\omega})}} \right|, \tag{20}$$

where $N_t^{\omega} = \sum_{m=1}^{M} \sum_{j=1}^{t} \omega_j I_{jm}$, and the estimated change-point $\hat{t}_0$ is the value t where K^{ω} attains its maximum. K^{ω} is asymptotically Kolmogorov distributed as $T \to \infty$ if $\omega_t = 1$; otherwise, the p value has to be estimated by Monte Carlo simulations using `simulate.p.value = TRUE`. Weights equal to 1 or proportional to the inverse of the standard deviation of I_t are recommended. The test has been defined in terms of the number of upper or lower records N_t^{ω}, but it can also be defined in terms of $d_t^{\omega} = N_t^{\omega,(U)} - N_t^{\omega,(L)}$ or $s_t^{\omega} = N_t^{\omega,(U)} + N_t^{\omega,(L)}$, depending on `record = c("upper", "lower", "d", "s")`.

4. An example: Daily maximum temperature in Zaragoza

This section illustrates how package **RecordTest** can be used to analyze the effect of global warming on the records and extremes of a daily maximum temperature series, the series in Zaragoza, Spain. It is shown how the functions in the package cover all the steps of the analysis: data preparation, exploratory analysis, and inference to study the non-stationary behavior of the extremes and to identify the time, the periods of the year, and the features where the non-stationary behavior appears.

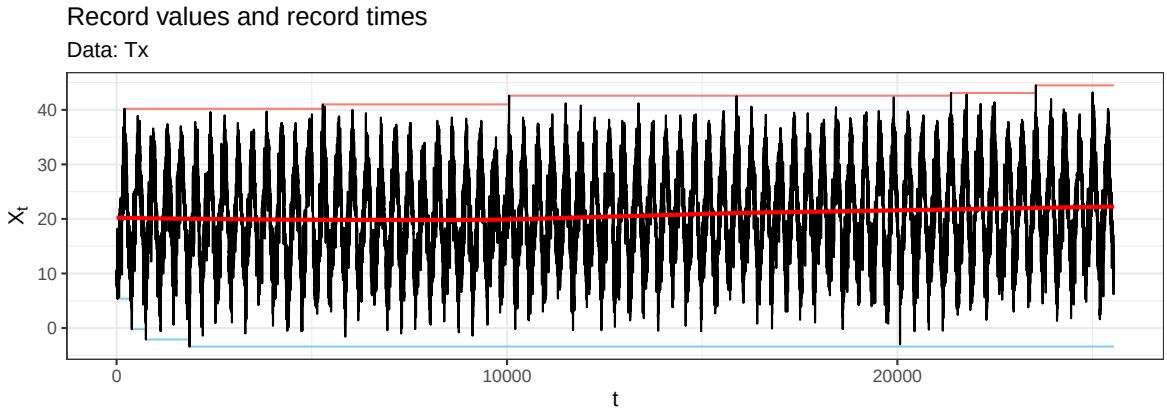

Figure 1: Daily maximum temperature at Zaragoza, Spain (1951–2020). LOESS (solid red), and upper (red) and lower (blue) records.

4.1. Dataset

The '`data.frame`' `TX_Zaragoza` included in **RecordTest** contains two columns: `DATE`, the dates in '`Date`' format, spanning from `"1951-01-01"` to `"2020-12-31"`; and `TX`, the daily maximum temperature series at Zaragoza Airport (Spain), in tenths of degree Celsius (°C). The dataset has been downloaded from the European Climate Assessment & Dataset (ECA&D; Klein Tank *et al.* 2002) and modified by eliminating all the observations from the 29th of February. This is because when the series is split, these days would yield a four-year time series that is difficult to join to the analysis of the other yearly time series. The series with the 29th of February is also available as `TX_Zaragoza29F`. The series has three missing observations indicated by `NA` corresponding to `"1951-03-31"`, `"1965-01-04"`, and `"1965-10-05"`. The dataset can be accessed after loading the package:

```
R> library("RecordTest")
R> Tx <- TX_Zaragoza$TX / 10
```

4.2. Data preparation and exploratory analysis

Most daily temperature series present a clear seasonal component and a high serial correlation. That also is the case for the Zaragoza series, which can easily be seen by plotting the series using the function `records()` (see Figure 1). The output of this and all the plot functions in **RecordTest** are '`ggplot`' objects. Consequently, the plots can be easily improved using **ggplot2** (Wickham 2016) functions; an example of how to add a LOESS is shown in the following chunk,

```
R> records(Tx) +
+    ggplot2::geom_smooth(formula = y ~ x, method = stats::loess,
+      mapping = ggplot2::aes(y = Tx), se = FALSE, col = "red")
```

The plot reveals the seasonal behavior and a weak long-term time trend in the mean, summarized by the LOESS. The upper and lower records in the forward direction are also plotted, but their behavior is difficult to be analyzed due to the seasonality of the series.

Since the series shows annual seasonality, we use the function `series_split()` with `Mcols = 365` to apply the split procedure described in Section 3.1, and it returns 365 subseries of length $T = 70$. Then, we use `series_uncor()` to extract from them a subset of uncorrelated subseries, and it returns $M = 76$ subseries of length $T = 70$. We do not change the default option `test.fun = stats::cor.test` given that dependence in temperature data is usually linked to correlation. The default significance level is `alpha = 0.05`. Multiple testing corrections are not implemented in this example since, in order to apply the tests, it is more important to guarantee that the selected series are uncorrelated, while the problem of rejecting series which are uncorrelated is less problematic. The indexes of the subset of the selected series are shown with the argument `return.value = "indexes"`. As we can see, a separation between 2 and 14 days yields uncorrelated series.

```
R> TxZ365 <- series_split(Tx, Mcols = 365)
R> TxZ <- series_uncor(TxZ365)
R> dim(TxZ)

[1] 70 76

R> series_uncor(TxZ365, return.value = "indexes")

 [1]   1   5  13  16  20  25  29  34  40  48  54  62  76  80  83  88  95
[18] 100 104 108 111 118 122 126 133 138 143 152 156 162 169 176 179 188
[35] 194 197 199 202 205 207 209 212 215 220 223 226 229 238 241 245 251
[52] 257 261 267 270 274 277 286 292 298 306 308 312 314 317 321 325 333
[69] 337 339 342 345 350 357 361 364
```

The temperature series is rounded to the tenth of °C. Given that the rounding is small with respect to the value of the temperature, the assumption of continuous RVs is reasonable. The existence of a high percentage of ties may affect the validity of the tests, that are developed for continuous RVs, so it is advisable to apply `series_ties()` to check this percentage. The output is a '`list`' with three elements: (i) `number`, a vector with the total number of records, the number of "strong" and weak records, and the expected number of records under IID series, (ii) `percentage`, the percentage of record ties out of the total number of records, and (iii) `percentage.position`, the percentage of record ties by time t, e.g., 4.76% of records were ties at $t = 5$. `NaN`s indicate that there were no records of any type, e.g., at $t = 19$.

```
R> lapply(series_ties(TxZ, record = "upper"), round, digits = 2)

$number
        Total        Strong          Weak          Expected under IID
        391.0         375.0          16.0                       367.3

$percentage
[1] 4.09

$percentage.position
     1      2      3      4      5      6      7      8      9     10
```

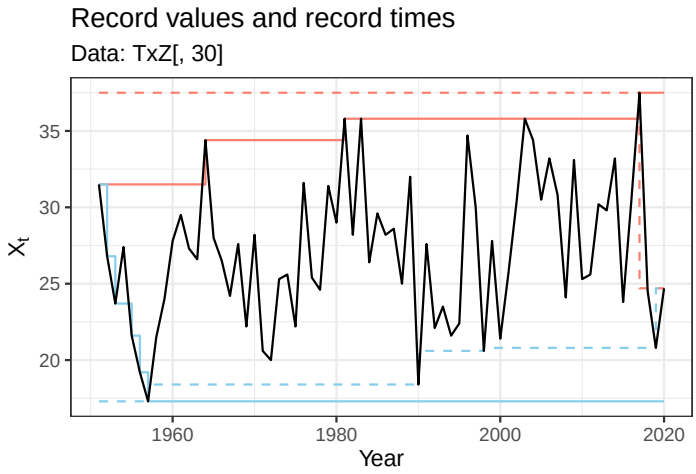

Figure 2: Daily maximum temperature of the 11th of June at Zaragoza, Spain (1951–2020). Upper (red) and lower (blue) records in the forward (solid) and backward (dashed) directions.

```
  0.00   2.44   0.00   0.00   4.76   0.00   0.00   0.00   0.00   0.00
    11     12     13     14     15     16     17     18     19     20
 11.11  14.29  20.00   0.00  16.67   0.00  50.00   0.00    NaN   0.00
    21     22     23     24     25     26     27     28     29     30
  0.00   0.00   0.00    NaN   0.00   0.00   0.00   0.00   0.00   0.00
    31     32     33     34     35     36     37     38     39     40
 20.00   0.00  33.33  50.00  40.00 100.00   0.00   0.00   0.00   0.00
    41     42     43     44     45     46     47     48     49     50
  0.00   0.00   0.00   0.00  16.67   0.00    NaN   0.00   0.00    NaN
    51     52     53     54     55     56     57     58     59     60
  0.00   0.00  25.00   0.00   0.00   0.00    NaN    NaN   0.00   0.00
    61     62     63     64     65     66     67     68     69     70
  0.00   0.00   0.00  20.00   0.00   0.00   0.00   0.00   0.00   0.00
```

```
R> lapply(series_ties(TxZ, record = "lower"), round, digits = 2)
```

Since the percentage of ties is about 4.09% for upper records and 3.74% for lower records (output omitted), it does not seem to be necessary to apply the function `series_untie()` to break the ties.

In series without seasonality, the analysis of records is easier. As an example, Figure 2 shows the plot obtained with the chunk below, which includes the upper and lower records in the forward and backward directions of the temperature measured on the 11th of June (30th column in `TxZ`).

```
R> records(TxZ[, 30], direction = "both") +
+    ggplot2::scale_x_continuous(name = "Year", breaks = c(10, 30, 50, 70),
+      labels = c("1960", "1980", "2000", "2020"))
```

The plot shows evidence of an increasing trend, since no lower records occur after 7 time units in the forward series, and the last upper record occurs at time point 3 in the backward series.

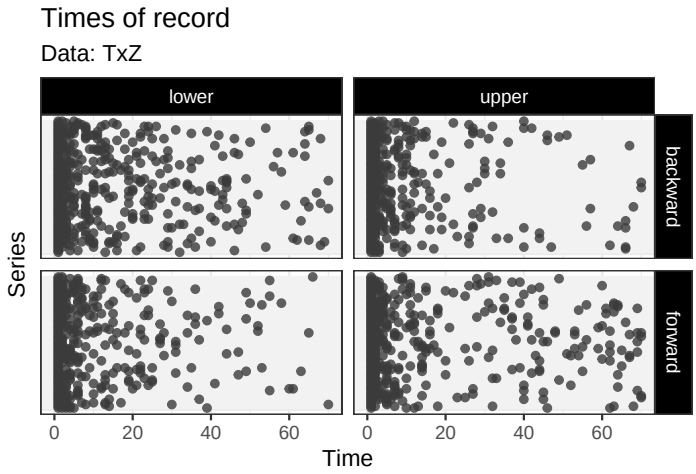

Figure 3: Plot of the times of record for the 76 uncorrelated subseries at Zaragoza, Spain (1951–2020).

The function `L.plot()` is used to summarize the times of the four types of records in each of the 76 uncorrelated subseries by means of the plot of the times of record,

```
R> L.plot(TxZ)
```

Figure 3 shows that as time evolves there are less FL and BU records than FU and BL records. This suggests non-stationary behavior since the pattern in the four plots should be similar in IID series. This effect is difficult to be observed in only one subseries; however, when the 76 subseries are plotted together the evidence is clearer.

4.3. Inference tools to study non-stationarity in the extremes

Tests to detect deviations from stationarity

The effect of global warming on the extremes of the temperature series may appear in the upper, the lower, or in both tails. We are interested in analyzing those hypotheses both jointly and individually, but in all cases by means of the null hypothesis H_0 in (7). The $M = 76$ subseries available correspond to different days within a year so they are not identically distributed. Consequently, under the alternative hypothesis the probabilities of record p_{tm} may be different in the M subseries.

Analysis of one tail. To analyze the behavior of the upper tail, we study the upper records. In the context of global warming, the alternative hypothesis of interest is

$$H_1 : p_{tm}^{(FU)} > 1/t, \qquad \text{for at least one } t = 1, \ldots, T, \text{ and } m = 1, \ldots, M. \tag{21}$$

This alternative hypothesis is quite general since it includes the existence of a monotonous positive trend in the mean, but also other types of non-stationarity, such as some non-monotonous trends. To test this hypothesis, we implement the weighted test in `N.test()` using the simple linear weights $\omega_t = t - 1$ and default arguments,

```
R> N.test(TxZ, weights = function(t) t - 1)

        Test on the weighted number of upper records with weights = t - 1

data:  TxZ
Z = 3.3138, p-value = 0.0004602
alternative hypothesis: true 'N' is greater than 4952.704
sample estimates:
         N          E        VAR
  6335.000   4952.704 173877.950
```

The hypothesis of stationarity is rejected at any usual significance level α. `N.test()` calculates the value of the statistic `N`, its standardized version `Z` with a continuity correction by default, and its expected value and variance under the null hypothesis H_0, `E` and `VAR`, respectively. The Monte Carlo p value obtained with the argument `simulate.p.value = TRUE` is very similar (not shown), indicating that the normal approximation is good.

To analyze the behavior of the lower tail, we also use `N.test()`. Under the alternative hypothesis of a positive trend in the mean, the probability of lower records is less than under the null hypothesis H_0, and we have to use the arguments `record = "lower"` and `alternative = "less"`. A significant behavior against stationarity is also observed in the lower tail, since the following yields a p value equal to `0.001044` (output omitted).

```
R> N.test(TxZ, weights = function(t) t - 1, record = "lower",
+    alternative = "less")
```

Given that the complete series is available, we can add more information to the study of one tail using the backward series and the more powerful tests implemented in `foster.test()`. Here, we apply the statistic U^{ω} defined in Section 3.4, based on the forward and backward upper records. The alternative for a positive trend in the mean must be the default `alternative = "greater"`. The p value is lower than that obtained with `N.test()` so more evidence to reject the null hypothesis H_0 is found,

```
R> foster.test(TxZ, weights = function(t) t - 1, statistic = "U")

        Forward - backward upper records test with weights = t - 1

data:  TxZ
Z = 4.0641, p-value = 2.411e-05
alternative hypothesis: true 'statistic' is greater than 0
sample estimates:
statistic         E       VAR
   3110.0       0.0  585579.8
```

Analysis of both tails. To carry out a joint analysis of both tails, we use the D^{ω} statistic in (16) based on the four types of records. The alternative for a positive trend is again the default `alternative = "greater"`,

```
R> foster.test(TxZ, weights = function(t) t - 1, statistic = "D")

        Foster-Stuart D-statistic test with weights = t - 1

data:  TxZ
Z = 5.1889, p-value = 1.058e-07
alternative hypothesis: true 'statistic' is greater than 0
sample estimates:
statistic         E       VAR
     5692         0   1203318
```

The more robust version of the statistic against serial correlation as defined in (12) but for D^{ω} can be calculated using the argument `distribution = "t"`. The null hypothesis H_0 is also rejected at any usual significance level α,

```
R> foster.test(TxZ, weights = function(t) t - 1, statistic = "D",
+    distribution = "t")

        Foster-Stuart D-statistic test with weights = t - 1

data:  TxZ
t = 5.263, df = 75, p-value = 6.507e-07
alternative hypothesis: true 't' is greater than 0
```

Another option to carry out a joint analysis is to apply Brown's method using the default option that combines the p values of `N.test()` for the four types of records. Although it is the default option, we specify the alternative hypothesis for the four types of records with `alternative`, as an example of use,

```
R> brown.method(TxZ, weights = function(t) t - 1,
+    alternative = c("FU" = "greater", "FL" = "less", "BU" = "less",
+    "BL" = "greater"))

    Brown's method on the weighted number of records with weights = t - 1

data:  TxZ
X-squared = 38.669, df = 4.7592, c = 1.6810, p-value = 2.088e-07
```

It is noteworthy that the tests joining the information of the four types of records give the lowest p values, on the order of 10^{-7}. They lead to conclude, at any usual significance level α, that the probabilities of record are greater for FU and BL records, and less for FL and BU records, than expected under the CRM. This gives evidence of non-stationarity in the occurrence of records in the subseries and, consequently, the existence of an increasing positive trend in daily maximum temperature that affects the occurrence of extremes.

Graphical tools to detect deviations from stationarity

We have formally tested the existence of a significant non-stationary behavior both in the upper and lower tails of temperature. Our next aim is to characterize that behavior using graphical tools, to identify when it appears, which features are affected, etc.

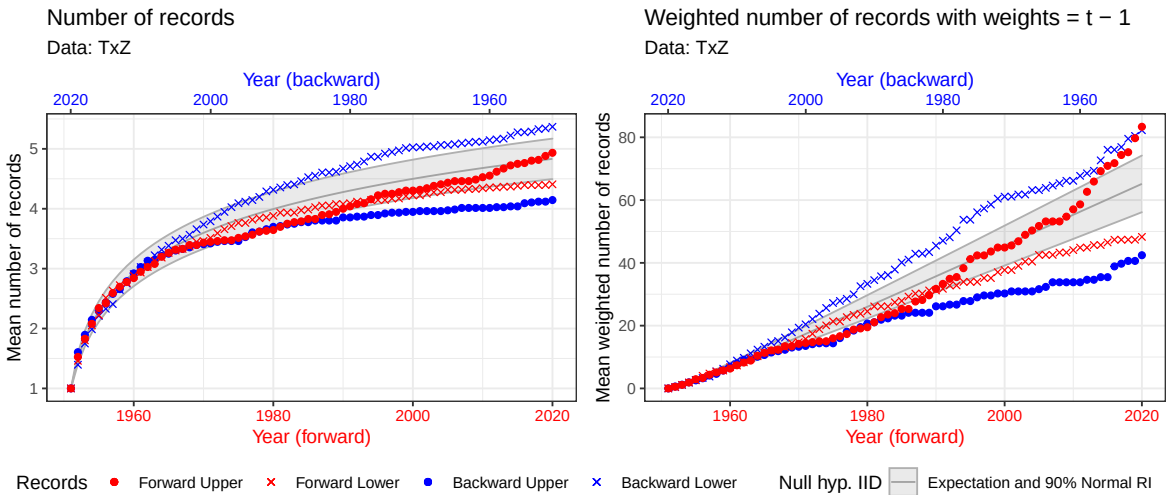

Figure 4: Plot of the number of records for the 76 uncorrelated subseries at Zaragoza, Spain (1951–2020). Expected values and 90% RIs (mean, 5th and 95th percentiles of the distribution of $\bar{N}_t^{\omega}$ under the null hypothesis H_0) for the four types of records (gray shaded area). Left: Unweighted statistics. Right: Weighted statistics with linear weights.

First, we analyze the behavior of the four types of records with the plot of the number of records using `N.plot()`. To facilitate the comparison, all types of records are displayed together using the default argument `record`.

```
R> N.plot(TxZ)
R> N.plot(TxZ, weights = function(t) t - 1)
```

The chunk above gives the plots by default, but Figure 4 is obtained adding **ggplot2** functions to draw the time axis for the forward and backward series; the complete code is available in the supplementary material. The left plot shows that the number of FU records in the 80s is slightly lower than expected in a stationary series. From that point onward, the number of records increases until the end, although it does not become significantly high. FL records have a stationary behavior up to the 90s, but its number starts to be lower than expected thereafter. Backward records show more clear deviations of stationarity, and this suggests that non-stationary behavior is stronger in the last part of the observed period. Both types of backward records are outside the RIs from the first 30 observations, which correspond to the period spanning from 1991 to 2020. The non-stationary behavior observed in the four types of records is the behavior expected in a series with a positive trend in the mean. The right plot is obtained using linear weights to give more importance to the occurrence of records in high values of t, where the probability of record is lower. It shows that the use of weights leads to clearer evidence of non-stationarity: the deviation of the forward records is now significant and the deviation in the backward series is detected even earlier.

To analyze both tails jointly, we combine the information of the four types of records in one signal. We can show a plot equivalent to the previous one based on the Foster-Stuart D^{ω} statistic in (16) with `foster.plot()`, whose expected value under the null hypothesis H_0 is zero. We do not show the **ggplot2** functions for simplicity,

```
R> foster.plot(TxZ, weights = function(t) t - 1)
```

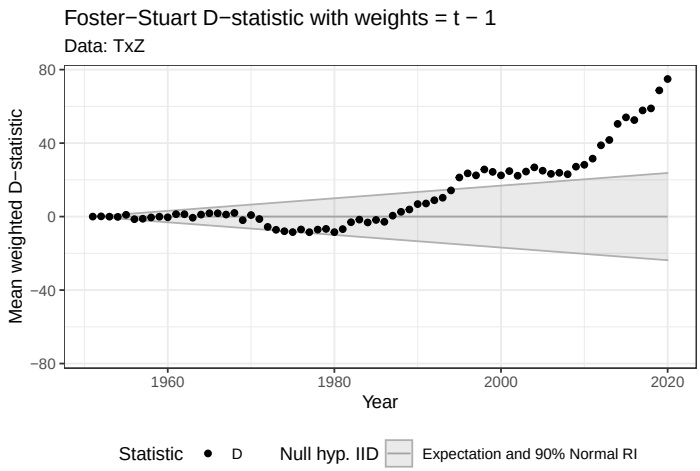

Figure 5: Plot of the mean value of the Foster-Stuart D^{ω} statistic with linear weights for the 76 uncorrelated subseries at Zaragoza, Spain (1951–2020), observed up to time t, $t = 1, \ldots, T$. Expected values and 90% RIs under the null hypothesis H_0 (gray shaded area).

Figure 5 shows a significant non-stationary behavior from the latest 90s onward and the statistic shows a strong increasing trend starting around 2010.

Another approach to characterize non-stationarity is the analysis of the probabilities of records. We plot $t\hat{p}_t$ against t using `p.plot()` with the default argument `plot = "1"`, for FU records. Under the null hypothesis H_0, the fitted regression to those points should be a horizontal line, but different alternatives may be fitted; here, a quadratic trend is considered,

```
R> p.plot(TxZ, record = c("FU" = 1, "FL" = 0, "BU" = 0, "BL" = 0),
+    smooth.formula = y ~ poly(x, degree = 2))
```

The top plot in Figure 6 shows that the fitted curve is clearly different from zero and many values $t\hat{p}_t$ from the late 90s onward are outside the RIs. This plot helps us to identify the years where the probability of record is much higher than expected. To characterize the lower tail, the FL and BL records are shown in the same plot but with different colors using `point.col`,

```
R> p.plot(TxZ, record = c("FU" = 0, "FL" = 1, "BU" = 0, "BL" = 1),
+    point.col = c("FU" = NA, "FL" = "blue", "BU" = NA, "BL" = "red"))
```

The bottom plot in Figure 6 shows that FL records are less informative in the case of an increasing trend. In effect, in that case, FL probabilities tend to decrease, but given that they are always bounded by zero, points lower than the low interval bound cannot appear.

To formally check if the deviation from the CRM is significant, we apply the F test in `p.regression.test()` to study $\mathsf{E}(t\hat{p}_t) = 1$. Since the previous function `p.plot()` suggests a quadratic trend as an alternative, we use

```
R> p.regression.test(TxZ, formula = y ~ poly(x, degree = 2))

        Regression test on the upper records probabilities
```

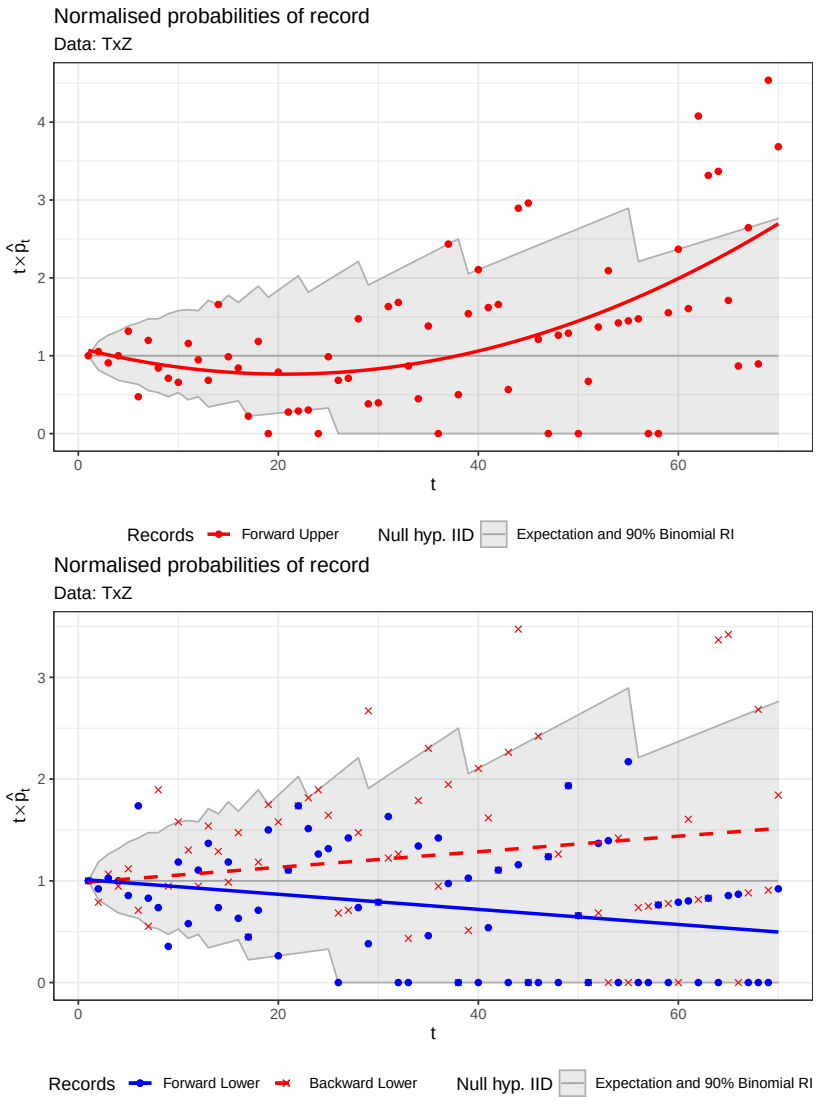

Figure 6: Plot of the normalized probabilities of record for the 76 uncorrelated subseries at Zaragoza, Spain (1951–2020). Expected values and 90% RIs for $t \times \hat{p}_t$ (gray shaded area). Top: FU records with quadratic time trend. Bottom: FL and BL records with linear time trend.

```
data:  TxZ
F = 9.0496, df1 = 3, df2 = 66, p-value = 4.225e-05
alternative hypothesis: two-sided for record probabilities
null values:
         (Intercept) poly(x, degree = 2)1 poly(x, degree = 2)2
                   1                    0                    0
sample estimates:
         (Intercept) poly(x, degree = 2)1 poly(x, degree = 2)2
            1.265065             4.028521             2.337520
```

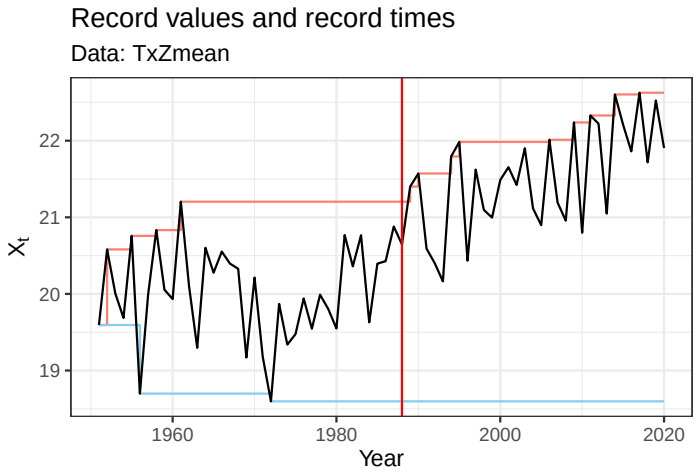

Figure 7: Annual mean temperature at Zaragoza, Spain (1951–2020). Change-point estimate (vertical solid red), and upper (red) and lower (blue) records.

We also apply the F test to the BL records using `series_rev()` to calculate the backward series, and the argument `records = "lower"` (output omitted),

```
R> p.regression.test(series_rev(TxZ), record = "lower")
```

The resulting p value is `0.039`. Both tests suggest that the p_{tm}'s are significantly different from $1/t$. The evidence in BL records is not so strong but still significant, as it is observed in the number of points outside the RIs in the bottom plot in Figure 6.

Tests for change-point detection

Once we have found evidence of a trend in the tails of the temperature distribution, our aim is to identify the time point where this trend starts. First, we consider a series without seasonal behavior, the annual mean temperature. Figure 7 shows the annual mean temperature at Zaragoza, together with the change-point estimate, and its upper and lower records, resulting from

```
R> TxZmean <- rowMeans(TxZ365, na.rm = TRUE)
R> records(TxZmean) +
+    ggplot2::scale_x_continuous(name = "Year", breaks = c(10, 30, 50, 70),
+      labels = c("1960", "1980", "2000", "2020")) +
+    ggplot2::geom_vline(xintercept = change.point(TxZmean)$estimate,
+      color = "red")
```

It seems reasonable to check the null hypothesis H_0 against the alternative hypothesis H_1 in (19) using a change-point test based on upper records without weights,

```
R> change.point(TxZmean)

        Records test for single changepoint detection
```

```
data:  TxZmean
Kolmogorov = 3.7425, p-value = 1.366e-12
alternative hypothesis: two.sided
sample estimates:
probable changepoint time
                      38
```

The p value of order 10^{-12} yields to reject the null hypothesis H_0 at any usual significance level α, and the estimated change-point $\hat{t}_0 = 38$ corresponds to the year 1988.

The change-point test can also be applied to the 76 uncorrelated subseries as given in the chunk below (output omitted). The estimated change-point on a daily scale is $\hat{t}_0 = 36$ (1986), and the p value `0.0003547` is significant at any usual significance level α.

```
R> change.point(TxZ)
```

5. Summary and future work

The study of non-stationary behavior in the extremes and the tails of a distribution is important in data analysis in many fields, such as environmental sciences, climate, finance, or sports. However, most of the available software packages to analyze non-stationarity focuses on the study of the mean. As far as we know, the R package **RecordTest** is the only available software package for the analysis of record-breaking events. In addition, the use of records provides a useful general framework for a fully non-parametric analysis of non-stationary behavior in the extremes. The underlying idea of all the inference tools implemented in the package is to use the distribution of the record occurrences under the classical record model, and study if the observed records are compatible with that behavior.

The package offers functions that cover all the steps in this type of analysis. This includes functions to prepare the data, obtaining a set of uncorrelated series with no seasonal behavior from the original series, identify the variables for characterizing the record occurrence, and implement graphical tools for exploratory analysis. The main functionality of the package is the implementation of all the tests to detect non-stationarity based on records currently available in the literature, and complementary graphical tools. The null hypothesis H_0 of all the tests is that the series are sequences of IID continuous RVs, expressed in terms of the probabilities of record, i.e., $p_t = 1/t$. There are two main families implemented, the first one can be applied even when the only information available are the times of record, and this includes tests based on the number of records, the probabilities of record, and the likelihood of the record indicators. The second family requires to know the entire series but it includes the most powerful tests. The underlying idea is to combine the information from four types of records, the upper and lower records in the forward and backward series, using joint statistics or joint p values. Different alternative hypothesis, one-sided, two-sided or even the existence of a change-point can be studied with the wide range of available tests.

The applicability of the package to analyze real data is illustrated with the analysis of the effect of global warming on the extremes of the daily maximum temperature in Zaragoza, Spain. The availability of the tools implemented in **RecordTest** will favor the realization of studies for analyzing records and non-stationarity in the extremes in many fields.

Future work will focus on the implementation of permutation tests, although this approach requires further development in the literature. This procedure will capture the dependence between the M series, so the tests will not require independent series. It will be especially useful to jointly analyze series with spatio-temporal dependence.

Acknowledgments

This work has been partially supported by the Grant PID2020-116873GB-I00 funded by MCIN/AEI/10.13039/501100011033; the Research Group E46_20R: Modelos Estocásticos funded by Gobierno de Aragón; and Jorge Castillo-Mateo was supported by the Doctoral Scholarship ORDEN CUS/581/2020 funded by Gobierno de Aragón. The authors thank the ECA&D project for providing the data.

References

Arnold BC, Balakrishnan N, Nagaraja HN (1998). *Records*. Wiley Series in Probability and Statistics. John Wiley & Sons, New York. doi:10.1002/9781118150412.

Benestad RE (2003). "How Often Can We Expect a Record Event?" *Climate Research*, **25**(1), 3–13. doi:10.3354/cr025003.

Benestad RE (2004). "Record-Values, Nonstationarity Tests and Extreme Value Distributions." *Global and Planetary Change*, **44**(1–4), 11–26. doi:10.1016/j.gloplacha.2004.06.002.

Bocharov G (2022). ***pyextremes**: Extreme Value Analysis (EVA) in Python*. Python package version 2.2.5, URL https://github.com/georgebv/pyextremes.

Castillo-Mateo J (2022a). "Distribution-Free Changepoint Detection Tests Based on the Breaking of Records." *Environmental and Ecological Statistics*, **29**(3), 655–676. doi:10.1007/s10651-022-00539-2.

Castillo-Mateo J (2022b). ***RecordTest**: Inference Tools in Time Series Based on Record Statistics*. R package version 2.1.1, URL https://CRAN.R-project.org/package=RecordTest.

Cebrián AC, Abaurrea J, Asín J (2015). "**NHPoisson**: An R Package for Fitting and Validating Nonhomogeneous Poisson Processes." *Journal of Statistical Software*, **64**(6), 1–25. doi:10.18637/jss.v064.i06.

Cebrián AC, Castillo-Mateo J, Asín J (2022). "Record Tests to Detect Non-Stationarity in the Tails with an Application to Climate Change." *Stochastic Environmental Research and Risk Assessment*, **36**(2), 313–330. doi:10.1007/s00477-021-02122-w.

Cheng L, AghaKouchak A, Gilleland E, Katz RW (2014). "Non-Stationary Extreme Value Analysis in a Changing Climate." *Climatic Change*, **127**(2), 353–369. doi:10.1007/s10584-014-1254-5.

Coles S (2001). *An Introduction to Statistical Modeling of Extreme Values*. Springer Series in Statistics. Springer-Verlag, London. doi:10.1007/978-1-4471-3675-0.

Coumou D, Robinson A, Rahmstorf S (2013). "Global Increase in Record-Breaking Monthly-Mean Temperatures." *Climatic Change*, **118**(3–4), 771–782. doi:10.1007/s10584-012-0668-1.

Diersen J, Trenkler G (1996). "Records Tests for Trend in Location." *Statistics*, **28**(1), 1–12. doi:10.1080/02331889708802543.

Foster FG, Stuart A (1954). "Distribution-Free Tests in Time-Series Based on the Breaking of Records." *Journal of the Royal Statistical Society B*, **16**(1), 1–22. doi:10.1111/j.2517-6161.1954.tb00143.x.

Gembris D, Taylor JG, Suter D (2002). "Trends and Random Fluctuations in Athletics." *Nature*, **417**(6888), 506. doi:10.1038/417506a.

Gembris D, Taylor JG, Suter D (2007). "Evolution of Athletic Records: Statistical Effects versus Real Improvements." *Journal of Applied Statistics*, **34**(5), 529–545. doi:10.1080/02664760701234850.

Gilleland E, Katz RW (2016). "**extRemes** 2.0: An Extreme Value Analysis Package in R." *Journal of Statistical Software*, **72**(8), 1–39. doi:10.18637/jss.v072.i08.

Heffernan JE, Stephenson AG (2018). ***ismev**: An Introduction to Statistical Modeling of Extreme Values*. R package version 1.42, URL https://CRAN.R-project.org/package=ismev.

Hirsch RM, Slack JR, Smith RA (1982). "Techniques of Trend Analysis for Monthly Water Quality Data." *Water Resources Research*, **18**(1), 107–121. doi:10.1029/wr018i001p00107.

Hong Y (2013). "On Computing the Distribution Function for the Poisson Binomial Distribution." *Computational Statistics & Data Analysis*, **59**, 41–51. doi:10.1016/j.csda.2012.10.006.

Hussain M, Mahmud I (2019). "**pyMannKendall**: A Python Package for Non Parametric Mann Kendall Family of Trend Tests." *Journal of Open Source Software*, **4**(39), 1556. doi:10.21105/joss.01556.

Kendall MG, Gibbons JD (1990). *Rank Correlation Methods*. A Charles Griffin Title, 5th edition. Oxford University Press, London.

Klein Tank AMG, Wijngaard JB, Können GP, Böhm R, Demarée G, Gocheva A, Mileta M, Pashiardis S, Hejkrlik L, Kern-Hansen C, Heino R, Bessemoulin P, Müller-Westermeier G, Tzanakou M, Szalai S, Pálsdóttir T, Fitzgerald D, Rubin S, Capaldo M, Maugeri M, Leitass A, Bukantis A, Aberfeld R, van Engelen AFV, Forland E, Mietus M, Coelho F, Mares C, Razuvaev V, Nieplova E, Cegnar T, Antonio López J, Dahlström B, Moberg A, Kirchhofer W, Ceylan A, Pachaliuk O, Alexander LV, Petrovic P (2002). "Daily Dataset of 20th-Century Surface Air Temperature and Precipitation Series for the European Climate Assessment." *International Journal of Climatology*, **22**(12), 1441–1453. doi:10.1002/joc.773.

Kojadinovic I (2023). ***npcp**: Some Nonparametric CUSUM Tests for Change-Point Detection in Possibly Multivariate Observations*. R package version 0.2-5, URL https://CRAN.R-project.org/package=npcp.

Lehmann J, Coumou D, Frieler K (2015). "Increased Record-Breaking Precipitation Events under Global Warming." *Climatic Change*, **132**(4), 501–515. doi:10.1007/s10584-015-1434-y.

Lehmann J, Mempel F, Coumou D (2018). "Increased Occurrence of Record-Wet and Record-Dry Months Reflect Changes in Mean Rainfall." *Geophysical Research Letters*, **45**(24), 13468–13476. doi:10.1029/2018gl079439.

Mann HB (1945). "Nonparametric Tests against Trend." *Econometrica*, **13**(3), 245–259. doi:10.2307/1907187.

McLeod AI (2022). ***Kendall**: Kendall Rank Correlation and Mann-Kendall Trend Test*. R package version 2.2.1, URL https://CRAN.R-project.org/package=Kendall.

Patakamuri SK, O'Brien N (2021). ***modifiedmk**: Modified Versions of Mann Kendall and Spearman's Rho Trend Tests*. R package version 1.6, URL https://CRAN.R-project.org/package=modifiedmk.

Pfaff B, McNeil A (2018). ***evir**: Extreme Values in R*. R package version 1.7-4, URL https://CRAN.R-project.org/package=evir.

Pohlert T (2020). ***trend**: Non-Parametric Trend Tests and Change-Point Detection*. R package version 1.1.4, URL https://CRAN.R-project.org/package=trend.

R Core Team (2022). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria. URL https://www.R-project.org/.

Ribatet M (2022). ***SpatialExtremes**: Modelling Spatial Extremes*. R package version 2.1-0, URL https://CRAN.R-project.org/package=SpatialExtremes.

Shcherbakov R, Davidsen J, Tiampo KF (2013). "Record-Breaking Avalanches in Driven Threshold Systems." *Physical Review E*, **87**(5), 052811. doi:10.1103/physreve.87.052811.

Southworth H, Heffernan JE, Metcalfe PD (2020). ***texmex**: Statistical Modelling of Extreme Values*. R package version 2.4.8, URL https://CRAN.R-project.org/package=texmex.

Stephenson AG (2002). "**evd**: Extreme Value Distributions." *R News*, **2**(2), 31–32. URL https://journal.R-project.org/articles/RN-2002-015/.

The MathWorks Inc (2022). *MATLAB – The Language of Technical Computing, Version R2022b*. Natick. URL https://www.mathworks.com/products/matlab/.

Van Aalsburg J, Newman WI, Turcotte DL, Rundle JB (2010). "Record-Breaking Earthquakes." *Bulletin of the Seismological Society of America*, **100**(4), 1800–1805. doi:10.1785/0120090015.

Van Rossum G, *et al.* (2011). *Python Programming Language*. URL https://www.python.org/.

Vogel RM, Zafirakou-Koulouris A, Matalas NC (2001). "Frequency of Record-Breaking Floods in the United States." *Water Resources Research*, **37**(6), 1723–1731. doi:10.1029/2001wr900019.

Wergen G (2013). "Records in Stochastic Processes – Theory and Applications." *Journal of Physics A: Mathematical and Theoretical*, **46**(22), 223001. doi:10.1088/1751-8113/46/22/223001.

Wergen G, Krug J (2010). "Record-Breaking Temperatures Reveal a Warming Climate." *EPL (Europhysics Letters)*, **92**(3), 30008. doi:10.1209/0295-5075/92/30008.

Wergen G, Volovik D, Redner S, Krug J (2012). "Rounding Effects in Record Statistics." *Physical Review Letters*, **109**(16), 164102. doi:10.1103/physrevlett.109.164102.

Wickham H (2016). ***ggplot2**: Elegant Graphics for Data Analysis*. 2nd edition. Springer-Verlag, New York. doi:10.1007/978-3-319-24277-4.

Yoder MR, Turcotte DL, Rundle JB (2010). "Record-Breaking Earthquake Intervals in a Global Catalogue and an Aftershock Sequence." *Nonlinear Processes in Geophysics*, **17**(2), 169–176. doi:10.5194/npg-17-169-2010.

Affiliation:

Jorge Castillo-Mateo, Ana C. Cebrián, Jesús Asín
Department of Statistical Methods
University of Zaragoza
Pedro Cerbuna 12
50009 Zaragoza, Spain
E-mail: jorgecm@unizar.es, acebrian@unizar.es, jasin@unizar.es

Journal of Statistical Software https://www.jstatsoft.org/
published by the Foundation for Open Access Statistics https://www.foastat.org/
March 2023, Volume 106, Issue 5 *Submitted:* 2021-11-24
doi:10.18637/jss.v106.i05 *Accepted:* 2022-12-13