Computers & Geosciences 32 (2006) 141–144
www.elsevier.com/locate/cageo

Short note

CLIM-X-DETECT: A Fortran 90 program for robust detection of extremes against a time-dependent background in climate records☆

Manfred Mudelsee[a,b,*]

[a]Climate Risk Analysis, Wasserweg 2, 06114 Halle (S), Germany
[b]Institute of Meteorology, University of Leipzig, Stephanstr. 3, 04103 Leipzig, Germany

Received 8 October 2004; received in revised form 6 May 2005; accepted 6 May 2005

1. Introduction

Extreme events have become one focus of climate research (Meehl et al., 2000). This type of natural hazard has potential societal effects (McCarthy et al., 2001). Detecting extremes in records of past climates is important for projecting future risks of such events.

The task of detection may be difficult when background signal and climatic variability are time-dependent, which is common in many types of records and climate archives. For example, in the application analyzed here, we search for peaks in the sulfate time series from a Greenland ice core. The peaks (sulfate extremes) come from volcanic eruptions, whereas the background and variability reflect changing oceanic and other input. A reliable detection requires robustness: number and size of the extremes, which are indeed assumed to be in the data, should have little influence on background and variability estimates (Lanzante, 1996). See for an extensive review of theory and methods for the statistics of extremes the following books: Beirlant et al. (2004), Coles (2001), Embrechts et al. (1997), and Reiss and Thomas (1997).

CLIM-X-DETECT is a Fortran 90 program that estimates robustly time-dependent background by running median smoothing (Härdle and Steiger, 1995) and time-dependent variability by the running median of absolute distances to the median (MAD) (Tukey, 1977). It uses an efficient updating scheme of the window data that allows processing of data sizes in the order of many thousands at PC level. CLIM-X-DETECT is a separate part of the XTREND package for analyzing trends in the occurrence of extreme climatic events (Mudelsee et al., 2003). CLIM-X-DETECT can readily be applied in other fields where extremes are searched in time series data, such as geophysics, astronomy, or econometrics.

2. Background and variability estimation

Hampel (1985) considers robust outlier or extremes detection on a data sample (no time dependence): if

$$x(i) > \mathrm{MED}(x) + z\,\mathrm{MAD}(x), \qquad (1)$$

$i = 1, \ldots, n$, then $x(i)$ may be considered as an extreme value. $x(i)$ are the data (size n); $\mathrm{MED}(x)$ and $\mathrm{MAD}(x)$ are the median and MAD of the sample, respectively. (That is, $\mathrm{MAD}(x) =$ median of $\{|x(i) - \mathrm{MED}(x)|; i = 1, \ldots, n\}$.) z is a threshold to be selected. z too small produces a liberal detection (too many extreme events), z too large a conservative detection (too few events); Hampel (1985) used extensive Monte Carlo simulation

☆Code available at http://www.iamg.org/CGEditor/index.htm or http://www.climate-risk-analysis.com.

*Corresponding author at: Institute of Meteorology, University of Leipzig, Stephanstr. 3, 04103 Leipzig, Germany. Tel.: +49 345 5323860.

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

0098-3004/$ - see front matter
doi:10.1016/j.cageo.2005.05.010

experiments and concluded that $z = 3.5$ yields a good compromise in many cases of distributions of x.

The extension to a time series, $\{x(i), t(i); i = 1, \ldots, n\}$, time $t(i)$ strictly monotonically increasing, is straightforward: if

$$x(i) > \mathrm{MED}_{j=i-k}^{j=i+k}(x(j)) + z\,\mathrm{MAD}_{j=i-k}^{j=i+k}(x(j)), \tag{2}$$

$i = k+1, \ldots, n-k$, then $x(i)$ is the detected extreme at time $t(i)$. $\mathrm{MED}_{j=i-k}^{j=i+k}(x(j))$ is the running median ($2k+1$ window points, with $k \leqslant (n-1)/2$), and serves as estimate of time-dependent background; $\mathrm{MAD}_{j=i-k}^{j=i+k}(x(j))$ is the median of absolute distances to the running median, and serves as estimate of time-dependent variability. When also the size of an extreme is important, it might be useful to consider the scaled quantity

$$x^*(i) = [x(i) - \mathrm{MED}_{j=i-k}^{j=i+k}(x(j))]/\mathrm{MAD}_{j=i-k}^{j=i+k}(x(j)), \tag{3}$$

$i = k+1, \ldots, n-k$.

Cross-validation can be used to solve the smoothing problem (choice of k). We use two such criterions: L_1-norm (Marron, 1987) and median criterion (Zheng and Yang, 1998):

$$CV_1(k) = \left[\sum_{i=1}^{n} |x(i) - \mathrm{MED}_{j=i-k, j\neq i}^{j=i+k}(x(j))|\right] \bigg/ n, \tag{4}$$

$$CV_{\mathrm{m}}(k) = \mathrm{median}\{|x(i) - \mathrm{MED}_{j=i-k, j\neq i}^{j=i+k}(x(j))|\}, \tag{5}$$

where $\mathrm{MED}_{j=i-k, j\neq i}^{j=i+k}(x(j))$ is the delete-one background estimate. The cross-validation functions measure the average performance of the delete-one estimate to predict the observation $x(i)$. Optimal k values minimize $CV_1(k)$ or $CV_{\mathrm{m}}(k)$. (One leaves out the point $j = i$ to exclude the trivial solution $k = 0$.)

Using two criterions allows a more balanced look at the background and variability properties of the data than using just one criterion. Note that also local minima of the cross-validation functions may indicate some relevant structure in the data (Marron, 1988).

Further extensions of the method, considered to be beyond the scope of this note, are: adaptive smoothing (k time-dependent) and automatic safeguarding against autocorrelation effects. In autocorrelated time series, an extreme event at time $t(i)$ may be the mere result of some previous event, for example, at $t(i-1)$. In such cases, the cross-validation functions suggest under-smoothing ($k = 1$ or 2). It might be advisable then to resample the original time series at a lower time resolution to reduce autocorrelation effects.

3. Optimal median smoothing

Running median smoothing and running MAD determination are computationally expensive because these procedures require some sorting operations on the window points.

Härdle and Steiger (1995) devised an algorithm for running median smoothing based on a double-heap order of window points, which is optimal in the sense that no faster other algorithm apparently exists. The double-heap is an array $\{y(l); l = -k, \ldots, k\}$ that stores the window points $\{x(j); j = i-k, \ldots, i+k\}$ as follows (see Härdle and Steiger (1995), page 260).

1. $y(0) =$ median of $\{x(j); j = i-k, \ldots, i+k\}$;
2. $\max[y(-2m), y(-2m-1)] \leqslant y(-m); \quad m \leqslant (k-1)/2$;
3. $\min[y(2m+1), y(2m)] \geqslant y(m); \quad m \leqslant (k-1)/2$.

Fig. 1 in the paper by Härdle and Steiger (1995) illustrates this structure. Härdle and Steiger (1995; p. 261) explain the updating of the double-heap order when the window runs: 'When $x(i-k)$ is removed from the window the data structure has an empty place ... To update, we propagate the 'hole' to the apex of the relevant heap [i.e., $y(1)$ or $y(-1)$] ... Inserting the new value $x(i+k+1)$ into the data structure is analogous.' Running median calculation thus does not require more expensive algorithms like those for bringing all window points into numerical order; usage of Härdle and Steiger's (1995) double-heap algorithm is sufficient.

CLIM-X-DETECT (subroutine runmed) uses two adaptations of the sort_heap algorithm (Press et al., 1996) for updating the double-heap. First, generation of the initial double-heap structure when the window starts is performed by sorting (routine indexx (Press et al., 1996)). Second, CLIM-X-DETECT's subroutine runmed performs also the delete-one running median calculations for the cross-validation (Eqs. (4) and (5)) as well as the running MAD calculations (see subroutine background, which calls runmed).

4. Examples

Fig. 1A shows robust extremes detection and background estimation for an artificial time series. The threshold selection ($z = 4.0$) successfully achieves detection of all pre-defined extremes. ($z = 3.5$ might be slightly too liberal for these data.) The cross-validated window width ($k = 21$) captures the pre-defined sinusoidal background variation. On the other hand, the non-robust method using running mean and standard deviation (Fig. 1B) fails to detect nine of the 18 pre-defined extremes.

The GISP2 ice core from Greenland is an archive of environmental conditions in the North Atlantic region over the past hundred thousand years (Hammer et al., 1997). The GISP2 sulfate time series, here analyzed using the early Holocene data (Zielinski et al., 1994),

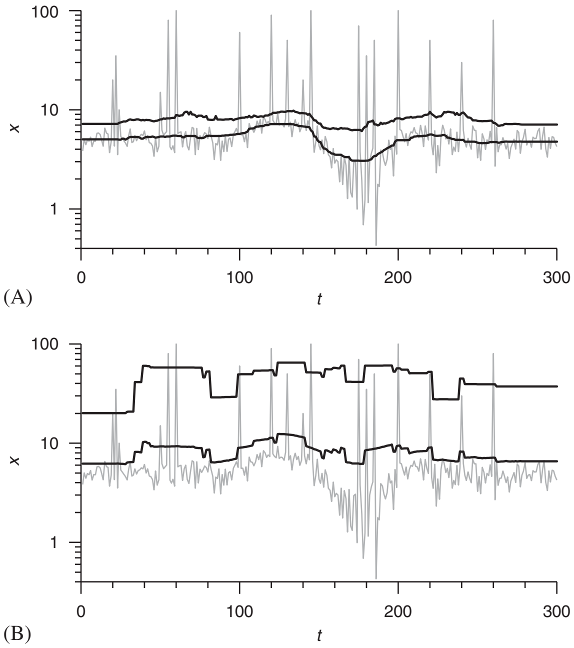

Fig. 1. Detection of extremes in artificial data. The time series $t(i), x(i)$ (gray lines in A and B) was generated as follows. $t(i) = i, i = 1, \ldots, 300; x(i) = 5 + \varepsilon(i) + 3s(i)$, where $\varepsilon(i) \sim \mathrm{N}(0;1)$, and $s(j) = \sin(2\pi(t(j) - 100)/100)$, $j = 100, \ldots, 200$, $s(j) = 0$, elsewhere. Subsequently, 18 predefined extremes were set as $x(20) = 20$, $x(22) = 35$, $x(24) = 10$, $x(50) = 15$, $x(55) = 80$, $x(60) = 100$, $x(100) = 60$, $x(120) = 90$, $x(130) = 50$, $x(140) = 20$, $x(145) = 100$, $x(175) = 70$, $x(180) = 35$, $x(185) = 50$, $x(200) = 100$, $x(220) = 50$, $x(240) = 30$, and $x(260) = 80$. (A) Robust detection using running median as estimate of time-dependent background (lower heavy line) and running median $+z \cdot \mathrm{MAD}$ as time-dependent threshold (upper heavy line). Cross-validated number of window points (Eq. (5)) is $k = 21$. $z = 4.0$ achieves detection of the 18 extremes; $z = 3.5$ finds two additional, spurious events, at $t = 115$ and $t = 124$. (B) Non-robust detection: running mean ($k = 21$) as background estimate (lower heavy line) and running mean $+2.7$ standard deviations as detection threshold (upper heavy line). (A normal distribution with standard deviation unity has an MAD of around $0.67 \simeq 2.7/4.0$.) Only nine events are detected. The non-robust method is corrupted by the presence of the extremes (overestimations of background and variability) and therefore unsuited for detecting extremes (Lanzante, 1996).

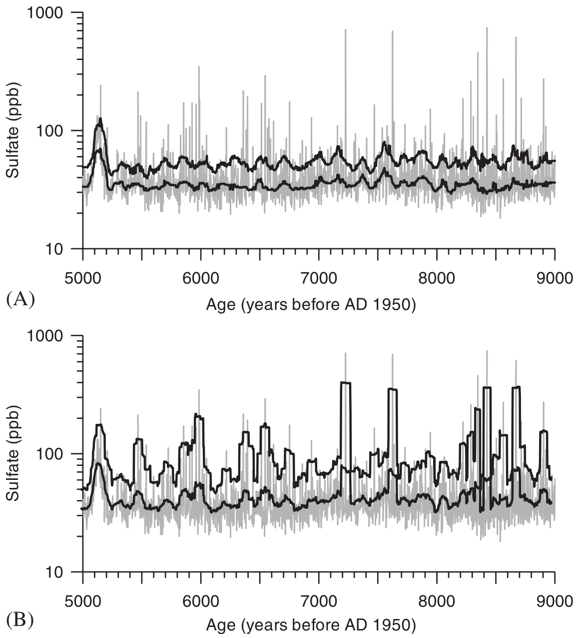

Fig. 2. Detection of extremes in GISP2 sulfate record (gray lines in A and B), $n = 1806$. (A) Robust detection: running median with a cross-validated (Eq. (5)) number of window points, $k = 15$, estimates time-dependent background (lower heavy line); running median $+3.5 \cdot \mathrm{MAD}$ is time-dependent detection threshold (upper heavy line). (B) Non-robust detection: running mean ($k = 15$) as background estimate (lower heavy line) and running mean $+ 2.36$ standard deviations as detection threshold (upper heavy line).

records a steady input from oceanic and other sources (background signal). Superimposed on this signal are occasional peaks stemming from volcanic eruptions. Besides to determine volcanic activity during the Holocene, the record further allows to reconstruct the climate evolution in the North Atlantic region (time-dependent background and variability).

Fig. 2A shows robust extremes detection and background estimation for the GISP2 sulfate record. The running MAD detection uses Hampel's (1985) rule ($z = 3.5$), the window width ($k = 15$) is from cross-validation and corresponds to an average window length of 66 years, thus allowing for decadal/centennial-scale background variations. The number of detected sulfate peaks is 201, which means that on average every 20 years a peak occurs; the time-dependence of those occurrences can be further analyzed using kernel estimation techniques, see Mudelsee et al. (2003, 2004). Notably, the cooling event at around 8200 years before present (Hammer et al., 1997) seems not to have been caused by elevated volcanic activity.

Fig. 2B shows non-robust extremes detection using running mean and standard deviation. Only 71 peaks are "detected" because of overestimated background and variability values. Such overestimations are clearly visible at around the peaks at ~7200 years before present (defined as AD 1950) and ~7600 years. Although Lanzante (1996) warned of using non-robust detection methods, this method has been frequently used. An unfortunate example is Cuomo et al. (2000), who even denoted usage of mean and standard deviation as a "robust" method in the title of their paper.

Acknowledgements

The constructive review comments are gratefully acknowledged. The present work was supported by a Research Grant (MU 1595/1) from the Deutsche Forschungsgemeinschaft, Bonn, Germany. A Windows executable of CLIM-X-DETECT and instructions for installation are available from http://www.climate-risk-analysis.com.

Appendix. Short CLIM-X-DETECT manual

CLIM-X-DETECT uses time series $t(i), x(i)$ in ASCII format, one pair of values per line. Pass file names for data input and output via configuration file 'CLIM-X-DETECT.cfg'. Data size (minimum: $n = 25$) is detected automatically. The time series is shown on the screen. In Part 1 of the program, you may extract a time interval for the analysis. Continuing with Part 2 (Extreme events detection), you first enter the k range for which the cross-validation functions (Eqs. (4) and (5)) are calculated. This is advisable for long time series and prior knowledge about the timescale of background variations, to reduce computing costs. CLIM-X-DETECT then plots $CV_1(k)$ and $CV_{\mathrm{m}}(k)$ and gives the minimizing k values to guide selection of k. Then appear the graphs of $x(i)$, $\mathrm{MED}_{j=i-k}^{j=i+k}(x(j))$, and $\mathrm{MED}_{j=i-k}^{j=i+k}(x(j)) + z\,\mathrm{MAD}_{j=i-k}^{j=i+k}(x(j))$ against $t(i)$. Initially, the curves for two z values ($z = 2$ and $z = 4$) are shown simultaneously, but you may test other values. Note that the curves are simply extrapolated ($i = 1, \ldots, k; i = n - k + 1, \ldots, n$) by constants to cover the full time interval; more adaptive approaches (k decreasing at the boundaries) can be implemented using the source code. The plot setting can be changed (zooms, logarithmic x-axis). When the treshold (z) is finally set and you continue, the scaled extremes time series, $x^*(i)$, is shown. You may either write data, background and variability estimates, and scaled extremes to output files, or go back to Parts 1 or 2. The source code of CLIM-X-DETECT, the configuration file, and the installation instructions give further information.

References

Beirlant, J., Goegebeur, Y., Segers, J., Teugels, J., De Waal, D., Ferro, C., 2004. Statistics of Extremes. Wiley, Chichester, 522pp.

Coles, S., 2001. An Introduction to Statistical Modeling of Extreme Values. Springer, London, 208pp.

Cuomo, V., Di Bello, G., Lapenna, V., Piscitelli, S., Telesca, L., Macchiato, M., Serio, C., 2000. Robust statistical methods to discriminate extreme events in geoelectrical precursory signals: Implications with earthquake prediction. Natural Hazards 21 (2–3), 247–261.

Embrechts, P., Klüppelberg, C., Mikosch, T., 1997. Modelling Extremal Events for Insurance and Finance. Springer, Berlin, 648pp.

Hammer, C., Mayewski, P.A., Peel, D., Stuiver, M. (Eds.), 1997. The Greenland summit ice cores. Journal of Geophysical Research 102 (C12) (special issue).

Hampel, F.R., 1985. The breakdown points of the mean combined with some rejection rules. Technometrics 27 (2), 95–107.

Härdle, W., Steiger, W., 1995. Optimal median smoothing. Applied Statistics 44 (2), 258–264.

Lanzante, J.R., 1996. Resistant, robust and non-parametric techniques for the analysis of climate data: theory and examples, including applications to historical radiosonde station data. International Journal of Climatology 16 (11), 1197–1226.

Marron, J.S., 1987. What does optimal bandwidth selection mean for nonparametric regression estimation? In: Dodge, Y. (Ed.), Statistical Data Analysis Based on the L_1-Norm and Related Methods. North-Holland, Amsterdam, pp. 379–392.

Marron, J.S., 1988. Automatic smoothing parameter selection: a survey. Empirical Economics 13 (3–4), 187–208.

McCarthy, J.J., Canziani, O.F., Leary, N.A., Dokken, D.J., White, K.S., 2001. Climate Change 2001: Impacts, Adaptation, and Vulnerability. Contribution of Working Group II to the Third Assessment Report of the Intergovernmental Panel on Climate Change. Cambridge University Press, Cambridge, 1032pp.

Meehl, G.A., Zwiers, F.W., Evans, J., Knutson, T., Mearns, L., Whetton, P., 2000. Trends in extreme weather and climate events: issues related to modeling extremes in projections of future climate change. Bulletin of the American Meteorological Society 81 (3), 427–436.

Mudelsee, M., Börngen, M., Tetzlaff, G., Grünewald, U., 2003. No upward trends in the occurrence of extreme floods in central Europe. Nature 425 (6954), 166–169.

Mudelsee, M., Börngen, M., Tetzlaff, G., Grünewald, U., 2004. Extreme floods in central Europe over the past 500 years: role of cyclone pathway 'Zugstrasse Vb'. Journal of Geophysical Research 109 (D23) doi:10.1029/2004JD005034.

Press, W.H., Teukolsky, S.A., Vetterling, W.T., Flannery, B.P., 1996. Numerical Recipes in Fortran 90. Cambridge University Press, Cambridge, 1486pp.

Reiss, R.-D., Thomas, M., 1997. Statistical Analysis of Extreme Values. Birkhäuser, Basel, 316pp.

Tukey, J.W., 1977. Exploratory Data Analysis. Addison-Wesley, Reading, MA, 688pp.

Zheng, Z.g., Yang, Y., 1998. Cross-validation and median criterion. Statistica Sinica 8 (3), 907–921.

Zielinski, G.A., Mayewski, P.A., Meeker, L.D., Whitlow, S., Twickler, M.S., Morrison, M., Meese, D.A., Gow, A.J., Alley, R.B., 1994. Record of volcanism since 7000 B.C. from the GISP2 Greenland ice core and implications for the volcano–climate system. Science 264 (5161), 948–952.