Long memory and the aggregation of AR(1) processes: Some Monte Carlo results

M. Mudelsee

Summary

Granger (1980) found that the aggregation of m short-memory AR(1) processes yields a long-memory process. Thereby he assumed $m \to \infty$, Gaussian shape and beta-distributed AR(1) parameters over $(0; 1)$. To test hypotheses that long memory in climate time series comes from aggregation, the finding of Granger (1980) cannot be directly applied. First, the number of "microclimatic" processes to be aggregated is finite. Second, climatic processes often produce right-skewed data. Third, the AR(1) parameters of the microclimatic processes could be restricted to a narrower interval than $(0; 1)$. We therefore perform Monte Carlo simulations to study aggregation in climate time series under realistic conditions. The long-memory parameter, H, is estimated by fitting an ARFIMA model to various types of aggregations. Our results are as follows. First, for m above a few hundred, H approaches saturation. Second, the distributional shape has little influence, as noted by Granger (1980). Third, the upper limit of the interval for the AR(1) parameter has a strong influence on the saturation value of H, as noted by Granger (1980).

Zusammenfassung

Granger (1980) fand heraus, dass die Summe von m schwach seriell abhängigen AR(1)-Prozessen einen stark seriell abhängigen Prozess ergibt. Er nahm dabei an, dass $m \to \infty$ geht, die Verteilungen Gaußsch sind und die AR(1)-Parameter beta-verteilt über $(0; 1)$ sind. Um die Hypothese zu testen, daß starke serielle Abhängigkeit in Klimazeitreihen von dieser "Aggregation" rührt, kann das Ergebnis von Granger (1980) jedoch nicht direkt angewendet werden. Erstens: die Anzahl "mikroklimatischer", zu summierender Prozesse is endlich. Zweitens: Klimaprozesse erzeugen oft rechtsschief verteilte Daten. Drittens: die AR(1)-Parameter der mikroklimatischen Prozesse mögen auf ein engeres Intervall begrenzt sein als $(0; 1)$. Wir führen deshalb Monte-Carlo-Simulationen durch, um die Aggregation in Klimazeitreihen für realistische Bedingungen zu studieren. Der Parameter H, der die starke serielle Abhängigkeit beschreibt, wird geschätzt durch die Anpassung eines ARFIMA-Modelles an unterschiedliche Aggregations-Typen. Unsere Ergebnisse sind wie folgt. Erstens: für m oberhalb einiger hundert erreicht H Sättigung. Zweitens: die Verteilungsform hat geringen Einfluß, wie von Granger (1980) bemerkt. Drittens: die obere Grenze des Intervalles für den AR(1)-Parameter hat einen starken Einfluß auf den Sättigungwert von H, wie von Granger (1980) bemerkt.

1. Introduction

Runoff time series from rivers are an important contribution to the hydrological database. Such records, measured since the early 19th century, document dry and wet conditions in the catchment area of a particular river station. For example, one of the longest records in the database of the Global Runoff Data Centre (Koblenz, Germany) comes from the Rhine station Cologne, starting in 1807. Runoff (volume water per time unit) is usually inferred from water stage measurements via calibration formulas (Mudelsee et al. 2003: Fig. 1 therein). Although this introduces proxy error into the data, runoff time series are considered to serve well in the evaluation of past precipitation integrated over a considerable area (Milly et al. 2002). These hydrological records are therefore essential to be studied also within the context of climatic changes during the instrumental period (Houghton et al. 2001).

Persistence, also termed positive autocorrelation, serial dependence or memory, is a property shared by most climatic processes. The ultimate reason lies within the physics of the climate system. Nonzero lengths, masses and heat capacities let processes in the atmosphere run at a finite speed. Extending the climate system to other compartments such as the cryosphere, biosphere or hydrosphere likely increases the strength of the persistence.

Consider as an example the collection of precipitation over a “microclimatic” unit area. Rainfall has a high spatial variability (Liljequist and Cehak 1984: Chapter 15 therein), so the size of such an idealized unit is small, say, of the order of 100 km^2. A proportion is “lost” owing to interception and evapotranspiration, what remains goes into the soil. The storage capacity of the soil varies spatially and temporally to some degree. That means, this “microhydrological” storage unit may be of a similar size. If we now assume that precipitation is a random process and the microhydrological storage unit exhibits a simple linear release rule (the relation between input and output), then the output (“microrunoff”) is an AR(1) process (see next paragraph), with an autocorrelation parameter dependent on the release rule parameter (Klemeš 1983).

The simplest mathematical persistence model is the first-order autoregressive or AR(1) process (Priestley 1981),

$$\begin{aligned} X(1) &= \mathcal{E}_{\mathrm{N}(0,\,1)}(1), \\ X(i) &= a \cdot X(i-1) + \mathcal{E}_{\mathrm{N}(0,\,1-a^2)}(i), \qquad i = 2, \ldots, n. \end{aligned} \tag{1}$$

Herein, X is a climate variable (e.g., runoff), i counts the time steps, n is the data size, $0 \leq a < 1$ is a constant and $\mathcal{E}_{\mathrm{N}(\mu,\,\sigma^2)}(i)$ is a Gaussian random process with mean μ, variance σ^2 and no serial dependence, that means, $E\left[\mathcal{E}_{\mathrm{N}(\mu,\,\sigma^2)}(j) \cdot \mathcal{E}_{\mathrm{N}(\mu,\,\sigma^2)}(k)\right] = 0$ for $j \neq k$. It readily follows that $X(i)$ is a strictly stationary process with zero mean and unity variance. The AR(1) model has an exponentially decreasing autocorrelation function (Priestley 1981: Section 3.5 therein),

$$\rho(h) = a^{|h|}, \qquad h = 0, \pm 1, \pm 2, \ldots, \tag{2}$$

where h is the time lag. This behaviour might be referred to as “exponentially decreasing memory.” Because of the fast (exponential) decay, the AR(1) model is said to be of short memory or short-range dependence. Hasselmann (1976) showed theoretically that the AR(1) model is a suitable description of climate processes.

Starting with Hurst (1951), who analysed water stage records from the River Nile, a number of authors, including Mandelbrot and Wallis (1969), Hosking (1984) and Kantelhardt et al. (2003), have concluded that for hydrological runoff time series, instead of the AR(1) persistence model a long-memory or long-range dependence model would be appropriate. Such processes (Beran 1994; Robinson 2003) have the property

$$\rho(h) \rightarrow C\, h^{2H-1} \qquad \text{as } h \rightarrow \infty, \tag{3}$$

where $C \neq 0$ and $H < 0.5$. This decrease is slower than in the AR(1) case, hence it is said to exhibit long-range serial dependence or long memory. If runoff records exhibited long memory, the consequences would be far reaching. First, prediction of hydrological extremes such as floods over longer timescales might, at least in principle, become possible. Second, the noise property long-memory would have an impact on the methodology to evaluate the accuracy of hydrological estimates. For example, bootstrap methods belong to the most powerful approach to achieve this because they are not distributionally restricted. The bootstrap can further be successfully adapted to the case of serial dependence (Lahiri 2003). Long memory would therefore indicate how to carry out the adaption.

Granger (1980) showed how the aggregation of short-memory AR(1) processes,

$$Y(i) = \sum_{j=1}^{m} X_j(i), \qquad i = 1, \ldots, n, \tag{4}$$

can produce a long-memory process, $Y(i)$. The assumptions made thereby are:

1. large number of aggregated processes ($m \rightarrow \infty$),
2. processes $X_j(i)$ of type Gaussian AR(1) and
3. beta-distributed AR(1) parameters a_j over the interval $(0; 1)$.

The natural question about long memory in hydrological time series arises whether this may be the result of aggregation of individual microhydrological units over the catchment area. A positive answer would mean a remarkably simple explanation and also corroborate the empirical findings. However, a direct application of Granger's (1980) result is prohibited by the following points:

1. a limited number of aggregated processes,
2. right-skewed distributions and
3. a restricted range of the AR(1) parameter a_j.

As regards point 1, it is *a priori* not clear whether for a given river station and its catchment area the number of aggregated precipitation processes is large enough to apply the limit value found by Granger (1980). This is the case especially for small catchment areas and a lower spatial variability of precipitation. As regards point 2, it is generally known that runoff and precipitation data exhibit right-skewness. The heavy tails are subject of extreme value analyses, return period estimation and climate risk analysis (Embrechts et al. 1997; Mudelsee et al. 2003, 2004). Point 3 is rather difficult to assess. There is certainly spatial precipitation variability over a large enough catchment area, and there may also be spatial variability in soil and storage properties. That is, we can safely assume that the AR(1) parameter a shows variation. But the upper bound of this a distribution over the microhydrological units is rather difficult to quantify.

Because analytical results for the long-memory parameter seem not to be obtainable in the "realistic case" (finite number of aggregated processes, right-skewed distributions, bounds below 1 of the AR(1) parameter), we perform mathematical simulations. (We remark that Linden (1999) obtained analytical results for $n \to \infty$ and uniformly distributed a_j.) These Monte Carlo experiments produce random numbers from AR(1) processes (Gaussian and right-skewed), add those numbers and estimate the long-memory parameter H from the sum. The results obtained are curves of estimated H versus the number of aggregated processes, for a range of upper bounds of the AR(1) parameters, a_j.

This paper forms the first part of our study of long memory in hydrological time series. Other parts analyse H for various, spatially distributed river stations.

2. Microhydrological processes

We analyse a time series from the catchment area of the station Dresden at the River Elbe (Germany) to infer how good the assumption of randomness (Klemeš 1983) is fulfilled. The data are monthly means from January 1900 to December 1998. They come from an updated version of the gridded database of observations from Hulme et al. (1998). The resolution is 2.5° latitude by 3.75° longitude. The box centred at 50° N, 15° E contains the major portion of the catchment area. The series itself was obtained by averaging measurements of (maximum) seven stations. The series data, originally strongly right-skewed, was logarithmically transformed and the strong annual signal (and also its harmonic at a period of 0.5 years) subsequently removed by a harmonic analysis tool (Schulz and Stattegger 1997). The resulting series is shown in Figure 1.

A fit of the AR(1) model to the series yielded an estimated autocorrelation parameter of $\widehat{a} = 0.06$, corresponding to a persistence time (decay period of the autocorrelation function) of 10.8 days (Mudelsee 2002). Note that at a daily resolution, this persistence time value corresponds to an autocorrelation parameter of $\exp(-1\,\mathrm{d}/10.8\,\mathrm{d}) \approx 0.91$. The AR(1) model fit gave a value for Akaike's (1973) information criterion of AIC = 1707.4. We also tried fitting an ARFIMA$(1, \delta, 0)$ model, which yielded $\widehat{\delta} = 0.033$, an estimated AR(1) parameter of 0.024 and AIC = 1707.9. Because of the lower AIC, the AR(1) model was preferred. The analysed precipitation series thus has no long-memory dependence and only a small AR(1) parameter.

Note that the microclimatological process of precipitation is transformed into a microhydrological process by infiltration into the soil (see Introduction). Soil properties regarding storage are rather unknown over the analysed time interval. It might well be that the resulting microhydrological process has a severely stronger autocorrelation (Klemeš 1983), but it appears not possible to give an upper bound.

3. Simulation method

3.1. Aggregation

As first, simple distributional example, we take the strictly stationary Gaussian AR(1) process (Eq. 1).

As second, right-skewed distribution we take the lognormal:

$$X'_j(i) = \exp\left[X_j(i)\right], \qquad i = 1, \dots, n, \qquad j = 1, \dots, m, \tag{5}$$

where $X_j(i)$ is the strictly stationary Gaussian AR(1) process (Eq. 1). The lognormal is a distributional shape commonly found in data from the natural sciences (Aitchison and Brown 1957). The autocorrelation parameter a of the process $X(i)$ is drawn from a uniform distribution over $(0; a_{\max})$ (exclusive of the bounds), with $a_{\max} = 0.15$, 0.2, 0.4, 0.6, 0.8, 0.9, 0.95 and 1.00. A data size of $n = 1000$ is used throughout, which is the typical length of monthly time series covering the instrumental period. The number of summands (Eq. 4) is varied from $m = 1$ in logarithmically spaced steps up to $m = 10000$.

The data generation part is written in Fortran 90 at a precision comparable to FORTRAN 77's double precision. We use the uniform random number generator of Park and Miller (1988) and the routine gasdev from Press et al. (1996) to produce Gaussian deviates.

3.2. Estimation of the long-memory parameter

The long-memory parameter is estimated by fitting a fractional autoregressive integrated moving average or ARFIMA(p, δ, q) model. (In the case of lognormally distributed $X'_j(i)$, we use the logarithmic transformation $Y'(i) = \log\left[Y(i)\right], i = 1, \dots, n$.) Specifically, we set the moving-average order q as zero and the autoregressive parameter p equal to one. The relation between the ARFIMA$(1, \delta, 0)$ and the AR(1) model is as follows. δ defines (Brockwell and Davis 1991) a fractional difference operator, $(1-B)^\delta$, where $|\delta| < 0.5$

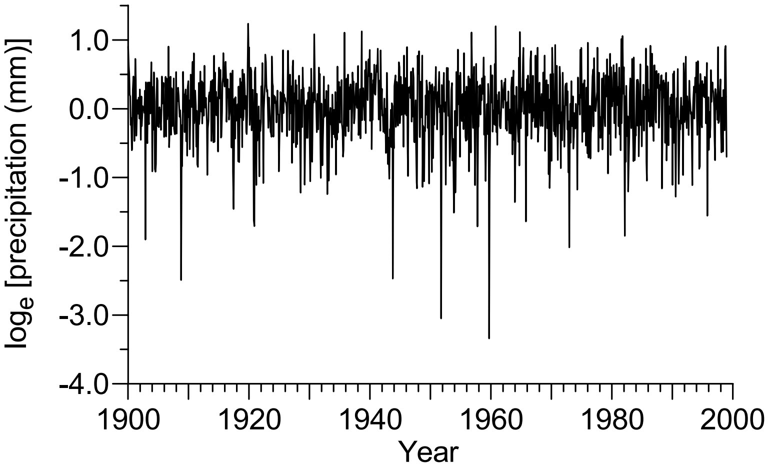

Figure 1: Precipitation time series for the Dresden (Elbe) catchment area.

and B is the backshift operator. The backshift operator shifts one step back in time, for example, $B\,X_j(i) = X_j(i-1)$. The ARFIMA$(1,\delta,0)$ model is then an AR(1) model (Eq. 1) where on the right-hand side $X_j(i)$ is replaced by $(1-B)^{\delta}\,X_j(i)$. For the trivial case $\delta = 0$, the ARFIMA$(1,\delta,0)$ model reduces to the AR(1) model. The ARFIMA(p,δ,q) model has $H = \delta$ (Brockwell and Davis 1991).

The logarithmic transformation of the $Y(i)$ data is necessary to achieve a (roughly) Gaussian shape of the simulation time series data. Fitting the ARFIMA$(1,\delta,0)$ model is achieved using an exact maximum likelihood criterion (Beran 1994). We use the software module ARFIMA, written in the Ox statistical computer language (Doornik and Ooms 2001). The numerical implementation is described by Doornik and Ooms (2003).

3.3. Monte Carlo simulations

One simulation loop consists of the following steps.

1. generate data, $X_j(i), X'_j(i), Y(i)$ (lognormal distribution, $Y'(i)$);
2. standardize $Y(i)$ (lognormal distribution, $Y'(i)$);
3. estimate H via the ARFIMA$(1,\delta,0)$ model fitted to $Y(i)$ (lognormal distribution, $Y'(i)$) as $H = \delta$.

These steps are carried out $n_{\mathrm{sim}} = 400$ times. The average of H over the simulations with its standard error is then plotted against m.

4. Results and conclusions

The first, major result is that H approaches saturation already for m above a few hundred. This behaviour apparently does not depend on the distributional shape (Gaussian, Fig. 2, versus lognormal, Fig. 3). This point on the m axis, from where saturation behaviour occurs, does not depend on $a_{\max}$, as long as $a_{\max}$ is above a value of about 0.4 to 0.6.

Second, the distributional shape has little influence, as noted already by Granger (1980). Third, the upper limit of the interval for the AR(1) parameter has a strong influence on the saturation value of H, as noted by Granger (1980). Below $a_{\max}$ =0.6 or 0.4, the aggregated series seem not to exhibit a nonzero long-memory parameter H.

Acknowledgements

This work was supported by the Deutsche Forschungsgemeinschaft (Bonn, Germany).

References

Aitchison J, Brown JAC (1957) *The Lognormal Distribution.* Cambridge University Press, Cambridge, 176 pp.

Akaike H (1973) Information theory and an extension of the maximum likelihood principle. In: Petrov BN, Csáki F (Eds) *Second International Symposium on Information Theory.* Akadémiai Kiadó, Budapest, pp 267–281.

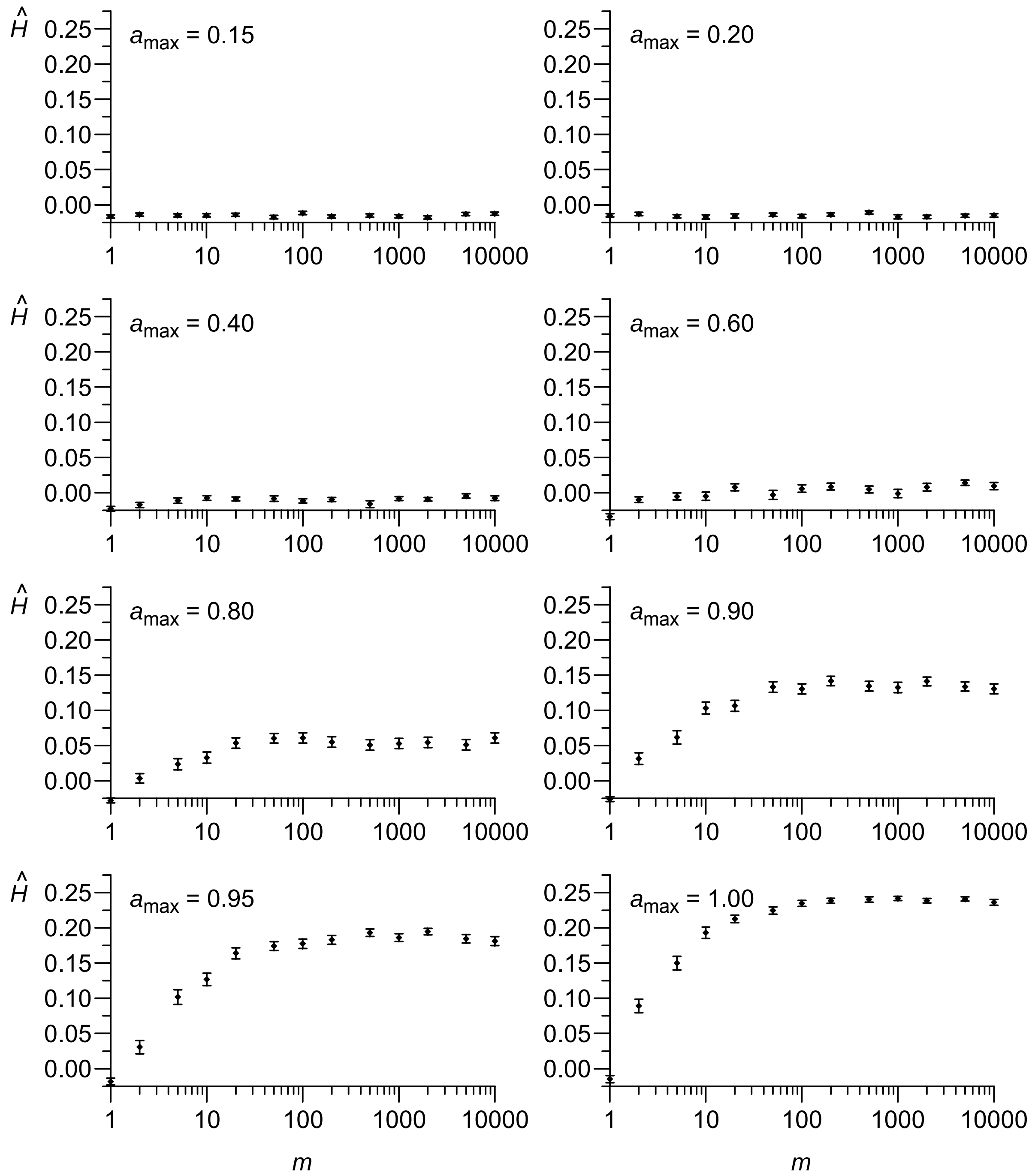

Figure 2. Results (normal distribution). Average of $\widehat{H}$ over $n_{\mathrm{sim}} = 400$ simulations (filled circles) with standard errors (vertical bars).

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

Brockwell PJ, Davis RA (1991) *Time Series: Theory and Methods.* Second edition. Springer, New York, 577 pp.

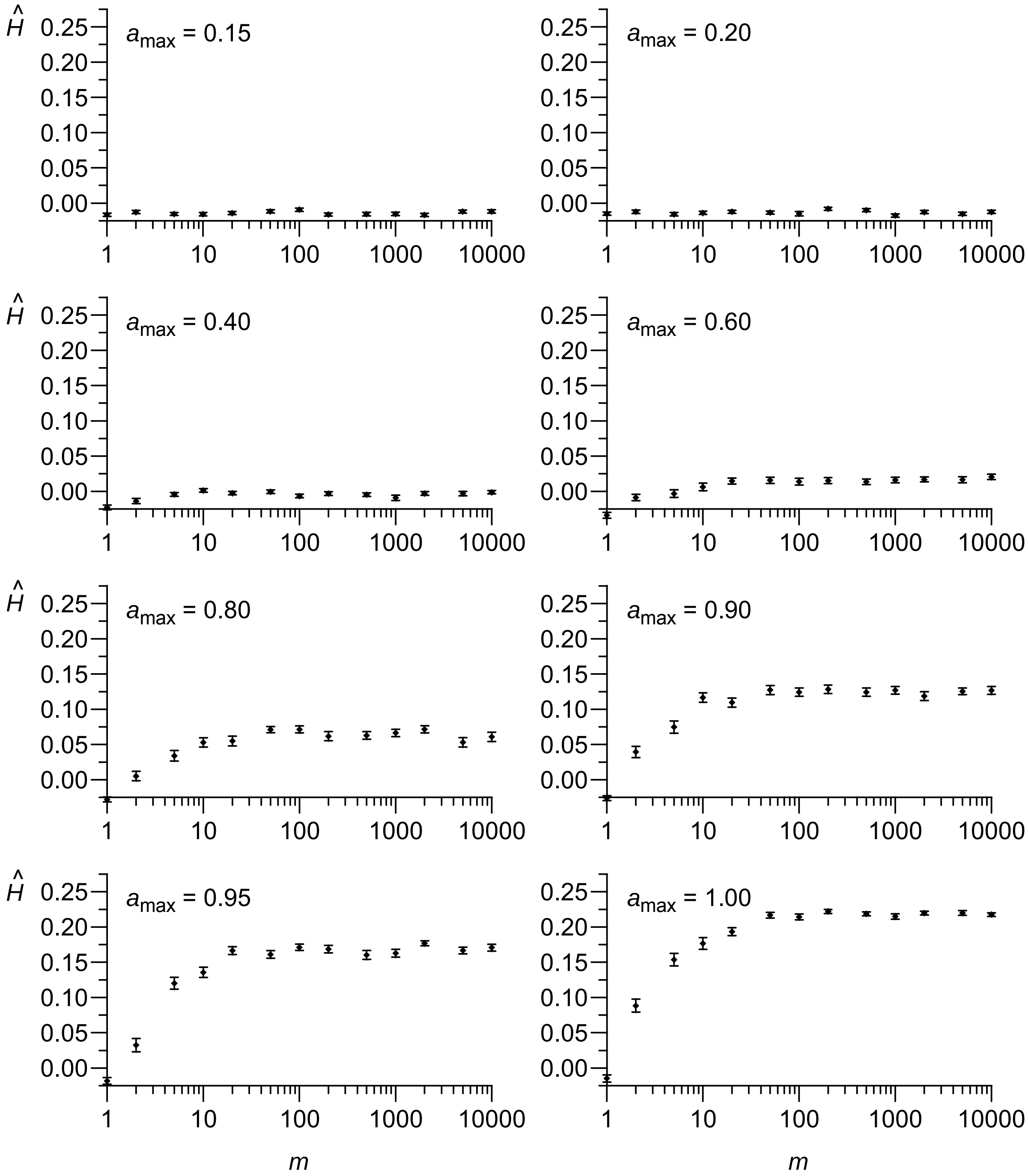

Figure 3. Results (lognormal distribution). Average of $\widehat{H}$ over $n_{\mathrm{sim}} = 400$ simulations (filled circles) with standard errors (vertical bars).

Doornik JA, Ooms M (2001) *A Package for Estimating, Forecasting and Simulating Arfima Models: Arfima package 1.01 for Ox.* Published oneself, Oxford, 32 pp. http://www.doornik.com, 18 December 2005

Doornik JA, Ooms M (2003) Computational aspects of maximum likelihood estimation of autoregressive fractionally integrated moving average models. *Computational Statistics and Data Analysis* 42(3): 333–348.

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

Granger CWJ (1980) Long memory relationships and the aggregation of dynamic models. *Journal of Econometrics* 14(2): 227–238.

Hasselmann K (1976) Stochastic climate models: Part I. Theory. *Tellus* 28(6): 473–485.

Hosking JRM (1984) Modeling persistence in hydrological time series using fractional differencing. *Water Resources Research* 20(12): 1898–1908.

Houghton JT, Ding Y, Griggs DJ, Noguer M, van der Linden PJ, Dai X, Maskell K, Johnson CA (Eds) (2001) *Climate Change 2001: The Scientific Basis. Contribution of Working Group I to the Third Assessment Report of the Intergovernmental Panel on Climate Change.* Cambridge University Press, Cambridge, 881 pp.

Hulme M, Osborn TJ, Johns TC (1998) Precipitation sensitivity to global warming: Comparison of observations with HadCM2 simulations. *Geophysical Research Letters* 25(17): 3379–3382.

Hurst HE (1951) Long-term storage capacity of reservoirs. *Transactions of the American Society of Civil Engineers* 116(??): 770.

Kantelhardt JW, Rybski D, Zschiegner SA, Braun P, Koscielny-Bunde E, Livina V, Havlin S, Bunde A (2003) Multifractality of river runoff and precipitation: Comparison of fluctuation analysis and wavelet method. *Physica A* 330(1–2): 240–245.

Klemeš V (1983) Hydrology, stochastic. In: Kotz S, Johnson NL, Read CB (Eds) *Encyclopedia of statistical sciences*, volume 3. Wiley, New York, pp 694–700.

Lahiri SN (2003) *Resampling Methods for Dependent Data.* Springer, New York, 374 pp.

Liljequist GH, Cehak K (1984) *Allgemeine Meteorologie.* 3 edition. Vieweg, Braunschweig, 396 pp.

Linden M (1999) Time series properties of aggregated AR(1) processes with uniformly distributed coefficients. *Economics Letters* 64(1): 31–36.

Mandelbrot BB, Wallis JR (1969) Some long-run properties of geophysical records. *Water Resources Research* 5(2): 321–340.

Milly PCD, Wetherald RT, Dunne KA, Delworth TL (2002) Increasing risk of great floods in a changing climate. *Nature* 415(6871): 514–517.

Mudelsee M (2002) TAUEST: A computer program for estimating persistence in unevenly spaced weather/climate time series. *Computers and Geosciences* 28(1): 69–72.

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): D23101 (doi:10.1029/2004JD005034).

Park SK, Miller KW (1988) Random number generators: Good ones are hard to find. *Communications of the ACM* 31(10): 1192–1201.

Press WH, Teukolsky SA, Vetterling WT, Flannery BP (1996) *Numerical Recipes in Fortran 90.* Cambridge University Press, Cambridge, 935-1446 pp.

Priestley MB (1981) *Spectral Analysis and Time Series.* Academic Press, London, 890 pp.

Robinson PM (Ed) (2003) *Time Series with Long Memory.* Oxford University Press, Oxford, 382 pp.

Schulz M, Stattegger K (1997) SPECTRUM: Spectral analysis of unevenly spaced paleoclimatic time series. *Computers and Geosciences* 23(9): 929–945.

Author's address:

Dr. Manfred Mudelsee
University of Leipzig
Institute of Meteorology
Stephanstrasse 3
04103 Leipzig
Germany