A statistical model to analyse natural catastrophe claims by means of record values

Prof. Dr. Dietmar Pfeifer
Institut für Mathematische Stochastik
Universität Hamburg
Germany

Abstract: We present a statistical model based on record values of non–i.i.d. observations to analyse and forecast claims arising out of natural catastrophes, and/or to detect trends over time. In particular, claims data from U.S. hurricanes and Japanese taifuns are discussed.

1. Introduction. It is a fairly evident fact that insurance claims due to the occurence of natural catastrophes have raised enormously over the past decades all over the world. Several arguments have been proposed in order to explain this develoment, besides economic factors (inflation rate and increase in insured value) mainly environmental influences (e.g. climatic change, cf. **sigma** 2/93). Although it has become popular to develop physical models for the analysis and forcasting of claims, in particular in the area of storm events, it seems that reliable conclusions with respect to general premium calculation especially for reinsurance contracts remain rather vague. On the other hand, mathematical modeling has already done a good job in the statistical analysis of catastrophe claims (cf. R. Schnieper: The insurance of catastrophe risks, C. Partrat and C. Huygues–Beaufond: Rate making for natural events coverages in the USA, J.M. Cozzolino and E.M. Gaydos: Measuring the probability of disastrous losses, in: **SCOR Notes** 1993, special issue on "catastrophe risks"). In this paper, we want to present a particular approach to the investigation of catastrophe claims in the presence of a trend, which is based on a combination of parametric and semi–parametric methods from the area of statistics of extremes. In a first step, the "type" of trend will be analysed using the number of record values in the times series of claims data, in the second step, a maximum–likelihood estimation is performed with the data taking into account what type of trend has been detected before. In order to check the validity of the model assumptions, the estimates for the trend parameter obtained from the parametric as well as the semi–parametric approach can be compared. Tests for trends based on the number of record values have been used earlier, see e.g. J. Diersen and G. Trenkler (1996) and the references given therein; however, the proposed combination of semi–parametric and parametric models seems to be novel, and has

seemingly not yet been applied to insurance data before.

The method will be illustrated by two data records from catastrophe insurance claims, namely U.S. hurricane events from 1949 to 1992 and Japanese taifun events from 1977 to 1991.

2. Nevzorov's record model. The semi–parametric part of the analysis relies on a record model that has in more detail been studied by V.B. Nevzorov (see e.g. his survey paper from 1988) and K. Borovkov and the author (1995). [It dates actually back to a paper by M.C.K. Yang (1975) who studied the frequent breaking of sports records in the light of a geometric growth of the world population.] In order to avoid too much technical difficulties, we suppose that the catastrophe claims considered here are realizations of an independent sequence $\{X_n\}_{n\in\mathbf{N}}$ of non–negative random variables with support $\mathbf{R}^+ = [0,\infty)$ and continuous cumulative distribution functions $\{F_n\}_{n\in\mathbf{N}}$ such that $F_n = F^{\gamma_n}$ with $\gamma_n > 0$, where F is a fixed cumulative distribution function with $F(0) = 0$. Define *record indicators* by

$$I_1 = 1, \quad I_n = \begin{cases} 1, & \text{if } X_n > \max\{X_1,\ldots,X_{n-1}\} \\ 0, & \text{otherwise} \end{cases} \quad \text{for } n > 1,$$

i.e. $I_n = 1$ iff observation X_n is a *record value* in the sequence. Under the above assumptions, the record indicators are independent random variables with

$$P(I_n = 1) = \frac{\gamma_n}{\gamma_1 + \ldots + \gamma_n}, \quad n \in \mathbf{N}.$$

In the standard case of an i.i.d. sequence , this reduces to a well–known result which was independently discovered by Dwass and Rényi:

$$P(I_n = 1) = \frac{1}{n}, \quad n \in \mathbf{N}.$$

For a statistical analysis of claims data, it is useful to consider the number of record values in a finite number of observations, i.e. we consider

$$S_n = \sum_{i=1}^{n} I_i, \quad n \in \mathbf{N}.$$

From what has been said above it is clear that we have

$$E(S_n) = \sum_{i=1}^{n} p_i, \quad Var(S_n) = \sum_{i=1}^{n} p_i(1-p_i), \quad n \in \mathbf{N}.$$

In the i.i.d. situation, this reduces to

$$E(S_n) = \sum_{i=1}^{n} \frac{1}{i} \approx \ln n, \quad Var(S_n) = \sum_{i=1}^{n} \frac{i-1}{i^2} \approx \ln n - \frac{\pi^2}{6}, \quad n \in \mathbf{N}.$$

For our purposes, we shall particularly consider the choice $\gamma_i = \gamma^{i-1}$ with a fixed (but possibly unknown) parameter $\gamma \geq 1$ (called *trend parameter*). Note that for $\gamma = 1$, we have the i.i.d. situation (no trend), while for $\gamma > 1$, the random variables $\{X_n\}_{n\in\mathbf{N}}$ are stochastically increasing (positive trend). In the latter case, we have asymptotically

$$p_i = p_i(\gamma) = \frac{\gamma - 1}{\gamma}(1 - \gamma^{-i})^{-1} \approx \frac{\gamma - 1}{\gamma}$$

for large i, hence

$$E(S_n) \approx \frac{\gamma - 1}{\gamma} n$$

for large n.

Associated with $I_1, \ldots, I_n$ and S_n are the so–called *record times* $T_1, \ldots, T_{S_n}$ which denote the observation times at which record values occur:

$$T_1 := 1, \quad T_{k+1} := \min\left\{i \leq n \mid X_i > X_{T_k}\right\} = \min\left\{T_k < i \leq n \mid I_i = 1\right\},\ k < S_n,$$

with the evident property that $I_{T_k} = 1$ for all $k = 1, 2, \ldots, S_n$.

Note that the exponential parameter case does not necessarily anticipate the type of trend in the data. For instance, if $F(x) \sim e^{-e^{-Ax}}$ for $x \to \infty$ (Gumbel distribution type, $A > 0$), then $F^{\gamma^{i-1}}(x) \sim \exp\left(-e^{-Ax+(i-1)\ln\gamma}\right)$ which corresponds asymptotically to a *linear* trend in the mean, while if we have the relation $F(x) \sim e^{-(Ax)^{-\alpha}}$ (Fréchet or Pareto distribution type, $A > 0$) then $F^{\gamma^{i-1}}(x) \sim \exp\left(-\left\{A\gamma^{(1-i)/\alpha}x\right\}^{-\alpha}\right)$ which corresponds asymptotically to an *exponential* trend in the mean indeed if $\alpha > 1$. Actually, by a suitable form of the cumulative distribution function F it is possible to obtain more or less arbitrary types of trend in the mean for the cases considered here.

From the above calculation it follows that given the observations $I_1, \ldots, I_n$ of record indicators in a sequence of n data, the log–likelihood function $L(\gamma)$ for $\gamma \geq 1$ is given by

$$\begin{aligned} L(\gamma) &= \ln\left(\prod_{i=1}^{n} p_i(\gamma)^{I_i}(1 - p_i(\gamma))^{(1-I_i)}\right) \\ &= S_n \ln(\gamma - 1) - \ln\left(\gamma^n - 1\right) - \sum_{k=2}^{S_n} \ln\left(1 - \gamma^{1-T_k}\right) \end{aligned}$$

with derivative

$$L'(\gamma) = \frac{S_n}{\gamma - 1} - \frac{n\gamma^{n-1}}{\gamma^n - 1} - \frac{1}{\gamma}\sum_{k=2}^{S_n} \frac{T_k - 1}{\gamma^{T_k - 1} - 1}, \quad \gamma > 1.$$

Unfortunately, it is in general not possible to solve $L'(\gamma) = 0$ explicitly to find the ML–estimator $\hat{\gamma} = \hat{\gamma}(I_1, \ldots, I_n)$ except for the trivial cases $n \in \{1, 2\}$. For the

explicit analysis of the data sets, the symbolic computer algebra system MAPLE© was used therefore.

3. The parametric statistical model. Since by economic arguments [e.g., inflation as one possible factor] it is reasonable to assume that a possible trend in the data is of exponential type and that natural catrastrophe claims are frequently the largest claims occuring over the year, we shall base the parametric statistical model on a combination of Nevzorov's record model and the parametric class of Fréchet distributions [one of the three extreme–value distribution classes], i.e. we shall assume that the cumulative distribution functions F_i for the yearly claims are of the form

$$F_i(x) = \exp\left(-\gamma^{i-1}(Ax)^{-\alpha}\right), \quad A, \alpha > 0,\ \gamma \geq 1.$$

Note that among the possible extreme value distributions, the Fréchet distribution family is the only one with a finite left endpoint, and that an arbitrary cumulative distribution function F is in the *domain of attraction* of the Fréchet distribution [i.e. the properly normalized maxima of independent observations generated by F converge in distribution to a Fréchet limit, see e.g. Leadbetter et al. (1983)] iff F has infinite right endpoint and

$$\lim_{t\to\infty} \frac{1-F(tx)}{1-F(t)} = x^{-\alpha}, \quad \text{for all} \quad x > 0,$$

i.e. iff the tail $1 - F$ is *regularly varying of index* $-\alpha$ (see e.g. Bingham et al. (1987)).

In order to avoid economically meaningless parameter constellations we restrict our considerations only to a *scale family* with scale parameter $A > 0$ rather than to a combined scale and location family.

For the above parametric family, the log–likelihood function $L_D(A, \alpha, \gamma)$ for the observed data set $X_1, \ldots, X_n$ is given by

$$L_D(A, \alpha, \gamma) = \frac{n(n-1)}{2} \ln \gamma - (\alpha + 1) \sum_{i=1}^{n} \ln X_i - \sum_{i=1}^{n} \gamma^{i-1}(AX_i)^{-\alpha} + n \ln\left(\alpha A^{-\alpha}\right)$$

for $A, \alpha > 0, \gamma \geq 1$. Since in general this function cannot be maximized by elementary calculations, a particular stochastic search procedure was performed for the explicit data analysis.

4. Data sets and data analysis. The following data sets were analysed by the methods outlined above:

Yearly claims in Million U.S. $ from U.S. hurricane events from 1949 to 1992 (record values shown in boldface; source: **Catastrophe Reinsurance Newsletter**, April 1993)

year	49	50	51	52	53	54	55	56	57	58	59
claims	**8.3**	**174**	7.7	7.3	14.3	136	25.2	20	32	5	13.1
year	60	61	62	63	64	65	66	67	68	69	70
claims	91	100	81	11	67.2	**515**	57	41.5	36.1	165.3	309.9
year	71	72	73	74	75	76	77	78	79	80	81
claims	31.6	100	76.6	454.4	119.2	34.5	42.6	79	**752.5**	100	201.6
year	82	83	84	85	86	87	88	89	90	91	92
claims	220	**880**	276.7	543	82	150	130	**4195**	625.6	1700	**15500**

Tab. 1: U.S. data

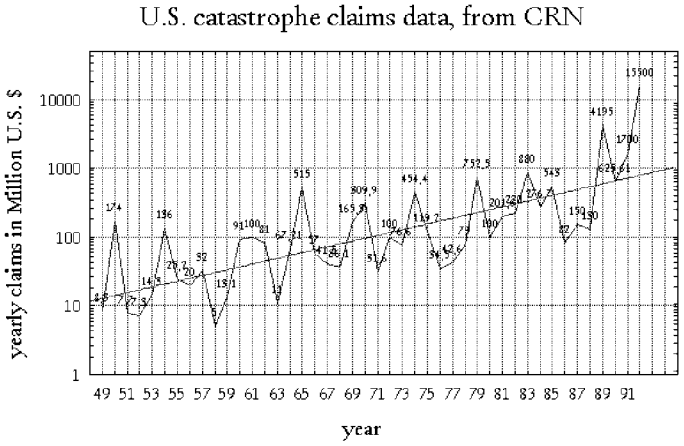

Fig. 1: U.S. data in logarithmic scale
linear fit with least squares

Yearly claims in 1000 JYen from Japanese taifun events from 1977 to 1991 (record values shown in boldface; source: personal communication)

year	77	78	79	80	81
claims	**298.112**	**448.981**	**4090.363**	1032.387	**10642.608**
year	82	83	84	85	86
claims	**27940.639**	7875.241	6680.829	**37127.459**	16498.341
year	87	88	89	90	91
claims	17107.156	2810.834	14281.424	**39013.490**	**484332.000**

Tab. 2: Japan data

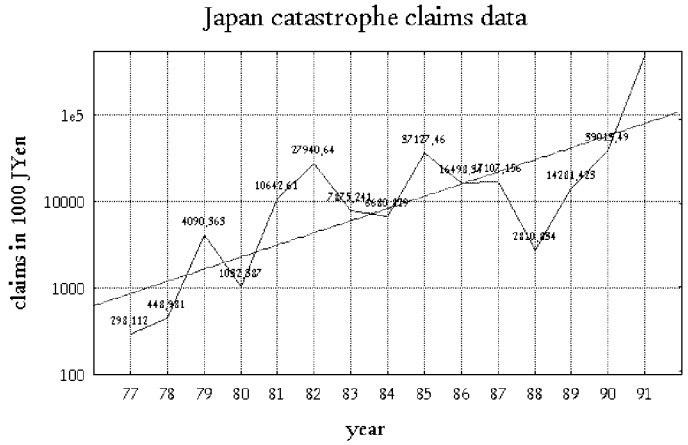

Fig. 2: Japan data in logarithmic scale
linear fit with least squares

The graphical data displays in logarithmic scale reflect quite well from another point of view that the assumption of an exponential trend in the data is reasonable. The following table gives the estimated trend parameters $\hat{\gamma}$ from the three approaches *semi–parametric (s.–par.)* [via record values], *joint maximum–likelihood (jML)*, and *least squares (l.–sq.)* [from the graphical analysis; here $\hat{\gamma} = e^{\hat{\alpha}\hat{m}}$ where $\hat{m}$ is the estimated slope for the regression line and $\hat{\alpha}$ an estimate of α, e.g. via jML].

U.S. data

Method	s.–par.	jML	l.–sq.
$\hat{\gamma}$	1.14	1.10	1.11

Japan data

Method	s.–par.	jML	l.–sq.
$\hat{\gamma}$	1.81	1.30	1.34

Tab. 3

While for the U.S. data, all approaches give nearly the same estimate for γ, the situation is not so clear for the Japan data. The following figures show the log–likelihood functions in the semi–parametric setting as well as the graphs of the number S_n (+) of record values up to observation no. n and record times (•).

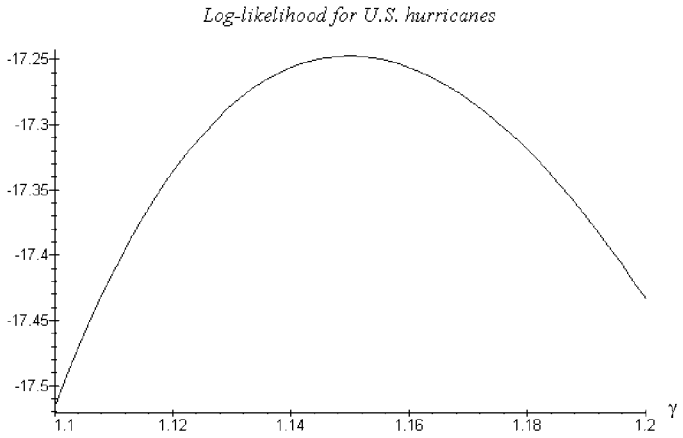

Fig. 3: graph of $L(\gamma)$

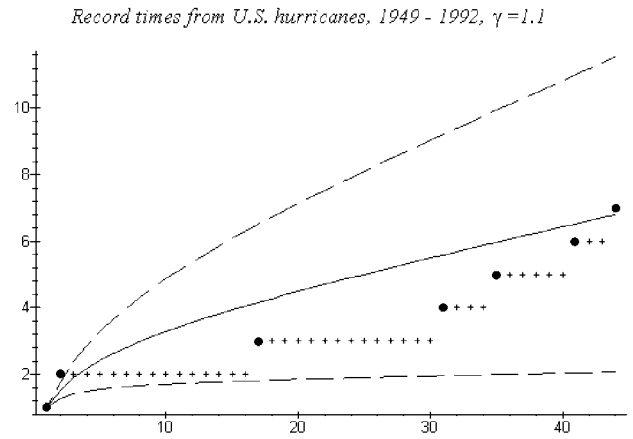

Fig. 4

solid line: $E(S_n)$

dashed lines: $E(S_n) \pm \sigma(S_n) \approx \dfrac{\gamma-1}{\gamma} n \pm \dfrac{1}{\gamma}\sqrt{(\gamma-1)n}$

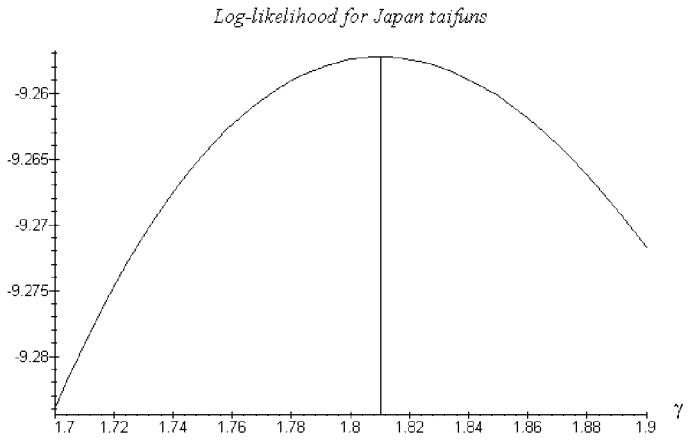

Fig. 5: graph of $L(\gamma)$

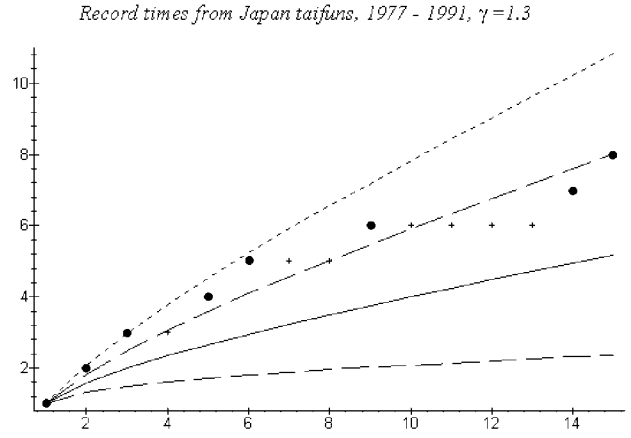

Fig. 6

solid line: $E(S_n)$

dashed lines: $E(S_n) \pm \sigma(S_n) \approx \dfrac{\gamma-1}{\gamma} n \pm \dfrac{1}{\gamma}\sqrt{(\gamma-1)n}$

dotted line: $E(S_n) + 2\sigma(S_n)$

The last figure shows that the jML estimate $\hat{\gamma} = 1.3$ for the Japan data is still acceptable within our statistical framework.

The following figures show graphs of simulated data from Fréchet type trend models with the jML estimes $\hat{\gamma}$ from Table 3, together with the corresponding original data sets. Note that the corresponding jML estimates for the scale parameter α for the Fréchet claim distributions are given by $\hat{\alpha} = 1.06$ for the U.S. data and $\hat{\alpha} = 0.9$ for

the Japan data, indicating that the trend–corrected claims data are coming from "dangerous" distributions (Fréchet distributions with $\alpha \leq 1$ do not possess a finite expectation, for $\alpha \leq 2$ no variance exists).

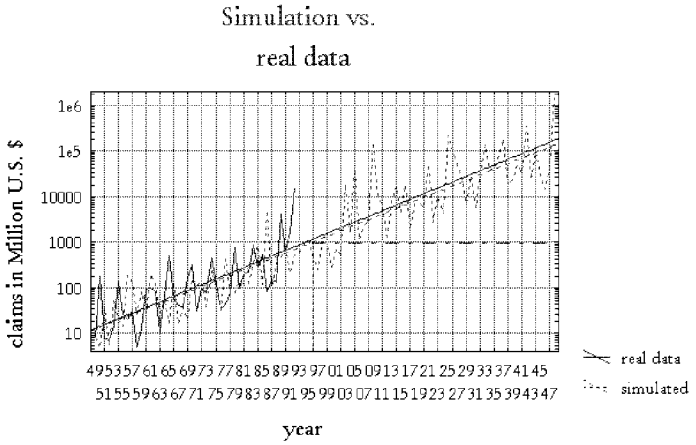

Fig. 7: U.S. data

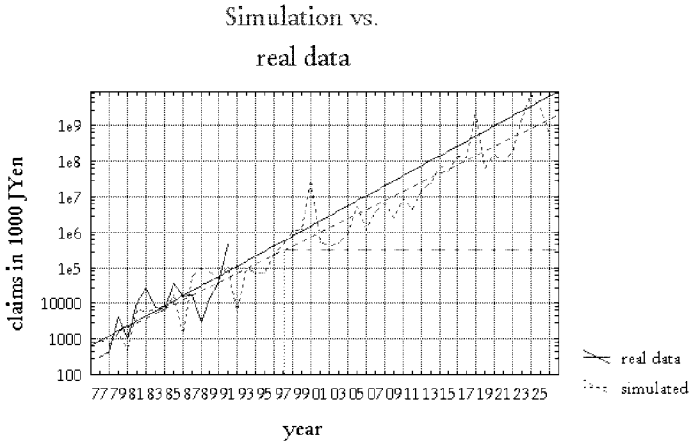

Fig. 8: Japan data

It is interesting to see that in the simulated data the next "big" catastrophe (due to the occurence of record values) is predicted for the year 2003 in the U.S. case [comparable in size to the damage caused by hurricane *Andrew* in 1992] and for the year 1998 in the Japan case.

We should like to point out here that a similar analysis of the U.S.hurricane data for only the first 22 years from 1949 to 1970 results in practically *the same* para-

meter estimates for γ and α as those given in Table 3. **This means that under the assumptions of model validity, a catastrophe like the one due to hurricane Andrew could have been forcasted for the years between 1987 and 2003 already in the year 1970!**

In order to test the goodness–of–fit for the Fréchet trend model, the data [original and simulated] were detrended and transformed with the inverse (estimated) cumulative distribution function, to give approximately uniformly distributed data in the case of model validity. The tests were performed with the STATSTICA© module *Nonparametric Statistics*. Note, however, that due to the estimation procedure before transformation of the data the rejection levels for the Kolmogorov–Smirnov test as well as the χ^2–goodness–of–fit test are lower than in the standard case.

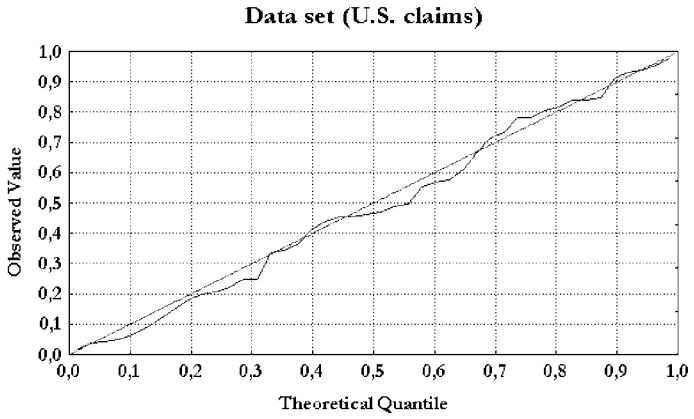

Fig. 9: Q–Q–plot for transformed data

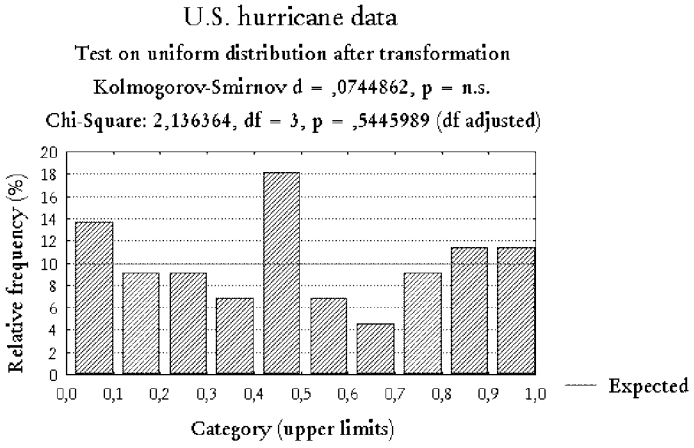

Fig. 10: U.S. data, original

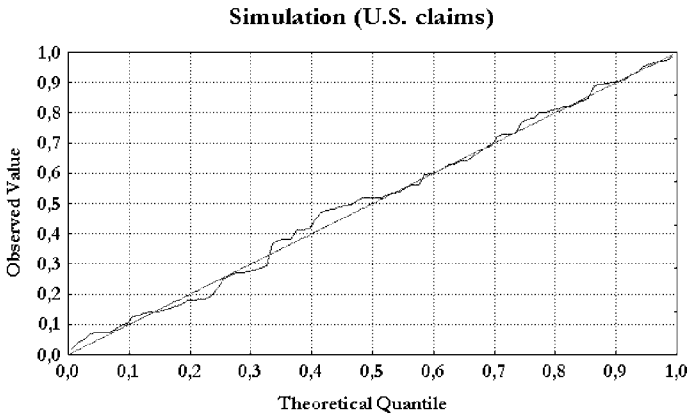

Fig. 10: Q-Q–plot for transformed data

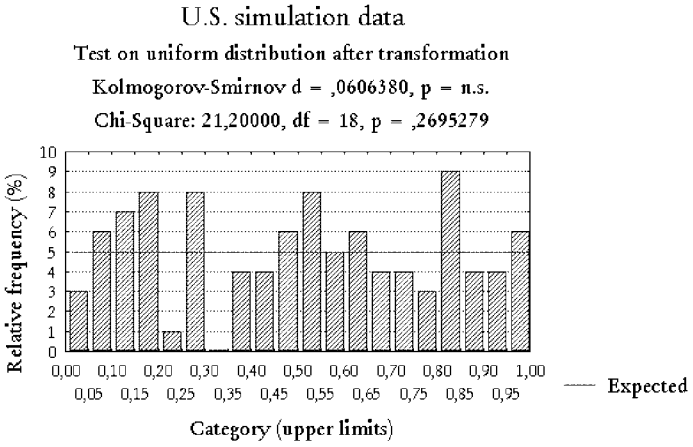

Fig. 11: U.S. data, simulation

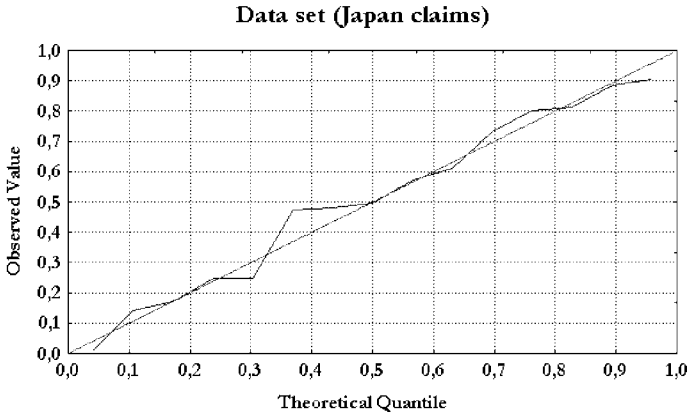

Fig. 12: Q-Q–plot for transformed data

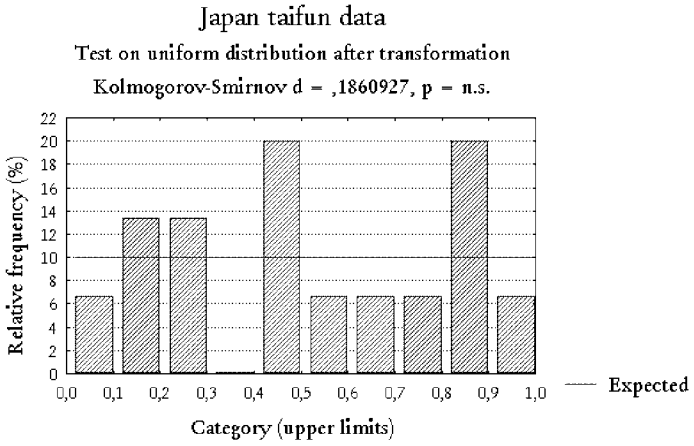

Fig. 13: Japan data, original

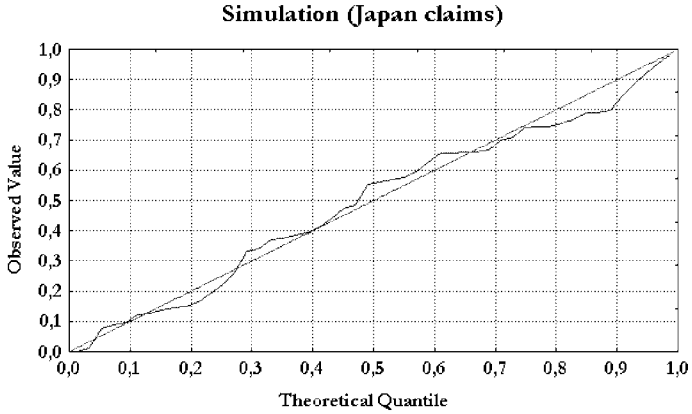

Fig. 14: Q–Q–plot for transformed data

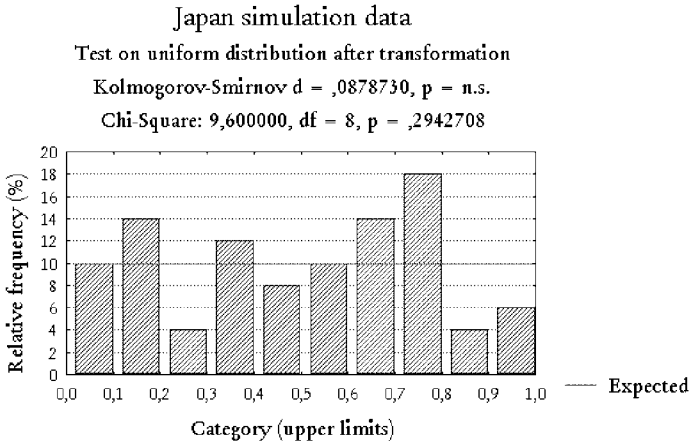

Fig. 15: Japan data, simulation

5. Conclusions. For the storm events analysed here it turns out that in the past a yearly increase in claimsize expectation of about $\left(\gamma^{1/\alpha} - 1\right) \times 100 = 9.4\%$ can be stated quite definitely for the U.S., whereas for Japan, a yearly increase in the claimsize median [the expectation does not exist for $\alpha = 0.9$] of about $\left(\gamma^{1/\alpha} - 1\right) = 33.2\%$ can be cautiously assumed [based on the jML estimate for γ]. It is of course questionable to what extend these rates of increase are due to inflation and other economic influences, however the calculated magnitude is large enough to admit that besides these factors also climatic change could perhaps play an important role here.

The foregoing analysis suggests that it might be promising to apply the combined parametric and semi–parametric method outlined above also to more general catastrophe claims data.

Acknowledgement. We would like to express our gratitude to the *Verein zur Förderung der Versicherungswissenschaften in Hamburg e.V.* for financial support, by which in particular the use of the necessary mathematical and statistical software was enabled, and to E. Sperling for providing the Japan taifun data.

6. References.

[1] Catastrophe Reinsurance Newsletter (1993) No. 2, p.8.

[2] N. Bingham, C.M. Goldie, J. Teugels (1987): *Regular Variation.* Camb. Univ. Press, Cambridge.

[3] K. Borovkov, D. Pfeifer (1995): On record indices and record times. J. Stat. Plann. Inf. 45, 65 – 79.

[4] J. Diersen, G. Trenkler (1996): Records tests for trend in location. Statistics 28, 1 – 12.

[5] M.R. Leadbetter, G. Lindgren, H. Rootzén (1983): *Extremes and Related Properties of Random Sequences and Processes.* Springer, N.Y.

[6] V.B. Nevzorov (1988): Records. Th. Probab. Appl. 32, 201 – 228.

[7] sigma (1993), Vol. 2.

[8] SCOR Notes, April 1993. International Prize in Actuarial Science: Catastrophe risks.

[9] E. Sperling (1993), Hannover Rück Eisen und Stahl. Personal Communication.

[10] M.C.K. Yang (1975): On the distribution of inter–record times in an increasing population. J.Appl.Prob. 12, 148 – 154.