Record statistics in time series with drift: theory and applications

Joachim Krug
Institute of Theoretical Physics, University of Cologne

- Introduction: What are records, and why do we care?
- Record statistics beyond i.i.d. RV's: The linear drift model
- Application: Record-breaking temperatures and global warming
- Correlations between record events

Joint work with Jasper Franke and Gregor Wergen

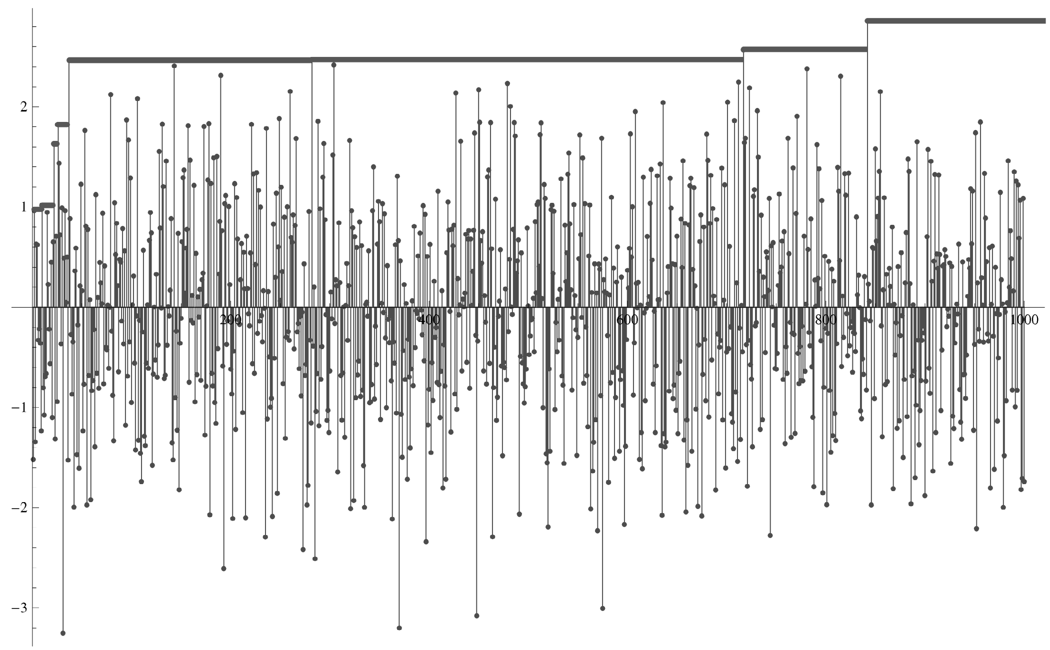
Records in popular culture

9.11.2006:
1188 Parisians kissing at La Defense

http://www.guinnessworldrecords.com/gwrday/frenchkiss.aspx

Basic facts about records I

- A record is an entry in a sequence of random variables (RV's) X_n which is larger (upper record) or smaller (lower records) than all previous entries



- If the RV's are independent and identically distributed (i.i.d.), the probability for a record at time n is $P_n = 1/n$ by symmetry

- This result is universal, i.e. independent of the underlying distribution (provided it is continuous)

Basic facts about records II

N. Glick, Am. Math. Mon. **85**, 2 (1978)

- The expected number of records up to time n is

$$\langle R_n \rangle = \sum_{k=1}^{n} \frac{1}{k} = \ln(n) + \gamma + \mathcal{O}(1/n)$$

where $\gamma \approx 0.5772156649....$ is the Euler-Mascheroni constant

- Record events are independent: The sequence of records is a Bernoulli process with success probability P_n, which converges to a Poisson process in logarithmic time for large n

- In particular, the variance of the number of records is

$$\langle (R_n - \langle R_n \rangle)^2 \rangle = \sum_{k=1}^{n} \left(\frac{1}{k} - \frac{1}{k^2} \right) = \ln(n) + \gamma - \frac{\pi^2}{6} + \mathcal{O}(1/n)$$

Beyond the i.i.d. model

Records in growing populations

M.C.K. Yang, J. Appl. Prob. **12**, 148 (1975)

- Motivation: Olympic records occur at an essentially constant (non-decreasing) rate

- Model: At each time n a new "generation" of N_n i.i.d. RV's becomes available simultaneously. By symmetry, the probability of a new record at time n is then

$$P_n = \frac{N_n}{\sum_{k=1}^{n} N_k}$$

- For an exponentially growing population, $N_n = a^n$, this yields

$$P_n = \frac{a^n(a-1)}{a(a^n-1)} \to \frac{a-1}{a} \text{ for } n \to \infty.$$

- The growth of the world population is insufficient to explain the occurrence rate of Olympic records under this model.

Records from broadening distributions

JK, JSTAT (2007) P07001

- Let X_n be drawn from $p_n(x) = n^{-\alpha} f(x/n^{\alpha})$ with $\alpha > 0$

- Asymptotic growth of the number of records depends on the universality class of f in the sense of extreme value statistics.

  Fréchet class: $f(x) \sim x^{-(\mu+1)} \quad \Rightarrow \quad \langle R_n \rangle \approx (1 + \alpha\mu) \ln(n)$

  Gumbel class: $f(x) \sim \exp[-x^{\beta}] \quad \Rightarrow \quad \langle R_n \rangle \sim \alpha \ln^2(n)$

  Weibull class: $f(x) \sim (x_{\max} - x)^{\delta-1}, \delta > 0 \quad \Rightarrow \quad \langle R_n \rangle \sim (\alpha^{\delta} n)^{1/(\delta+1)}$

- Effect of broadening is stronger for fast decaying tails

- Broadening generically induces correlations between record events (see later)

Records of random walks

S.N. Majumdar & R.M. Ziff, PRL **101**, 050601 (2008)

- Let X_n be an unbiased random walk:

$$X_n = \sum_{k=1}^{n} \eta_k$$

with i.i.d. RV's η_k drawn from a symmetric, continuous distribution $\phi(\eta)$

- The probability of having m records in n steps is given by

$$P(m,n) = \binom{2n-m+1}{n} 2^{-2n+m-1} \rightarrow \frac{1}{\sqrt{\pi n}} \exp[-m^2/4n]$$

- Mean number of records: $\langle R_n \rangle \approx \sqrt{4n/\pi}$

- This result does not require $\phi(\eta)$ to have finite variance

See also poster by Gregor Wergen!

The linear drift model

R. Ballerini & S. Resnick (1985); J. Franke, G. Wergen, JK, JSTAT (2010) P10013

- Let $X_n = Y_n + vn$ with i.i.d. RV's Y_n and a drift speed $v > 0$

- Let Y_n have probability density $p(y)$ and probability distribution function $q(x) = \int^x dy\, p(y)$. Then

$$P_n(v) = \int dx_n\, p(x_n - vn) \prod_{k=1}^{n-1} q(x_n - vk) = \int dx\, p(x) \prod_{k=1}^{n-1} q(x + vk)$$

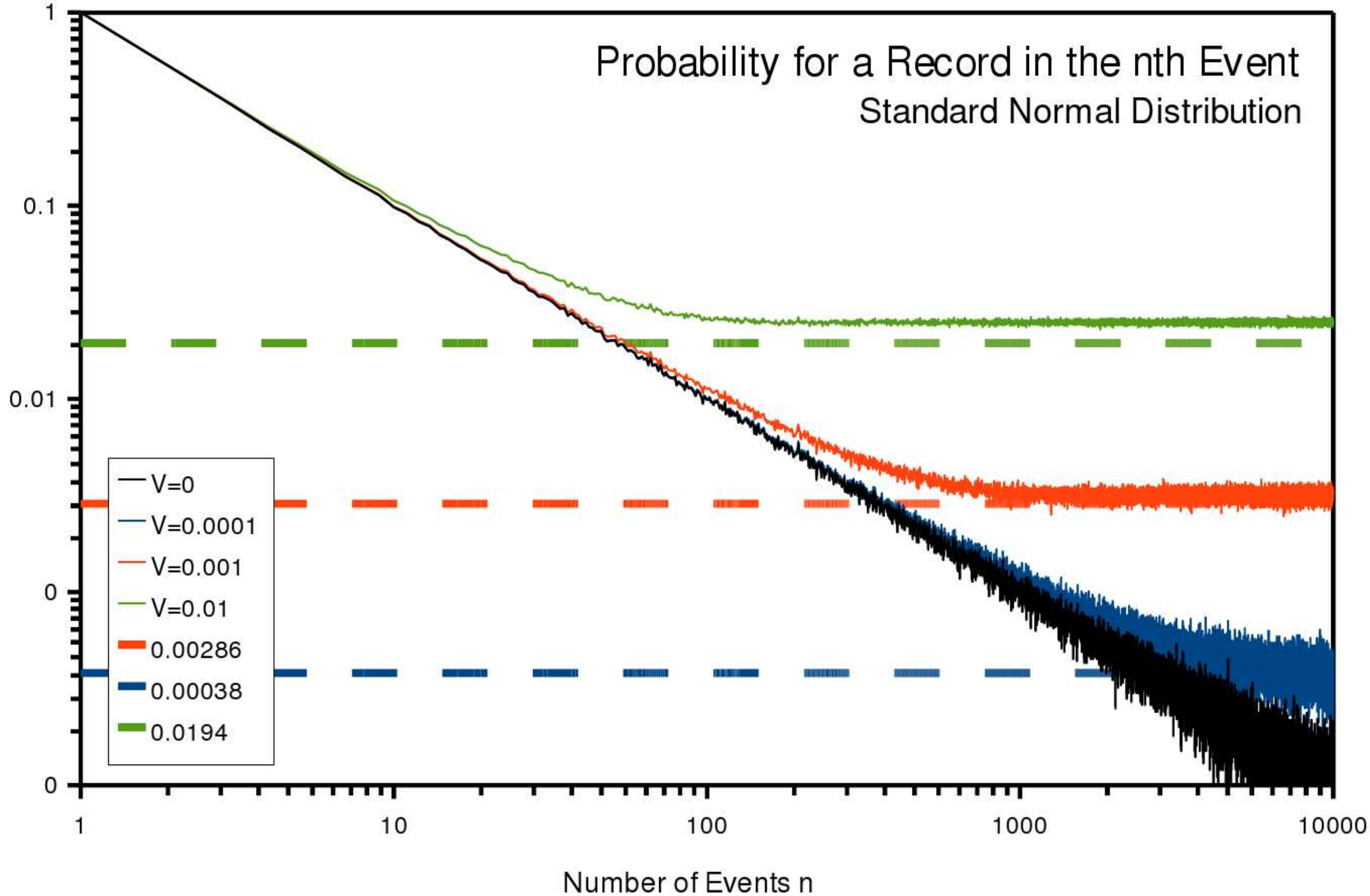
- Limiting record rate

$$\lim_{n\to\infty} P_n(v) \equiv P_\infty(v) > 0$$

for $v > 0$ provided $p(x)$ has a finite first moment.

- Model also appears in the context of elastic manifolds in random media (Le Doussal & Wiese, PRE 2009) and evolutionary pathways in random fitness landscapes (Franke et al., PLoS Comp. Biol., in press)

Simulation of the record rate for Gaussian RV's



Crossover time scale $n^*(v) \longrightarrow \infty$ for $v \longrightarrow 0$

An exactly solvable case

- For the Gumbel distribution $q(x) = \exp[-e^{-x/b}]$

$$\prod_{k=1}^{n-1} q(x - vk) = \exp[-e^{-x/b} \sum_{k=1}^{n-1} e^{-vk/b}] = q(x)^{\alpha_n} \quad \text{with} \quad \alpha_n = \sum_{k=1}^{n-1} (e^{-v/b})^k$$

$$\Rightarrow \; P_n(v) = \int_0^1 dq q^{\alpha_n} = \frac{1}{\alpha_n + 1} = \frac{1 - e^{-v/b}}{1 - e^{-nv/b}}$$

- Limiting record rate for $v > 0$ is $P_\infty(v) = 1 - e^{-v/b}$

- For $v < 0$ the record rate decays exponentially with n and the expected number of records remains finite.

- **Conjecture**: The expected number of records is finite for $v < 0$ for any distribution with a finite mean

- Gumbel distribution is the only case in which the stochastic independence of record events of the i.i.d. model is preserved.

Ordering probability

J. Franke, G. Wergen, JK, JSTAT (2010) P10013

- What is the probability Π_N that all N events are upper records, i.e. that

$$X_1 < X_2 < \ldots < X_N?$$

- For i.i.d. RV's we have $\Pi_N = \prod_{k=1}^{N} \frac{1}{k} = \frac{1}{N!}$

- For the linear drift model with Gumbel-distributed i.i.d. part one finds

$$\Pi_N = \frac{(1-e^{-\theta})^N}{\prod_{k=1}^{N}(1-e^{-\theta k})} \approx \sqrt{\frac{\theta}{2\pi}}\, e^{\pi^2/6\theta}\, (1-e^{-\theta})^N \text{ for } N \to \infty$$

with $\theta = v/b$

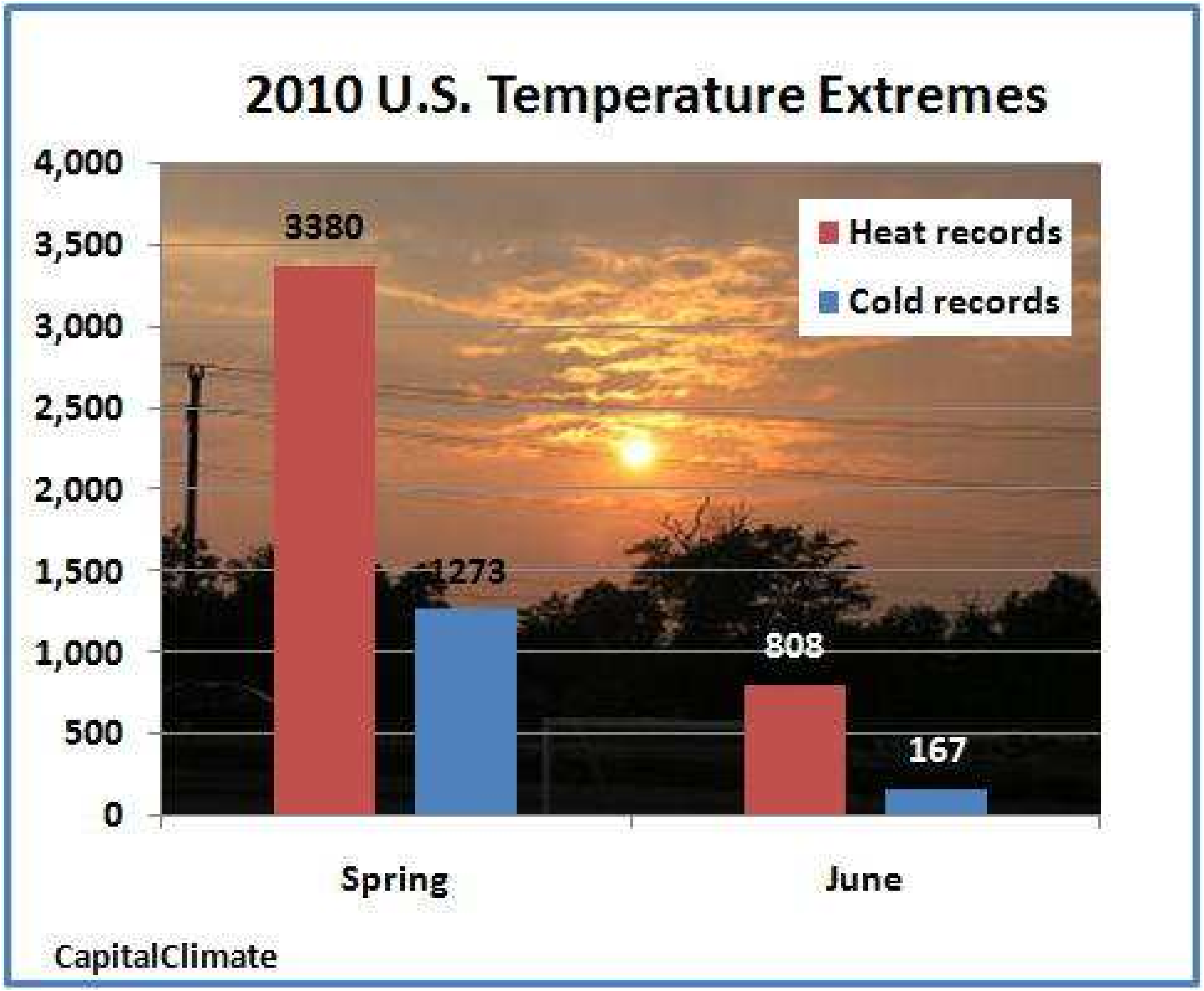
- **Conjecture**: The ordering probability Π_N decays exponentially (rather than factorially) with N for $v > 0$ and any distribution with a finite mean

Application to global warming

The 2010 summer heat wave

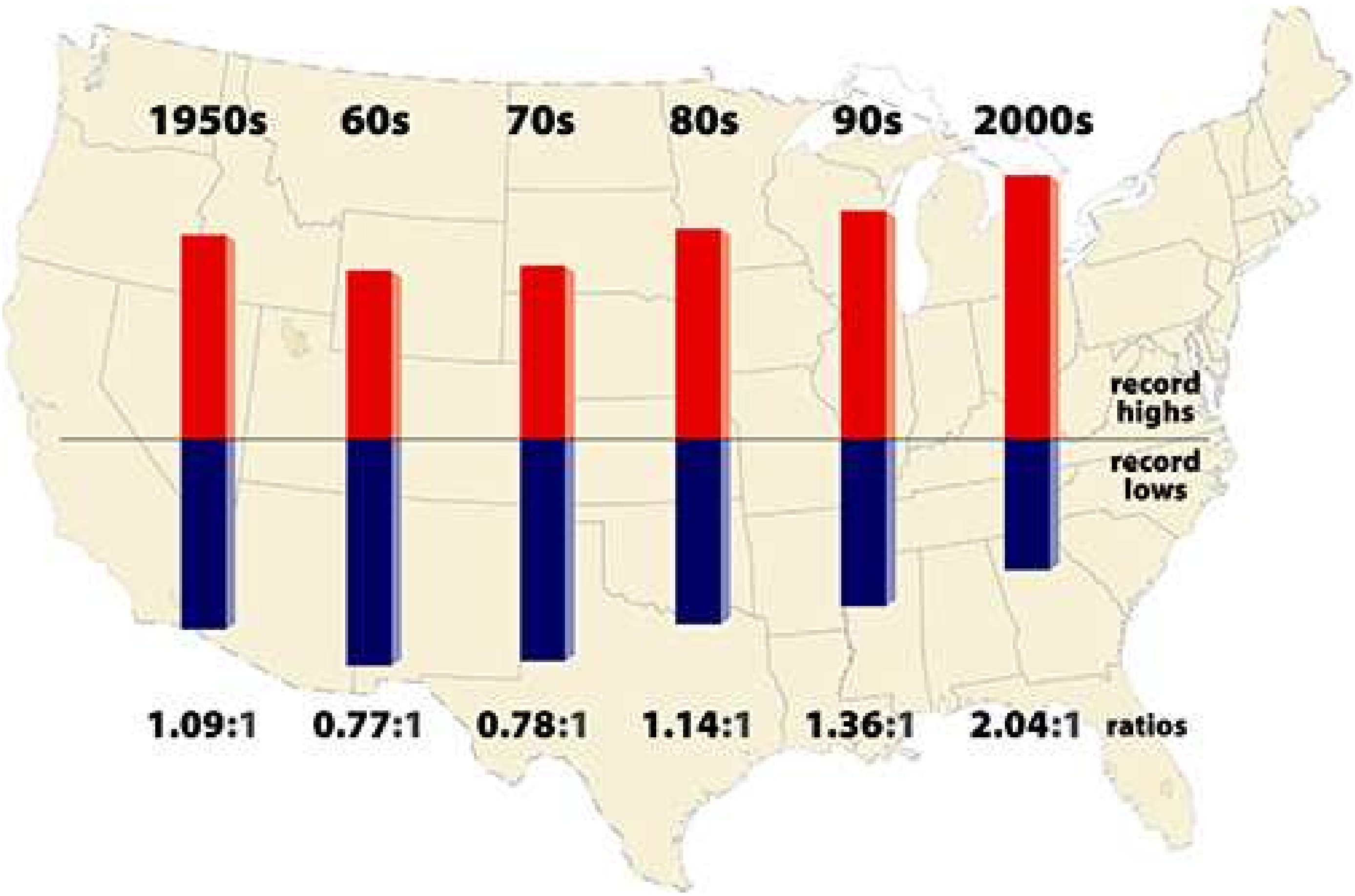
http://www.spiegel.de/

The 2010 summer heat wave



http://climateprogress.org/2010/07/05/heat-wave-global-warming/

Temperature records in the USA



http://www.ucar.edu/news/releases/2009/maxmin.jsp

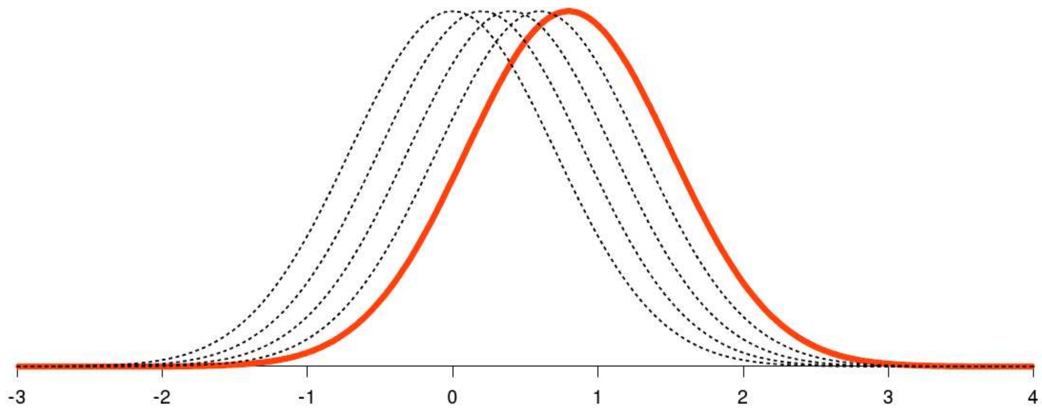
based on G.A. Meehl et al., Geophys. Res. Lett. **36** (2009) L23701

Record-breaking temperatures and global warming

R.E. Benestad (2003); S. Redner & M.R. Petersen (2006)

- Question: Does global warming significantly increase the occurrence of record-breaking high daily temperatures?

- Model: The temperature on a given calendar day of the year is an independent Gaussian RV with constant standard deviation σ and a mean that increases at speed v



- Typical values: $v \approx 0.03^{\mathrm{o}}$C/yr, $\sigma \approx 3.5^{\mathrm{o}}$C $\Rightarrow v/\sigma \ll 1$

Expansion for small drift speed

J. Franke, G. Wergen, JK, JSTAT (2010) P10013

- We want to compute the record rate $P_n(v) = \int dx\, p(x) \prod_{k=1}^{n-1} q(x+vk)$
- To leading order in v we have $q(x+vk) \approx q(x) + vkp(x)$

$$\Rightarrow P_n \approx \int dx\, p(x) q(x)^{n-1} + \frac{vn(n-1)}{2} \int dx\, p(x)^2 q(x)^{n-2} = \frac{1}{n} + vI_n$$

with $I_n = \frac{n(n-1)}{2} \int dx\, p(x)^2 q(x)^{n-2}$

- Asymptotic behavior of I_n depends on the universality class of p:

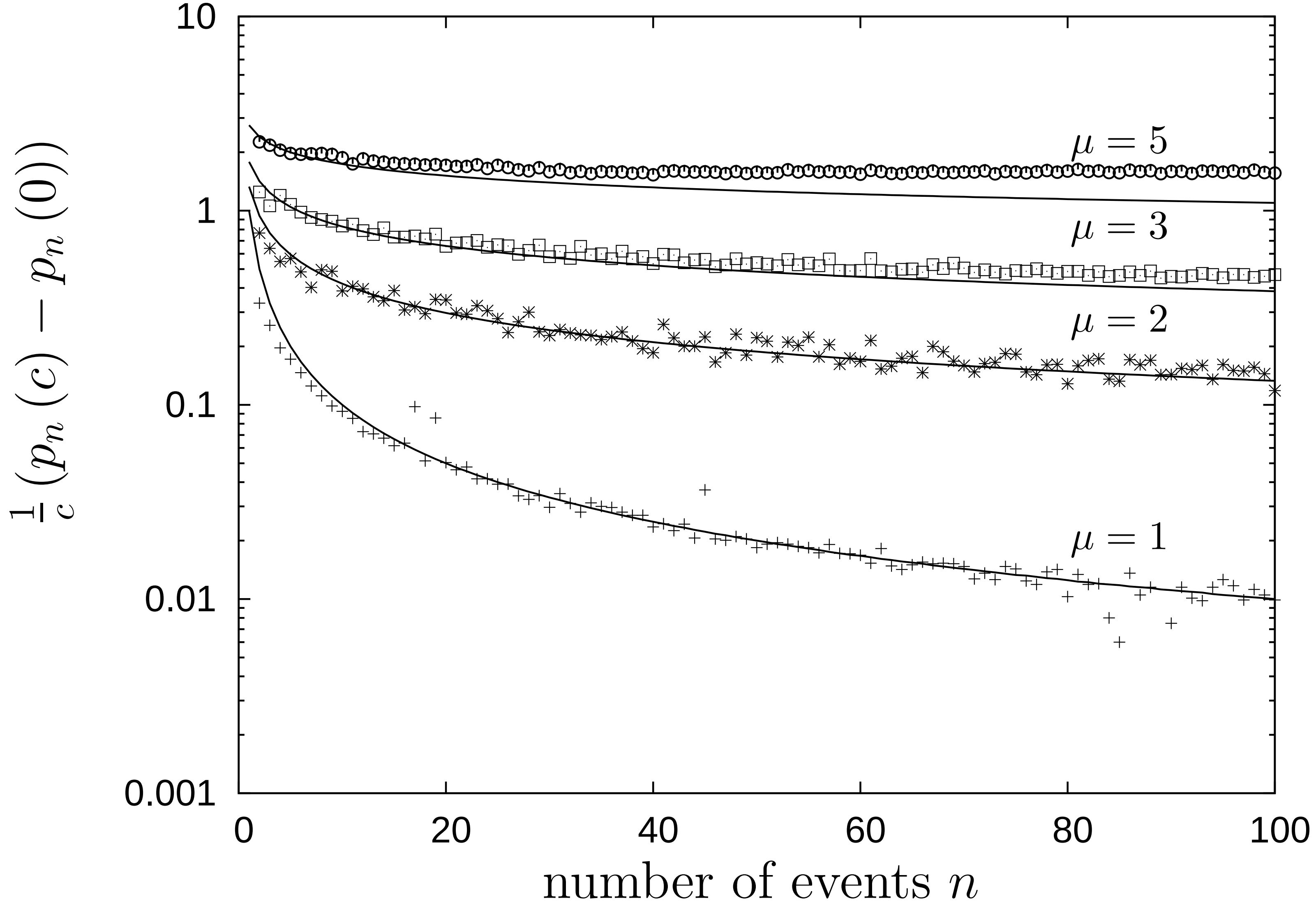
Fréchet class: $p(x) \sim x^{-(\mu+1)} \quad \Rightarrow \quad I_n \sim n^{-1/\mu} \to 0$

Weibull class: $p(x) \sim (x_{\max} - x)^{\delta - 1}, \delta > \frac{1}{2} \quad \Rightarrow \quad I_n \sim n^{1/\delta} \to \infty$

Gumbel class: $p(x) \sim e^{-x^\beta} \Rightarrow \quad I_n \sim (\ln n)^{1-\frac{1}{\beta}}$

- **Conjecture**: Expansion is singular for Weibull distributions with $\delta < \frac{1}{2}$

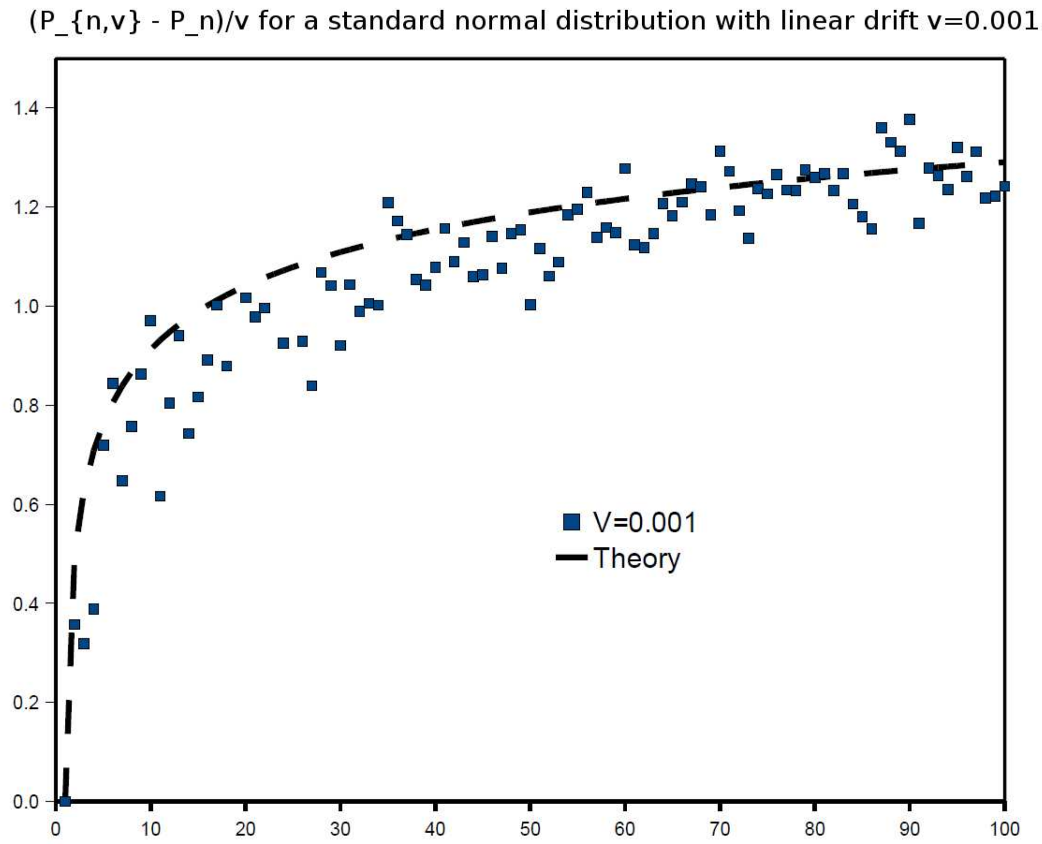
Comparison to simulations: Fréchet class



- In the Gaussian case I_n can be evaluated in closed form only for $n = 2, 3$

- A saddle point approximation for large n yields the result

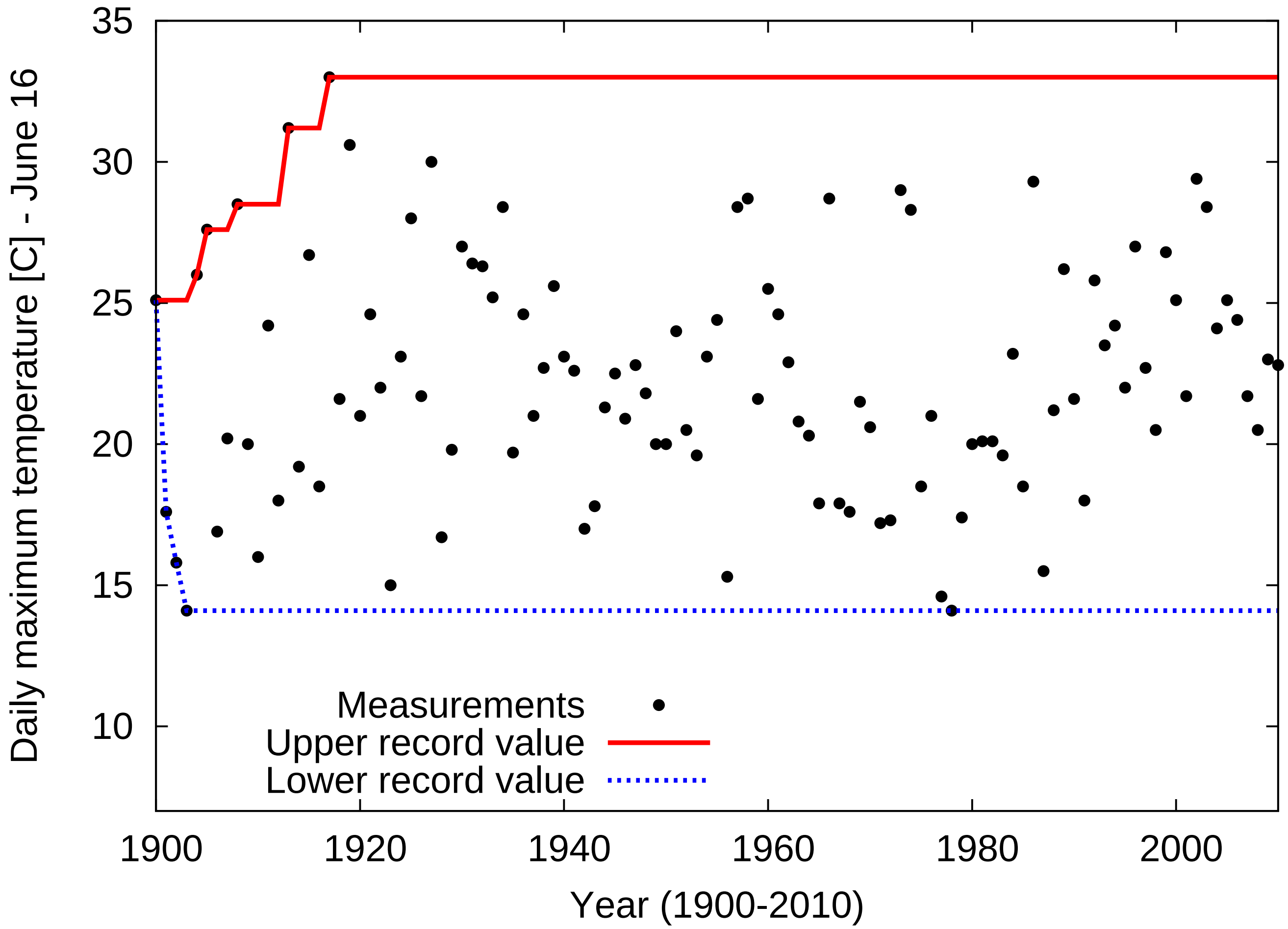
$$P_n(v) \approx \frac{1}{n} + \frac{v}{\sigma} \frac{(2\pi)^{3/2}}{e^2} \sqrt{\ln(n^2/8\pi)}$$



Analysis of temperature records

G. Wergen, JK, EPL **92** 30008 (2010)

Maximum temperature on June 16 in Parc Montsouris



Expected number of records in a stationary climate is 5.3 ± 1.9

Data sets for daily temperatures

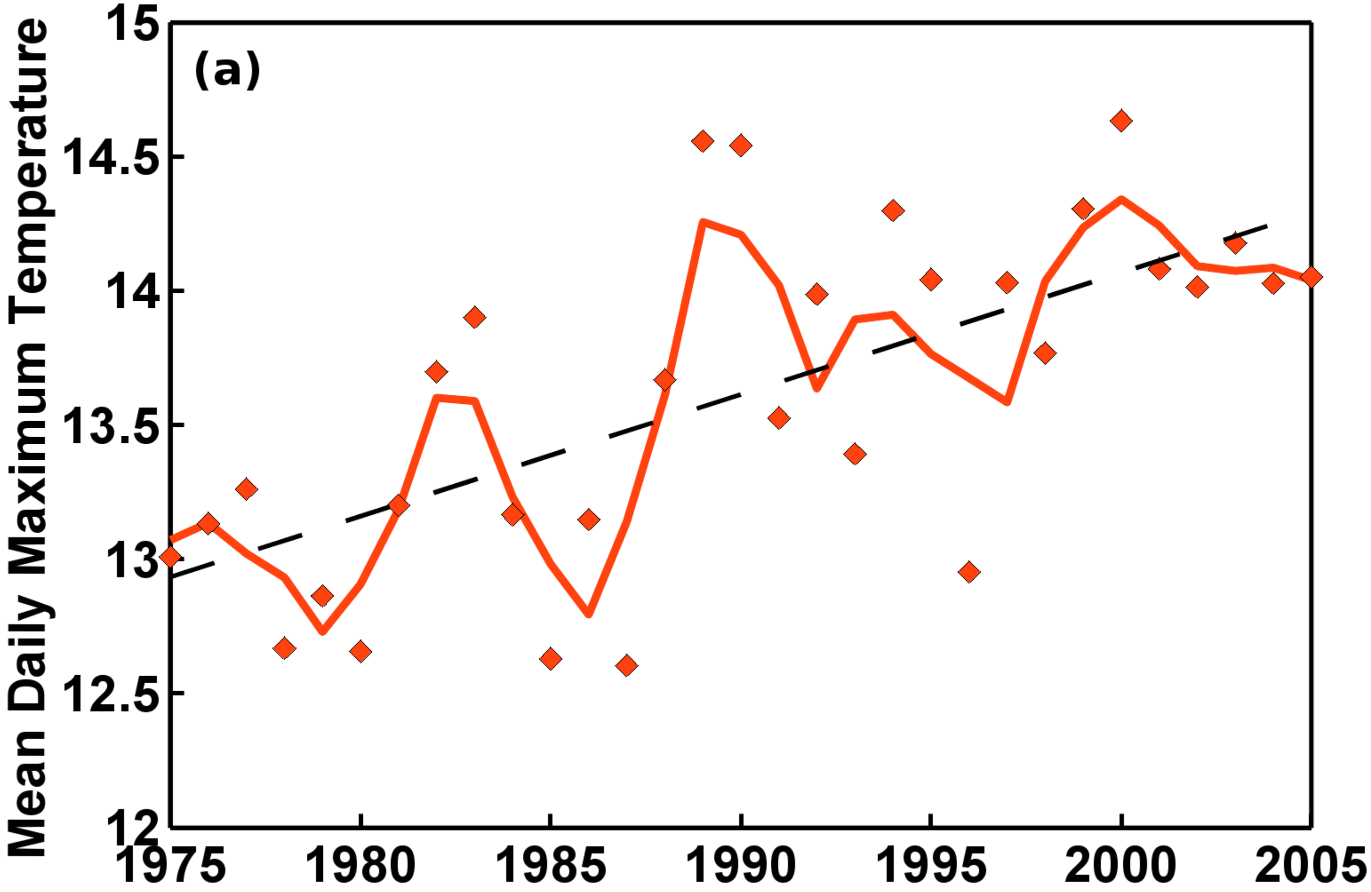
European data

- 43 stations over 100 year period 1906-2005
- 187 stations over 30 year period 1976-2005
- 30 year data: Constant warming rate $v \approx 0.047 \pm 0.003^{\circ}$C/yr, standard deviation $\sigma \approx 3.4 \pm 0.3^{\circ}$C $\Rightarrow v/\sigma \approx 0.014$

American data

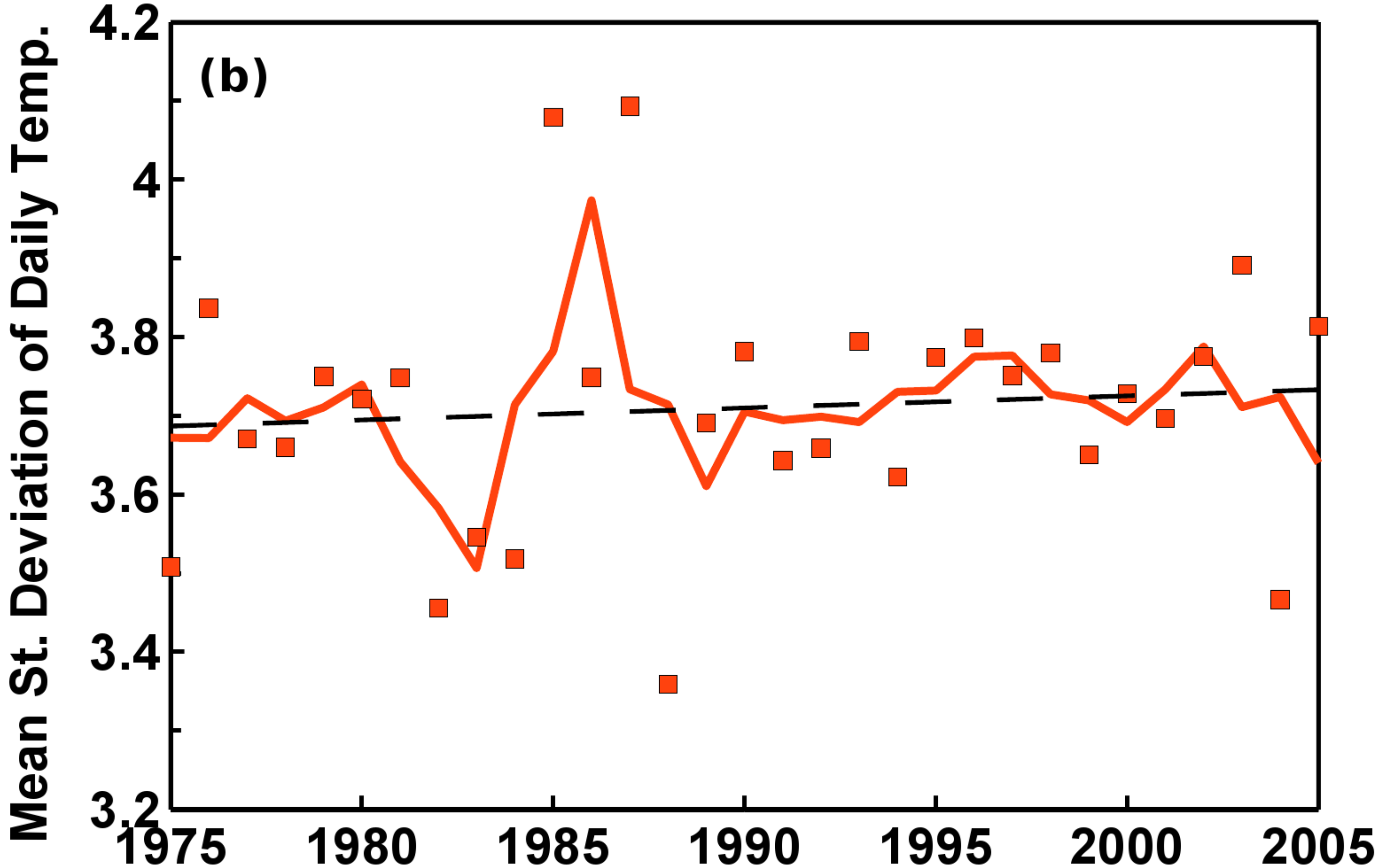
- 10 stations over 125 year period 1881-2005
- 207 stations over 30 year period 1976-2005
- Continental climate implies larger variability:
  $\sigma = 4.9 \pm 0.1^{\circ}$C, $v = 0.025 \pm 0.002^{\circ}$C/yr $\Rightarrow v/\sigma \approx 0.005$
- Significant effect of rounding to integer degrees Fahrenheit

European data: Mean daily maximum temperature

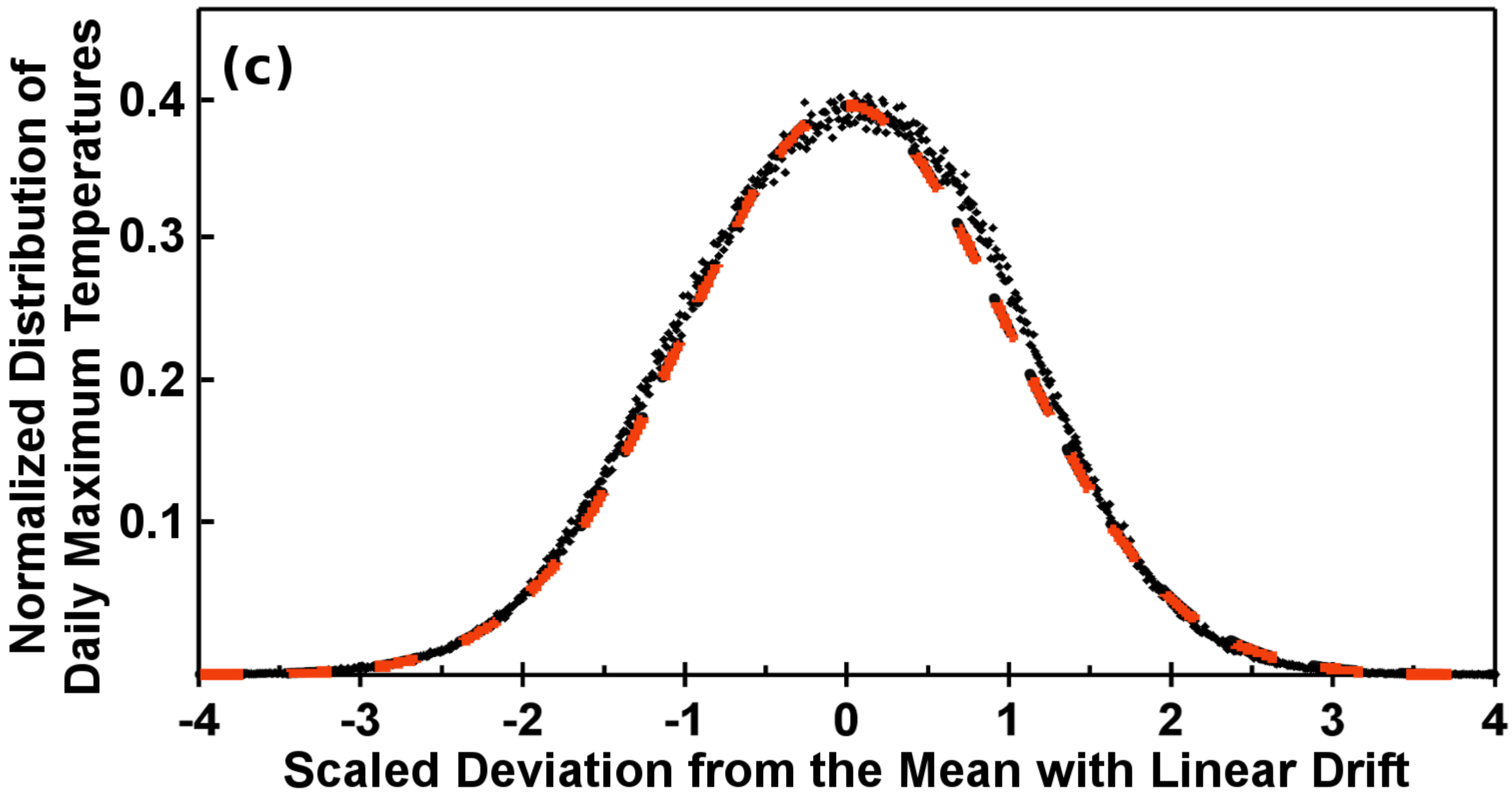


Full line: Sliding 3-year average

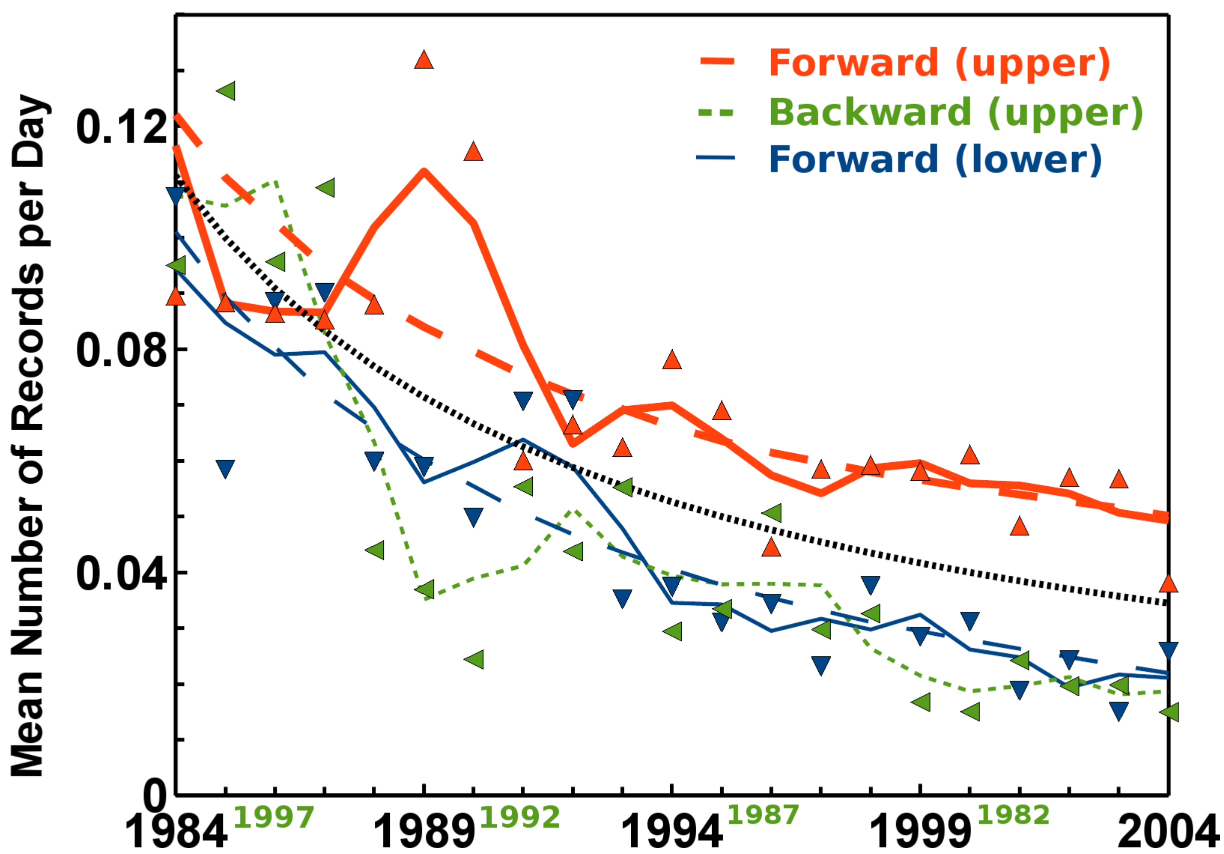
European data: No trend in the standard deviation



European data: Temperature fluctuations are Gaussian

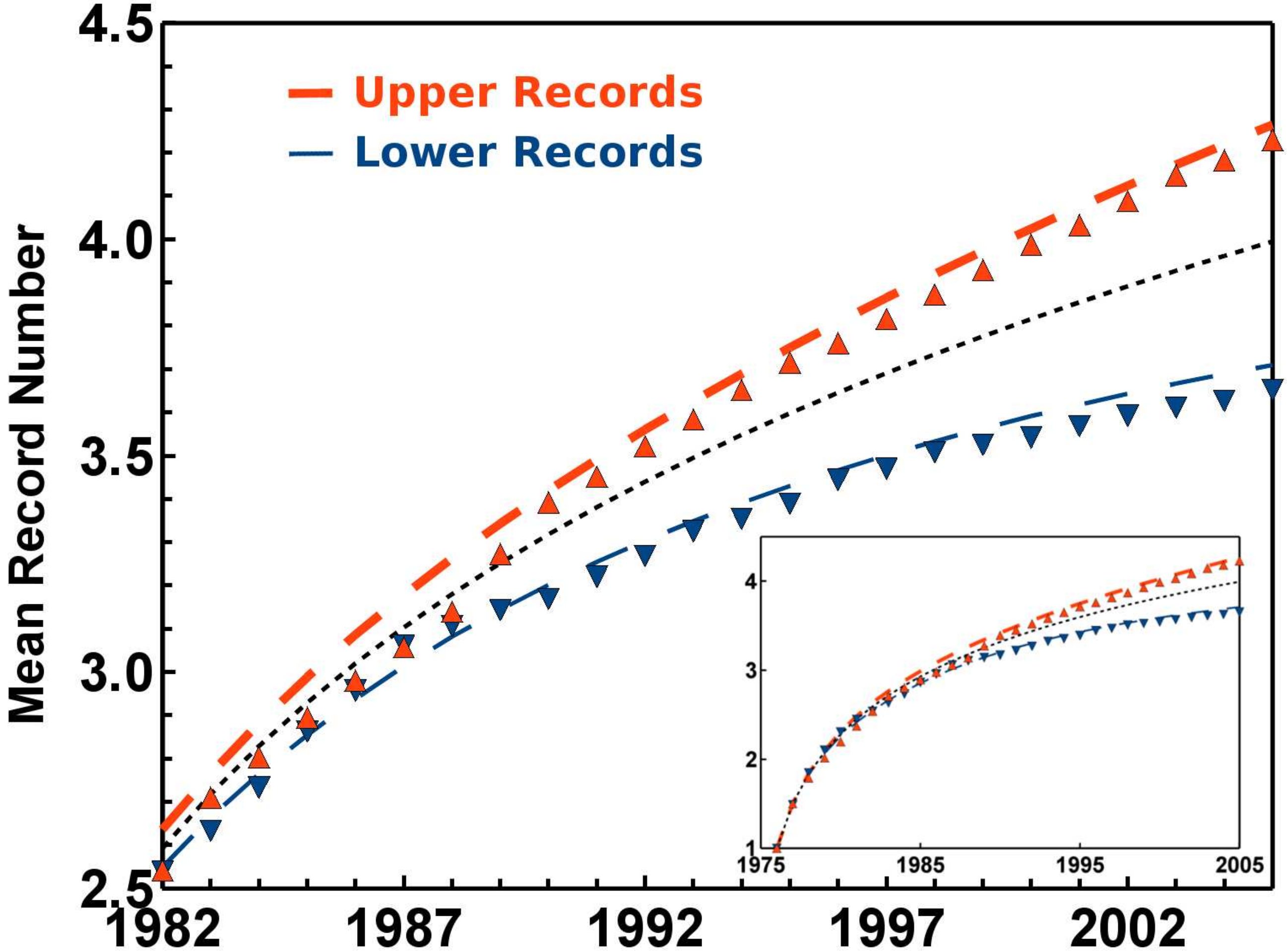


Record frequency in Europe: 1976-2005

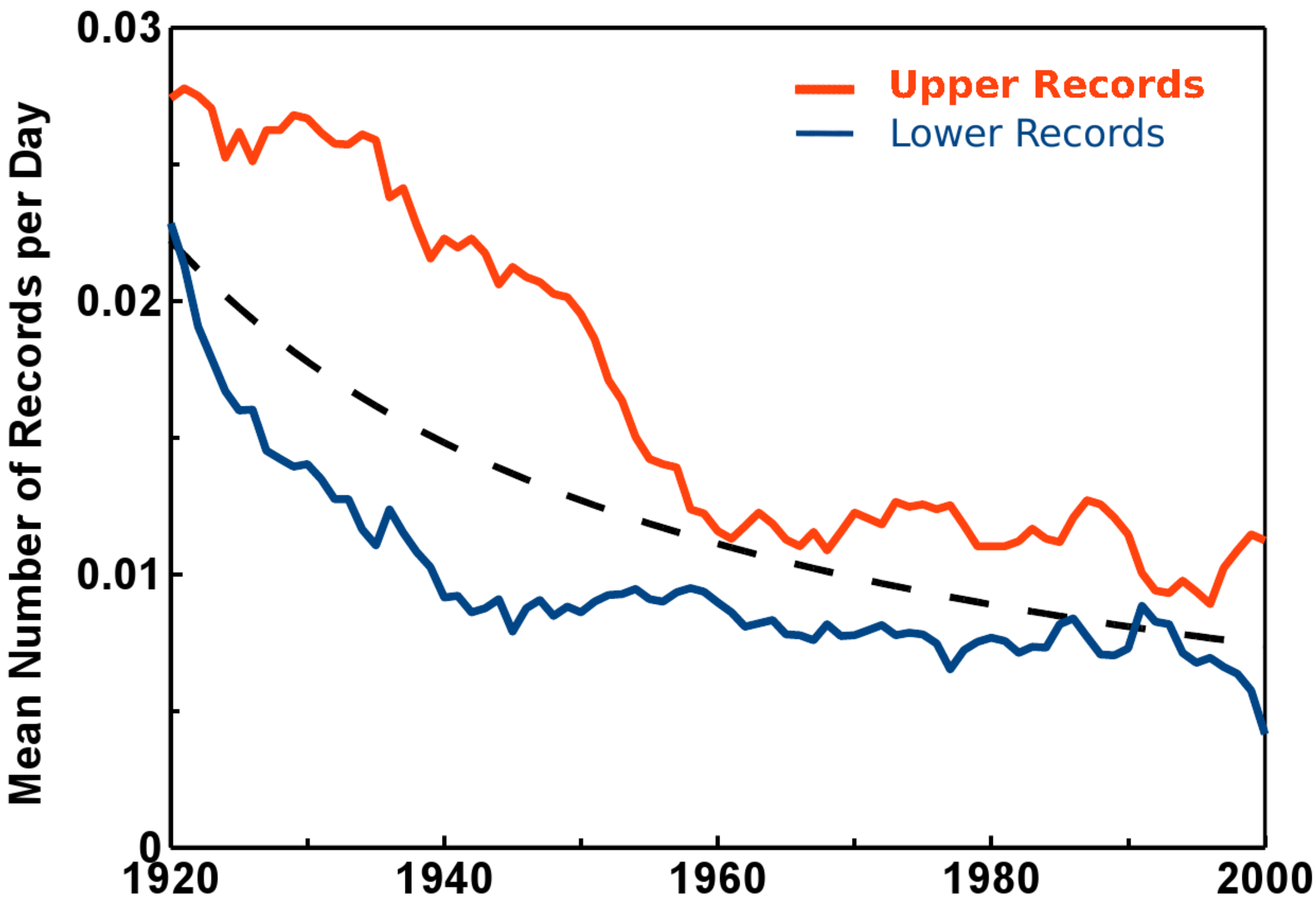


- Expected number of records in stationary climate: $\frac{365}{30} \approx 12$
- Observed record rate is increased by about 40 % $\Rightarrow$ 5 additional records

Mean record number: 1976-2005



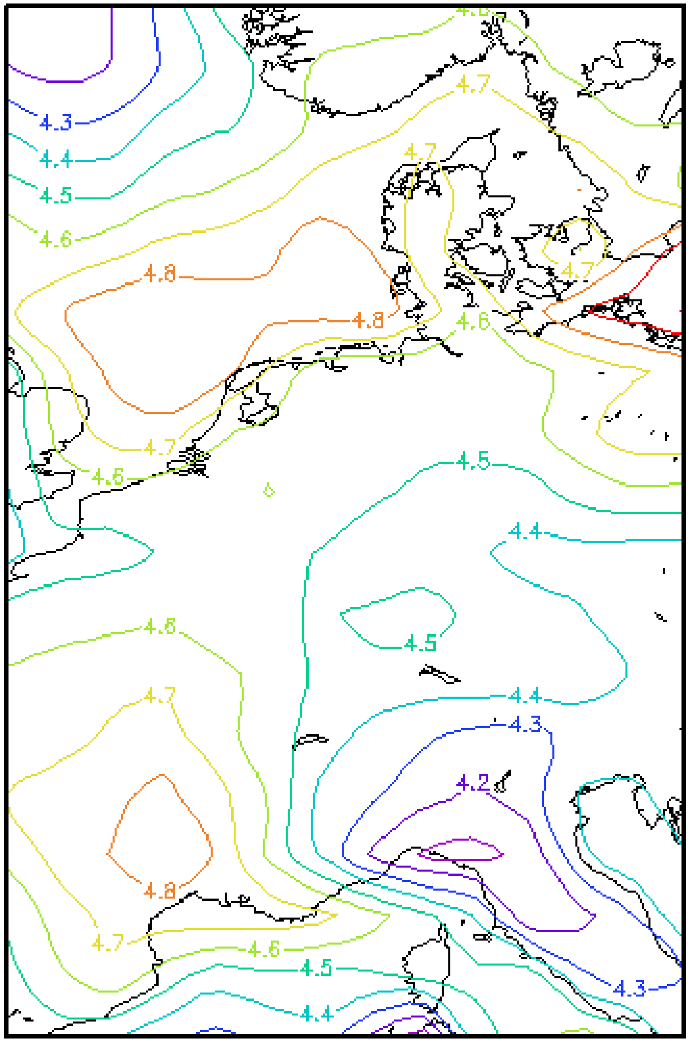
Record frequency in the US: 1881-2005

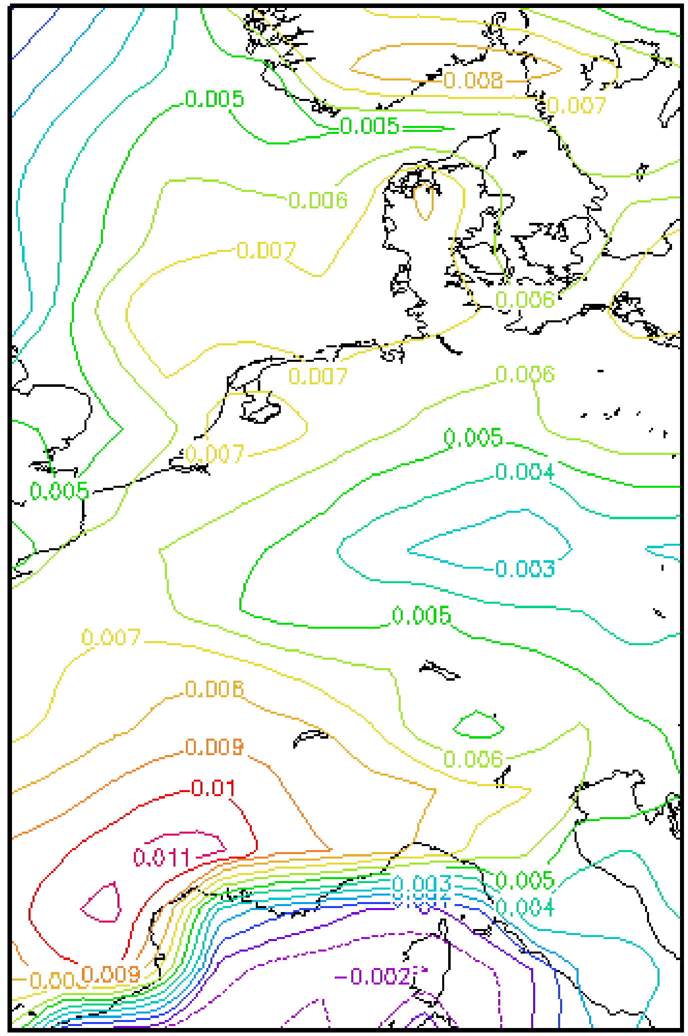


Dashed line: $P_n = (1 - d/\sigma)/n$ with discretization unit $d = 1^{\circ}$F $= (5/9)^{\circ}$C

Re-analysis data: Record maps

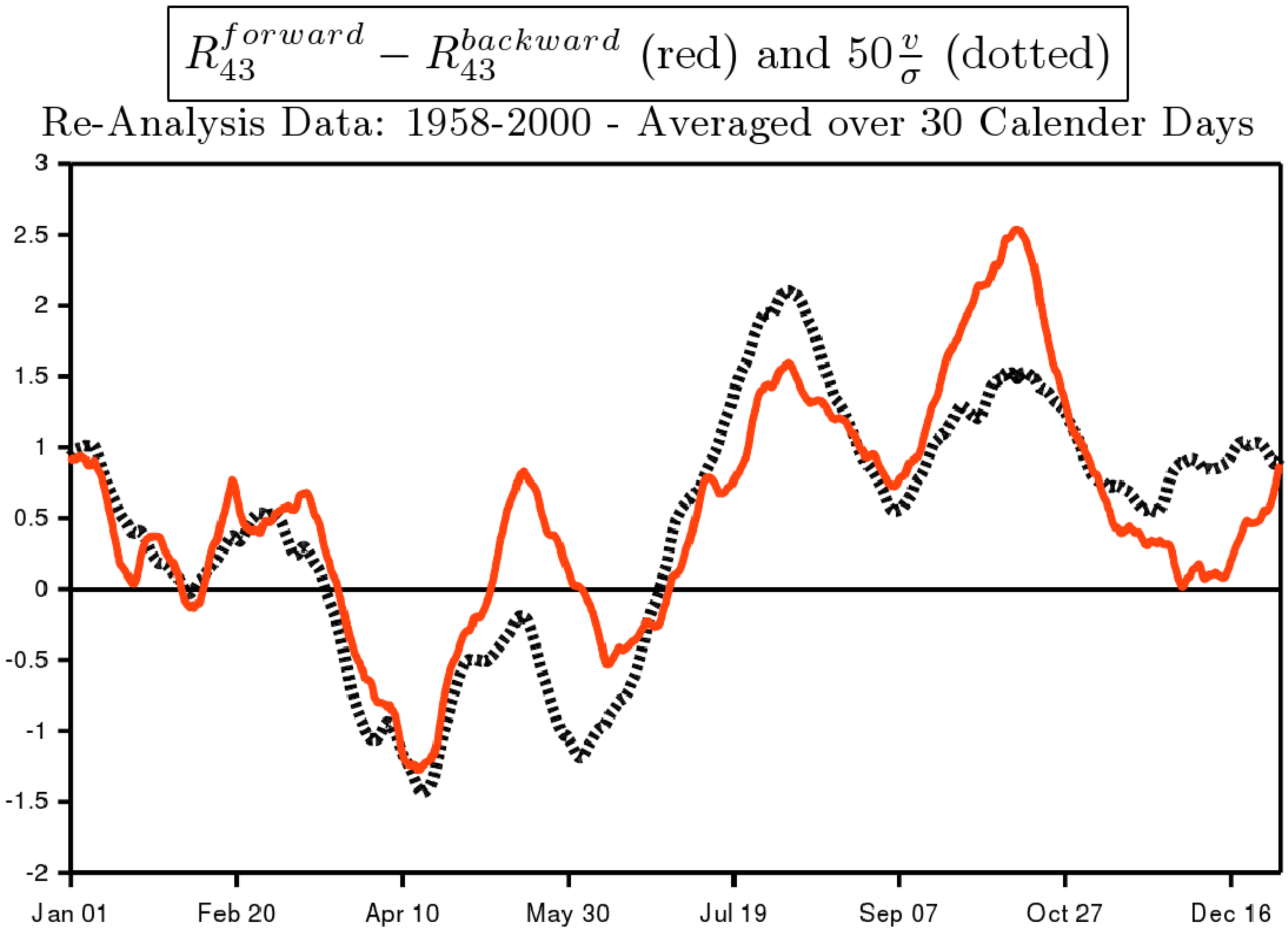
number of records 1957-2000 **normalized warming rate v/σ**





Expected record number in a stationary climate is 4.36

Re-analysis data: Seasonal variation



A record-based test of changing temperature variability

A. Anderson, A. Kostinski, J. Appl. Meteor. Climat. **49**, 1681 (2010)

- For a given temperature time series, consider the quantity

$$\mathscr{R} \equiv R_>^H - R_<^H + R_>^L - R_<^L$$

where $R_>^{H,L}$ is the number of high (H) and low (L) records of the forward time series and $R_<^{H,L}$ the corresponding numbers backward in time

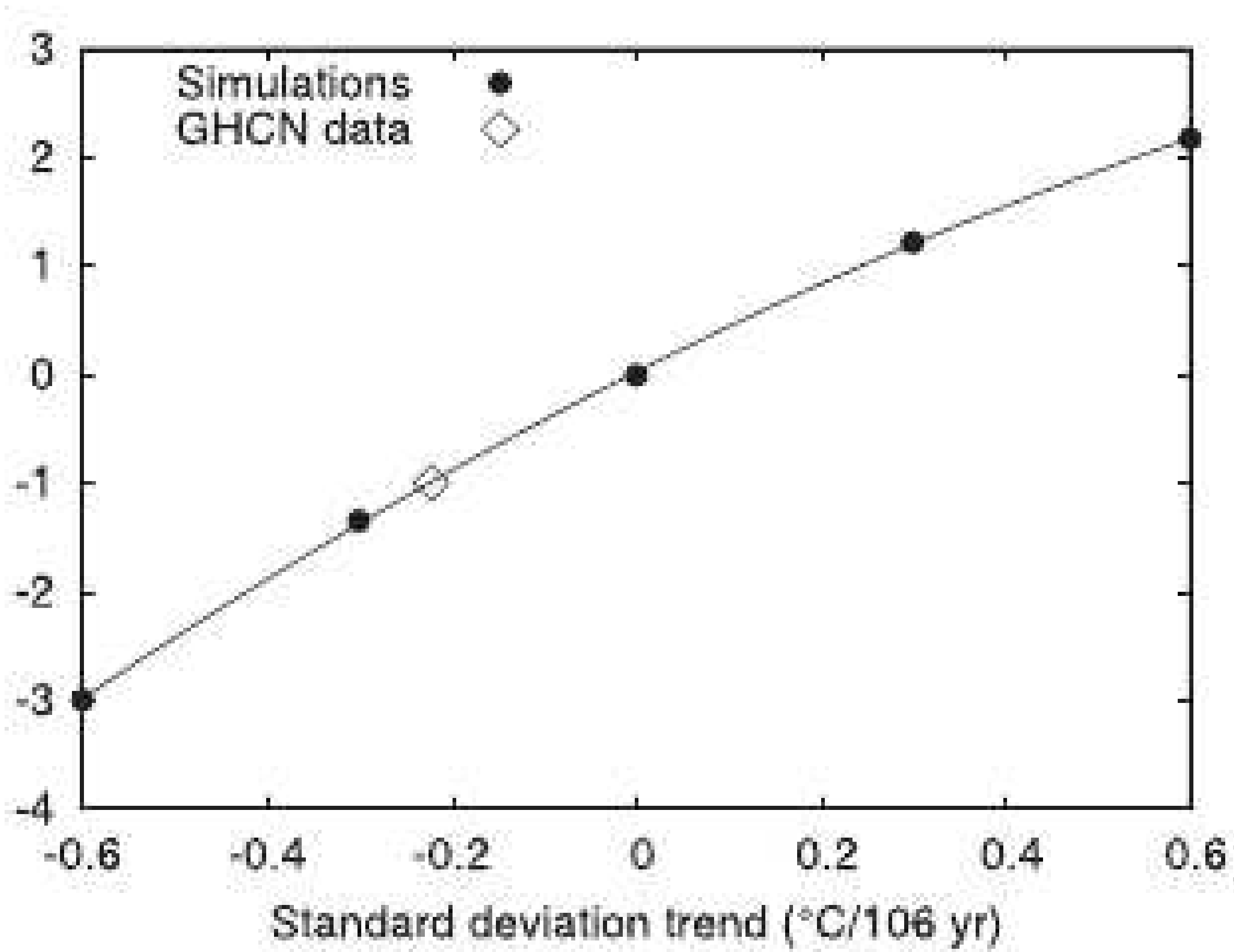
- $\mathscr{R}$ is insensitive to drift, because it vanishes to leading order in the drift speed, but can pick up small changes in the variance of the time series

- Based on a large worldwide data set of monthly temperatures, Anderson & Kostinski argue that $\langle \mathscr{R} \rangle < 0$, indicating decreasing temperature variability.

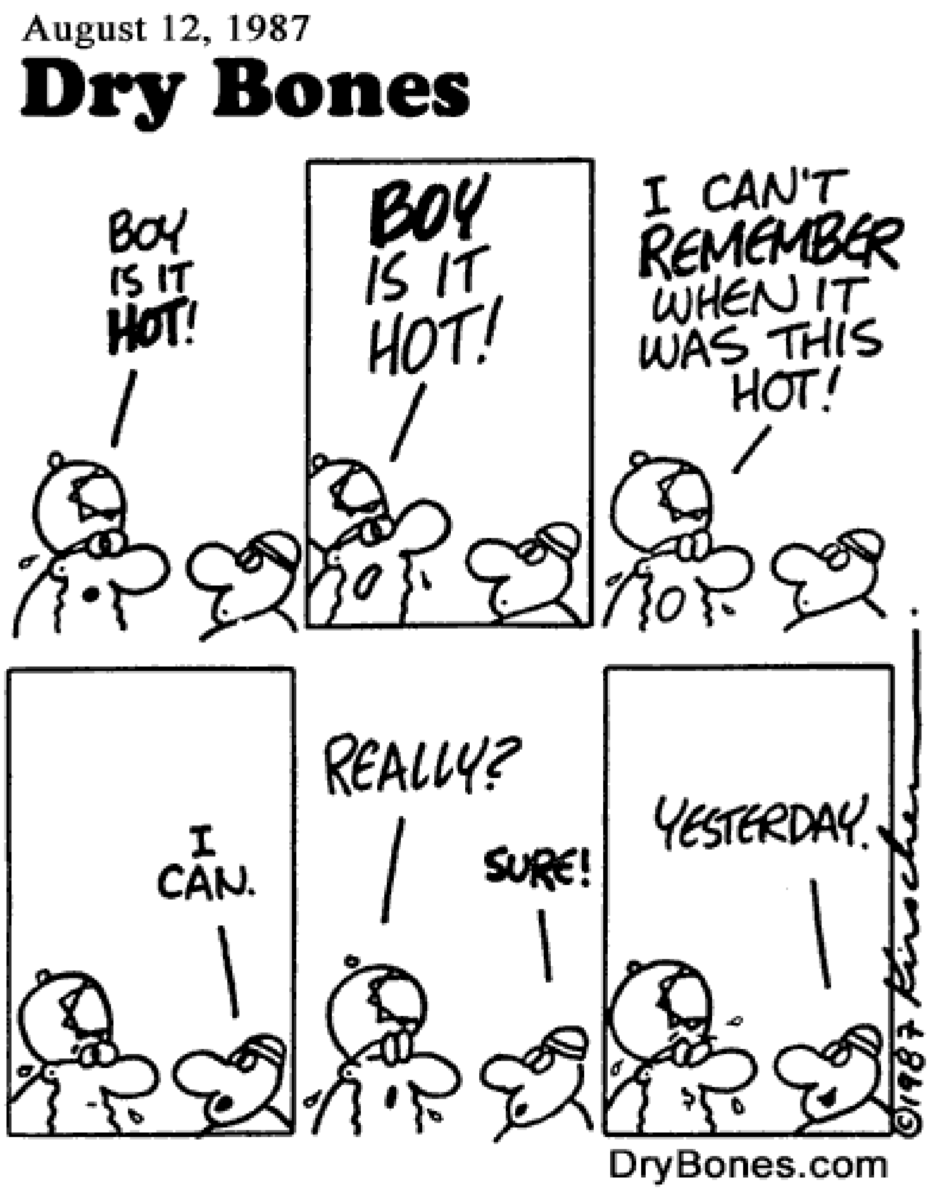
A record-based test of changing temperature variability

A. Anderson, A. Kostinski, J. Appl. Meteor. Climat. **49**, 1681 (2010)

$\langle \mathscr{R} \rangle$



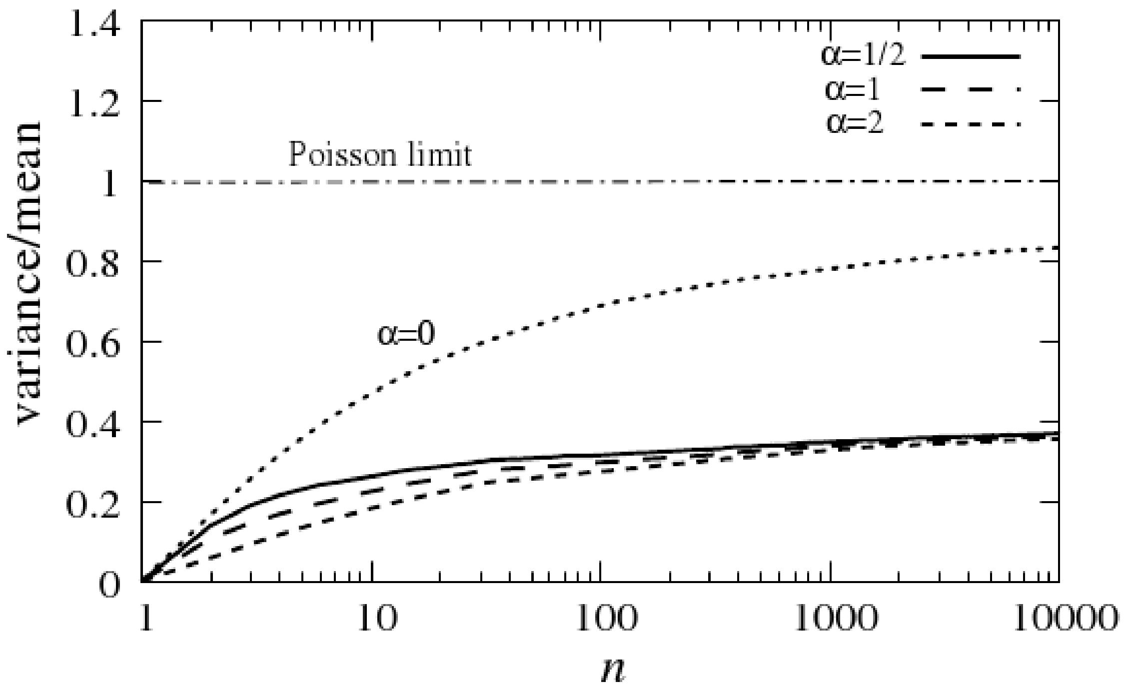
Correlations between record events



Records from broadening distributions

JK, JSTAT (2007) P07001

- RV's X_n drawn from $p_n(x) = n^{-\alpha} f(x/n^{\alpha})$ with $\alpha > 0$
- Simulations indicate sub-Poissonian fluctuations in the number of records, indicating that record events repel each other



Example: Uniform distribution

Record correlations in the linear drift model

G. Wergen, J. Franke, JK, arXiv:1105:3915

- Consider the quantity

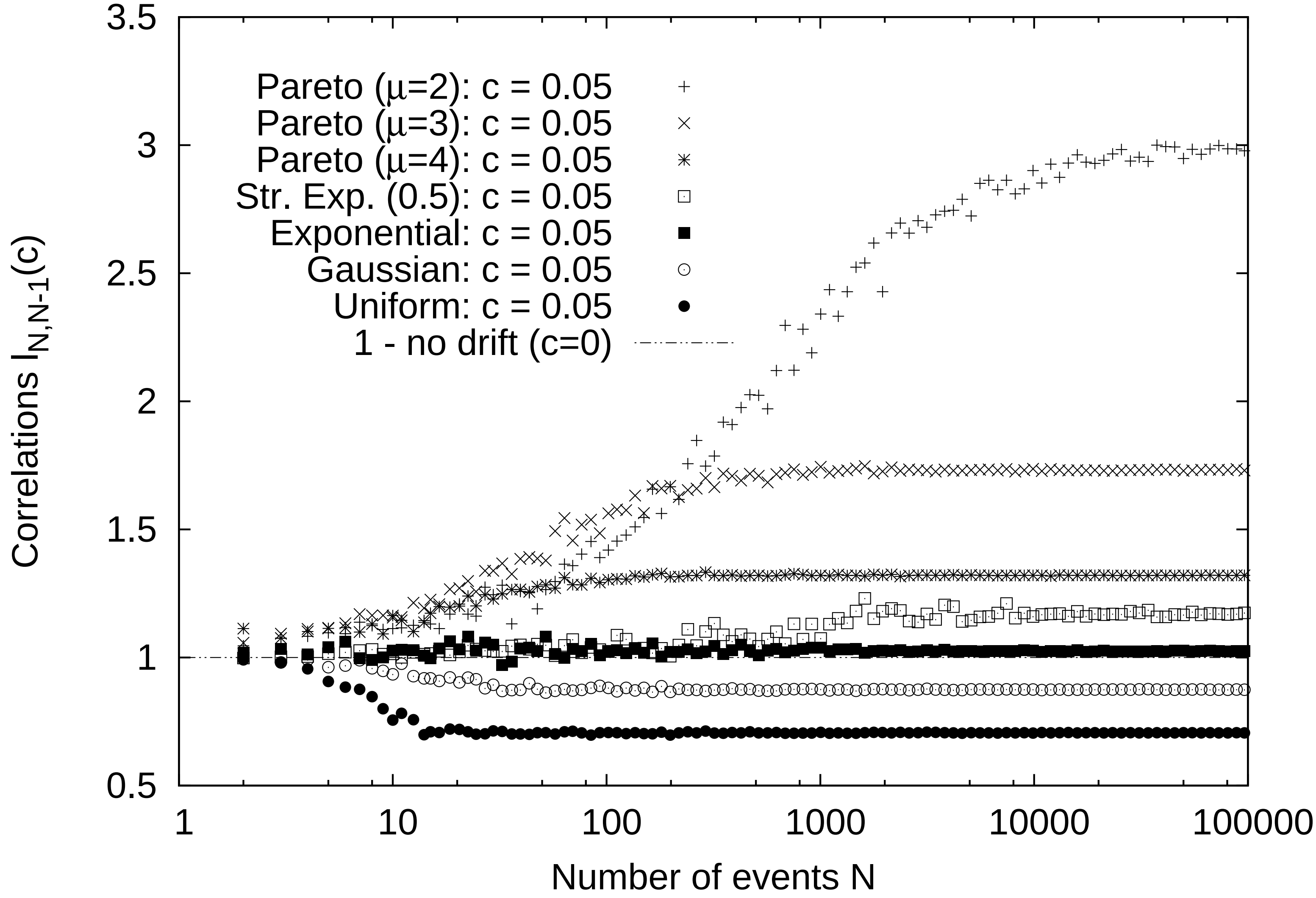
$$l_{N,N-1}(v) = \frac{P_{N,N-1}}{P_N P_{N-1}} \quad \text{with} \quad P_{N,N-1} = \mathrm{Prob}[X_N \text{ record and } X_{N-1} \text{ record}]$$

- $l_{N,N-1}(0) = 1$ and $l_{N,N-1}(v) \equiv 1$ for Gumbel-distributed i.i.d part
- $\lim_{N\to\infty} l_{N,N-1}(v)$ exists for $v > 0$ but not necessarily for $v < 0$
- Small v expansion yields $l_{N,N-1}(c) \approx 1 + vJ_N(v)$ with

$$J_N \approx -\frac{1}{2}N^4 \frac{dI_N}{dN} - N^3 I_N \approx \frac{\kappa}{2} N^3 I_N$$

where κ is the extreme value index of $p(x) \sim (1+\kappa x)^{-\frac{\kappa+1}{\kappa}}$

Record correlations in the linear drift model



For details see poster by Jasper Franke

Conclusions

- Records statistics as a paradigm of non-stationary dynamics of rare events
- Linear drift model a simple yet rich generalization of record statistics to non-i.i.d. RV’s
- Global warming affects the rate of record-breaking temperatures in a moderate but significant way
- Record events in the linear drift model can be positively or negatively correlated depending on the tail of the underlying distribution