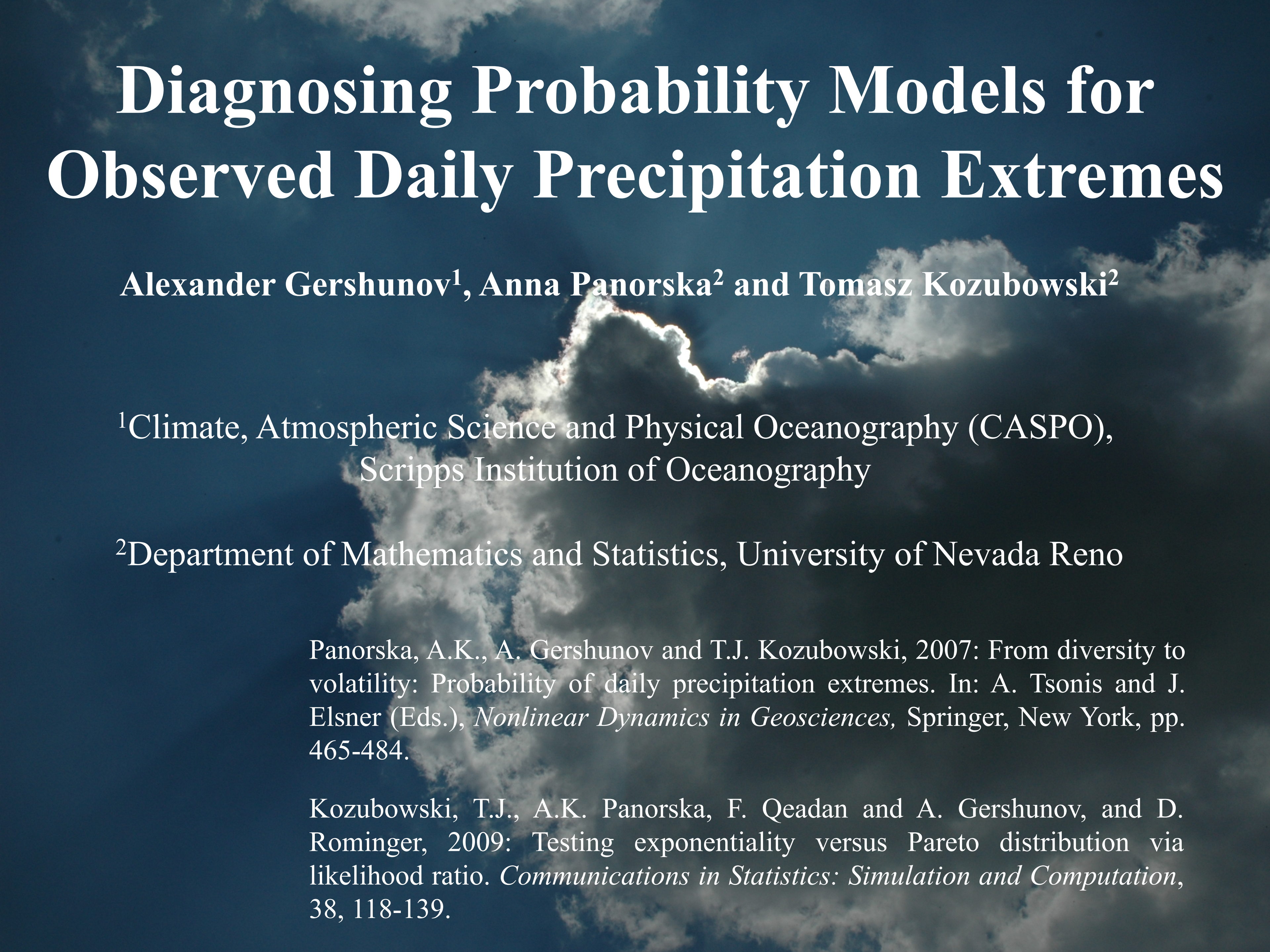


Diagnosing Probability Models for Observed Daily Precipitation Extremes

Alexander Gershunov[1], Anna Panorska[2] and Tomasz Kozubowski[2]

[1]Climate, Atmospheric Science and Physical Oceanography (CASPO),
Scripps Institution of Oceanography

[2]Department of Mathematics and Statistics, University of Nevada Reno

Panorska, A.K., A. Gershunov and T.J. Kozubowski, 2007: From diversity to volatility: Probability of daily precipitation extremes. In: A. Tsonis and J. Elsner (Eds.), *Nonlinear Dynamics in Geosciences,* Springer, New York, pp. 465-484.

Kozubowski, T.J., A.K. Panorska, F. Qeadan and A. Gershunov, and D. Rominger, 2009: Testing exponentiality versus Pareto distribution via likelihood ratio. *Communications in Statistics: Simulation and Computation*, 38, 118-139.

Motivations

- A rigorous view of extreme precipitation events in climate requires a parametric approach.
- Exponential tails may not be adequate to model probabilities of high-frequency precipitation extremes in volatile precipitation regions. We want to check how adequate or inadequate they are.
- Indications are that heavy tailed distributions may be more appropriate for this task.
- However, identifying and fitting heavy tails is difficult
  - » Typical methods involve graphical analysis
  - » There are many heavy-tailed distributions
  - » Identification and parameter estimation needs to be automated

Are precipitation extremes exponentially distributed?

If not, what is a reasonable distribution?

Heavy-tailed modeling of extreme event probabilities

»Heavy-tailed PDFs (power laws) allow for more extremes than traditional PDFs

»Arise naturally as limit sums of random variables

»Random variable X (e.g. precipitation) is "heavy tailed" if the probability (P) that it exceeds a value x is of power order $x^{-\alpha}$ for large x

$$P(X > x) \approx cx^{-\alpha}, \text{ as } x \rightarrow \infty, \text{ where } c, \alpha > 0$$

Approach

Peaks over threshold (POT) methodology

Three possible limiting PDFs for exceedances (approximations)

Balkema - de Haan - Pickands theorem (Balkema and de Haan 1974 and Pickands 1975) provides the limiting distribution of exceedances. The theorem says that when the threshold (u) increases, the distribution of the exceedance $X^{[u]}$ converges to a Generalized Pareto (GP) distribution. Any GP distribution has to be one of the following three kinds: exponential, Pareto or beta (finite, not applicable). So: no matter what the original distribution of X is, the exceedance $X^{[u]}$ over any threshold u is (approximately) one of only three distributions (effectively two).

Moreover, if original distribution has exponential tails, then the exceedance pdf will be exponential. If it has heavy tails, then the exceedance pdf will be Pareto.

Using this result, we seek a statistic, a decision rule, to classify observed daily precipitation tails into exponential and heavy

Exponential vs. Heavy Tails

H$_0$: data comes from an exponential distribution, *versus the alternative*
H$_1$: data comes from a Pareto distribution.

We approached this problem using ideas from the theory of likelihood ratio tests (Lehmann, 1997). The approach is to consider the ratio of the maxima of the likelihoods of the observed sample under the null (Pareto in the numerator) and alternative (exponential in the denominator) models. The logarithm of the likelihood ratio statistic is:

$$L = \log\left(\frac{\max\left(\sup_{\alpha>0,s>0} L_{Pareto}(\vec{x}|\alpha, s), \sup_{\sigma>0} L_{\exp}(\vec{x}\,|\,\sigma)\right)}{\sup_{\sigma>0} L_{\exp}(\vec{x}|\sigma)} \right),$$

Where $\vec{x}$ is the observed sample (of excesses); $L_{Pareto}(\vec{x}|\alpha,s)$ and $L_{\exp}(\vec{x}|\sigma)$ are the likelihood functions of the sample under Pareto and exponential hypotheses, respectively. We use a Pareto distribution with the survival function S(x) = P(X>x) = $(1/(1+1/s\alpha))^{\alpha}$ and exponential distribution with the survival function S(x) = P(X>x) = exp(-x/σ). In the Pareto case, the α parameter determines the thickness of its tail and is of primary importance. The scale parameter s is of secondary importance. In the exponential case σ is the scale parameter.

Computation of L: the maximum likelihood procedure

$$\sup_{\sigma>0} \log(L_{\exp}(\vec{x} \mid \sigma)) = n(-\log(\overline{x}) - 1),$$

where $\overline{x}$ is the sample mean. The natural logarithm of the supremum of the Pareto likelihood is

$$\sup_{\alpha>0, s>0} \log(L_{Pareto}(\vec{x} \mid \alpha, s)) = n(\log(\hat{\alpha}) - \log(\hat{s}) - 1 - 1/\hat{\alpha}),$$

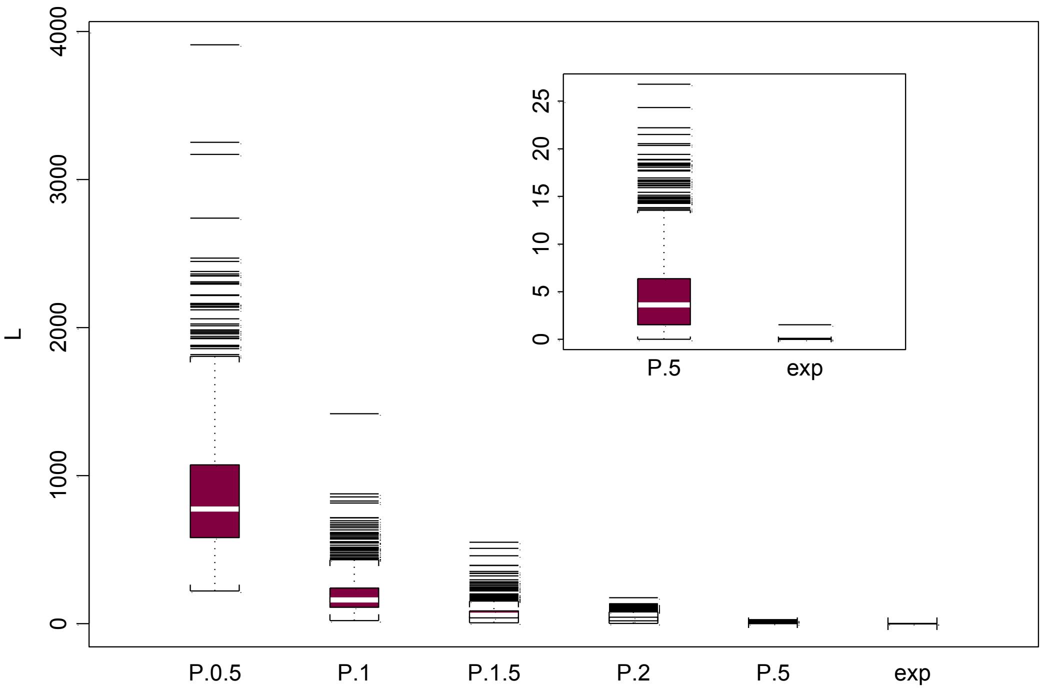
where $\hat{\alpha}$ and $\hat{s}$ are the MLEs of α and s for the Pareto likelihood. The computation of $\hat{s}$ requires numerical maximization of the function

$$u(t) = -\log\left(\frac{\sum_{i=1}^{n} \log(1+x_i t)}{n}\right) + \log(t) - \frac{1}{n}\sum_{i=1}^{n} \log(1+x_i t),$$

with respect to t > 0. If $\hat{t}$ is the maximum of u(t), then the MLE of s is $\hat{s} = 1/\hat{t}$. The MLE of α is

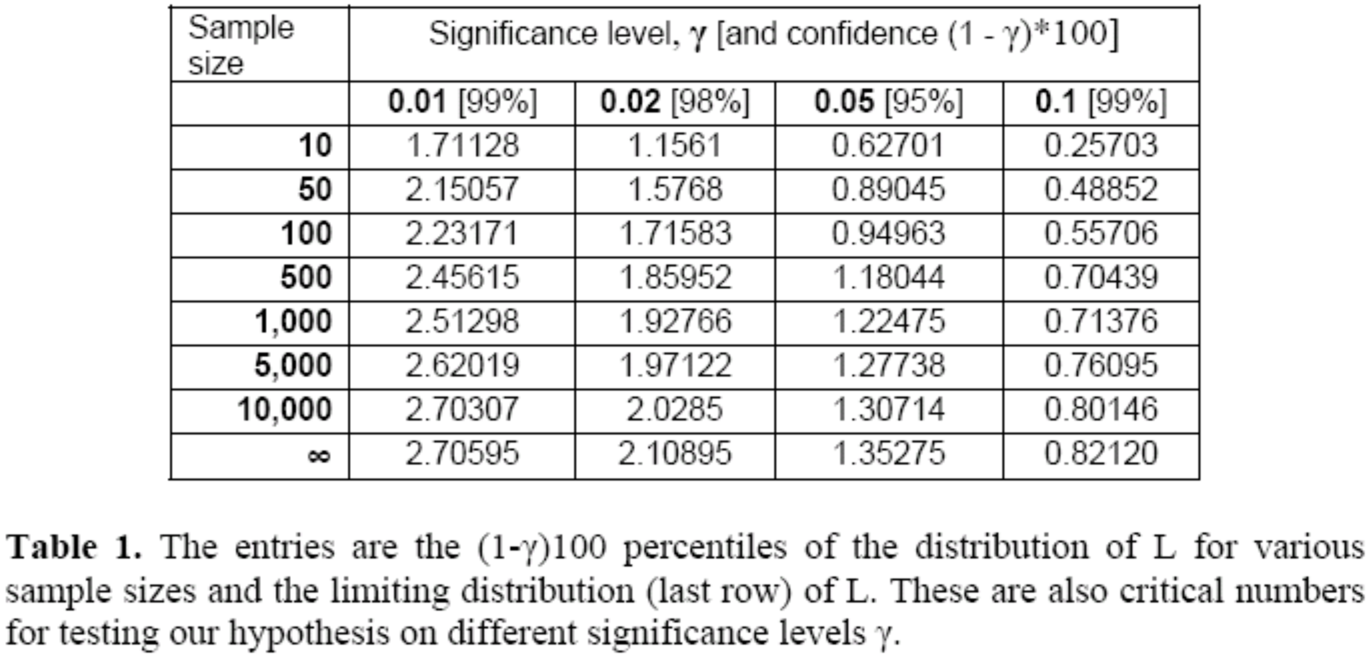
$$\hat{\alpha} = \frac{1}{(1/n)\sum_{i=1}^{n} \log(1+\frac{x_i}{\hat{s}})}.$$

Statistical Properties of L



Boxplots of simulated distributions of L. The first five boxplots were done using 10,000 observations of L from Pareto samples of size 1,000 with α varying from 0.5 (first boxplot) to 5 (second to the last boxplot). The last boxplot corresponds to 10,000 observations of L from exponential samples of size 1,000. The inset blows up the last two boxplots.

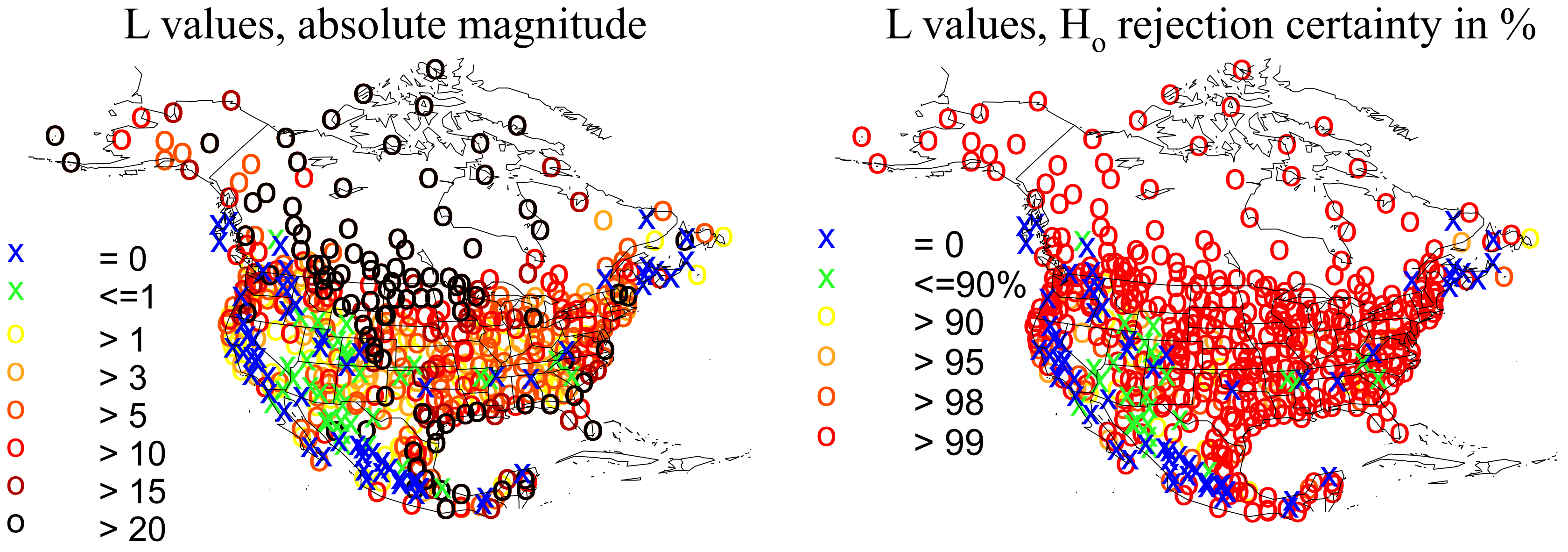
Critical Values of L



Sample size	Significance level, γ [and confidence (1 - γ)*100]			
	0.01 [99%]	**0.02** [98%]	**0.05** [95%]	**0.1** [99%]
10	1.71128	1.1561	0.62701	0.25703
50	2.15057	1.5768	0.89045	0.48852
100	2.23171	1.71583	0.94963	0.55706
500	2.45615	1.85952	1.18044	0.70439
1,000	2.51298	1.92766	1.22475	0.71376
5,000	2.62019	1.97122	1.27738	0.76095
10,000	2.70307	2.0285	1.30714	0.80146
∞	2.70595	2.10895	1.35275	0.82120

Table 1. The entries are the (1-γ)100 percentiles of the distribution of L for various sample sizes and the limiting distribution (last row) of L. These are also critical numbers for testing our hypothesis on different significance levels γ.

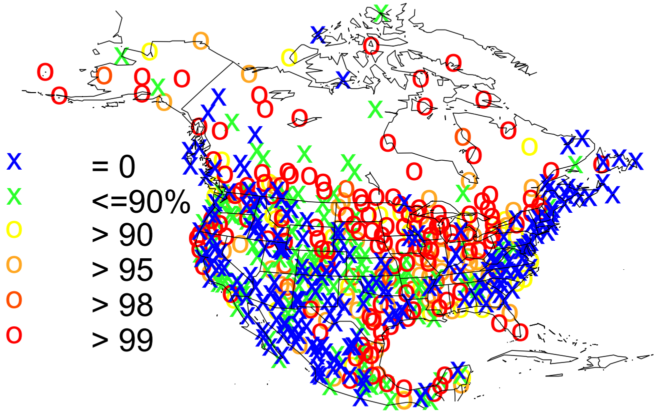
Log likelihood ratio: all data


L values, absolute magnitude

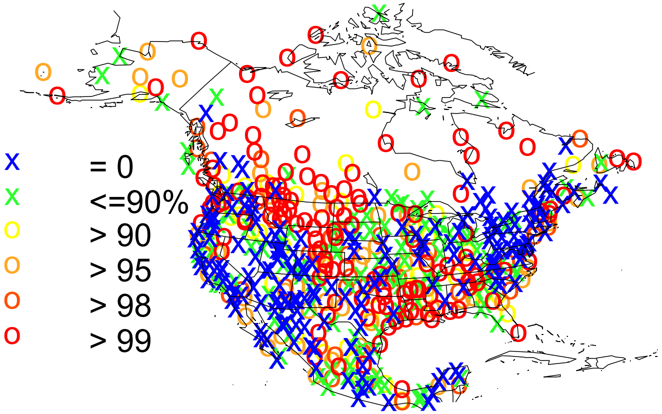
L values, H_o rejection certainty in %

Log likelihood ratio (L) computed for daily excesses over local 75th percentile at each of the 560 stations. **(a)** Values close to zero, L <= 1 (blue and green x' s) represent approximately exponential tails, while yellow, red and black circles represent progressively heavier tails. **(b)** Level of confidence, $(1 - \gamma)*100$, for rejecting the null hypothesis (H_o) of exponential tails. Blue x' s represent exponential tails, green x' s represent stations at which the H_o cannot be rejected with reasonable (90%) confidence. Yellow and progressively redder circles represent stations at which H_o can be rejected with 90, 95, 98 and 99% confidence in favor of the Pareto alternative. For example, H_o can be rejected at 81% of stations with 95% confidence.

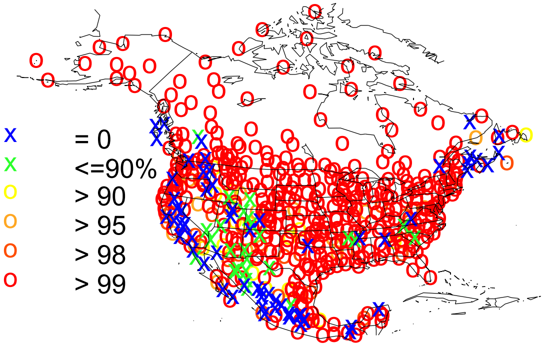
Log likelihood ratio: seasonal data


DJF

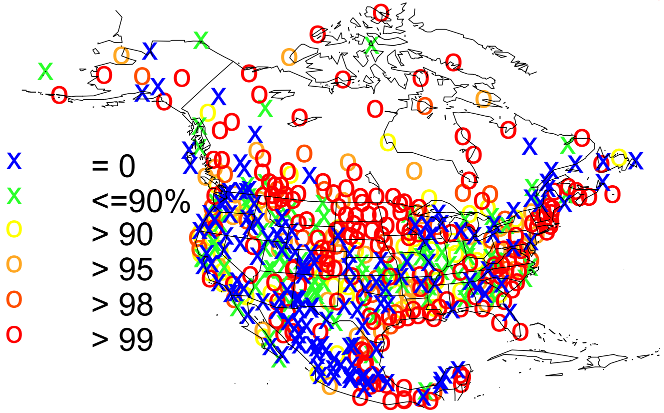
37% heavy with 95% confidence


MAM

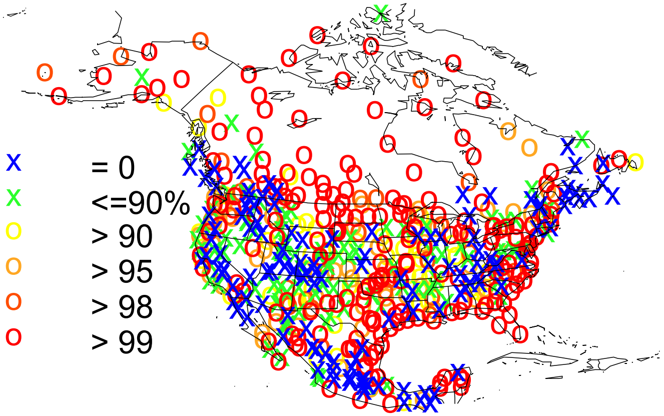
40% heavy with 95% confidence


Four Seasons

81% heavy

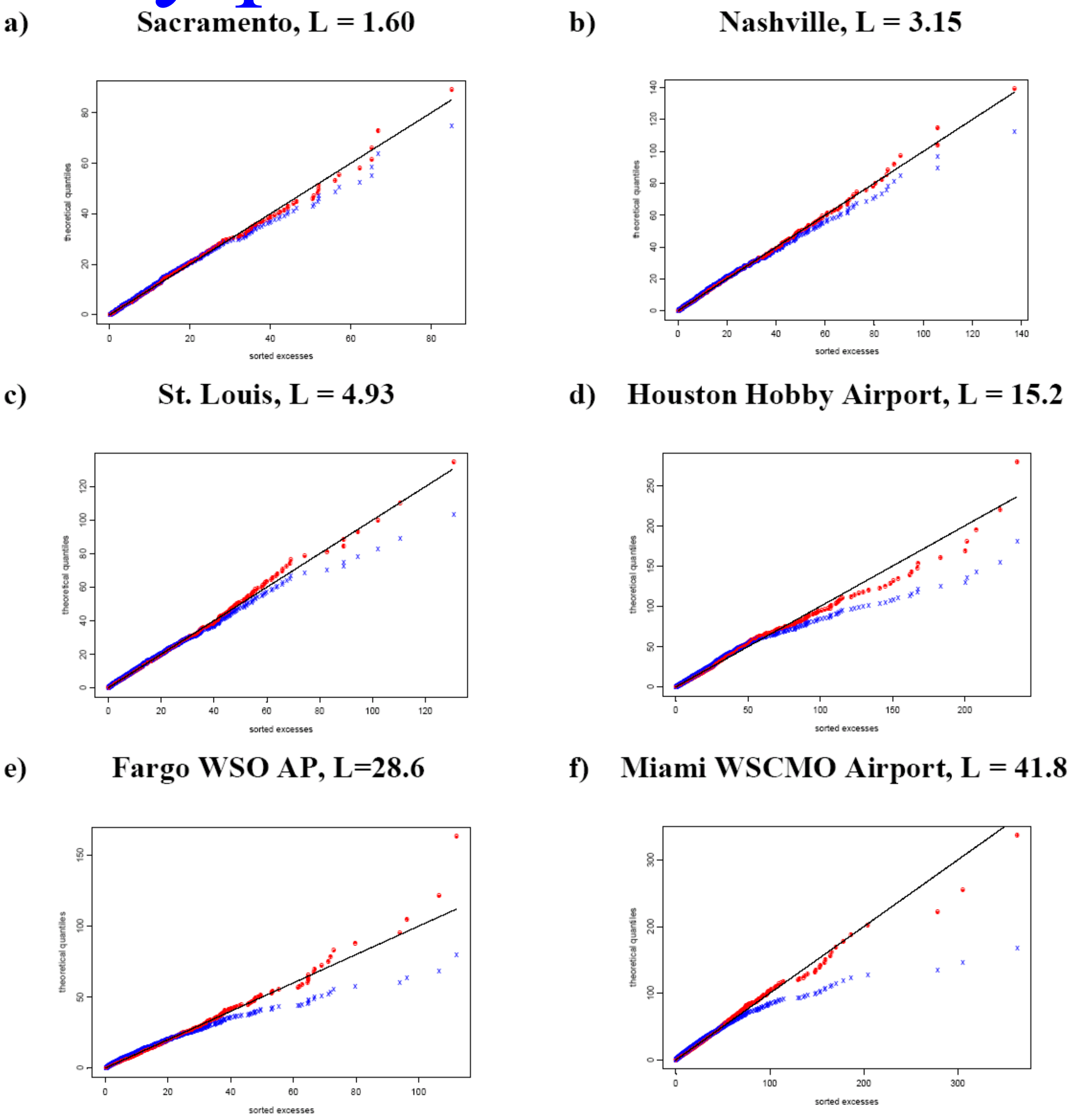

JJA

47% heavy with 95% confidence


SON

51% heavy with 95% confidence

Probability plots for selected stations



a) Sacramento, L = 1.60

b) Nashville, L = 3.15

c) St. Louis, L = 4.93

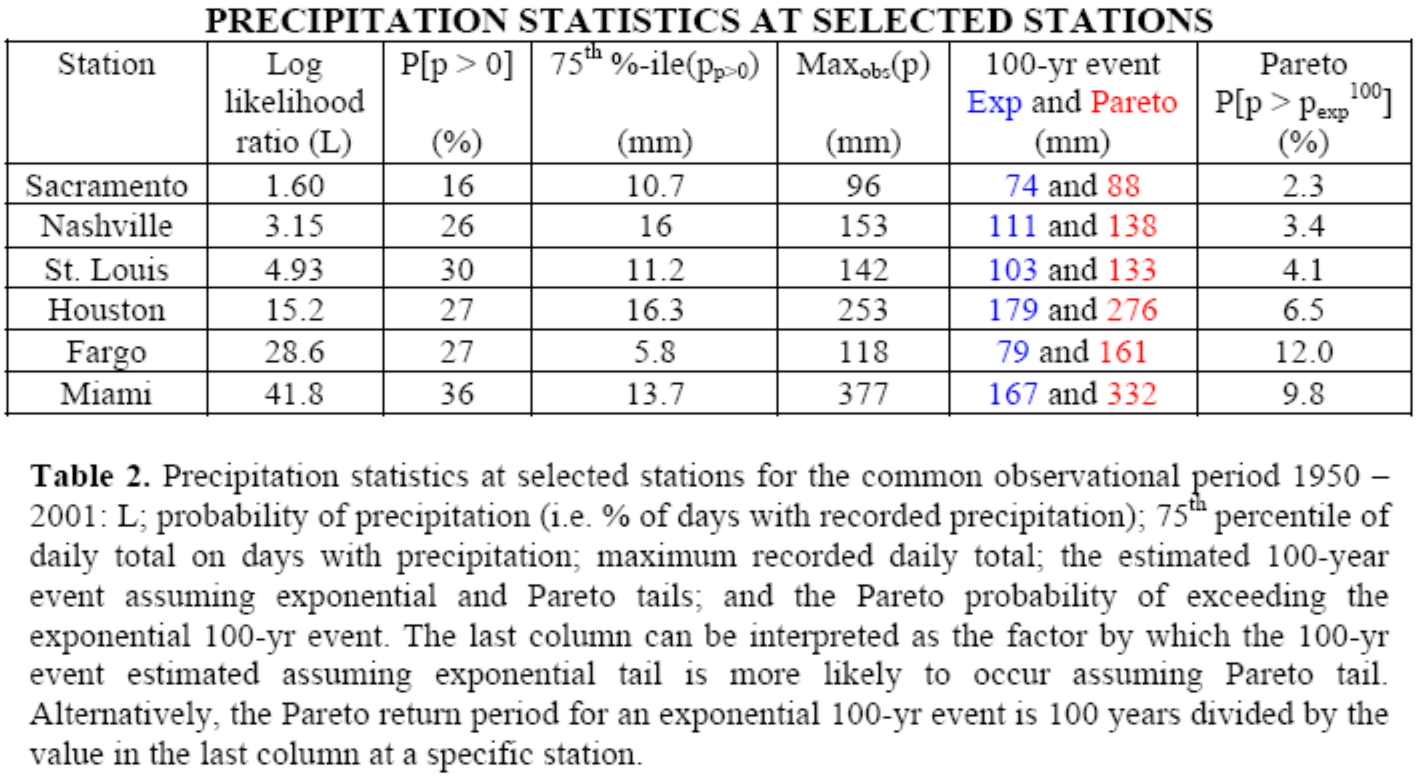
d) Houston Hobby Airport, L = 15.2

e) Fargo WSO AP, L=28.6

f) Miami WSCMO Airport, L = 41.8

Probability plots for excesses over threshold (75th percentile) at selected stations arranged in order of increasing L. Sorted observed excesses displayed in mm along the x-axis are plotted against the corresponding theoretical quantiles derived from the fitted exponential (blue x's) and Pareto (red o's) models.

Precipitation Stats at Selected Stations



PRECIPITATION STATISTICS AT SELECTED STATIONS

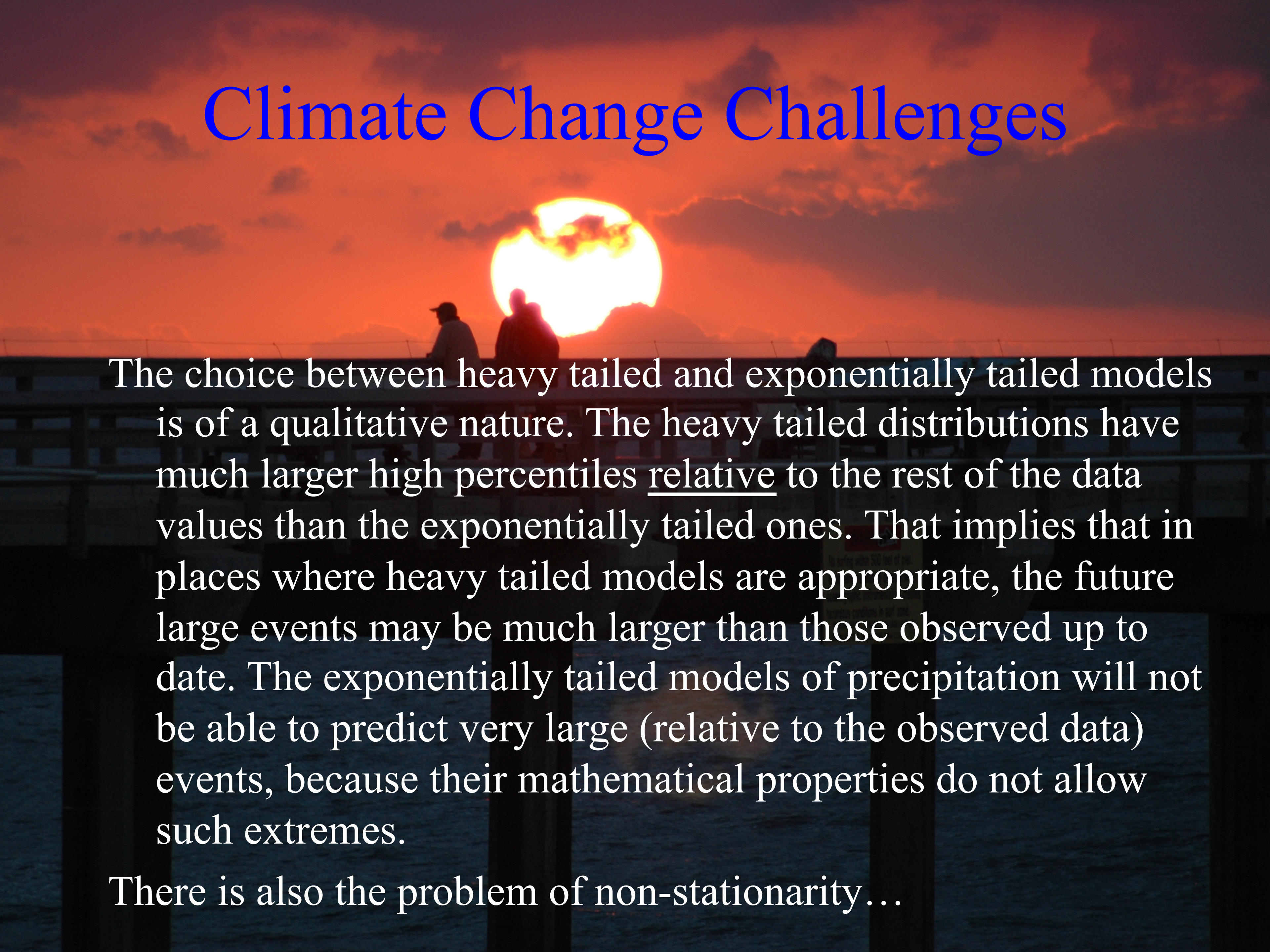
Station	Log	$P[p > 0]$	75th %-ile($p_{p>0}$)	$Max_{obs}(p)$	100-yr event	Pareto
	likelihood				Exp and Pareto	$P[p > p_{exp}^{100}]$
	ratio (L)	(%)	(mm)	(mm)	(mm)	(%)
Sacramento	1.60	16	10.7	96	74 and 88	2.3
Nashville	3.15	26	16	153	111 and 138	3.4
St. Louis	4.93	30	11.2	142	103 and 133	4.1
Houston	15.2	27	16.3	253	179 and 276	6.5
Fargo	28.6	27	5.8	118	79 and 161	12.0
Miami	41.8	36	13.7	377	167 and 332	9.8

Table 2. Precipitation statistics at selected stations for the common observational period 1950 – 2001: L; probability of precipitation (i.e. % of days with recorded precipitation); 75th percentile of daily total on days with precipitation; maximum recorded daily total; the estimated 100-year event assuming exponential and Pareto tails; and the Pareto probability of exceeding the exponential 100-yr event. The last column can be interpreted as the factor by which the 100-yr event estimated assuming exponential tail is more likely to occur assuming Pareto tail. Alternatively, the Pareto return period for an exponential 100-yr event is 100 years divided by the value in the last column at a specific station.

Summary

- Diversity is the mother of volatility
- Exponential tails are inadequate to model daily extremes in most regions of North America
- Heavy tailed models are appropriate on theoretical and empirical grounds
- These results can be directly extended to many climatic studies and applications
- What about climate change?
- How do climate models do?

Climate Change Challenges



- The choice between heavy tailed and exponentially tailed models is of a qualitative nature. The heavy tailed distributions have much larger high percentiles relative to the rest of the data values than the exponentially tailed ones. That implies that in places where heavy tailed models are appropriate, the future large events may be much larger than those observed up to date. The exponentially tailed models of precipitation will not be able to predict very large (relative to the observed data) events, because their mathematical properties do not allow such extremes.
- There is also the problem of non-stationarity…