# Extreme Value Theory: A Practical Introduction

Greg Herman


What were the odds?

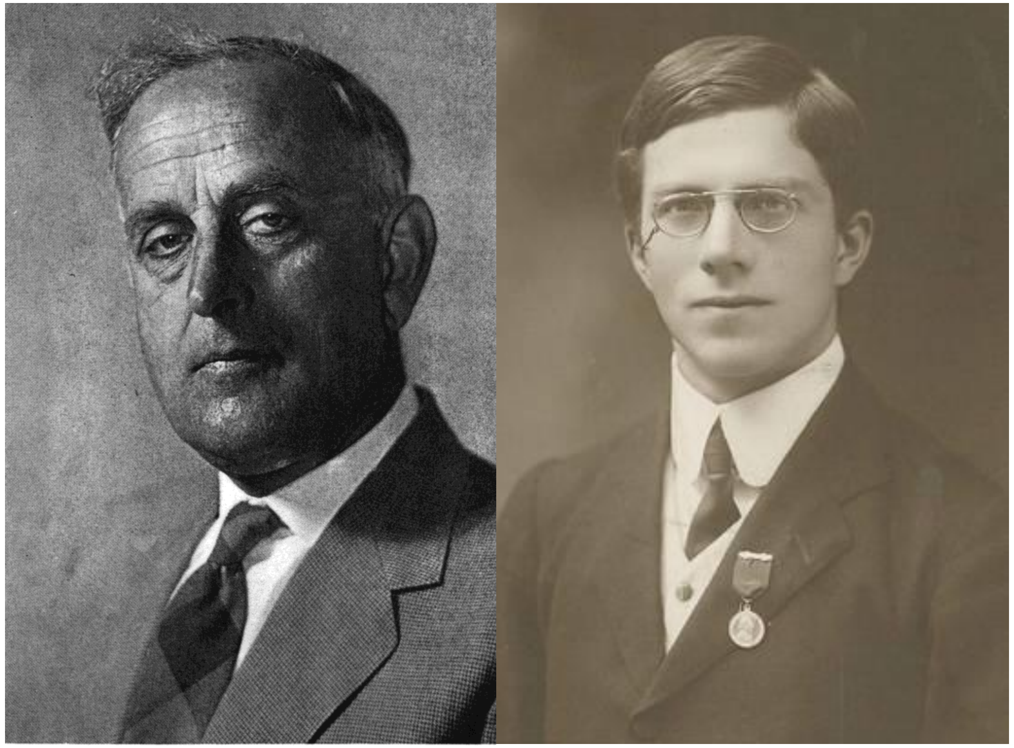
# What is Extreme Value Theory (EVT)?

- Statistical Theory concerning *extreme values*- values occurring at the tails of a probability distribution
- Society, ecosystems, etc. tend to adapt to routine, near-normal conditions: these conditions tend to produce fairly minimal impacts
- In contrast, unusual and extreme conditions tend to have much more substantial net impacts despite, by definition, occurring a much smaller proportion of the time.
- From perspective of impacts, important to be able to understand and quantify rare and extreme behavior, despite the challenges in doing so

# Why do we care about EVT?

- Many applications throughout all realms of the atmospheric sciences
- Weather
  - Quantifying extreme precipitation, streamflow, etc.
  - Quantifying windstorm and/or tropical cyclone intensity
  - Quantifying heat waves or cold snaps
  - Quantifying drought
- Climate
  - Quantifying any of the above in a changing/changed climate
  - Quantifying poleward/equatorward extremes in jet positioning
- Chemistry
  - Quantifying daily maximum concentrations of a harmful chemical species
- Lots more…

# Fisher-Tippett Theorem



L. H. C. TIPPETT

- "The Central Limit Theorem of EVT"
- Mathematical Formulation:

Consider the set of IID random variables:
$X_1, X_2, \ldots, X_n$

Each $X_i$ is sampled from some unknown underlying distribution function F

Let $M_n = MAX\{X_1, X_2, \ldots, X_n\}$

Given that certain limiting conditions are met, namely that there exists a set of real numbers $(a_n, b_n), a_n \geq 0$ such that $\frac{M_n - b_n}{a_n}$ has a *non-degenerate* limiting distribution in the limit of large n (analogous to CLT requirement of underlying distribution having finite variance)

Then: $M_n \sim GEV(\mu, \sigma, \xi)$

# Fisher-Tippett Theorem

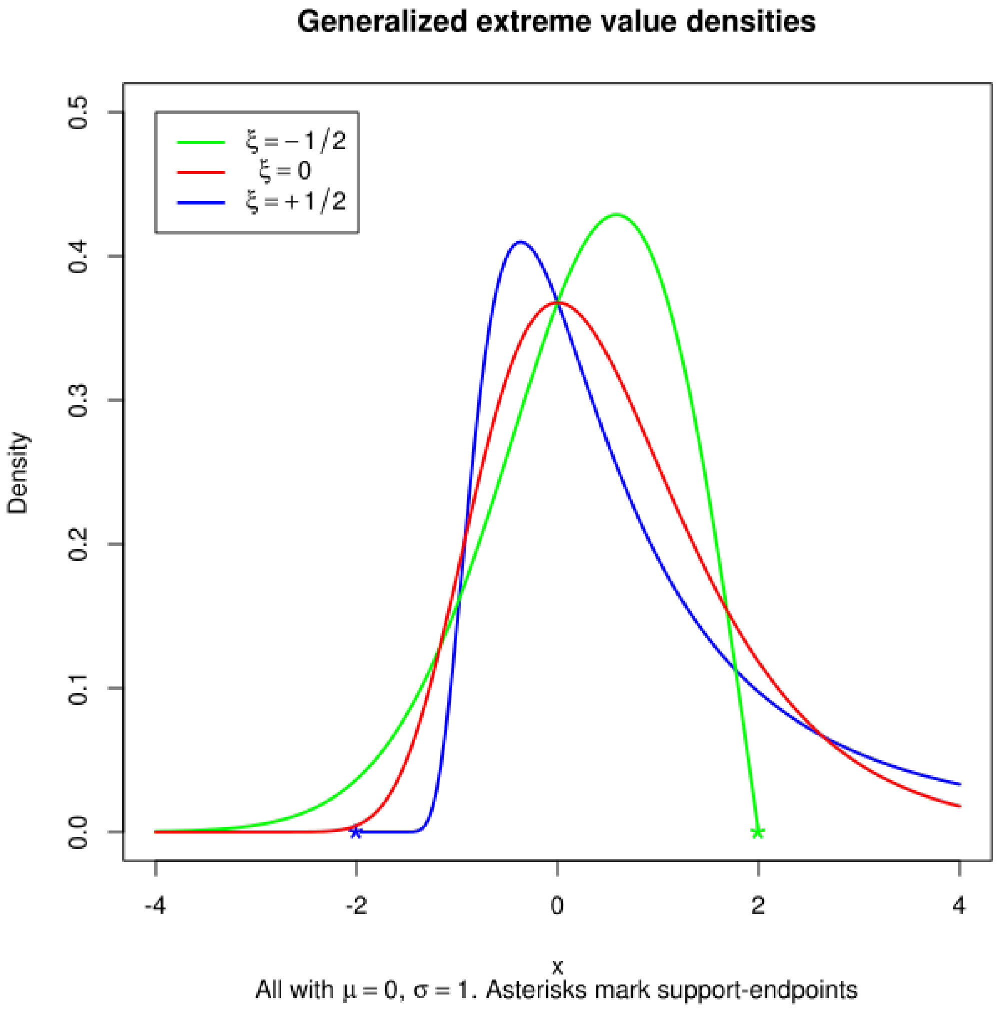
- What does this mean?
- Provided your underlying probability distribution D of a random variable X is not highly unusual (same as with CLT, though different conditions), regardless of what D is, and provided that n is sufficiently large, maxima M of samples of size n drawn from D will be distributed as the Generalized Extreme Value Distribution (GEV)
- (Central Limit Theorem is very similar…just replace ‘maxima’ with ‘mean’ and ‘Normal’ for ‘Generalized Extreme Value’)

# Generalized Extreme Value Distribution (GEV)

- Three parameter distribution:
  1. Location parameter μ
     - Shifts distribution left/right
  2. Scale parameter σ
     - Determines 'spread' of distribution
  3. Shape parameter ξ
     - Determines shape of distribution, and distribution which it belongs
     - $\begin{cases} Rev.Weibull & \xi < 0 \\ Gumbel & \xi = 0 \\ Fréchet & \xi > 0 \end{cases}$

$$PDF\colon\ f(x;\ \mu,\sigma,\xi) = \frac{1}{\sigma} t(x)^{\xi+1} e^{-t(x)}$$

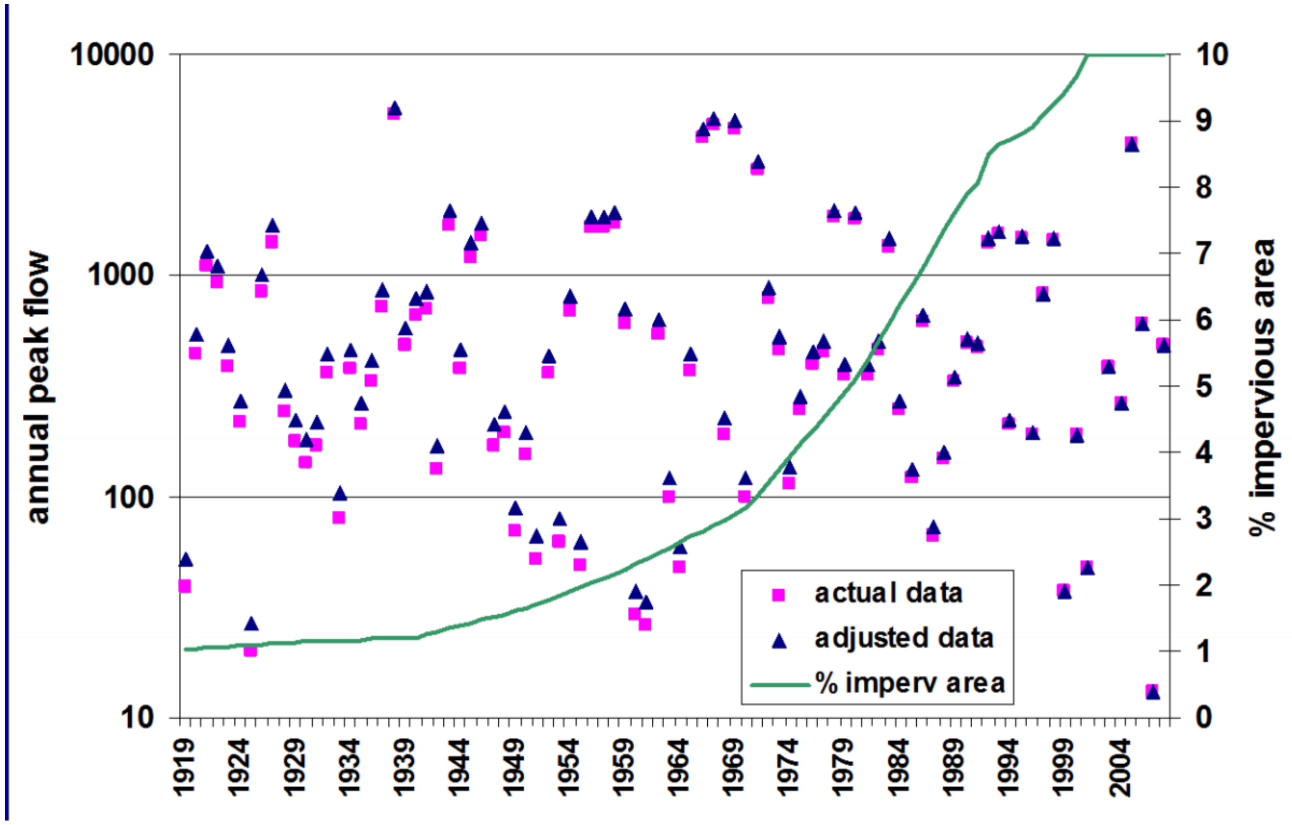
$$\text{Where: } t(x) = \begin{cases} \left(1 + \left(\frac{x-\mu}{\sigma}\right)\xi\right)^{\frac{-1}{\xi}} & \xi \neq 0 \\ e^{-\frac{x-\mu}{\sigma}} & \xi = 0 \end{cases}$$

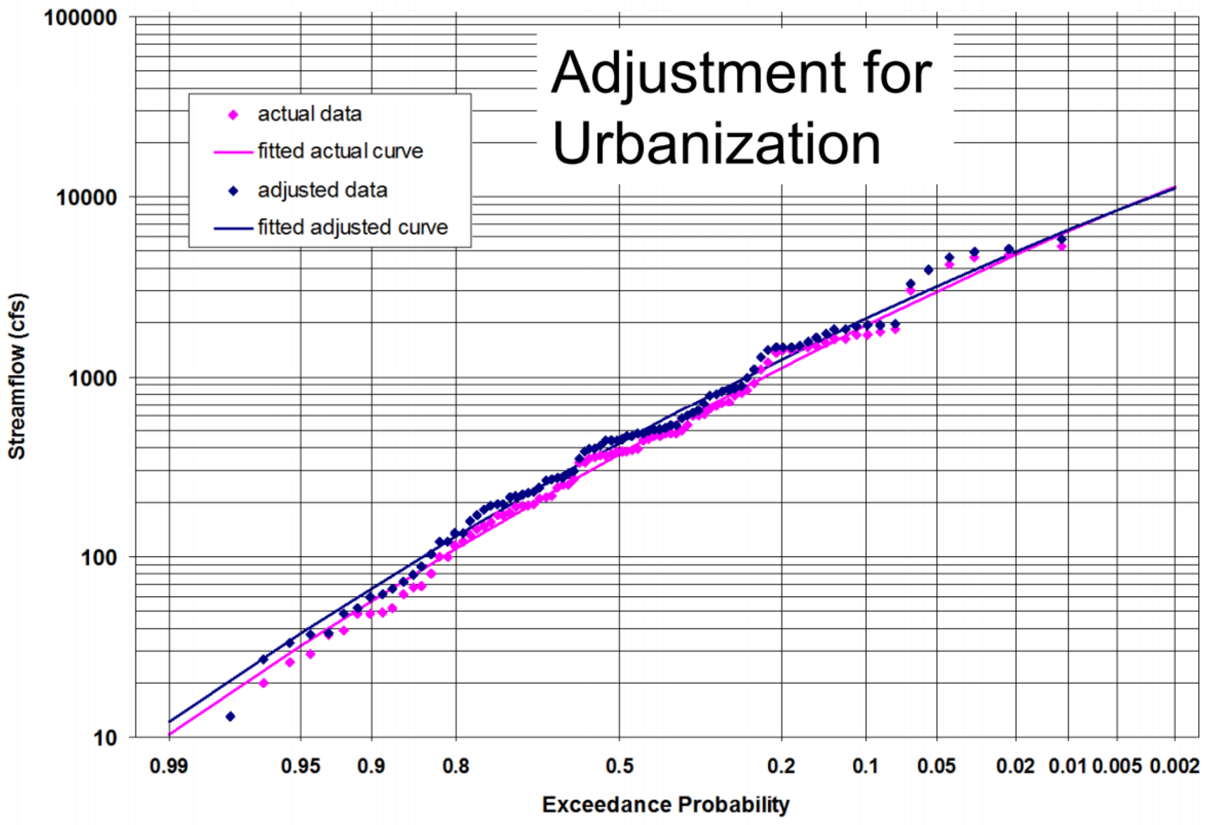


# A note on assumptions

- Assumes IIDRVs
  - Independent
    - Each sample is independent from all other samples (big assumption!)
  - Identically Distributed
    - Each sample is taken from the same underlying distribution
    - Assumes no *seasonality* in the time series (big assumption!)
      - Can correct for seasonality; not always a major concern
    - Assumes *stationarity* in the time series (big assumption!)
      - Can correct for known trends
  - Random
    - Value sampled from a probability distribution, varies due to chance
    - Measurements are unbiased estimators of the desired quantity

# Example: Urbanization Correction



# Data Issues: Ensuring data independence

- One approach: determine the *e*-folding time $\tau$ of the time series phenomenon being examined

AMS:

- Ensure that annual maximum values are separated by at least $\tau$. If not, take the block maxima of adjacent blocks with a separation of $\tau$

*PDS (discuss later):*

- *Require that retained exceedances of u be separated by a period of at least $\tau$ with values less than u to be considered independent*
- *Select the maximum value within this dependent period to be retained in the PDS*

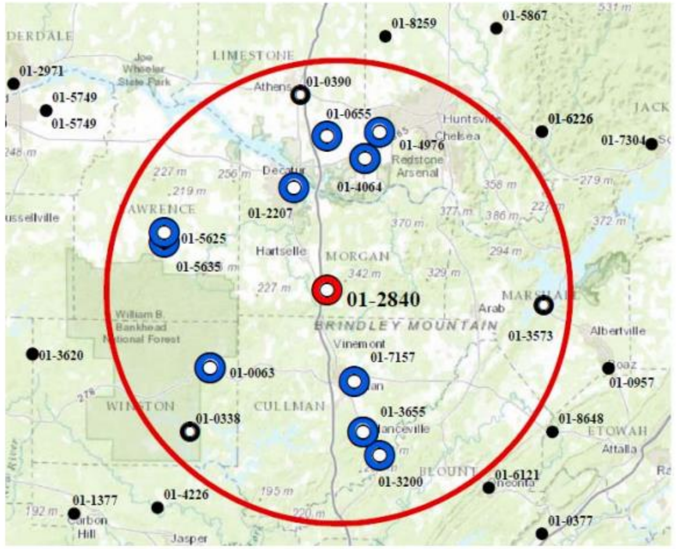
# Data Issues: Insufficient Data

- Can derive return period threshold estimates for any arbitrarily large return period
- The quality of recurrence interval estimates and their uncertainty is proportional/inversely proportional to the ratio of the recurrence interval being estimated to the data record length
  - With 100+ years of data, expect a reasonably accurate and confident estimate of a 100-year RP threshold
  - With 5 years of data, expect a highly uncertain estimate of the 100-year RP threshold
  - Shape parameter in particular highly sensitive to data sample
- Is it worthwhile?
  - Possibly; depends on tolerance to uncertainty

# Regionalization

- Often in atmospheric science, have observations at many different locations, but not for extended period (especially when dealing with observations)
- Can often use data from other locations to artificially increase data record length
  - Pick locations spatially distant enough so as to carry independent information (e.g. using FNL observations for quantifying FCL extremes may or may not carry useful additional information not already contained in the FCL obs)
  - Pick locations where you know, subjectively or (preferably) objectively, the climatology/underlying distribution is “close enough” to the location of interest so that observations from that location may be considered random samples from the actual location of interest
- This can help improve estimates for event frequencies substantially greater than the true data record length

**Regional L-moments calculation.** For a given duration, regional estimates of L-moment ratios (L-CV, L-skewness and L-kurtosis) were obtained by averaging corresponding station-specific estimates weighted by record lengths. Regional L-moment ratios were then used to estimate higher order L-moments at each station.



# EVT Applications

- Two approaches:
    1. Block Maxima/"Annual Maximum Series" (AMS)
    2. Threshold Exceedances/"Partial Duration Series" (PDS)
- Block Maxima
    - Pros:
        - Simple to apply and interpret
    - Cons:
        - Framework not as often directly useful
        - Not the most efficient use of time series
- Threshold Exceedance
    - Pros:
        - Relevant and efficient use of data
    - Cons:
        - Harder to implement
        - Harder to know conditions of theory have been satisfied

# Block Maxima/Annual Maximum Series

- What is this, and how/why would I use it?
- Types of questions being answered: in a given block size/(year), how likely is an exceedance of a specified threshold? What threshold can be expected to be exceeded, on average, once every N blocks (years)?
- Pick some sufficiently large block size *n*
  - A typical choice for daily data is one year
- Take the time series you want to analyze, break it into blocks, and take the maximum of each block to form a new series S
  - Using a block size of one-year, this is called an Annual Maximum Series
- Fit a GEV distribution to S, use as desired

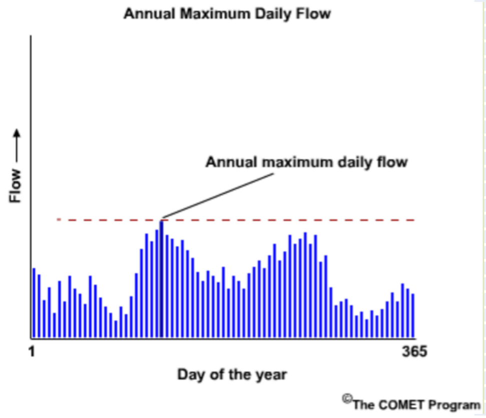
# Example: Ozone Concentrations

- Suppose you are interested in how ozone concentrations in Denver will change in a changing climate. In particular, you're worried about daily exceedances of 75 ppb, which, though rare, can cause serious adverse health effects for vulnerable populations.
- Science Question: Will these exceedances get more common under climate change, and, if so, by how much?
- Begin with your daily time series of maximum ozone concentrations

| | | | |
|---|---|---|---|
| 1860 | 6 | 1 | 32.495 |
| 1860 | 6 | 2 | 37.427 |
| 1860 | 6 | 3 | 29.24 |
| 1860 | 6 | 4 | 30.837 |
| 1860 | 6 | 5 | 29.719 |
| 1860 | 6 | 6 | 27.037 |
| 1860 | 6 | 7 | 33.903 |
| 1860 | 6 | 8 | 35.362 |
| 1860 | 6 | 9 | 34.927 |
| 1860 | 6 | 10 | 22.954 |
| 1860 | 6 | 11 | 26.075 |
| 1860 | 6 | 12 | 28.532 |
| 1860 | 6 | 13 | 29.161 |
| 1860 | 6 | 14 | 29.531 |
| 1860 | 6 | 15 | 29.613 |
| 1860 | 6 | 16 | 30.702 |
| 1860 | 6 | 17 | 29.703 |
| 1860 | 6 | 18 | 30.563 |
| 1860 | 6 | 19 | 31.036 |
| 1860 | 6 | 20 | 28.802 |
| 1860 | 6 | 21 | 24.643 |
| 1860 | 6 | 22 | 31.084 |
| 1860 | 6 | 23 | 31.075 |
| 1860 | 6 | 24 | 31.182 |
| 1860 | 6 | 25 | 37.582 |
| 1860 | 6 | 26 | 35.286 |
| 1860 | 6 | 27 | 35.914 |
| 1860 | 6 | 28 | 36.126 |
| 1860 | 6 | 29 | 28.332 |
| 1860 | 6 | 30 | 32.453 |
| 1860 | 7 | 1 | 33.869 |
| 1860 | 7 | 2 | 26.905 |
| 1860 | 7 | 3 | 29.78 |
| 1860 | 7 | 4 | 35.274 |
| 1860 | 7 | 5 | 32.321 |
| 1860 | 7 | 6 | 35.106 |
| 1860 | 7 | 7 | 31.582 |
| 1860 | 7 | 8 | 30.572 |
| 1860 | 7 | 9 | 32.495 |
| 1860 | 7 | 10 | 30.58 |
| 1860 | 7 | 11 | 34.455 |
| 1860 | 7 | 12 | 27.309 |
| 1860 | 7 | 13 | 24.175 |
| 1860 | 7 | 14 | 30.453 |
| 1860 | 7 | 15 | 29.947 |
| 1860 | 7 | 16 | 24.534 |
| 1860 | 7 | 17 | 23.63 |
| 1860 | 7 | 18 | 36.05 |
| 1860 | 7 | 19 | 39.7 |
| 1860 | 7 | 20 | 39.208 |
| 1860 | 7 | 21 | 32.948 |
| 1860 | 7 | 22 | 34.984 |
| 1860 | 7 | 23 | 32.29 |
| 1860 | 7 | 24 | 37.353 |
| 1860 | 7 | 25 | 32.7 |
| 1860 | 7 | 26 | 33.634 |
| 1860 | 7 | 27 | 29.212 |
| 1860 | 7 | 28 | 26.5 |
| 1860 | 7 | 29 | 24.713 |
| 1860 | 7 | 30 | 25.648 |
| 1860 | 7 | 31 | 26.823 |
| 1860 | 8 | 1 | 30.481 |
| 1860 | 8 | 2 | 30.973 |
| 1860 | 8 | 3 | 28.859 |
| 1860 | 8 | 4 | 27.447 |
| 1860 | 8 | 5 | 29.516 |
| 1860 | 8 | 6 | 26.027 |
| 1860 | 8 | 7 | 29.336 |
| 1860 | 8 | 8 | 21.883 |
| 1860 | 8 | 9 | 28.757 |
| 1860 | 8 | 10 | 26.993 |
| 1860 | 8 | 11 | 30.799 |
| 1860 | 8 | 12 | 32.742 |
| 1860 | 8 | 13 | 32.163 |
| 1860 | 8 | 14 | 30.956 |
| 1860 | 8 | 15 | 25.832 |
| 1860 | 8 | 16 | 23.949 |
| 1860 | 8 | 17 | 28.285 |
| 1860 | 8 | 18 | 28.2 |
| 1860 | 8 | 19 | 29.432 |
| 1860 | 8 | 20 | 35.052 |
| 1860 | 8 | 21 | 32.485 |
| 1860 | 8 | 22 | 29.658 |
| 1860 | 8 | 23 | 32.61 |
| 1860 | 8 | 24 | 29.51 |
| 1860 | 8 | 25 | 28.853 |
| 1860 | 8 | 26 | 27.874 |
| 1860 | 8 | 27 | 27.427 |
| 1860 | 8 | 28 | 30.213 |
| 1860 | 8 | 29 | 29.16 |
| 1860 | 8 | 30 | 32.93 |
| 1860 | 8 | 31 | 26.349 |

| Year | Annual Maximum Daily Concentration (ppb) |
|---|---|
| 1860 | 39.7 |
| 1861 | 38.537 |
| 1862 | 49.378 |
| 1863 | 42.7 |
| 1864 | 42.864 |
| 1865 | 45.252 |
| … | … |

- Begin with your daily time series of peak average ozone concentrations
- Extract the maximum value from each year
  - This forms an Annual Maximum Series (AMS)




| | | | |
|---|---|---|---|
| 1861 | 6 | 1 | 23.856 |
| 1861 | 6 | 2 | 30.696 |
| 1861 | 6 | 3 | 32.16 |
| 1861 | 6 | 4 | 28.004 |
| 1861 | 6 | 5 | 32.389 |
| 1861 | 6 | 6 | 32.03 |
| 1861 | 6 | 7 | 32.651 |
| 1861 | 6 | 8 | 31.264 |
| 1861 | 6 | 9 | 22.639 |
| 1861 | 6 | 10 | 26.97 |
| 1861 | 6 | 11 | 36.908 |
| 1861 | 6 | 12 | 33.8 |
| 1861 | 6 | 13 | 28.855 |
| 1861 | 6 | 14 | 31.678 |
| 1861 | 6 | 15 | 28.713 |
| 1861 | 6 | 16 | 28.585 |
| 1861 | 6 | 17 | 30.01 |
| 1861 | 6 | 18 | 32.867 |
| 1861 | 6 | 19 | 31.986 |
| 1861 | 6 | 20 | 31.478 |
| 1861 | 6 | 21 | 32.114 |
| 1861 | 6 | 22 | 35.853 |
| 1861 | 6 | 23 | 33.311 |
| 1861 | 6 | 24 | 25.496 |
| 1861 | 6 | 25 | 26.972 |
| 1861 | 6 | 26 | 27.247 |
| 1861 | 6 | 27 | 32.198 |
| 1861 | 6 | 28 | 26.713 |
| 1861 | 6 | 29 | 29.551 |
| 1861 | 6 | 30 | 24.999 |
| 1861 | 7 | 1 | 20.875 |
| 1861 | 7 | 2 | 29.508 |
| 1861 | 7 | 3 | 29.56 |
| 1861 | 7 | 4 | 27.951 |
| 1861 | 7 | 5 | 28.91 |
| 1861 | 7 | 6 | 29.517 |
| 1861 | 7 | 7 | 31.156 |
| 1861 | 7 | 8 | 32.588 |
| 1861 | 7 | 9 | 33.088 |
| 1861 | 7 | 10 | 29.334 |
| 1861 | 7 | 11 | 32.022 |
| 1861 | 7 | 12 | 36.745 |
| 1861 | 7 | 13 | 32.222 |
| 1861 | 7 | 14 | 38.537 |
| 1861 | 7 | 15 | 36.591 |
| 1861 | 7 | 16 | 31.549 |
| 1861 | 7 | 17 | 31.106 |
| 1861 | 7 | 18 | 30.073 |
| 1861 | 7 | 19 | 24.861 |
| 1861 | 7 | 20 | 25.783 |
| 1861 | 7 | 21 | 26.347 |
| 1861 | 7 | 22 | 28.833 |
| 1861 | 7 | 23 | 28.641 |
| 1861 | 7 | 24 | 26.1 |
| 1861 | 7 | 25 | 29.181 |
| 1861 | 7 | 26 | 31.795 |
| 1861 | 7 | 27 | 31.403 |
| 1861 | 7 | 28 | 29.92 |
| 1861 | 7 | 29 | 27.165 |
| 1861 | 7 | 30 | 24.74 |
| 1861 | 7 | 31 | 23.564 |
| 1861 | 8 | 1 | 23.987 |
| 1861 | 8 | 2 | 25.392 |
| 1861 | 8 | 3 | 28.897 |
| 1861 | 8 | 4 | 36.379 |
| 1861 | 8 | 5 | 37.512 |
| 1861 | 8 | 6 | 33.109 |
| 1861 | 8 | 7 | 29.225 |
| 1861 | 8 | 8 | 31.242 |
| 1861 | 8 | 9 | 33.191 |
| 1861 | 8 | 10 | 34.738 |
| 1861 | 8 | 11 | 33.48 |
| 1861 | 8 | 12 | 31.756 |
| 1861 | 8 | 13 | 31.728 |
| 1861 | 8 | 14 | 26.523 |
| 1861 | 8 | 15 | 22.243 |
| 1861 | 8 | 16 | 24.12 |
| 1861 | 8 | 17 | 25.15 |
| 1861 | 8 | 18 | 28.671 |
| 1861 | 8 | 19 | 28.224 |
| 1861 | 8 | 20 | 24.166 |
| 1861 | 8 | 21 | 24.443 |
| 1861 | 8 | 22 | 21.141 |
| 1861 | 8 | 23 | 20.105 |
| 1861 | 8 | 24 | 22.713 |
| 1861 | 8 | 25 | 24.244 |
| 1861 | 8 | 26 | 19.625 |
| 1861 | 8 | 27 | 20.888 |
| 1861 | 8 | 28 | 23.018 |
| 1861 | 8 | 29 | 26.904 |
| 1861 | 8 | 30 | 29.375 |
| 1861 | 8 | 31 | 28.404 |

# Some Python Code

```
import numpy #I never import numpy as np
import lmoments
#format per row is: [year, month, day, mda8]
hist_data = numpy.genfromtxt('historical_mda8.csv',delimiter=','
year_col, month_col, day_col, mda8_col = 0,1,2,3
DAYS_PER_YEAR = 30+31+31 #Only JJA data
NUM_YEARS = hist_data.shape[0]/DAYS_PER_YEAR
ams = numpy.zeros(NUM_YEARS)
for i in range(0,NUM_YEARS):
    ams[i] = numpy.max(hist_data[DAYS_PER_YEAR*i:DAYS_PER_YEAR*(i+1),mda8_col])
```

```
ams = array([ 39.7  ,  38.537,  49.378,  42.7  ,  42.864,  45.252,
41.516,  45.201,  44.725,  40.916,  37.896,  41.268,  47.356,
40.218,
       51.914,  39.179,  49.104,  40.688,  43.664,  57.477,  42.684,
       44.173,  44.981,  40.652,  40.201,  41.35 ,  41.878,  43.157,
       37.997,  52.527,  43.606,  48.933,  44.198,  42.538,  43.774,
       44.679,  54.046,  43.549,  40.843,  50.97 ,  38.518,  42.529,
       40.231,  44.136,  53.131,  41.754,  45.035,  41.678,  44.788,
       42.456,  42.526,  43.805,  41.547,  45.209,  41.795,  41.853,
       46.73 ,  40.788,  42.919,  41.734,  40.532,  42.192,  43.952,
       39.887,  44.805,  46.045,  44.854,  44.85 ,  40.341,  39.855,
       43.469,  45.708,  40.762,  43.631,  43.89 ,  44.225,  44.269,
       41.822,  43.18 ,  43.647,  48.838,  44.102,  46.04 ,  47.115,
       46.64 ,  46.512,  48.148,  48.624,  53.392,  50.559,  48.595,
       50.033,  49.847,  53.383,  50.436,  52.625,  49.836,  51.858,
       52.086,  50.463,  55.191,  52.52 ,  53.336,  57.24 ,  58.855,
       60.839,  58.76 ,  60.757,  59.233,  70.465,  60.48 ,  59.291,
       62.367,  62.559,  64.42 ,  66.503,  64.123,  66.42 ,  67.158,
       69.3  ,  68.274,  65.359,  66.468,  66.805,  68.198,  70.874,
       70.485,  68.611,  67.561,  71.122,  68.86 ,  69.813,  66.46 ,
       71.944,  69.933,  71.425,  71.682,  70.398,  69.033,  72.188,
       69.992,  74.937,  67.871,  65.661,  67.811])
```

# Distribution Fitting/Parameter Estimation

- After obtaining an AMS, ready to fit a GEV distribution to it
- Numerous methods of parameter estimation exist:
  1. Maximum Likelihood Estimation (MLE)
  2. **Method of L-Moments (MoLM)**
  3. Other less popular options

# L-Moments

- Analogous to traditional moments (mean, variance, skewness, kurtosis, etc.)
- Linear combinations of *ordered statistics* (ranked values)
- Suppose you have a set of measurements, $\{X_1, X_2, \ldots, X_{n-1}\}$, ordered smallest to largest. The n'th L-Moment may be defined as:

$\lambda_n(X) = \frac{1}{n}\sum_{i=0}^{n-1} -1^i \binom{n-1}{i} E[X_{n-i}]$, where E denotes expectation

- For a sample of size *s*, the sample L-moment may be computed as:

$$\widehat{\lambda_n}(X) = \frac{1}{n\binom{s}{n}} \sum_{i=1}^{s} \left( \sum_{b=0}^{n} -1^b \binom{n}{b} \binom{i-1}{n-b} \binom{s-1}{b} \right) X_i$$

- Probability Distribution parameters may be described as a function of the L-moments of the underlying distribution

# Python Code, Part II

lmoms = lmoments.samlmu(ams,4)

[51.324696551724152, 5.9213236590038347, 0.23805132564991779, 0.0062106119828451972]

params = lmoments.pelgev(lmoms)

[46.015121694247711, 7.692054511279907, -0.10329314112117975]

$\mu$ $\sigma$ $\xi$

- Voila! Ready to use fit to answer science questions.
- Science Question: Assuming the climate system has been (fairly) stationary over the historical period, how likely is a given year to experience a 75 ppb MDA8 exceedance?

lmoments.cdfgev(75,params)

0.95937276317621312

(1.0-lmoments.cdfgev(75,params))

0.040627236823786883 #The probability of exceeding 75 ppb MDA8 concentrations in Denver is about 4%

# Python Code, Part III

- Science Question: How does this compare with the Denver annual exceedance probability (AEP) of the climate system under RCP8.5?
- Simply repeat the process

```
rcp_data = numpy.genfromtxt('rcp85_mda8.csv',delimiter=',')
num_years_rcp = rcp_data.shape[0]/DAYS_PER_YEAR
rcp_ams = numpy.zeros(num_years_rcp)
for i in range(0,num_years_rcp):
    rcp_ams[i] = numpy.max(rcp_data[DAYS_PER_YEAR*i:DAYS_PER_YEAR*(i+1),mda8_col])
rcp_params = lmoments.pelgev(lmoments.samlmu(rcp_ams,4))
[69.052504208625095, 3.8698113941111165, -0.042061272523397253]
(1.0-lmoments.cdfgev(75,rcp_params))
0.20191405324688017 #Annual Exceedance Probability 20%
```

- Under RCP8.5, Denver Annual Exceedance Probability of $O_3$ MDA8 in excess of 75 ppb has increased from 4% to 20%

# Other Questions

- Can answer other related questions as well.  Often interested in what amount corresponds to a prescribed frequency.
- What ozone concentration has a probability of being exceeded in a given year of 1% in the historical climate?  In the RCP 8.5 climate?
- Historical:

lmoments.quagev(0.99,params)

91.312469413912154

- RCP8.5:

lmoments.quagev(0.99,rcp_params)

88.693087373127554

- The "100-year event", at least from the AMS perspective, actually goes down* under the RCP8.5 scenario, despite the increase in frequency of 75 ppb exceedances

# Threshold Exceedance/Partial Duration Series

- Often times, you may not be interested in the maximum values over a ‘block’.
- Consider the ozone problem; suppose any exceedances of 75 ppb are considered harmful, and these may occur multiple times in a given block/given year, and may go several years without any events of concern.
- In these scenarios, often most interested in PDS-based EVT statistics
- Also often termed “Peaks-over-Threshold” (PoT) method

# Pickands-Balkema-de Haan Theorem

- "The Second Theorem of Extreme Value Theory"
  - The one used for threshold exceedances
- Again, suppose you have some random variable X, sampled from an unknown underlying distribution F
- Define the conditional function $F_u$:

$$F_u(y) \equiv P((X-u) \leq y | X > u) = \frac{F(u+y) - F(u)}{1 - F(u)}$$

- $F_u$ describes the distribution of excesses over a given threshold *u*, given that the threshold has been exceeded
- Pickands-Balkema states:

$$\lim_{u \to \infty} F_u \sim GPD(\mu, \sigma, \xi)$$

where GPD denotes the *Generalized Pareto Distribution*

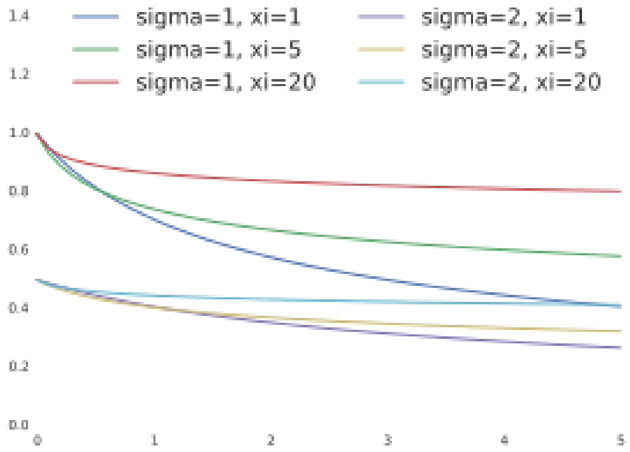
# Pickands-Balkema-de Haan Theorem

- What does this mean?
- Provided your underlying probability distribution F of a random variable X is not highly unusual, regardless of what F is, and provided that the threshold u is sufficiently large, exceedances of u will be distributed as the Generalized Pareto Distribution (GPD)



# Generalized Pareto Distribution (GPD)

- Three parameter distribution:
  1. Location parameter μ
     - Shifts distribution left/right
  2. Scale parameter σ
     - Determines 'spread' of distribution
  3. Shape parameter ξ
     - Determines shape of distribution, and distribution family to which it belongs
     - $$\begin{cases} Finite\ Tail & \xi < 0 \\ Exponential\ Decay\ Tail & \xi = 0\ (Exp.\, Dist.\, when\ \mu = 0) \\ Polynomial\ Decay\ Tail & \xi > 0 \left(Pareto\left({}^{\sigma}/_{\xi}, {}^{1}/_{\xi}\right)\right) \end{cases}$$




PDF: $f(x;\ \mu, \sigma, \xi) = \frac{1}{\sigma}\left(1 + \frac{x-\mu}{\sigma}\xi\right)^{-\frac{1}{\xi}-1}$

# Generalized Pareto Distribution- Other Properties

- Valid Interval: $\begin{cases} [\mu, \infty) & \xi \geq 0 \\ \left[\mu, \mu - \frac{\sigma}{\xi}\right] & \xi < 0 \end{cases}$
- CDF: $F(x;\ \mu, \sigma, \xi) = 1 - \left(1 + \frac{x-\mu}{\sigma}\xi\right)^{-\frac{1}{\xi}}$
- L-moments Estimators ($l_1$ mean, $l_2$ L-scale, $l_3$ L-skewness*):

$$\hat{\mu} = l_1 - l_2\left(2 + \frac{1 - 3l_3}{1 + l_3}\right)$$
$$\hat{\sigma} = l_2\left(1 + \frac{1 - 3l_3}{1 + l_3}\right)\left(2 + \frac{1 - 3l_3}{1 + l_3}\right)$$
$$\hat{\xi} = \frac{1 - 3l_3}{1 + l_3}$$

*$l_1 = \lambda_1;\ l_2 = \lambda_2;\ l_3 = \frac{\lambda_3}{\lambda_2}$

# PDS/PoT Methods: Application

- Usually more relevant than AMS-based statistics
- Also slightly more difficult to apply
  - Two Challenges:
    1. Selecting threshold *u*
    2. Ensuring data independence
- (1) a challenge because don't want to choose a threshold too low, else Pickands-Balkema won't apply. Don't want to choose a threshold too high, or you'll have a small sample and parameter estimates will be poor.
  - Common choice/lower bound: lowest AMS-value in the corresponding record (often choose slightly higher)
- (2) is a recurring problem with atmospheric data; adjacent samples in time often very correlated

# AMS/PDS Statistics

- Are AMS and PDS statistics related, and, if so, how?
- Yes.  Langbein (1949) related PDS-based average recurrence intervals (ARIs) to AMS-based annual exceedance probabilities (AEPs) through the simple equation:

$$AEP = 1 - e^{-1/ARI}$$

- Two approaches tend to converge at high return periods/rare events

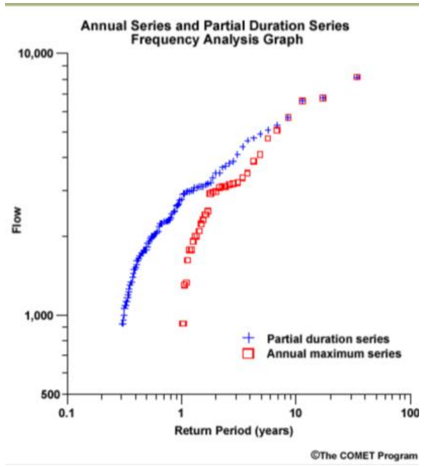



Table 2--Relationship of recurrence interval of floods on the two bases

| Partial-duration series | Annual floods |
|---|---|
| yr | yr |
| 0.5 | 1.16 |
| 1.0 | 1.58 |
| 1.45 | 2.00 |
| 2.0 | 2.54 |
| 5.0 | 5.52 |
| 10 | 10.5 |
| 20 | 20.5 |
| 50 | 50.5 |
| 100 | 100.5 |

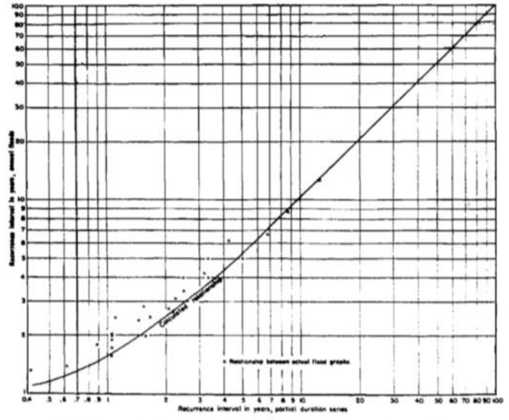


Fig. 1--Study of relationship between recurrence interval, partial-duration series and annual floods

# Uncertainty Quantification/Confidence Intervals

- Can always do by bootstrapping (most common, I think)
- Analytic alternative for GEV: $x(p) = \mu - (\sigma/\xi) \{1 - [- \log(1 - p)]\}{-}\xi$
- Confidence interval: Re-parameterize replacing location parameter $\mu$ with $x(p)$ & use profile likelihood

# Summary

- Extreme Value Theory has numerous applications throughout the atmospheric sciences and other fields
  - Many applications (outside of hydrology) have received limited attention or are just beginning to receive attention
- Very general: makes few assumptions about the underlying distribution of the time series data
- Relatively easy to apply: can be performed in ~10 lines of code
- Allows one to speak objectively and quantitatively about events which tend to have the largest societal impacts

# Python Code: E-Folding Time

```
#Determine e-folding time T
i = 1
corrcoef_sum = 0
DAYS_PER_YEAR = 30+31+31
NUM_YEARS = 145
while i < DAYS_PER_YEAR:
  for j in range(0,NUM_YEARS):
    hist_data_part = hist_data[j*DAYS_PER_YEAR:(j+1)*DAYS_PER_YEAR,mda8_col]
    corrcoef_sum += numpy.corrcoef(hist_data_part[i:],hist_data_part[:-1*i])[1,0]
  corrcoef_sum /= NUM_YEARS
  if corrcoef_sum < 1.0/math.e:
    break
  corrcoef_sum = 0
  i += 1
efold = i
#Result: efold = 2 (days); corr ~= 0.21 (corr(i=1) ~= 0.51)
```

# Python Code: Threshold Calculation

```
#Determine threshold u; use minimum AMS value
ams = numpy.zeros(NUM_YEARS)
for i in range(0,NUM_YEARS):
    ams[i] = numpy.max(hist_data[DAYS_PER_YEAR*i:DAYS_PER_YEAR*(i+1),mda8_col])
U = numpy.min(ams)
#Result: U = 37.896
```

# Python Code: Make Partial Duration Series

```
#Form PDS, given U and efold
pds = []
mda8_vals = hist_data[:,mda8_col]
pds_ufilt = mda8_vals[numpy.where(mda8_vals > U)]
pds_inds = numpy.where(mda8_vals > U)[0]
#Simple implementation that takes first independent exceedance
last_val = -1*efold
for i in range(0,mda8_vals.shape[0]):
  if i < last_val + efold:
    continue
  if mda8_vals[i] >= U:
    pds.append(mda8_vals[i])
    last_val = i
pds = numpy.array(pds)
#Result: pds.shape[0]/NUM_YEARS = 22.3 (x longer than AMS series…this is very high)
```

# Python Code: Fit and Apply

```
lmoms_pds = lmoments.samlmu(pds,4)
params_pds = lmoments.pelgpa(lmoms_pds)
```