Crash-Course in Extremes

Eric Gilleland

NATIONAL CENTER FOR ATMOSPHERIC RESEARCH

STATMOS Workshop on Spatial and Spatio-temporal EVA and Oceanography
21 to 23 May 2019

Modeling Block Maxima

Let $X_1, \dots, X_n \sim F_X(x)$ be IID.

Interest is in the distribution function for the maximum value of $X_i, i = 1, \dots, n$.

We know from our statistical theory class that:

$$\mathbb{P}[\max\{X_1, \dots, X_n\} \leq z] = \mathbb{P}[X_1 \leq z, \cdots, X_n \leq z] =$$

$$\prod_{i=1}^{n} \mathbb{P}[X \leq z] = \prod_{i=1}^{n} F_X(z) = F_X^n(z)$$

If n is large, $F^n \to 0$.

Also, F must be estimated, and small errors raised to the n-th power are exacerbated as n increases.

Modeling Block Maxima

ESPAÑOL ENTER A SCRATCH BONUS DRAW WHERE TO PLAY LOG IN TO MY LOTTERY

LOTTERY Games About Game Tools Winners News Giving Back

Powerball | It's Easy to Play | Drawing History | Claim Winnings

POWERBALL PAYOUTS & ODDS OF WINNING

MATCH	PRIZE	x2	x3	x4	x5	x10*	ODDS
5 + PB	Jackpot**	Power Play does not apply to jackpot**					1 in 292,201,338
5 of 5	$1,000,000	$2,000,000					1 in 11,688,054
4 + PB	$50,000	$100,000	$150,000	$200,000	$250,000	$500,000	1 in 913,129
4 of 5	$100	$200	$300	$400	$500	$1,000	1 in 36,525
3 + PB	$100	$200	$300	$400	$500	$1,000	1 in 14,494
3 of 5	$7	$14	$21	$28	$35	$70	1 in 580
2 + PB	$7	$14	$21	$28	$35	$70	1 in 701
1 + PB	$4	$8	$12	$16	$20	$40	1 in 92
PB	$4	$8	$12	$16	$20	$40	1 in 38

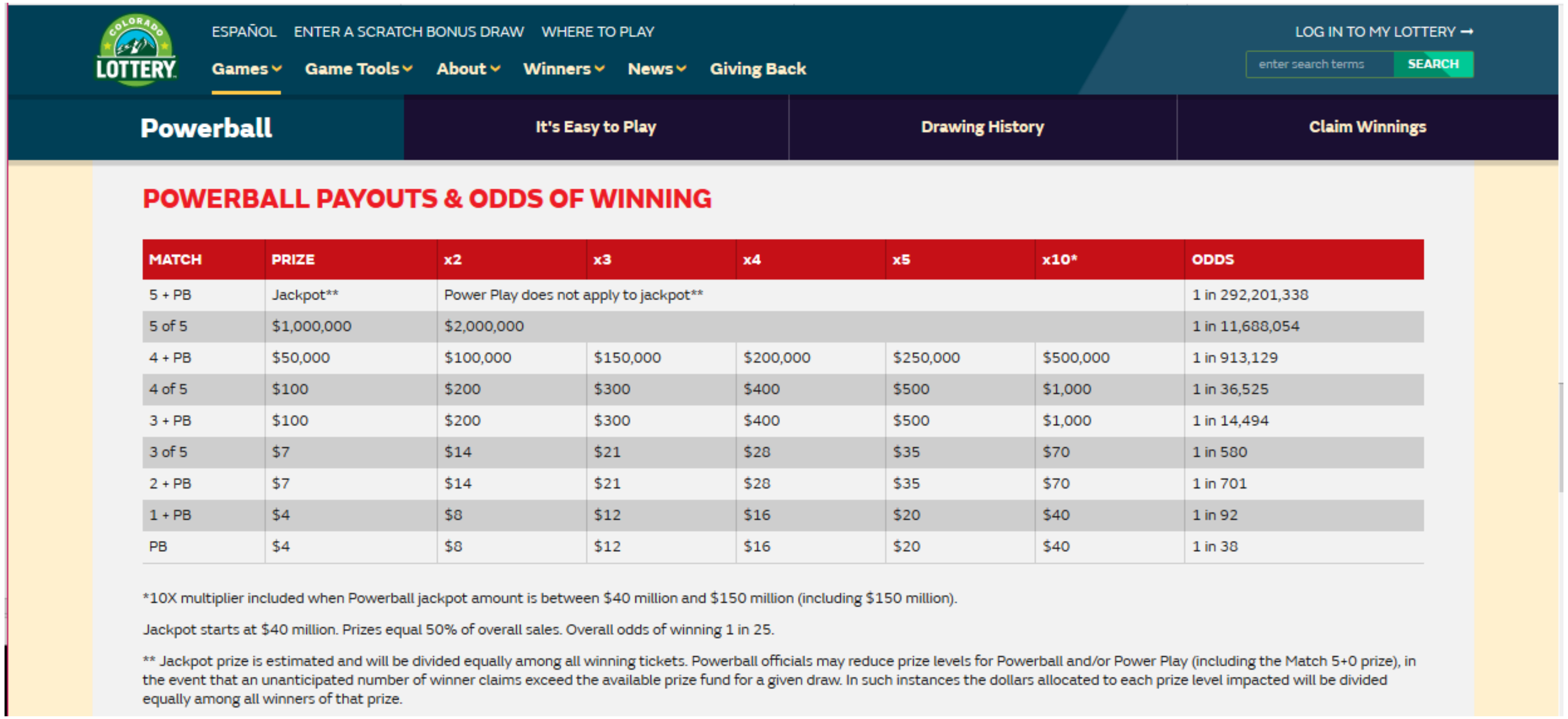
*10X multiplier included when Powerball jackpot amount is between $40 million and $150 million (including $150 million).

Jackpot starts at $40 million. Prizes equal 50% of overall sales. Overall odds of winning 1 in 25.

** Jackpot prize is estimated and will be divided equally among all winning tickets. Powerball officials may reduce prize levels for Powerball and/or Power Play (including the Match 5+0 prize), in the event that an unanticipated number of winner claims exceed the available prize fund for a given draw. In such instances the dollars allocated to each prize level impacted will be divided equally among all winners of that prize.

- To win the jackpot, need to get all five numbers sampled without replacement from 1 to 69 and the power ball, which is one number selected from 1 to 26.
- $\mathbb{P}26\ 42 \times -[\text{Jackpot}] = \frac{1}{\binom{69}{5}} \cdot \frac{1}{26} \approx 3.4 \times 10^{-9}$

Modeling Block Maxima



ESPAÑOL ENTER A SCRATCH BONUS DRAW WHERE TO PLAY

LOTTERY

Games Game Tools About Winners News Giving Back

LOG IN TO MY LOTTERY

enter search terms SEARCH

Powerball | It's Easy to Play | Drawing History | Claim Winnings

POWERBALL PAYOUTS & ODDS OF WINNING

MATCH	PRIZE	x2	x3	x4	x5	x10*	ODDS
5 + PB	Jackpot**	Power Play does not apply to jackpot**					1 in 292,201,338
5 of 5	$1,000,000	$2,000,000					1 in 11,688,054
4 + PB	$50,000	$100,000	$150,000	$200,000	$250,000	$500,000	1 in 913,129
4 of 5	$100	$200	$300	$400	$500	$1,000	1 in 36,525
3 + PB	$100	$200	$300	$400	$500	$1,000	1 in 14,494
3 of 5	$7	$14	$21	$28	$35	$70	1 in 580
2 + PB	$7	$14	$21	$28	$35	$70	1 in 701
1 + PB	$4	$8	$12	$16	$20	$40	1 in 92
PB	$4	$8	$12	$16	$20	$40	1 in 38

*10X multiplier included when Powerball jackpot amount is between $40 million and $150 million (including $150 million).

Jackpot starts at $40 million. Prizes equal 50% of overall sales. Overall odds of winning 1 in 25.

** Jackpot prize is estimated and will be divided equally among all winning tickets. Powerball officials may reduce prize levels for Powerball and/or Power Play (including the Match 5+0 prize), in the event that an unanticipated number of winner claims exceed the available prize fund for a given draw. In such instances the dollars allocated to each prize level impacted will be divided equally among all winners of that prize.

- $\mathbb{P}26\ 42 \times - [\text{Jackpot}] = \frac{1}{\binom{69}{5}} \cdot \frac{1}{26} \approx 3.4 \times 10^{-9}$
- Poisson distribution approximates the binomial when the success rate is constant as the sample size increases and the probability of success goes to zero.

Modeling Block Maxima

ESPAÑOL ENTER A SCRATCH BONUS DRAW WHERE TO PLAY

LOTTERY

Games About Game Tools Winners News Giving Back

LOG IN TO MY LOTTERY

enter search terms SEARCH

Powerball It's Easy to Play Drawing History Claim Winnings

POWERBALL PAYOUTS & ODDS OF WINNING

MATCH	PRIZE	x2	x3	x4	x5	x10*	ODDS
5 + PB	Jackpot**	Power Play does not apply to jackpot**					1 in 292,201,338
5 of 5	$1,000,000	$2,000,000					1 in 11,688,054
4 + PB	$50,000	$100,000	$150,000	$200,000	$250,000	$500,000	1 in 913,129
4 of 5	$100	$200	$300	$400	$500	$1,000	1 in 36,525
3 + PB	$100	$200	$300	$400	$500	$1,000	1 in 14,494
3 of 5	$7	$14	$21	$28	$35	$70	1 in 580
2 + PB	$7	$14	$21	$28	$35	$70	1 in 701
1 + PB	$4	$8	$12	$16	$20	$40	1 in 92
PB	$4	$8	$12	$16	$20	$40	1 in 38

*10X multiplier included when Powerball jackpot amount is between $40 million and $150 million (including $150 million).

Jackpot starts at $40 million. Prizes equal 50% of overall sales. Overall odds of winning 1 in 25.

** Jackpot prize is estimated and will be divided equally among all winning tickets. Powerball officials may reduce prize levels for Powerball and/or Power Play (including the Match 5+0 prize), in the event that an unanticipated number of winner claims exceed the available prize fund for a given draw. In such instances the dollars allocated to each prize level impacted will be divided equally among all winners of that prize.

- $\mathbb{P}26\ 42 \times - [\text{Jackpot}] = \frac{1}{\binom{69}{5}} \cdot \frac{1}{26} \approx 3.4 \times 10^{-9}$
- Waiting time between winning the jackpot (between Poisson events) is governed by the exponential distribution with rate equal to the reciprocal of the Poisson rate parameter.

Modeling Block Maxima

Let $X_1, \dots, X_n \sim F_X(x)$ be IID.

Connection between distribution for maxima and number of rare events (Poisson)

Let $N \sim \text{Poisson}(\lambda)$, represent the event that $X > u$ for some large constant threshold u.

$$\mathbb{P}[\max\{X_1, \dots, X_n\} \leq u] = \mathbb{P}[N = 0] = e^{-\lambda}$$

Modeling Block Maxima

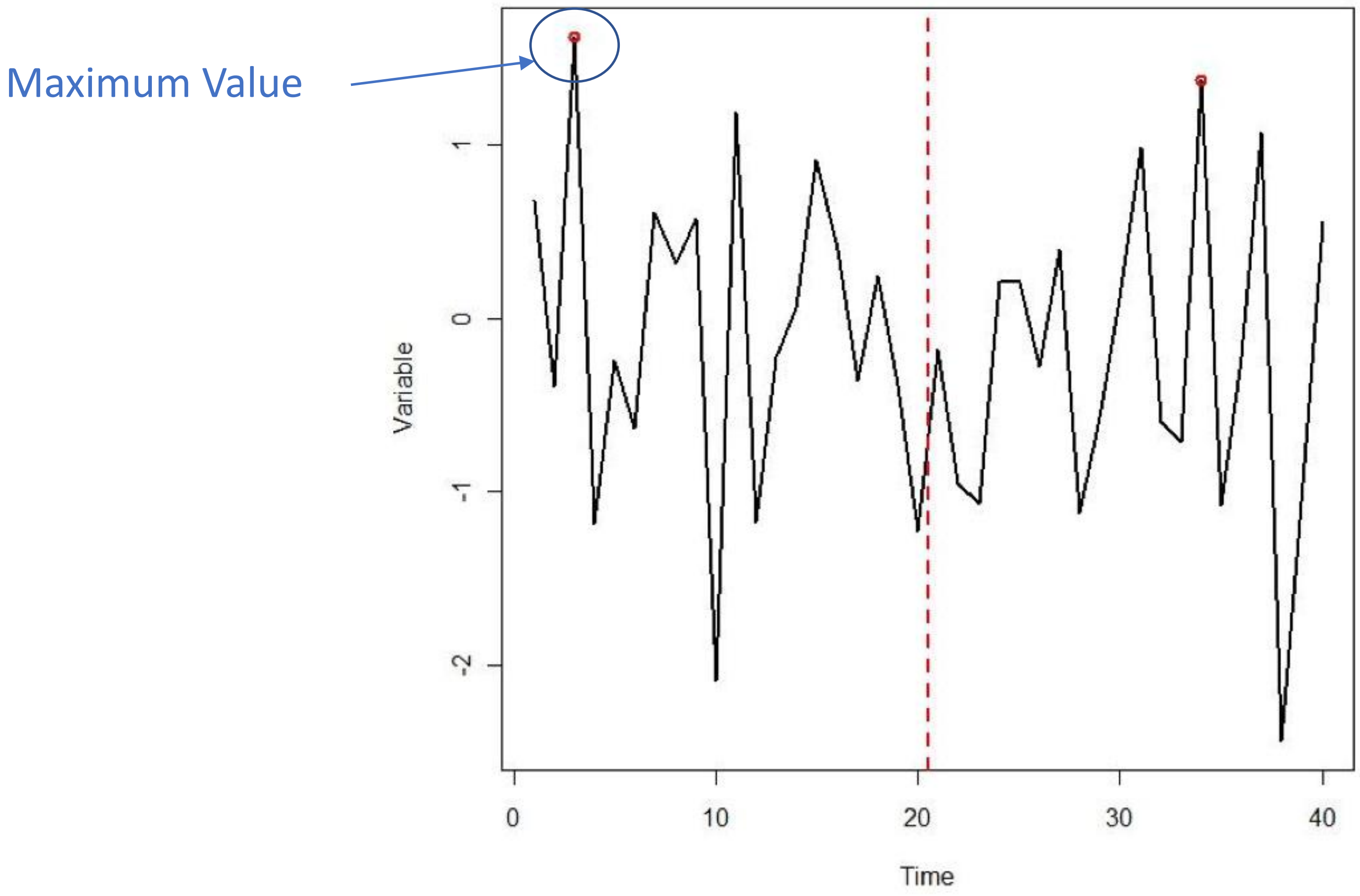
Sum stability: Let $X_1, \ldots, X_n \sim N\ (\mu, \sigma^2)$ be IID. Then, we have (exactly)

$$S_n = \sum_{i=1}^{n} X_i = X_1 + \cdots + X_n \sim N(n\mu, n\sigma^2\ )$$

That is,

$$\frac{S_n - n\mu}{\sqrt{n}\sigma} \sim N(0,1)$$

Modeling Block Maxima



Modeling Block Maxima

Max stability: Let $X_1, \dots, X_n \sim F_X$ be IID. Want to find a distribution for F_X^n that has the same form as F_X.

For $n = 2$, for example, notice that
$\max\{X_1, \dots, X_n\} = \max\{\max\{X_1, \dots, X_k\}, \max\{X_{k+1}, \dots, X_n\}\}$

Want a distribution, G, such that $G^2(z) = G(az + b)$ for constants $a > 0$ and b. More generally, that $G^n(z) = G(a_n z + b_n)$.

Set $M_n = \max\{X_1, \dots, X_n\}$, and suppose there exist sequences of constants $a_n > 0$ and b_n such that

$$\mathbb{P}\left[\frac{M_n - b_n}{a_n} \leq z\right] \to G(z) \text{ as } n \to \infty$$

Modeling Block Maxima

Max stability: If $G(\cdot)$ is non-degenerate, then it is called ***max-stable*** and has the form

$$G(z;\mu,\sigma,\xi) = \exp\left\{-\left[1+\frac{\xi}{\sigma}(z-\mu)\right]_+^{-1/\xi}\right\}$$

where $\sigma > 0$, $-\infty < \mu, \xi < \infty$ and $[x]_+ = \max(0, x)$.

Disguises a lot!

Modeling Block Maxima

$G(z;\mu,\sigma,\xi) =$

$$\exp\left\{-\left[1+\frac{\xi}{\sigma}(z-\mu)\right]^{-1/\xi}\right\}\cdot I_{(-\infty,\mu-\sigma/\xi]}(z)\cdot I_{(-\infty,0)}(\xi) +$$

$$\exp\left\{-\exp\left(\frac{z-\mu}{\sigma}\right)\right\}\cdot I_{\{0\}}(\xi) +$$

$$\exp\left\{-\left[1+\frac{\xi}{\sigma}(z-\mu)\right]^{-1/\xi}\right\}\cdot I_{[\mu-\sigma/\xi,\infty)}(z)\cdot I_{(0,\infty)}(\xi)$$

Modeling Block Maxima

$G(z;\mu,\sigma,\xi) =$

Upper bound
(reverse) Weibull distribution ($\xi < 0$)

$$\exp\left\{-\left[1+\frac{\xi}{\sigma}(z-\mu)\right]^{-1/\xi}\right\}\cdot I_{(-\infty,\mu-\sigma/\xi]}(z)\cdot I_{(-\infty,0)}(\xi)+$$

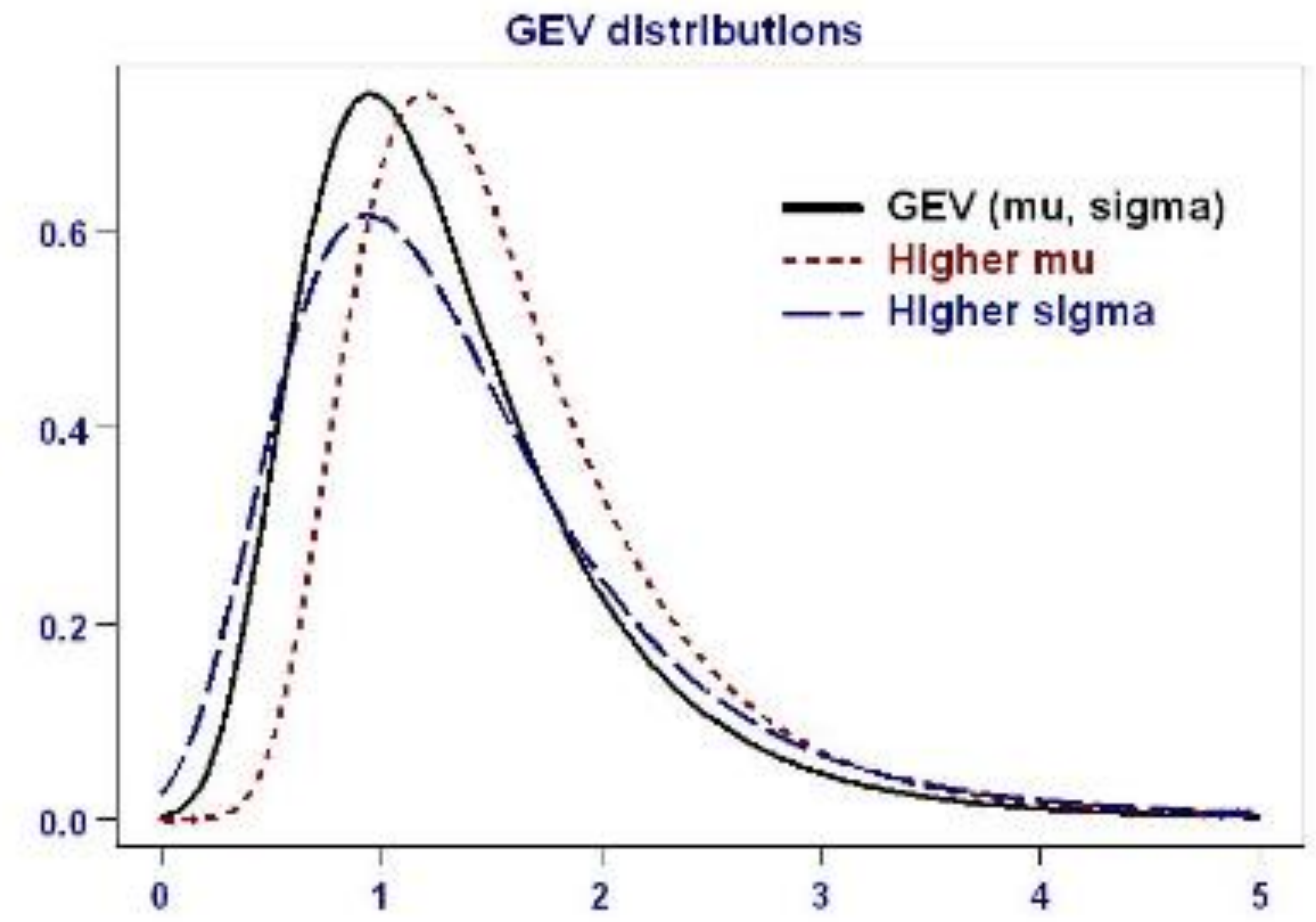
Light tail Gumbel distribution
defined by continuity as $\xi \to 0$

$$\exp\left\{-\exp\left(\frac{z-\mu}{\sigma}\right)\right\}\cdot I_{\{0\}}(\xi)+$$

$$\exp\left\{-\left[1+\frac{\xi}{\sigma}(z-\mu)\right]^{-1/\xi}\right\}\cdot I_{[\mu-\sigma/\xi,\infty)}(z)\cdot I_{(0,\infty)}(\xi)$$

Heavy-tail Fréchet distribution ($\xi > 0$)

Modeling Block Maxima



$\mathbb{E}[X] = \mu - \frac{\sigma(1-\Gamma(1-\xi))}{\xi}$, for $\xi < 1$

$\mathbb{V}[X] = \frac{\sigma^2\left(\Gamma(1-2\xi)-\Gamma^2(1-\xi)\right)}{\xi^2}$, for $\xi < \frac{1}{2}$

GEV Return Levels

Usually it is desired to estimate the T-year return level, which is easy to calculate for the GEV when fit to annual maxima as they are equivalent to the quantiles of the GEV, which can be found analytically.

The T-year return period is the $1 - \frac{1}{T}$ quantile of the GEV, or

$$G^{-1}\left(1 - \frac{1}{T}\right)$$

GEV Return Levels

Usually it is desired to estimate the T-year return level, which is easy to calculate for the GEV when fit to annual maxima as they are equivalent to the quantiles of the GEV, which can be found analytically.

The T-year return period is the $1 - \frac{1}{T}$ quantile of the GEV, or

$$G^{-1}\left(1 - \frac{1}{T}\right)$$

It is the value expected to occur, on average, once every T years. Typically plotted on a transformed scale so that if $\xi < 0$, the graph is concave with an asymptote at the upper bound, if $\xi = 0$, it is a straight line, and if $\xi > 0$ it is convex. Values of ξ that are close to zero will be close to a straight line.

$(1 - \alpha) \cdot 100\%$ CI's can be obtained from the delta method if maximum-likelihood (ML) estimation is used. Profile-likelihood CI's are generally more accurate, and realistic, but difficult to automate. Parametric bootstrap CI's are generally suitable.

Modeling Block Maxima

GEV distribution

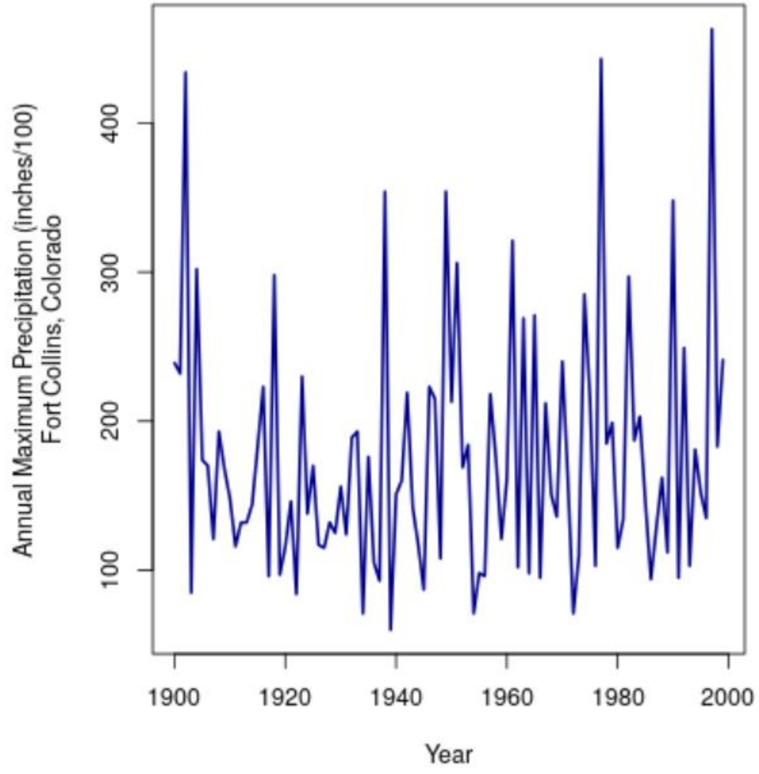
- Fit GEV directly to maxima over long blocks (e.g., annual)

Advantages of modeling block maxima

- Generally do not need to explicitly model annual or diurnal cycles
- Generally do not need to worry about temporal dependence
- Quantiles are easy to find and are equivalent to the T-period return level

Fitting the GEV to data

Fort Collins annual maximum precipitation (inches)

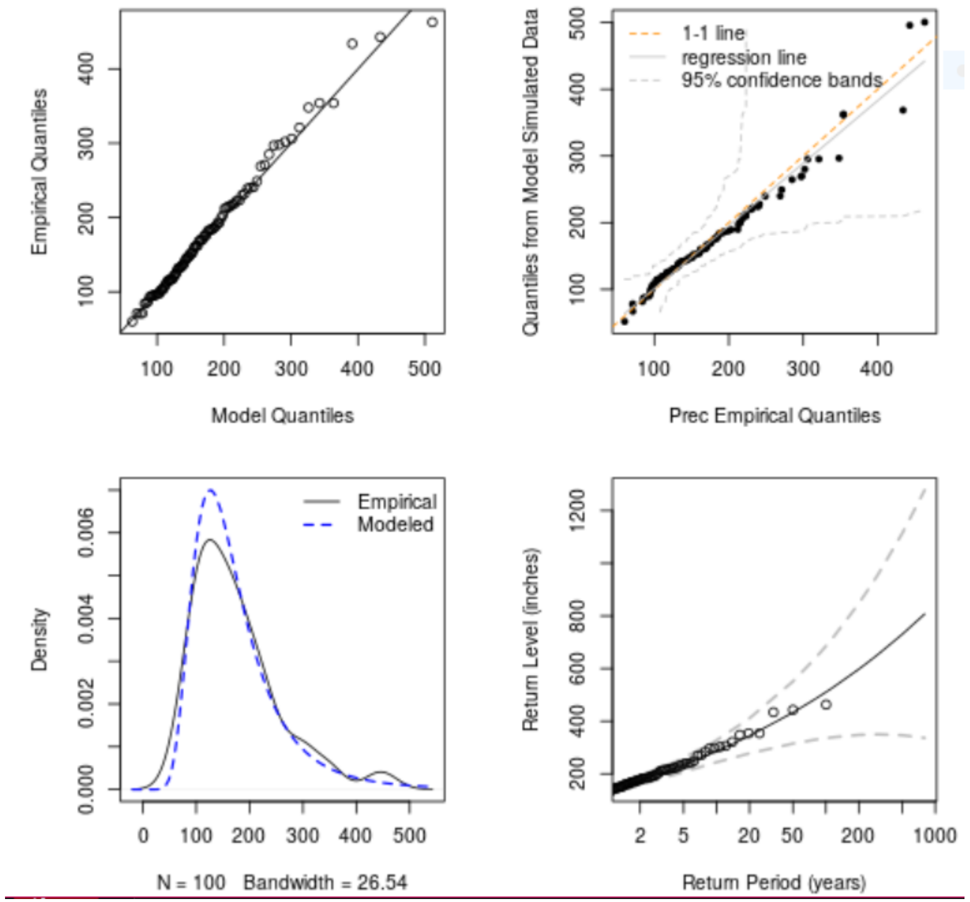



```
library( “extRemes” )
data( “ftcanmax” )

plot(ftcanmax, type="l", lwd=2,
   ylab = "Annual Maximum Precipitation (inches/100)\nFort Collins, Colorado", col = "darkblue")
```

Fitting the GEV to data

Fort Collins annual maximum precipitation (inches)

```
fit <- fevd( Prec, data = ftcanmax, units = “inches/100" )
fit
plot( fit )
[1] "Estimation Method used: MLE"


 Negative Log-Likelihood Value:  565.4816


 Estimated parameters:
 location     scale     shape 
134.66520  53.28089   0.17363 

 Standard Error Estimates:
  location      scale      shape 
6.16877130 4.87901653 0.09195688 

 Estimated parameter covariance matrix.
           location       scale        shape
location 38.0537393 17.06709037 -0.208376947
scale    17.0670904 23.80480234 -0.086937257
shape    -0.2083769 -0.08693726  0.008456068

 AIC = 1136.963 

 BIC = 1144.779
```



Fitting the GEV to data

Fort Collins annual maximum precipitation (inches)

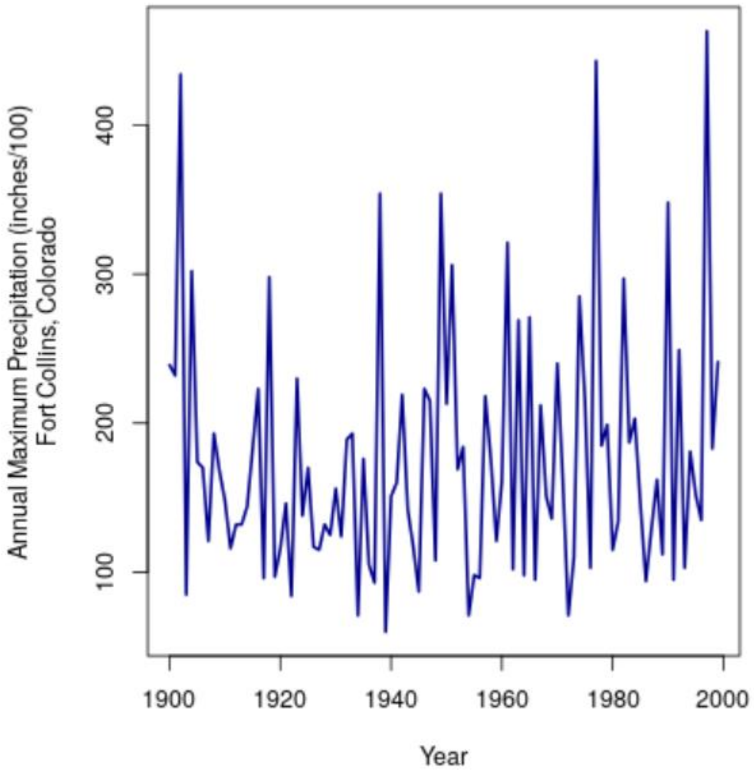
Data appear to be stationary, but can be difficult to tell for extremes. Can test using the likelihood-ratio test.

```
fit2 <- fevd( Prec, data = ftcanmax,
   location.fun = ~Year, units = "inches/100" )
```

```
lr.test( fit, fit2 )
```

likelihood-ratio $\approx -0.001 < \chi_1^2 = 3.8415$

Can also compare AIC/BIC values (lower is better)
stationary model: AIC = 1136.96, BIC = 1144.78
temporal trend model: AIC = 1138.96, BIC = 1149.39



Models $\mu(\text{Year}) = \mu_0 + \mu_1 \cdot \text{Year}$
Should use location.fun = ~ I(Year – 1900), but results are analogous in this case.

Modeling Excesses over a Threshold

Excesses over the dam (threshold)

- Interest might be in $Y = X - u$ conditioned on $X > u$.
- Analogue to max-stability (POT-stability)
- Y has an approximate generalized Pareto (GP) distribution given by

$$\mathbb{P}[Y \leq y | Y > 0] = H(y; \sigma_u, \xi) = 1 - \left[1 + \frac{\xi}{\sigma_u} y\right]_+^{-1/\xi}$$

Modeling Excesses over a Threshold

Peaks-Over-Threshold (POT) Stability: Suppose excesses, $Y = X - u$, have an exact GP distribution with parameters σ_u and ξ. Then, the excesses over a higher threshold, $\nu > u$, has a GP distribution with parameters σ_ν and ξ, where

$$\sigma_\nu = \sigma_u + \xi(\nu - u), \text{for } \nu > u$$

Connection between GP and GEV

Suppose $M_n = Z \sim \mathrm{GEV}(\mu, \sigma, \xi)$.

- We already saw the connection between M_n and the Poisson
- Note that the binomial success probability, that $X > u$, is given by $\mathbb{P}[X > u] = 1 - F_X(u)$.
- Using the Poisson approximation to the binomial and the fact that $\mathbb{P}[M_n \leq \cdot] = F_X^n(\cdot)$ gives $\mathbb{P}[M_n \leq u] \approx \exp\left[-n\left(1 - F_X(u)\right)\right]$ for large n and u such that $n\left(1 - F_X(u)\right)$ is approximately constant.

Connection between GP and GEV

Connection between GP and GEV df. Suppose $M_n = Z \sim \mathrm{GEV}(\mu, \sigma, \xi)$.

- Note further that $M_n = \max\{Y_1, \dots, Y_n\} + u$, so that only the upper tail of the distribution, $F_X(\cdot)$, determines the distribution of M_n.
- We seek a distribution for $Y = X - u$, conditional on $Y > 0$ (i.e., $X > u$)

$$F_{Y|Y>0}(y) = \mathbb{P}[Y > y|Y > 0] = \mathbb{P}[X > y + u|X > u] = \frac{\mathbb{P}[X > y + u, X > u]}{\mathbb{P}[X > u]} = \frac{\mathbb{P}[X > y + u]}{\mathbb{P}[X > u]} =$$

$$\frac{1 - F_X(y + u)}{1 - F_X(u)} \text{ for some } y > u$$

Connection between GP and GEV

Connection between GP and GEV df. Suppose $M_n = Z \sim \mathrm{GEV}(\mu, \sigma, \xi)$.

Now, $F_X^n(z) \approx G(z)$, where $G(z)$ is a GEV df with parameters μ, σ and ξ.

So,

$$n \cdot \log F_X(z) \approx -\left[1 + \frac{\xi}{\sigma}(z - \mu)\right]^{-1/\xi}$$

Connection between GP and GEV

For large values of z, the Taylor's series approximation gives that $\log F_X(z) \approx -\{1 - F_X(z)\}$, so that

$$-n \cdot \left(1 - F_X(z)\right) \approx -\left[1 + \frac{\xi}{\sigma}(z - \mu)\right]^{-1/\xi}$$

Thus,

$$1 - F_X(u) \approx \frac{1}{n}\left[1 + \frac{\xi}{\sigma}(u - \mu)\right]^{-1/\xi}$$

Connection between GP and GEV

Putting it all together

$$\mathbb{P}[Y > y | Y > 0] = \frac{1 - F_X(y+u)}{1 - F_X(u)} \approx$$

$$1 - F_X(u) \approx \frac{\frac{1}{n}\left[1 + \frac{\xi}{\sigma}(y + u - \mu)\right]^{-1/\xi}}{\frac{1}{n}\left[1 + \frac{\xi}{\sigma}(u - \mu)\right]^{-1/\xi}}$$

Connection between GP and GEV

$$\mathbb{P}[Y > y | Y > 0] = \frac{1 - F_X(y + u)}{1 - F_X(u)} \approx$$

$$1 - F_X(u) \ \approx \frac{\frac{1}{n}\left[1 + \frac{\xi}{\sigma}(y + u - \mu)\right]^{-1/\xi}}{\frac{1}{n}\left[1 + \frac{\xi}{\sigma}(u - \mu)\right]^{-1/\xi}} =$$

$$\left[\frac{1 + \xi(y + u - \mu)/\sigma}{1 + \xi(u - \mu)/\sigma}\right]^{-1/\xi} = \left[1 + \frac{\xi y}{\sigma + \xi(u - \mu)}\right]^{-1/\xi}$$

Connection between GP and GEV

Finally, $\mathbb{P}[Y \leq y | Y > 0] = 1 - \mathbb{P}[Y > y | Y > 0]$. Thus, $\mathrm{Y} \sim H(y; \sigma_u, \xi)$ with

$$\sigma_u = \sigma + \xi(u - \mu)$$

GPD Return Levels

The GPD quantiles, like those for the GEV, are easy to find. However, they do not correspond directly to the T-year return level. Need to first estimate $\mathbb{P}[X > u]$, and then it is straightforward to estimate return levels.

Frequency of Extremes

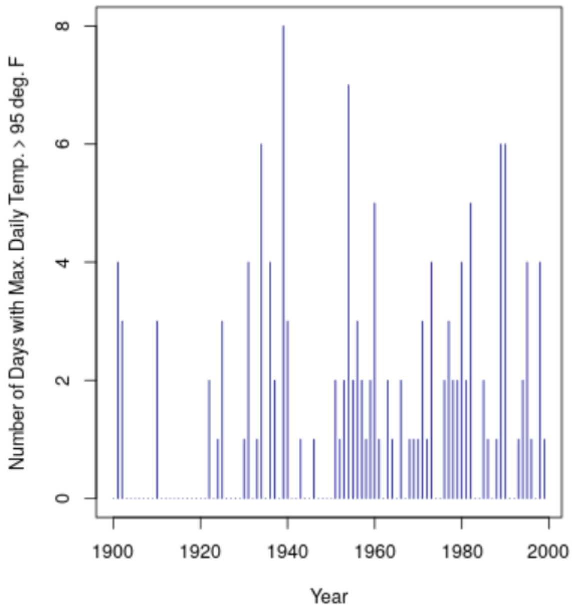
Number of days that maximum daily temperature (deg. F) in Fort Collins, Colorado exceeds 95 degrees F.

```
data(FCwx)

tempGT95 <- c(aggregate(FCwx$MxT,
    by = list(FCwx$Year),
    function(x) sum(x > 95, na.rm = TRUE))$x)

yr <- unique(FCwx$Year)
```

Frequency of Extremes

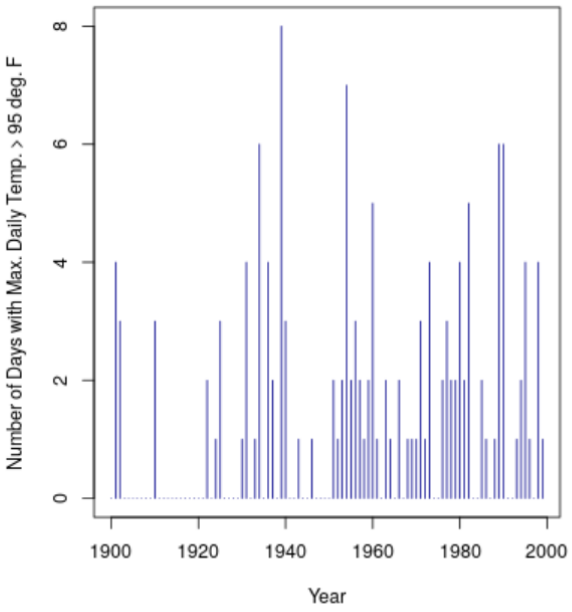
Number of days that maximum daily temperature (deg. F) in Fort Collins, Colorado exceeds 95 degrees F.

```
plot(yr, tempGT95,
    type = "h",
    col = "darkblue",
    xlab = "Year",
    ylab =
"Number of Days with Max. Daily Temp. > 95 deg. F")
```



Frequency of Extremes

Number of days that maximum daily temperature (deg. F) in Fort Collins, Colorado exceeds 95 degrees F.



Test for equality of mean and variance (Poisson distribution has mean and variance that are equal).

fpois(tempGT95)

mean = 1.4, variance = 3.3

$$234.29 > \chi^2_{n-1=99} = 123.23$$

Frequency of Extremes

Number of days that maximum daily temperature (deg. F) in Fort Collins, Colorado exceeds 95 degrees F.

Perhaps distribution is inhomogeneous. Does seem to have higher frequency with year.

```
fit <- glm(tempGT95~yr, family = poisson())
summary(fit)
```

```
Call:
glm(formula = tempGT95 ~ yr, family = poisson())

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-2.1390  -1.4250  -0.8085   0.4722   4.1276

Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept) -22.103934   5.924385  -3.731 0.000191 ***
yr            0.011483   0.003024   3.797 0.000146 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for poisson family taken to be 1)

    Null deviance: 228.36  on 99  degrees of freedom
Residual deviance: 213.46  on 98  degrees of freedom
AIC: 362.15

Number of Fisher Scoring iterations: 6
```

For more information…

- extRemes tutorial: Gilleland, E. and R. W. Katz, 2016. extRemes 2.0: An Extreme Value Analysis Package in R. *Journal of Statistical Software*, **72** (8), 1 - 39, DOI: 10.18637/jss.v072.i08.
- Rick Katz's web page preserved: https://ral.ucar.edu/projects/extremes/Extremes/