Using R to Analyze Extremes

Eric Gilleland,
National Center for
Atmospheric Research,
Boulder, Colorado, U.S.A.
http://www.ral.ucar.edu/staff/ericg

Statistical Assessment of Extreme Weather Phenomena Under Climate Change. *NCAR Advanced Study Program Summer Colloquium 2011*, 6–24 June.

The R programming language

R Development Core Team (2008). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0, http://www.r-project.org

Vance A, 2009. Data analysts captivated by R's power. *New York Times*, 6 January 2009. Available at: http://www.nytimes.com/2009/01/07/technology/business-computing/07program.html?_r=2

Already familiar with R?

Advanced (potentially useful) topics:

- Reading and Writing NetCDF file formats:
  http://www.image.ucar.edu/Software/Netcdf/
- A Climate Related Precipitation Example for Colorado:
  http://www.image.ucar.edu/~nychka/FrontrangePrecip/

Example

Fort Collins, Colorado daily precipitation amount

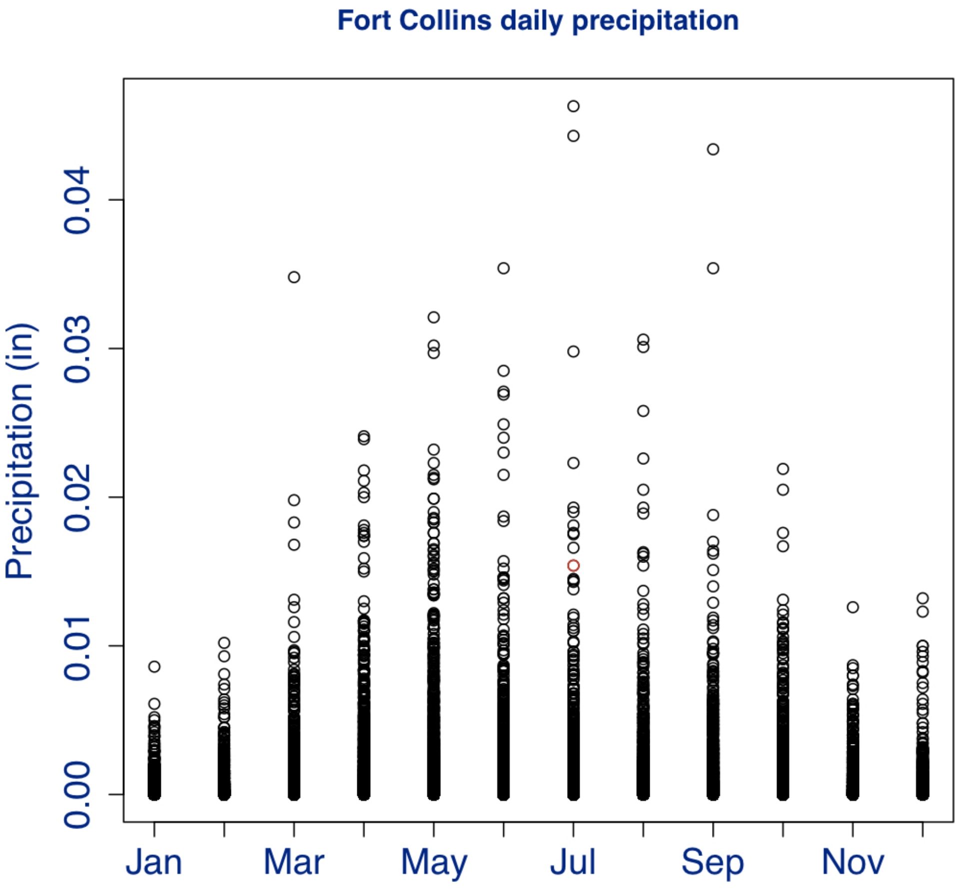
`http://ccc.atmos.colostate.edu/~odie/rain.html`

- Time series of daily precipitation amount (in), 1900–1999.
- Semi-arid region.
- Marked annual cycle in precipitation
  (wettest in late spring/early summer, driest in winter).
- No obvious long-term trend.
- Recent flood, 28 July 1997.
  (substantial damage to Colorado State University)

See, Katz et al. (2002), Adv. Water Res., 25:1287–1304 for more on these data. Source: Colorado Climate Center, Colorado State University (<URL: http://ulysses.atmos.colostate.edu>).

Example

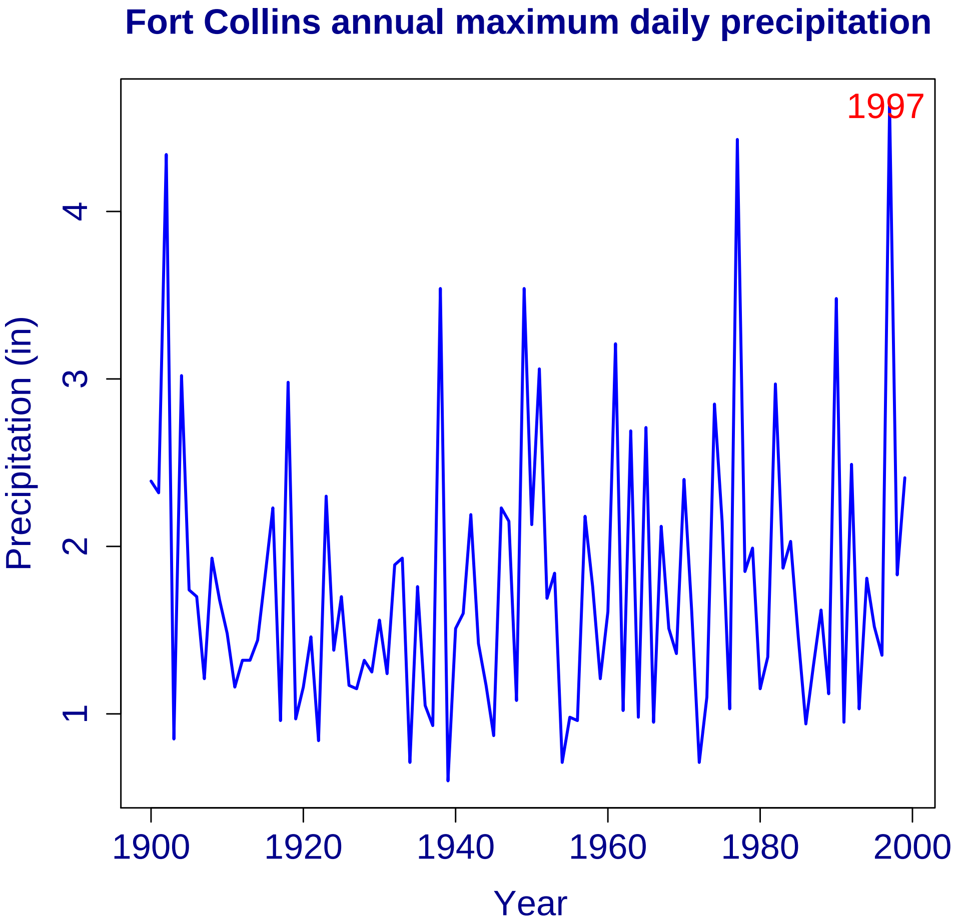
Fort Collins, Colorado precipitation



Example

Fort Collins, Colorado precipitation

Annual Maxima



Example

Fort Collins, Colorado precipitation

How often is such an extreme expected?

- Assuming no long-term trend emerges;
- Using annual maxima removes effects of seasonal trend in analysis.

```
require( extRemes)
data( ftcanmax)

# Fit GEV to Fort Collins annual maximum precipitation.
fit <- gev.fit( ftcanmax$Prec/100)

# Check the quality of the fit.
gev.diag( fit)
```

Example

Fort Collins, Colorado precipitation

Fit looks good (from diagnostic plots).

Parameter	Estimate	(Std. Error)
Location (μ)	1.347	(0.617)
Scale (σ)	0.533	(0.488)
Shape (ξ)	0.174	(0.092)

Heavy tail!

Example

Fort Collins, Colorado precipitation

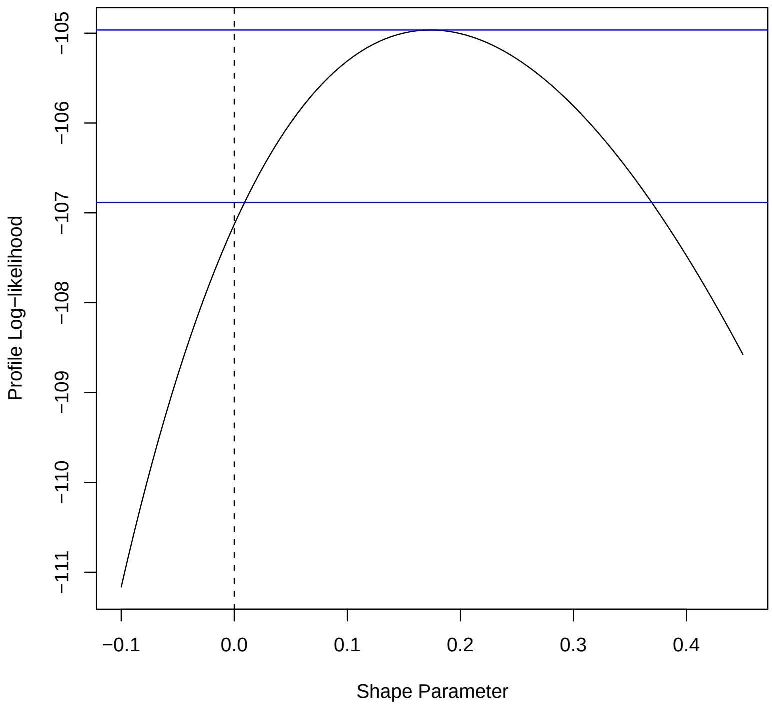
```
# Is the shape parameter really not zero?
# Perform likelihood ratio test against Gumbel type.
fit0 <- gum.fit( ftcanmax$Prec/100)
Dev <- 2*(fit0$nllh - fit$nllh)
pchisq( Dev, 1, lower.tail=FALSE)
```

Likelihood ratio test for $\xi = 0$ rejects hypothesis of Gumbel type (p-value ≈ 0.038).

Example

Fort Collins, Colorado precipitation

95% Confidence intervals for ξ, using profile likelihood, are: (0.009, 0.369).



Use `gev.profxi` and `locator(2)` to find CI's.

Example

Fort Collins, Colorado precipitation

Return Levels

```
fit.rl <- return.level( fit)
```

Return Period	Estimated Return Level (in)	95% CI
10	2.81	(2.41, 3.21)
100	5.10	?(3.35, 6.84)
$\vdots$	$\vdots$	$\vdots$

Example

Fort Collins, Colorado precipitation

Return Levels

CI's from `return.level` are based on the delta method, which assumes normality for the return levels. For longer return periods (e.g., beyond the range of the data), this assumption may not be valid. Can check by looking at the profile likelihood.

```
gev.prof( fit, m=100, xlow=2, xup=8)
```

Highly skewed! Using `locator(2)`, a better approximation for the (95%) 100-year return level CI is about (3.9, 8.0).

Example

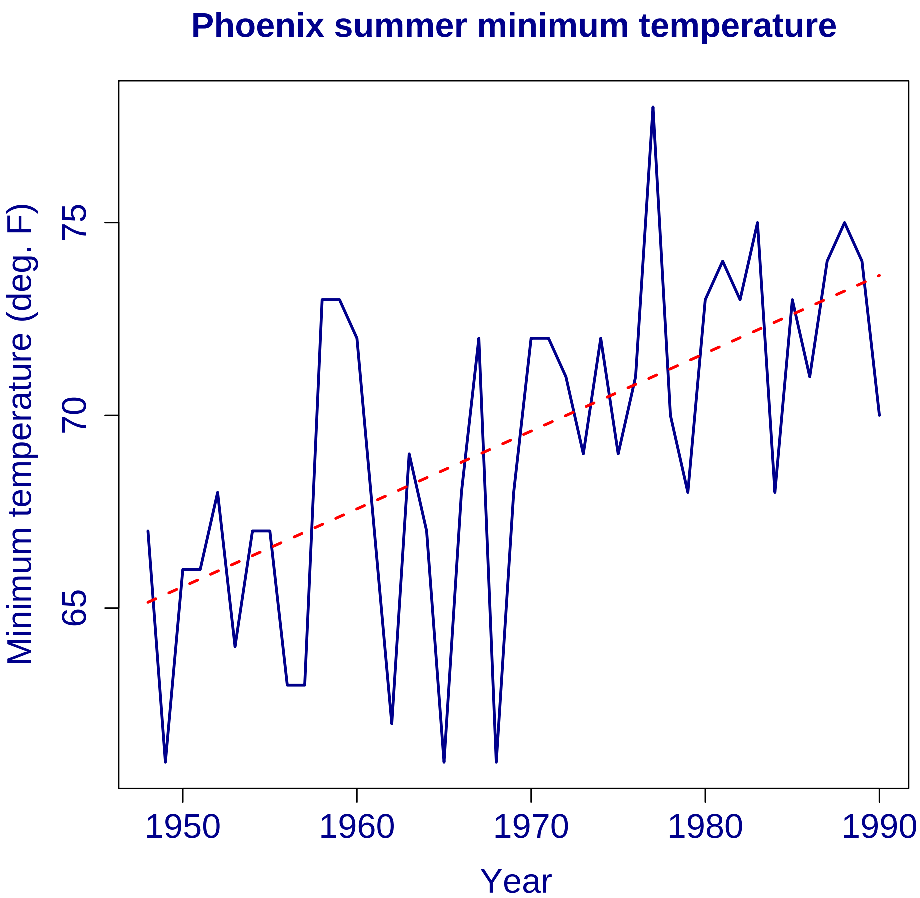
Fort Collins, Colorado precipitation

Probability of annual maximum precipitation at least as large as that during the 28 July 1997 flood (i.e., $\Pr\{\max(X) \geq 1.54 \text{ in.}\}$).

```
# Using the 'pgev' function from the "evd" package.
pgev( 1.54, loc=fit$mle[1],
          scale=fit$mle[2],
          shape=fit$mle[3],
          lower.tail=FALSE)

pgev( 4.6, loc=fit$mle[1],
          scale=fit$mle[2],
          shape=fit$mle[3],
          lower.tail=FALSE)
```

Long-term trend

Phoenix minimum temperature



Source: U.S. National Weather Service Forecast office at the Phoenix Sky Harbor Airport. For more info., see Balling et al. (1990), J. Climate, **3**, 1491–1494.

Long-term trend

Phoenix minimum temperature

Recall: $\min\{X_1, \ldots, X_n\} = -\max\{-X_1, \ldots, -X_n\}$.

Assume summer minimum temperature in year $t = 1, 2, \ldots$ has GEV distribution with:

$$\mu(t) = \mu_0 + \mu_1 \cdot t$$

$$\log \sigma(t) = \sigma_0 + \sigma_1 \cdot t$$

$$\xi(t) = \xi$$

Long-term trend

Phoenix minimum temperature

```
data( HEAT)
plot( HEAT$Tmin, type="l")
fit0 <- gev.fit( -HEAT$Tmin)
fit1 <- gev.fit( -HEAT$Tmin,
         ydat=matrix( 1:dim( HEAT)[1], ncol=1), mul=1)
fit2 <- gev.fit( -HEAT$Tmin,
         ydat=matrix( 1:dim( HEAT)[1], ncol=1), mul=1,
         sigl=1, siglink=exp)
deviancestat( fit0$nllh, fit1$nllh, v=1)
deviancestat( fit0$nllh, fit2$nllh, v=2)
```

Long-term trend

Phoenix minimum temperature

Note: To convert back to $\min\{X_1, \ldots, X_n\}$,
change sign of location parameters. But note that model is
$\Pr\{-X \leq x\} = \Pr\{X \geq -x\} = 1 - F(-x)$.

$$\hat{\mu}(t) \approx 66.170 + 0.196t$$

$$\log \hat{\sigma}(t) \approx 1.338 - 0.009t$$

$$\hat{\xi} \approx -0.21$$

Likelihood ratio test

for $\mu_1 = 0$ (p-value $< 10^{-5}$),

for $\sigma_1 = 0$ (p-value ≈ 0.366).

Long-term trend

Phoenix minimum temperature

Model Checking. Found the best model from a range of models, but is it a good representation of the data? Problem, what are the quantiles when the distribution is changing with a covariate? Transform data to a common distribution, and check the qq-plot.

1. GEV to exponential

$$\varepsilon_t = \left\{ 1 + \frac{\hat{\xi}(t)}{\hat{\sigma}(t)} [X_t - \hat{\mu}(t)] \right\}^{-1/\hat{\xi}(t)}$$

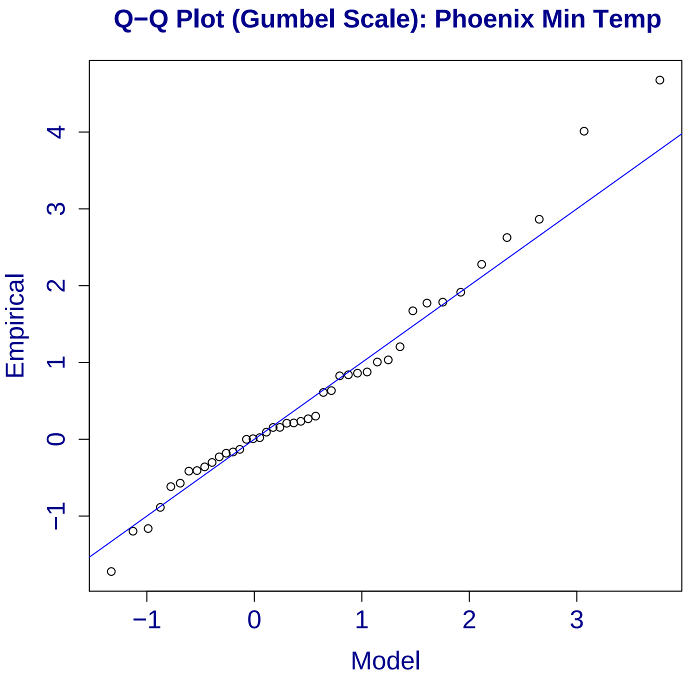
2. GEV to Gumbel (used by `ismev`/`extRemes`)

$$\varepsilon_t = \frac{1}{\hat{\xi}(t)} \log \left\{ 1 + \hat{\xi}(t) \left(\frac{X_t - \hat{\mu}(t)}{\hat{\sigma}(t)} \right) \right\}$$

Long-term trend

Phoenix minimum temperature

Model Checking. Found the best model from a range of models, but is it a good representation of the data? Problem, what are the quantiles when the distribution is changing with a covariate? Transform data to a common distribution, and check the qq-plot.




```
gev.diag( fit2)
```

Long-term trend

Phoenix minimum temperature

See help file for `gev.effective.rl` to see how to compute *effective* return levels.

Long-term trend

Physically based covariates

Winter maximum daily temperature at Port Jervis, New York

Let $X_1, \ldots, X_n$ be the winter maximum temperatures, and $Z_1, \ldots, Z_n$ the associated Arctic Oscillation (AO) winter index. Given $Z = z$, assume conditional distribution of winter maximum temperature is GEV with parameters

$$\mu(z) = \mu_0 + \mu_1 \cdot z$$

$$\log \sigma(z) = \sigma_0 + \sigma_1 \cdot z$$

$$\xi(z) = \xi$$

Data source: National Oceanic and Atmospheric Administration/National Climate Data Center (NOAA/NCDC). For more, see Thompson and Wallace (1998), Geophys. Res. Lett., 25, 1297–1300.

Long-term trend

Physically based covariates

Winter maximum daily temperature at Port Jervis, New York

```
data( PORTw)
names( PORTw)
dim( PORTw)
?PORTw # Get more information about these data.
plot( PORTw$Year, PORTw$TMX1, type="l",
         xlab="Year", ylab="Winter Max Temp (deg C)")
fit0 <- gev.fit( PORTw$TMX1)
fit1 <- gev.fit( PORTw$TMX1, ydat=PORTw, mul=9)
fit2 <-gev.fit( PORTw$TMX1, ydat=PORTw, sigl=9, siglink=exp)
fit12 <- gev.fit( PORTw$TMX1,
         ydat=PORTw, mul=9, sigl=9, siglink=exp)
```

Long-term trend

Physically based covariates

Winter maximum daily temperature at Port Jervis, New York

```
deviancestat( fit0$nllh, fit1$nllh, v=1)
deviancestat( fit0$nllh, fit2$nllh, v=1)
deviancestat( fit0$nllh, fit12$nllh, v=2)
deviancestat( fit1$nllh, fit12$nllh, v=1)
deviancestat( fit2$nllh, fit12$nllh, v=1)
```

Note: cannot use likelihood-ratio test (`deviancestat`) to directly test `fit1` vs. `fit2`. Why?

Long-term trend

Physically based covariates

$$\hat{\mu}(z) \approx 15.26 + 1.175 \cdot z$$

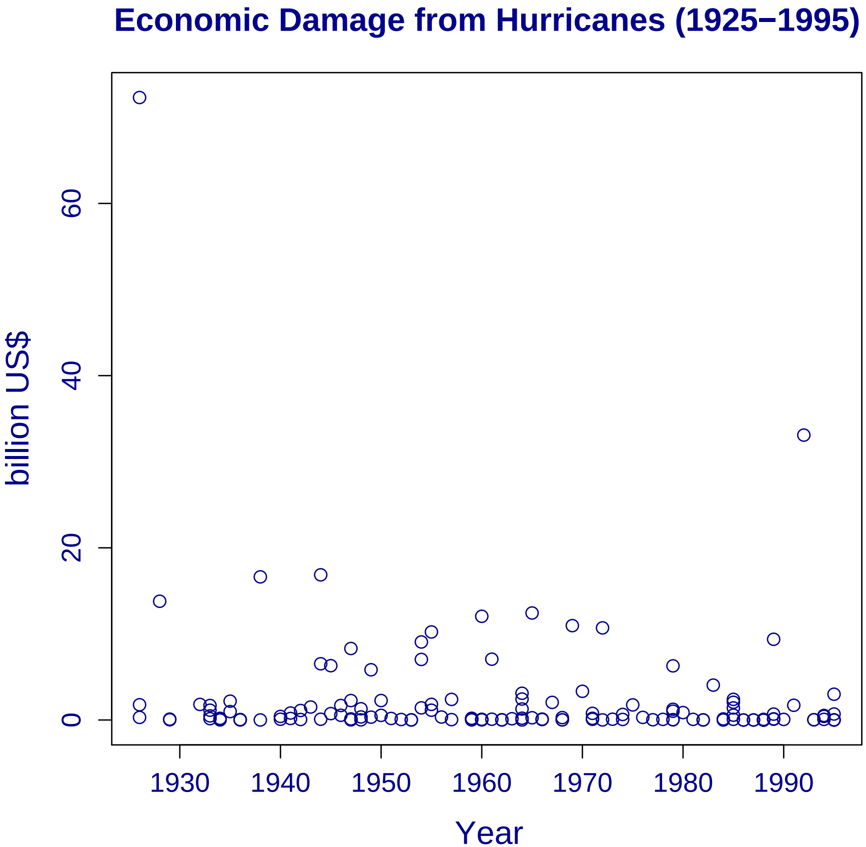
$$\log \hat{\sigma}(z) = 0.984 - 0.044 \cdot z$$

$$\xi(z) = -0.186$$

Likelihood-ratio test for $\mu_1 = 0$ (p-value < 0.001)
Likelihood-ratio test for $\sigma_1 = 0$ (p-value ≈ 0)
Likelihood-ratio test for $\mu_1 = 0$ and $\sigma_1 = 0$ (p-value ≈ 0.002)
Likelihood-ratio test for $\sigma_1 = 0$, given `fit1` (p-value ≈ 0.635)
Likelihood-ratio test for $\mu_1 = 0$ given `fit2` (p-value ≈ 0)

Peaks Over Thresholds (POT) Approach

Hurricane damage



Economic damage caused by hurricanes from 1926 to 1995.

Trends in societal vulnerability removed.

Excess over threshold of $u = 6$ billion US$.

For more, see Pielke and Landsea (1998), Wea. Forecasting, 13, 621–631. Data source:
http://sciencepolicy.colorado.edu/pielke/hp_roger/hurr_norm/data.html

Peaks Over Thresholds (POT) Approach

Hurricane damage

```
data( damage)
?damage
plot( damage[,1], damage[,3],
          xlab="", ylab="Economic Damage", type="l", lwd=2)
```

```
gpd.fitrange( damage$Dam, umin=1, umax=15, nint=15)
```

Choose a threshold low enough (lower variance), but high enough that the assumptions for the GPD are valid (lower bias). Looks like 6 billion USD would work; maybe something lower could also work.

Peaks Over Thresholds (POT) Approach

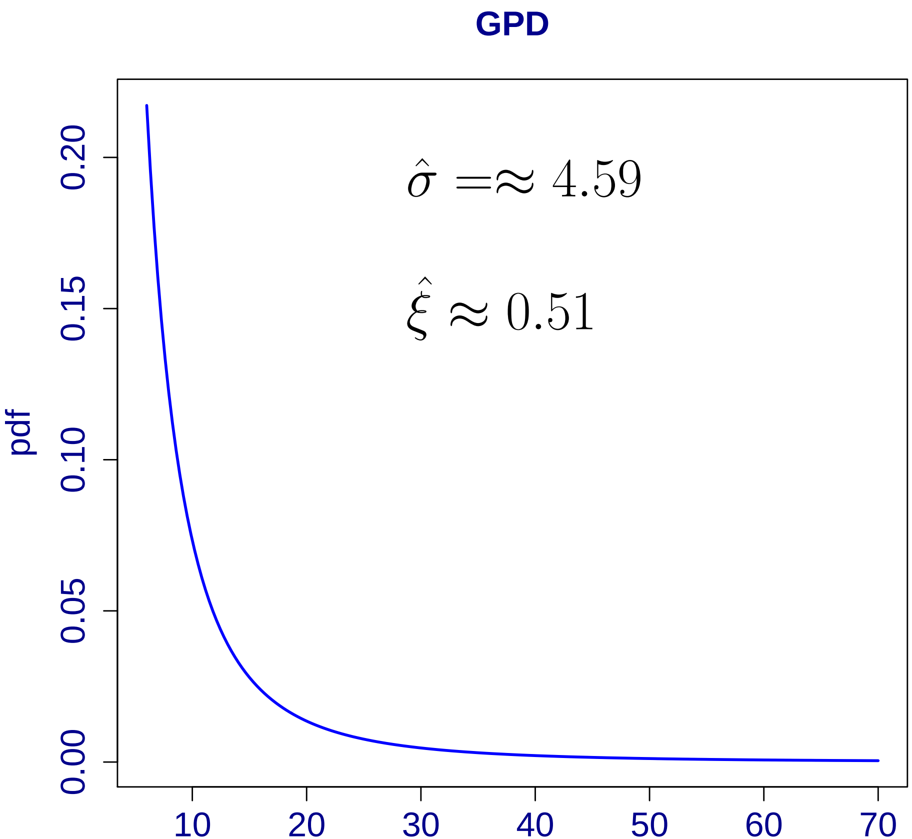
Hurricane damage

Hurricane Dennis (2005)
Caused at least 89 deaths and
2.23 billion USD in damage.

Impactfull despite being under the 6 billion USD threshold!

Peaks Over Thresholds (POT) Approach

Hurricane damage



Likelihood ratio test for

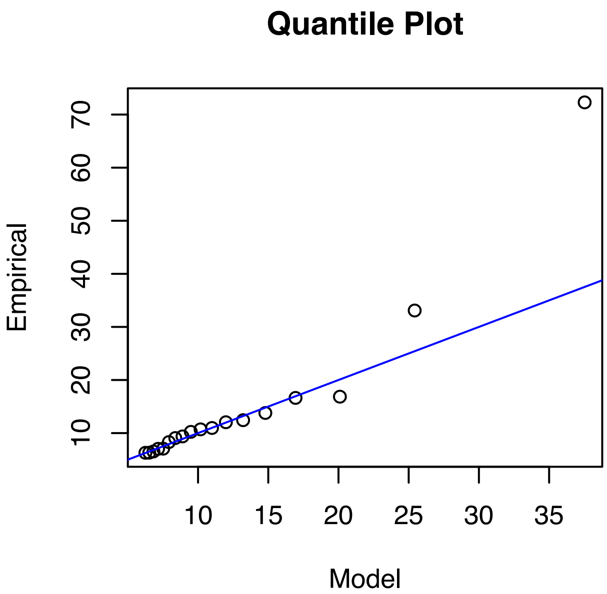
$\xi = 0$ (p-value≈ 0.018)

95% CI for shape parameter using profile likelihood. $0.05 < \xi < 1.56$

Heavy tail!

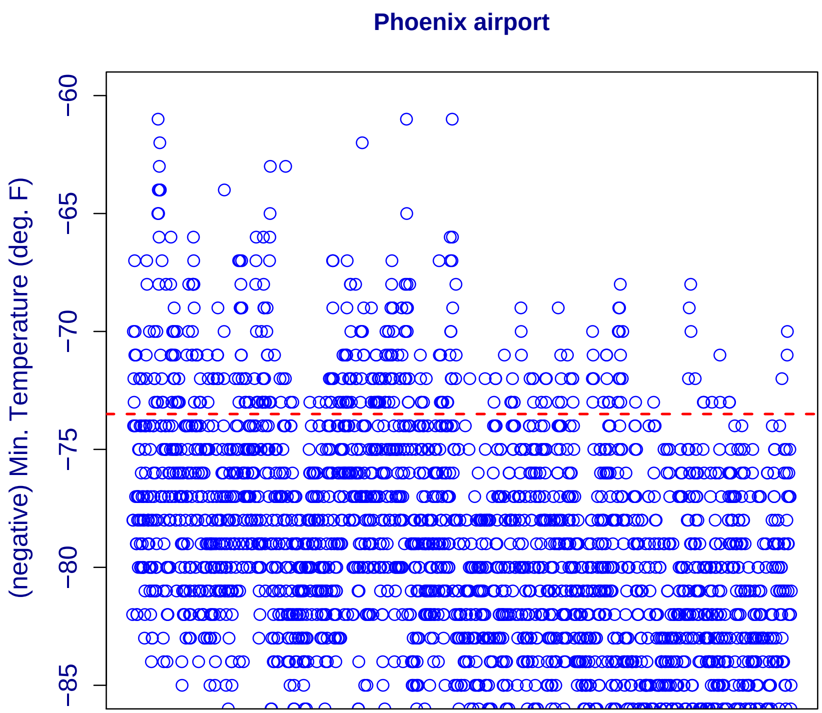
Peaks Over Thresholds (POT) Approach

Hurricane damage



Peaks Over Thresholds (POT) Approach

Dependence above threshold



Phoenix (airport) minimum temperature (^{o}F).

July and August 1948–1990.

Urban heat island (warming trend as cities grow).

Model lower tail as upper tail after negation.

Peaks Over Thresholds (POT) Approach

Dependence above threshold

```
# Fit without de-clustering.
plot( -Tphap$MinT, type="l")
abline(h=-73, col="darkred")
phx.fit0 <- gpd.fit( -Tphap$MinT, -73)
gpd.diag( phx.fit0)

# With runs de-clustering (r=1).
phx.dc1 <- dclust( -Tphap$MinT, u=-73, r=1,
                   cluster.by=Tphap$Year)
phxdc1.fit0 <- gpd.fit( phx.dc1$xdat.dc, -73)
plot( phx.dc1$xdat.dc, type="l")
abline(h=-73, col="darkred")
gpd.diag( phxdc1.fit0)
```

Peaks Over Thresholds (POT) Approach

Dependence above threshold

```
eiAnalyze( -Tphap$MinT, -73)
phx.dc11 <- dclust( -Tphap$MinT, u=-73, r=11,
                   cluster.by=Tphap$Year)
plot( phx.dc11$xdat.dc, type="l")
phxdc11.fit0 <- gpd.fit( phx.dc11$xdat.dc, -73)
abline(h=-73, col="darkred")
gpd.diag( phxdc11.fit0)
```

Peaks Over Thresholds (POT) Approach

Long-term warming trend

Varying threshold

```
lm( -MinT Year, data=Tphap)
plot( -Tphap$MinT, type="l")
lines( 1:dim( Tphap)[1], -60-0.1764*Tphap$Year,
          col="darkorange")
phx.fit1 <- gpd.fit( -Tphap$MinT, -60-0.1764*Tphap$Year)
gpd.diag( phx.fit1)
```

Peaks Over Thresholds (POT) Approach

Long-term warming trend

parameter covariate: $\log(\sigma) = \sigma_0 + \sigma_1 t$, $t = 0, \ldots, 0, 1, \ldots, 1, 2, \ldots$

```
yr <- matrix( Tphap$Year - 48, ncol=1)
phx.fit2 <- gpd.fit( -Tphap$MinT, -73,
          ydat=yr, sigl=1, siglink=exp)
gpd.diag( phx.fit2)

# Both ...
phxfit.both <- gpd.fit( -Tphap$MinT, -60-0.1764*Tphap$Year,
          ydat=yr, sigl=1, siglink=exp)
gpd.diag( phxfit.both)
```

Peaks Over Thresholds (POT) Approach

Point Process: *frequency and intensity of threshold excesses*

Fort Collins, Colorado daily precipitation

Analyze daily data instead of just annual maxima
(ignoring annual cycle for now).

Orthogonal Approach

$$\hat{\lambda} = 365.25 \cdot \frac{\text{No. } X_i > 0.395}{\text{No. } X_i} \approx 10.6 \text{ per year}$$

$$\hat{\sigma}^* \approx 0.323, \, \hat{\xi} \approx 0.212$$

Using `gpd.fit` to get $\hat{\sigma}^*$ and $\hat{\xi}$.

Peaks Over Thresholds (POT) Approach

Point Process: *frequency and intensity of threshold excesses*

Fort Collins, Colorado daily precipitation

```
data( FtCoPrec)
class( FtCoPrec)
colnames( FtCoPrec)
fit0 <- gpd.fit( FtCoPrec[,"Prec"], 0.395)

# Now fit Poisson Process (PP) to these data.
fit1 <- pp.fit( FtCoPrec[,"Prec"], 0.395)
pp.diag( fit1)
```

Peaks Over Thresholds (POT) Approach

Point Process: *frequency and intensity of threshold excesses*

Fort Collins, Colorado daily precipitation

Analyze daily data instead of just annual maxima
(ignoring annual cycle for now).

Point Process

$$\hat{\mu} \approx 1.384$$

$$\hat{\sigma} = 0.533$$

$$\hat{\xi} \approx 0.213$$

$$\hat{\lambda} = \left[1 + \frac{\hat{\xi}}{\hat{\sigma}}(u - \hat{\mu})\right]^{-1/\hat{\xi}} \approx 10.6 \text{ per year}$$

Risk Communication Under Stationarity

Unchanging climate

Compare previous GPD fits (with and without de-clustering).

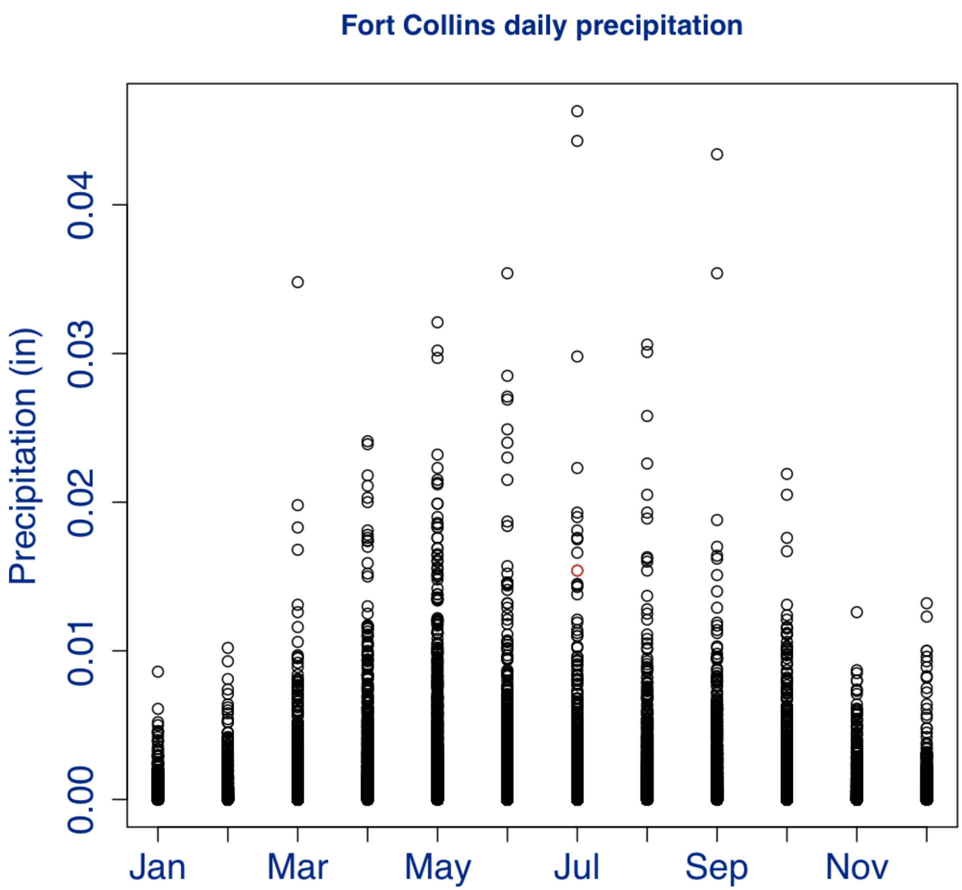
```
# Without de-clustering.
return.level( phx.fit0)

# With de-clustering (r=1).
return.level( phxdc.fit0)

# With de-clustering (r=11).
return.level( phxdc11.fit0)
```

Note: little difference in estimates, but relatively large difference in confidence intervals. Less difference between r=1 and r=11 runs-declustered fits.

Non-Stationarity

Cyclic variation

Fort Collins, Colorado precipitation



Annual Cycle

Fort Collins, Colorado precipitation Orthogonal approach. First fit annual cycle to Poisson rate parameter ($T = 365.25$):

$$\log \lambda(t) = \lambda_0 + \lambda_1 \sin\left(\frac{2\pi t}{T}\right) + \lambda_2 \cos\left(\frac{2\pi t}{T}\right)$$

```
prec <- FtCoPrec[,"Prec"]
ind <- prec > 0.395
trend1 <- sin(2*pi*(1:length(prec))/365.25)
trend2 <- cos(2*pi*(1:length(prec))/365.25)
ycov <- cbind( trend1, trend2)
lamfit <- glm( ind~ trend1+trend2, family=poisson())
summary( lamfit)
```

Annual Cycle

Fort Collins, Colorado precipitation

$$\log \hat{\lambda}(t) \approx -3.72 + 0.22 \sin\left(\frac{2\pi t}{T}\right) - 0.85 \cos\left(\frac{2\pi t}{T}\right)$$

Likelihood ratio test for $\lambda_1 = \lambda_2 = 0$ (p-value ≈ 0).

Annual Cycle

Fort Collins, Colorado precipitation Orthogonal approach. Next fit GPD with annual cycle in scale parameter.

$$\log \sigma^*(t) = \sigma_0^* + \sigma_1^* \sin\left(\frac{2\pi t}{T}\right) + \sigma_2^* \cos\left(\frac{2\pi t}{T}\right)$$

```
fitOrth <- gpd.fit( prec, 0.395, ydat=ycov, sigl=c(1,2), sigli
fitOrth0 <- gpd.fit( prec, 0.395)
pchisq( -2*(fitOrth$nllh - fitOrth0$nllh), 2, lower.tail=FALSE
```

Annual Cycle

Fort Collins, Colorado precipitation

$$\log \hat{\sigma}^*(t) \approx -1.24 + 0.09 \sin\left(\frac{2\pi t}{T}\right) - 0.30 \cos\left(\frac{2\pi t}{T}\right), \hat{\xi} \approx 0.18$$

Likelihood ratio test for $\sigma_1^* = \sigma *_2 = 0$ (p-value $< 10^{-5}$)

Annual Cycle

Fort Collins, Colorado precipitation

Annual cycle in location and scale parameters of the GEV re-parameterization approach PP model with $t = 1, 2, ...,$ and $T = 365.25$.

$$\mu(t) = \mu_0 + \mu_1 \sin\left(\frac{2\pi t}{T}\right) + \mu_2 \cos\left(\frac{2\pi t}{T}\right)$$

$$\log\sigma(t) = \sigma_0 + \sigma_1 \sin\left(\frac{2\pi t}{T}\right) + \sigma_2 \cos\left(\frac{2\pi t}{T}\right)$$

$$\xi(t) = \xi$$

Annual Cycle

Fort Collins, Colorado precipitation

```
fit0 <- pp.fit( prec, 0.395)
fit1 <- pp.fit( xdat=prec, threshold=0.395, npy=365.25,
          ydat=ycov, mul=c(1,2))
fit2 <- pp.fit( xdat=prec, threshold=0.395, npy=365.25,
          ydat=ycov, sigl=c(1,2), siglink=exp)
fit <- pp.fit( xdat=prec, threshold=0.395, npy=365.25,
          ydat=ycov, mul=c(1,2), sigl=c(1,2), siglink=exp)


# Likelihood ratio test of mu1=mu2=0.
pchisq( -2*(fit$nllh-fit2$nllh), 2, lower.tail=FALSE)

# sigma1=sigma2=0.
pchisq( -2*(fit$nllh - fit1$nllh), 2, lower.tail=FALSE)
```

Annual Cycle

Fort Collins, Colorado precipitation

$$\hat{\mu}(t) \approx 1.281 - 0.085 \sin\left(\frac{2\pi t}{T}\right) - 0.806 \cos\left(\frac{2\pi t}{T}\right)$$

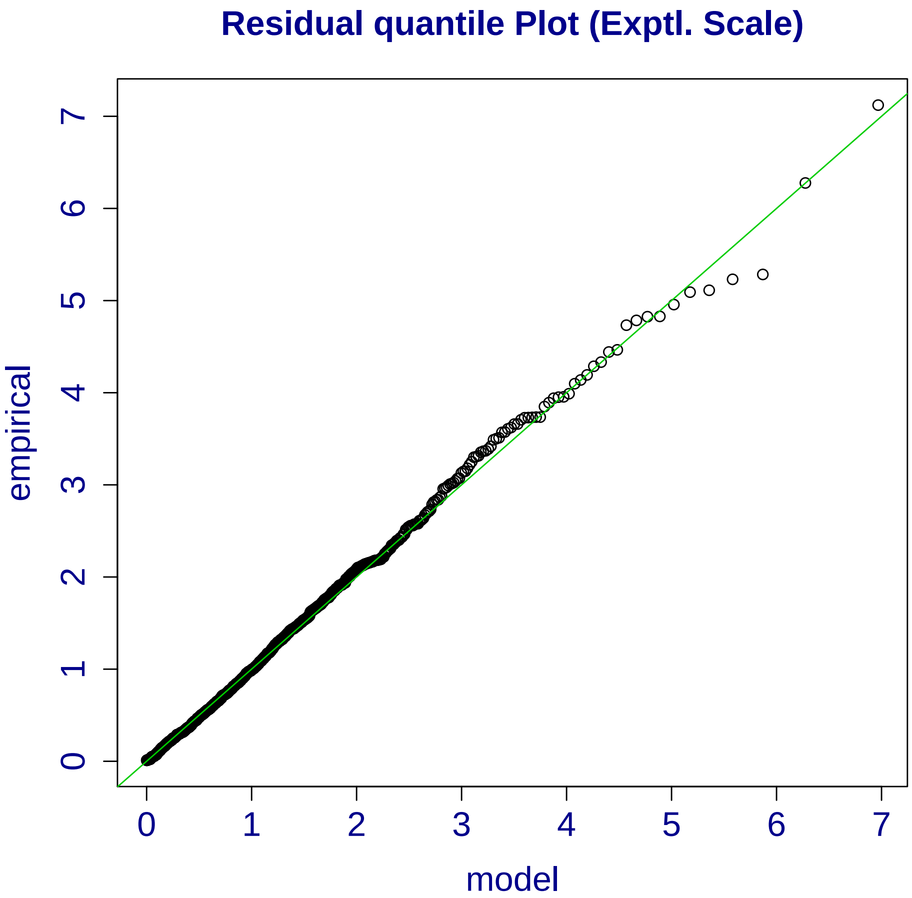
$$\log \hat{\sigma}(t) \approx -0.847 - 0.123 \sin\left(\frac{2\pi t}{T}\right) - 0.602 \cos\left(\frac{2\pi t}{T}\right)$$

$$\hat{\xi} \approx 0.182$$

Likelihood ratio test for $\mu_1 = \mu_2 = 0$ (p-value ≈ 0).
Likelihood ratio test for $\sigma_1 = \sigma_2 = 0$ (p-value ≈ 0).

Annual Cycle

Fort Collins, Colorado precipitation



`pp.diag( fit)`

References and Acknowledgements

Coles S, 2001. An introduction to statistical modeling of extreme values. Springer, London. 208 pp.

Katz RW, MB Parlange, and P Naveau, 2002. Statistics of extremes in hydrology. *Adv. Water Resources*, **25**:1287–1304.

Stephenson A and E Gilleland, 2006. Software for the analysis of extreme events: The current state and future directions. *Extremes*, **8**:87–109.

Weather and Climate Impacts Assessment Science (WCIAS) Program
http://www.assessment.ucar.edu