Flooding to Financial Disaster: An Introduction to Extreme Value Theory

Grant Weller
Department of Statistics
Colorado State University

Joint work with:
Dan Cooley, CSU
Steve Sain, NCAR

Extreme Events

Q: What does it mean to be an 'extreme event'?

Extreme Events

Q: What does it mean to be an 'extreme event'?

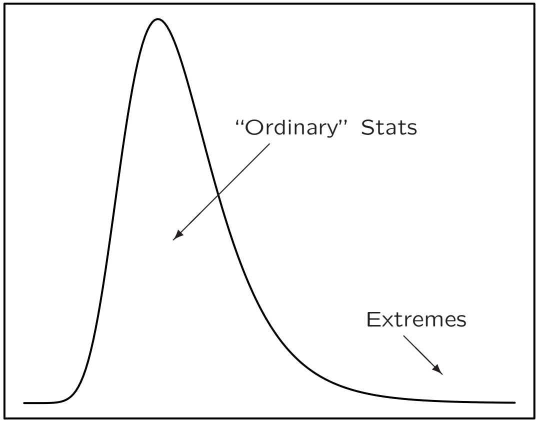
A: It depends whom you ask:

- Financial analysts: 'shocks' to a time series (market crash)
- Insurance companies: costly (though perhaps not rare) phenomena (hurricanes)
- Meteorologists: rare weather phenomena (a brown Christmas in Moorhead?)
- Applied mathematicians: extremes of Gaussian/other processes, large deviations
- Statisticians (like me!): Extreme Value Theory (this talk: data perspective)
- Many possible answers!

Extreme Value Theory

"Ordinary" Statistics: try to describe a distribution of data; possibly ignore very large or small values (outliers)

Extreme Value Theory: try to characterize the tail of the distribution; uses only the extreme observations



Why study extremes?

While infrequent, extremes often have large human impact.

Goal of an extreme value analysis: to quantify the magnitude of a ~~worst-case~~ really-bad-case scenario.

Application areas:

- hydrology (stream/river flows)
- climate variables: precipitation, wind, heat waves, ...
- finance
- insurance/reinsurance
- engineering (structural design, failure)
- not much done (yet) in medicine, biology, ecology

Why study extremes?

In fact, you don't have to go very far to find an example...

Larry Hansen; Fargo Flood homepage

Why study extremes?

2009 Red River of the North flood

- Highest ever recorded water level at Fargo (40.8 ft on 3/28/09) - flood stage is 18 ft
- Tens of millions of dollars in damages (1997: $3.5 billion)
- No classes at Concordia for two weeks!
- Related question: how unlikely was this event?
- We'll come back to this later...

...first, let's go back in time

Concordia College 2004-2008

Cobber football

- 2007 - school records for total points, touchdowns, yards in a season
- 2005 & 2007 teams - #1 and #2 in season rush yds and tds

2007 vs. Carleton

Mathematics & Economics major

- Advised by Dr. Zeng
- Other influences: Dr. Doug Anderson, Dr. Jim Forde, Dr. Haimeng Zhang, Dr. Dan Biebighauser

I had the best undergraduate advisor!

Then on to graduate school...

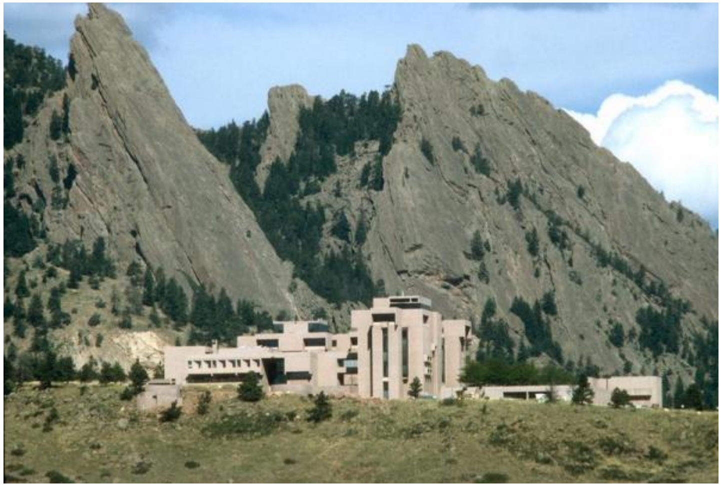
Department of Statistics, CSU Fort Collins, 2008-present

- M.S. in Statistics, 2010
- Ph.D. expected 2013
- Advisor: Dr. Dan Cooley
- learned to ski!

Research opportunities

NCAR, Boulder, CO

- Graduate Student Visitor (Summer 2011)
- Climate change research
- Extreme precipitation from climate models - more later



Mesa Lab

SAMSI, RTP, NC

- Visitor for Fall 2011
- Participant in 2011-2012 Uncertainty Quantification program
- Lived in Chapel Hill, NC

2012 National Champions?

Outline

This talk will mainly focus on applications and examples, with very basic introduction to theoretical results.

- Univariate extremes - Red River flooding example
- Brief introduction to bivariate extremes - describing dependence
- Pineapple Express project
- Suggestions for current Concordia math majors

So what about the 2009 Red River flood?

The Fargo Red River station has daily peak flow measurements (cfs) going back to 1901.

Q: How 'rare' was the 2009 event, in terms of daily flow?

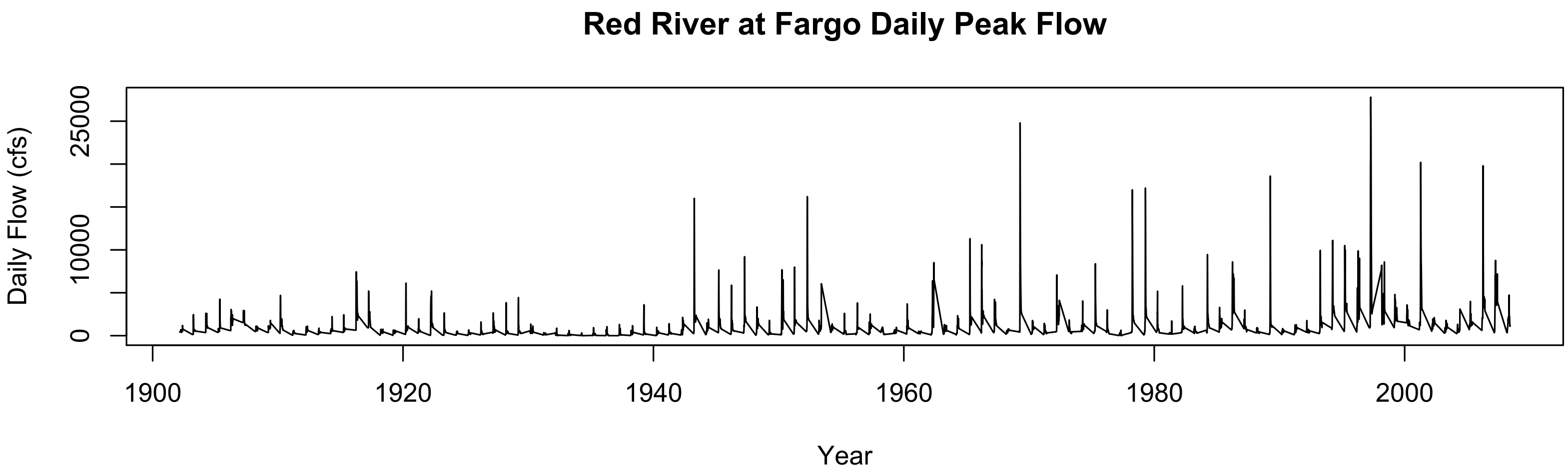
Let's use springtime (MAM) data from 1901-2008 to try to answer this.

Two approaches:

1. Analyze all the data; fit a gamma distribution
2. Use only the annual max data; fit a GEV

Warning: *Both* analyses are likely incorrect, for different reasons. Bonus points if you can tell me why later.

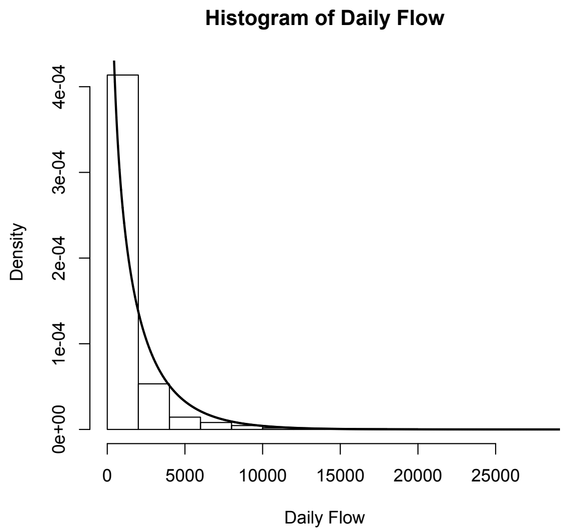
Modeling all springtime daily data



Let X_t be the daily peak flow on day t. Assume that $X_t \sim$ Gamma(α, β) (ignore zero-flow days).

Estimate the parameters via maximum likelihood: $\widehat{\alpha} = 0.62$, $\widehat{\beta} = 2765.08$.

Modeling all springtime daily data



Fit looks ok.

Maximum flow in 2009 was 29100 cfs on March 28^{th}. What is the probability that the maximum daily flow in 2009 would be *at least* this much, assuming this is the right model?

Estimate using all daily data

$$\mathbb{P}(X_t > 29100) = 1 - F_X(29100) = 7.400995 \times 10^{-6}$$

$$\begin{aligned}
\mathbb{P}(\text{ann. max} > 29100) &= 1 - \mathbb{P}(\text{entire year's obs} < 29100) \\
&= 1 - (1 - \mathbb{P}(\text{one obs} > 29100))^{92} \\
&= 1 - (1 - 7.400995 \times 10^{-6})^{92} \\
&= 6.807 \times 10^{-4}
\end{aligned}$$

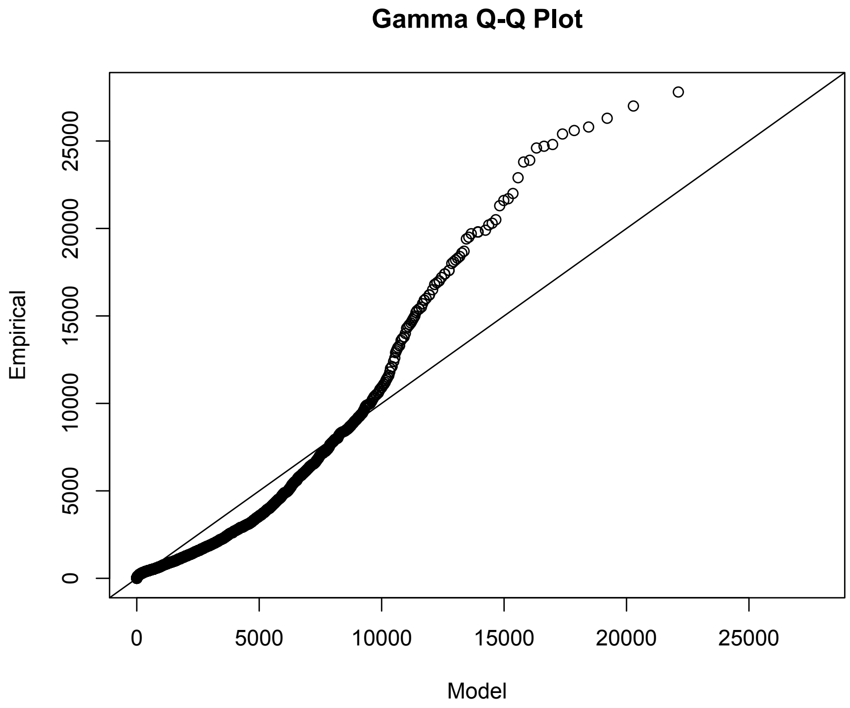
Associated "return period": $(6.807 \times 10^{-4})^{-1} = 1469$ years.

But is this the right way to analyze the data?

All daily data model

Two main problems with this model:

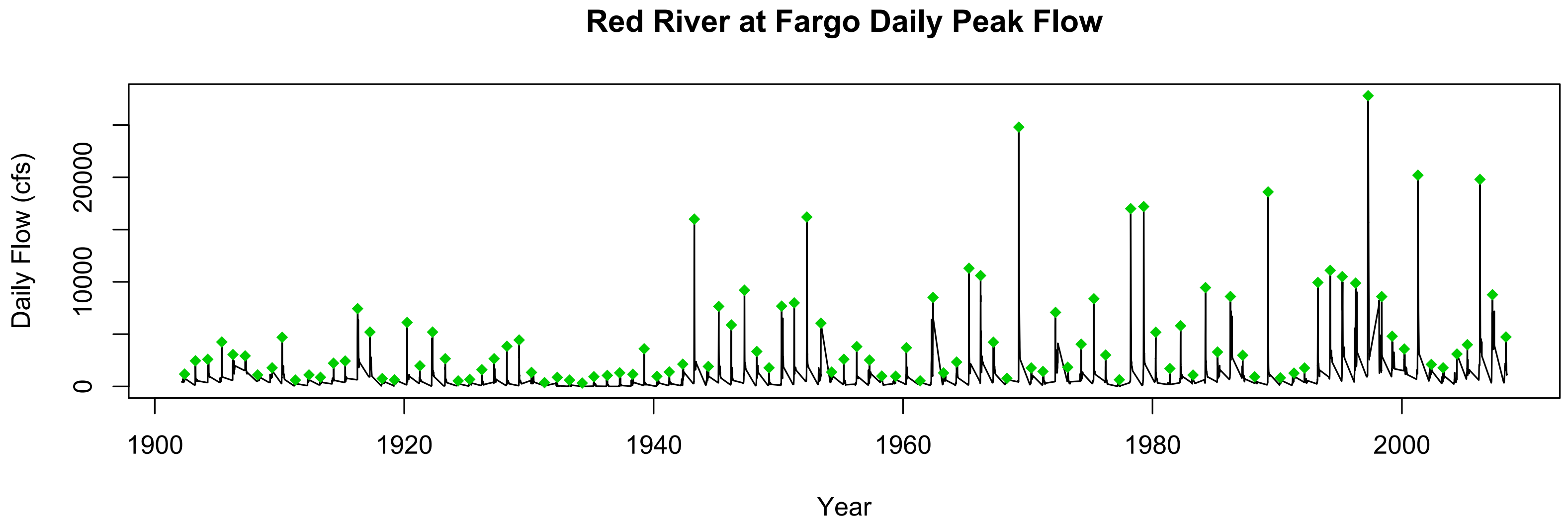
1. Assumes daily river flow rates are iid
2. Underestimates the tail of the distribution



99.4% of data and 99.8% of model's mass are < 15000.

Modeling Annual Maxima

Alternative approach: retain only the largest value from each year, and fit a generalized extreme value distribution



Why?

1. Intuitively, river usually only floods once each year
2. Mathematically, justified by theory

Fitting a GEV

Let $M_n = \max(X_t), t = 1, ..., n$. Assume $M_n \sim \text{GEV}(\mu, \sigma, \xi)$.

$$F_{M_n}(x) = \mathbb{P}(M_n \leq x) = \exp\left\{-\left[1 + \xi\left(\frac{x-\mu}{\sigma}\right)\right]^{-1/\xi}\right\}$$

PWM estimates: $\widehat{\mu} = 2367.54$, $\widehat{\sigma} = 2496.05$, $\widehat{\xi} = 0.32$.

$\mathbb{P}(\text{ann. max} > 29100) = 1 - F_{M_n}(29100) = 0.0095$

Associated return period: $(0.0095)^{-1} = 105$ years

100-year return level: 28561

95% confidence interval (delta method): $(8909, 48213)$

500-year return level: 51536

95% confidence interval (delta method): $(1718, 101354)$

A word of caution

Because we use one observation per year, uncertainty associated with return period/level estimates are *very high*.

Q: Why is uncertainty so important?

A word of caution

Because we use one observation per year, uncertainty associated with the previous return period estimate is *very high*.

Q: Why is uncertainty so important?

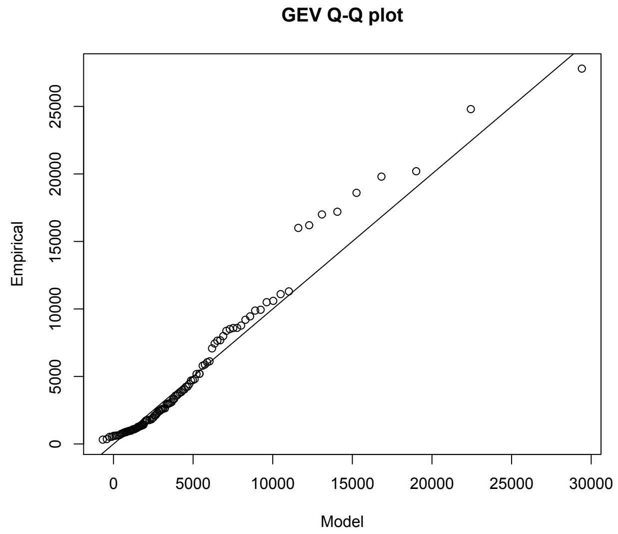
> "There are known knowns; there are things we know we know. We also know there are known unknowns; that is to say we know there are some things we do not know. But there are also unknown unknowns - there are things we do not know we don't know."
>
> Former Secretary of Defense Donald Rumsfeld

Think of uncertainty as a 'known unknown'. Important to acknowledge that this exists and quantify it.

This is the crux of Mathematical Statistics.

GEV model fit

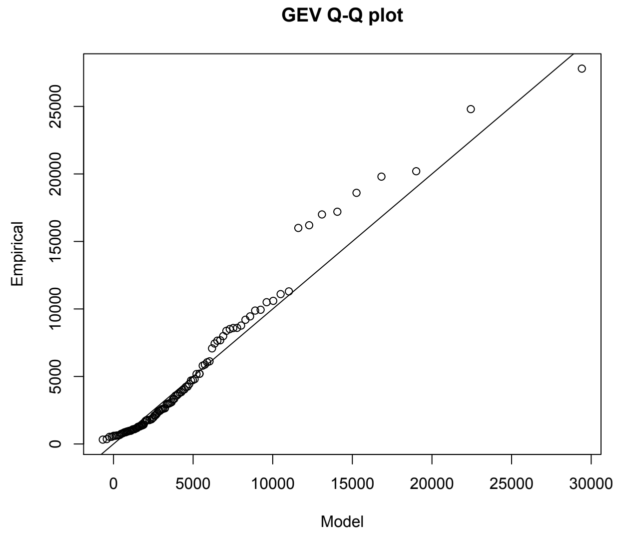
Plot shows annual maxima 1902-2008.



This model is a better fit. But are we missing something?

GEV model fit

Plot shows annual maxima 1902-2008.



This model is a better fit. But are we missing something?

Non-stationarity? Six of the seven largest flooding events occurred in the last 15 years.

Could be handled by a regression-type approach.

Why use only 'extreme' observations?

Two approaches for extracting extreme observations:
1. Block-maximum approach (done above)
2. Threshold-exceedance approach (skipped today)

Heuristic explanation: Phenomena which generate extreme observations are different than those which generate typical observations (Red River floods?).

Mathematical explanation: Assume X_t has cdf $F_X(x)$.

$$\begin{aligned} F_{M_n}(x) = P(M_n \le x) &= P(X_t \le x \text{ for all } t = 1, \ldots, n) \\ &= P^n(X_t \le x) \\ &= F_X^n(x) \end{aligned}$$

If we know F_X exactly, then we know F_{M_n} exactly. But if we have to estimate F_X, any errors in estimating the tail get amplified by a power of n.

Generalized Extreme Value distribution

The GEV distribution captures three types of tail behavior.

$$F(x) = \exp\left\{ - \left[1 + \xi \left(\frac{x - \mu}{\sigma} \right) \right]^{-1/\xi} \right\}.$$

The parameter ξ determines tail behavior and is difficult to estimate in practice.

- $\xi < 0$: Weibull case (bounded tail)
- $\xi = 0$: Gumbel case (light tail), interpreted as limit
- $\xi > 0$: Fréchet case (heavy tail)

Q: Why is the GEV the right distribution to fit to annual maximum data?

Recall from your introductory statistics course: the Central Limit Theorem says that the Normal distribution is the right distribution for sums of iid random variables...

Why the GEV?

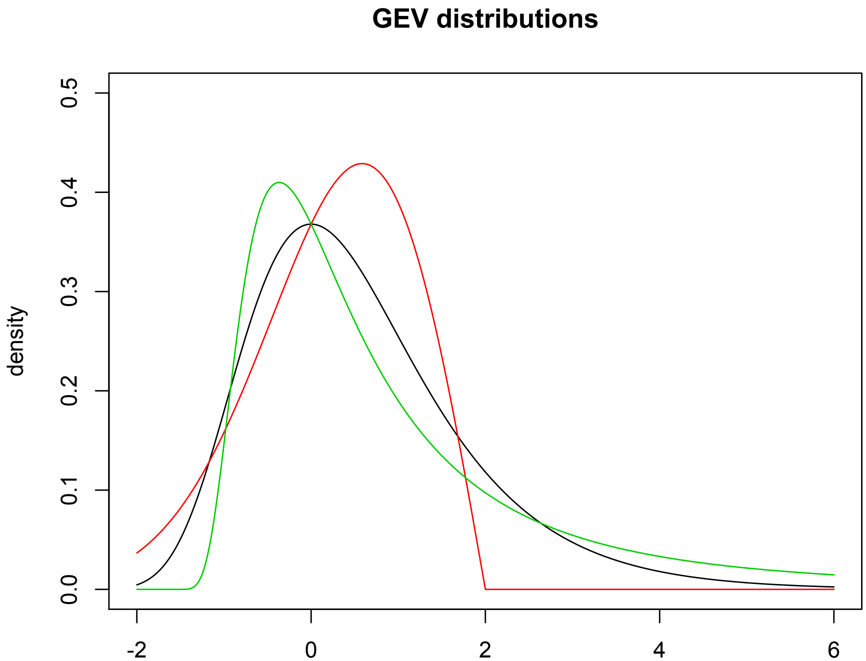
...there is a 'CLT-like' result for block maxima.

Let $M_n = \max_{t=1,\ldots,n} X_t$, where X_t are iid. If there exist normalizing sequences a_n and b_n such that $P\left(\frac{M_n - b_n}{a_n} \leq x\right) \to G(x)$ (nondegenerate) as $n \to \infty$, then

$$G(x) = \exp\left\{-\left[1 + \xi x\right]^{-1/\xi}\right\}.$$

We don't need information about the distribution of X_t to know about the distribution of M_n.

Limiting Distributions



Weibull ($\xi = -0.5$)

Gumbel

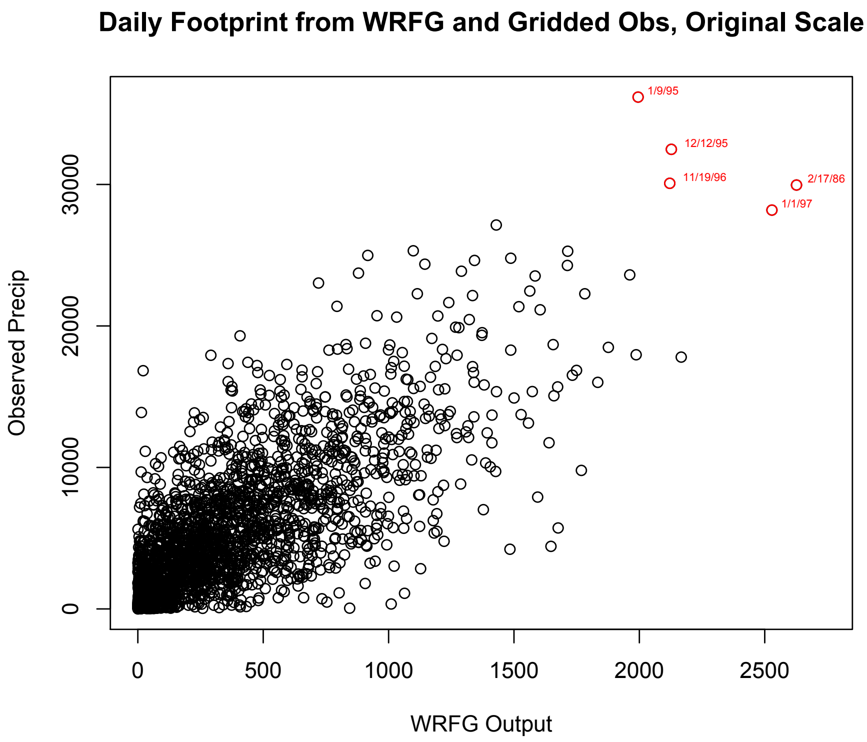
Fréchet ($\xi = 0.5$)

Bivariate Extremes

Assume $\mathbf{X}_t = (X_{t,1}, X_{t,2})$ iid.

Idea: Describe the joint tail of the distribution

Goal: Often extrapolation



Bivariate Extremes: Example

Portfolio consisting of two securities. What is the probability that one goes bust, given that the other does?

The Gaussian copula was a popular model for portfolio risks *at least until 2008:*

Bivariate Extremes: Example

Portfolio consisting of two securities. What is the probability that one goes bust, given that the other does?

The Gaussian copula was a popular model for portfolio risks *at least until 2008:*

$$\Pr[T_A<1, T_B<1] = \phi_2(\phi^{-1}(F_A(1)), \phi^{-1}(F_B(1)), \gamma)$$

Wired magazine: “Recipe for Disaster: The Formula that Killed Wall Street”. February 23, 2009.

Basic problem: this model doesn’t account for multiple securities experiencing an ‘extreme event’ at the same time.

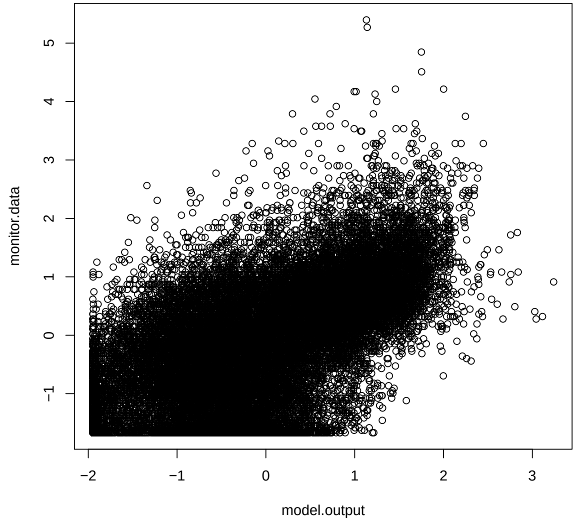
Some forward-thinking statisticians in extreme value theory were warning of this problem years ahead of time.

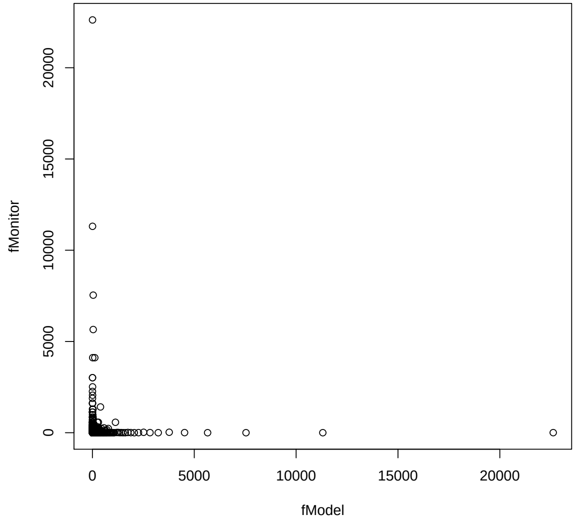
Their warnings went largely unheeded.

Bivariate Extremes

Marginal and dependence effects are typically handled separately.

Approach:
1. Transform marginal to something nice.
2. Describe dependence.

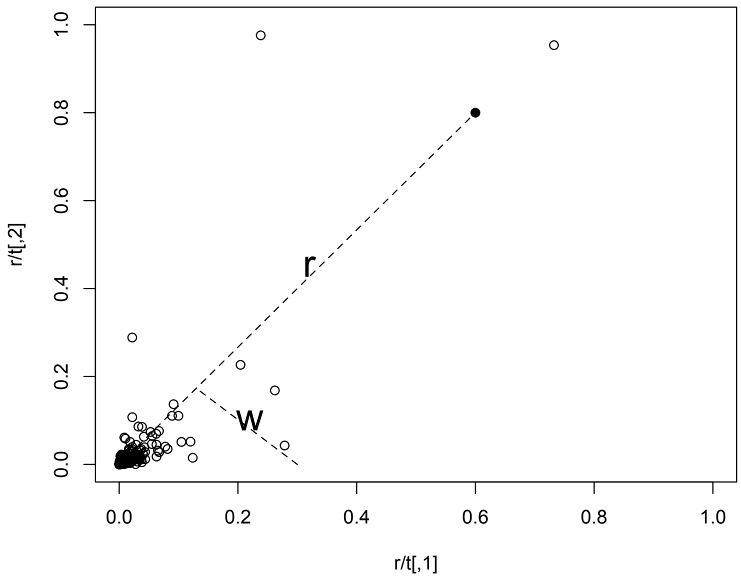




Note: high correlation does not imply tail dependence!

Describing tail dependence

Limit result: extremes occur according to a Poisson process.



Intensity measure factors into 'radial' component (r) and 'angular' component (w).

Dependence is described by a probability measure on w.

Pineapple Express project

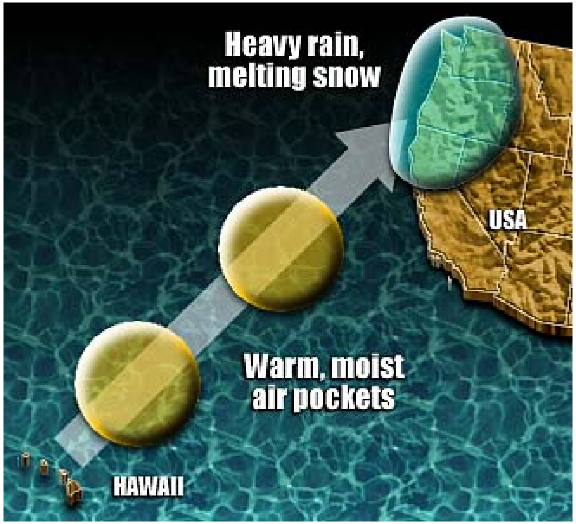
Pineapple Express project

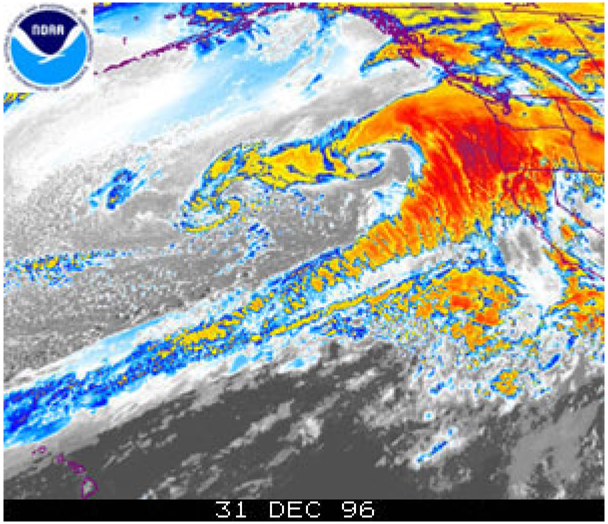
What this section is *not* about

Pineapple Express project

PE storms: caused by atmospheric rivers hitting the west coast in winter

- Often bring heavy rain and warm temperatures
- Great impact on water resources of western US

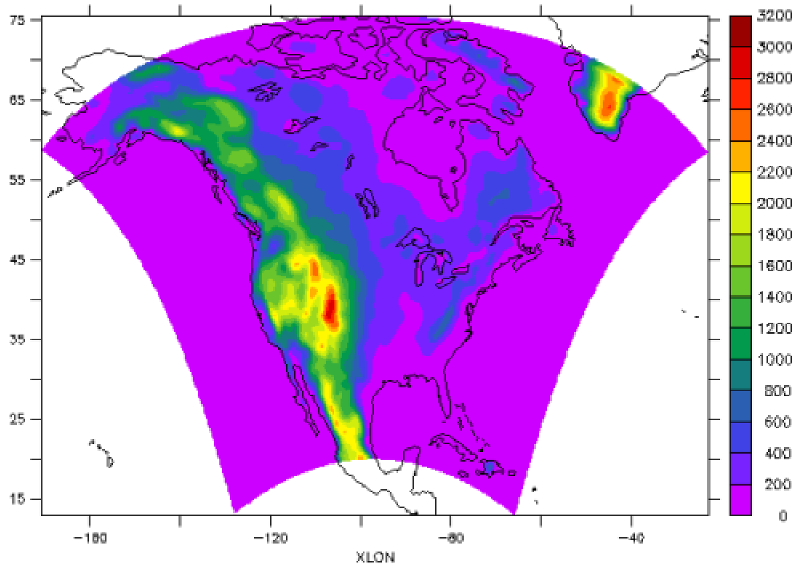




Question of Interest: How well are Regional Climate Models able to represent extreme precipitation caused by this phenomenon?

Regional Climate Models

Use input from a low-resolution global model and known physics of the Earth system to produce simulated weather over long periods of time at finer scales (∼50km).



NARCCAP domain

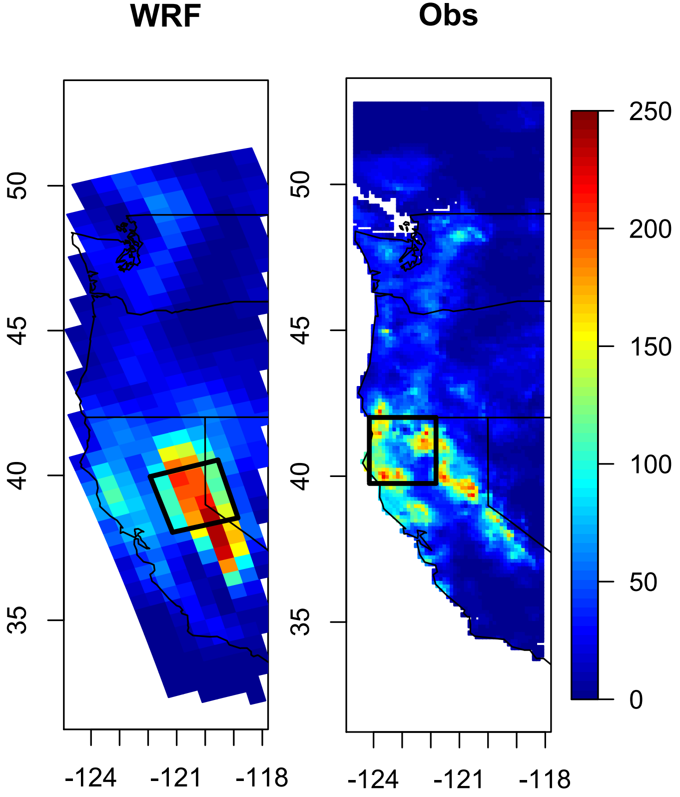
Aim: study regional impacts of climate change scenarios

For evaluation, RCMs forced by reanalysis for 1979-2004.

RCM output vs. observations

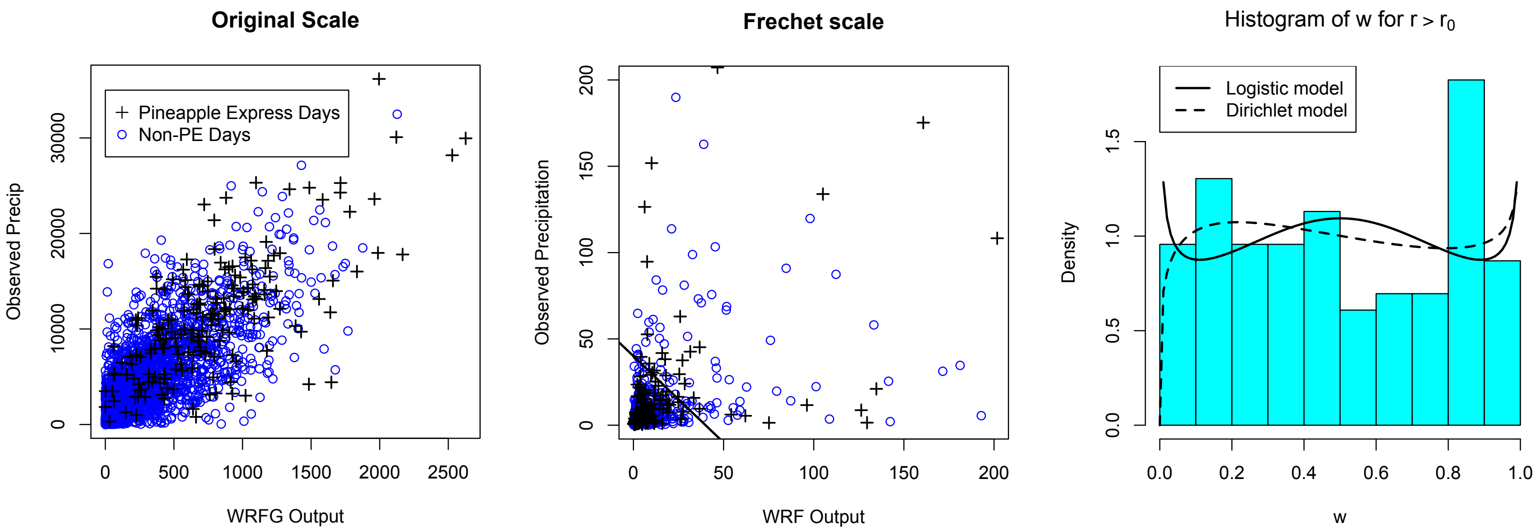
Idea: compare daily precipitation amounts from RCM to observed data

- Identify region and quantity that capture PE events
- NDJF days 1981-1999
- Right: January 1, 1997 (big PE event)
- Method: bivariate extreme value analysis



Bivariate extremes analysis

Transform each marginal to unit Fréchet and examine dependence

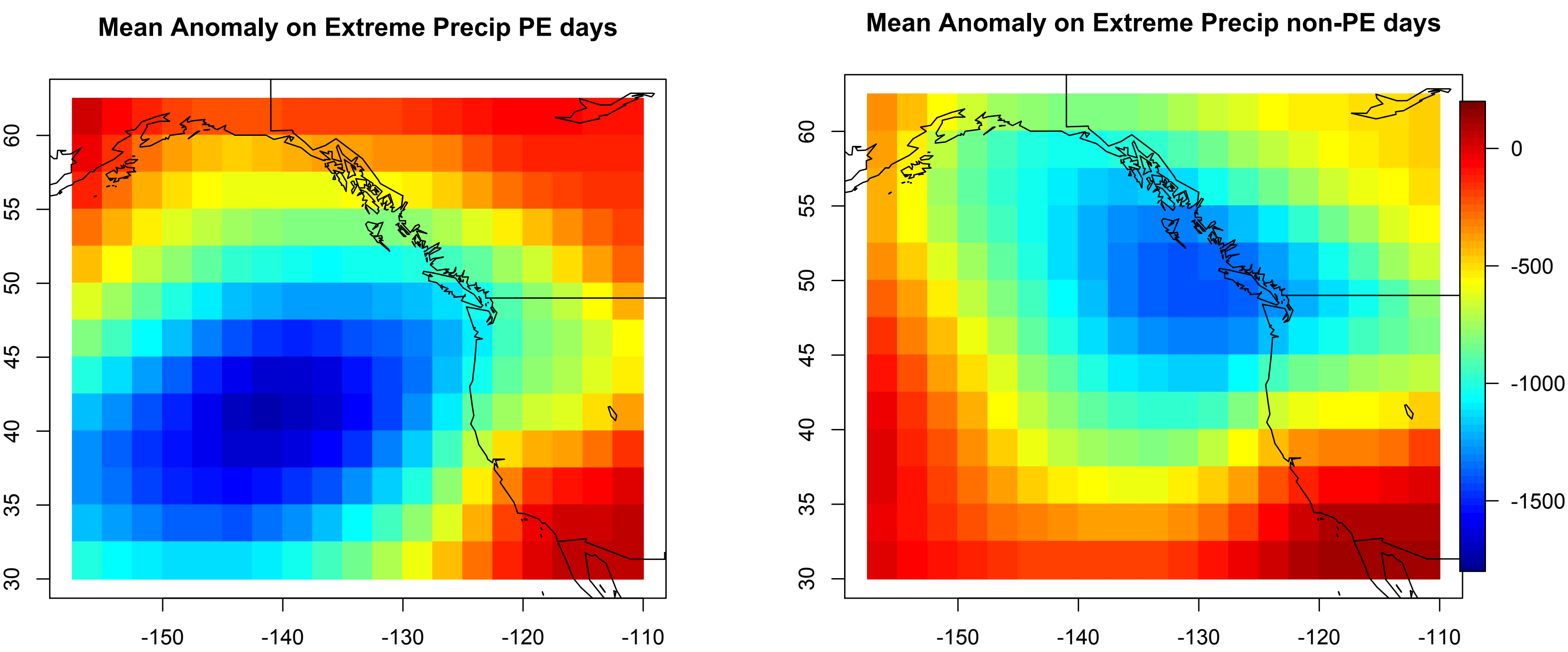


Find strong tail dependence - the WRF model represents the largest events quite well

Not all 'extreme' events are PE - aim to link to processes

Pineapple Express precipitation index

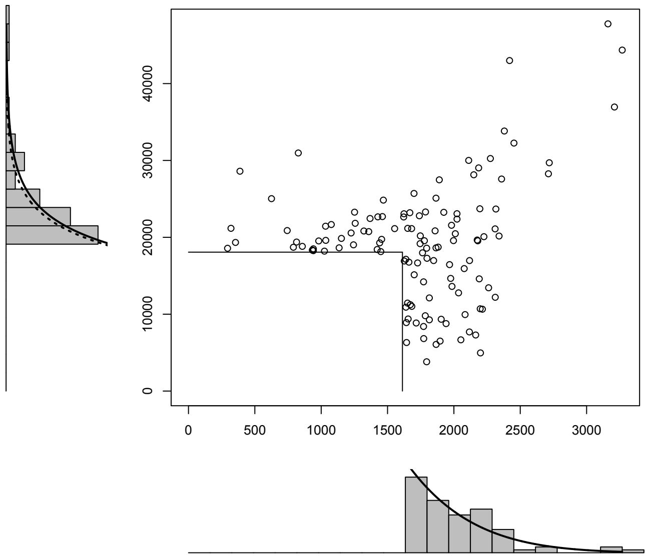
Developed a PE index from daily sea-level pressure fields

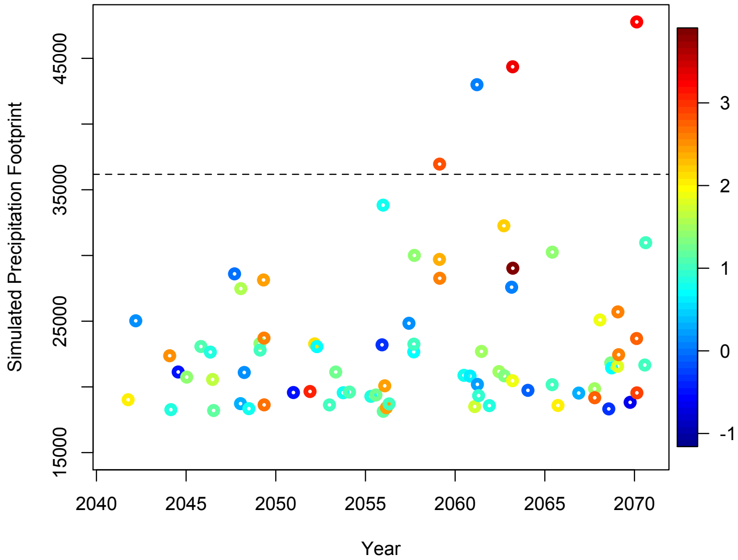


A first step toward linking regional extreme precipitation to large-scale processes

Future PE events from Regional Climate Models

Use fitted dependence model to study future extreme events





Findings: increase in both frequency and intensity of PE storms as produced by the WRF regional model forced by CCSM global model

Summary

- ‘Usual’ distributions don’t always capture tails correctly.
- Aim of EVT: study the tail of a distribution.
- An extreme value analysis utilizes only data considered ‘extreme’.
- Goal of the analysis is often extrapolation.
- Foundation provided by results from probability theory.
- Communication of uncertainty is critical.

- Dependence *not* described with correlations/covariances.
- Application areas: climate, finance, engineering, ...

Why graduate school?

- More and better job opportunities
  - Often more flexible and interesting
  - Not just a ‘number cruncher’
- Almost every day presents a new challenge
- Flexible (although busy) schedule
- You still get to be a student!

2009: Colorado State 23, Colorado 17 (in Boulder)

Why Statistics?

For Today's Graduate, Just One Word: Statistics

By STEVE LOHR
Published: August 5, 2009

- NY Times article: "the sexy job in the next ten years"
- Data is abundant, but people who can analyze it properly are not
- Computing power allows for new solutions to difficult problems
  - Need to quantify uncertainty in computer simulations
  - SAMSI 2011-2012 UQ program
- Wide variety of applications - can 'play in everyone else's backyard'

Preparing for grad school while at Concordia

As a Concordia mathematics major...

- Talk to your professors - they're always willing to help!
- Explore research opportunities
  - Research experiences for undergraduates (REU)
  - Undergraduate research with Concordia professors
  - SAMSI Undergraduate Workshop: February 24-25
- Take real analysis & other proof-based courses
- Get experience teaching/tutoring

When looking at schools...

- Ask a lot of questions!
- Money is important too - funding, fees, insurance, etc.
- If possible, make a visit

Reference and contact

Weller G., Cooley D., Sain S. An investigation of the pineapple express phenomenon via bivariate extreme value theory. *Submitted*.

Website: www.stat.colostate.edu/~weller

Email: gbweller@cord.edu