Parallel statistical extreme value analysis of climate

Chris Paciorek
Department of Statistics; University of California, Berkeley
with
Michael Wehner (LBNL), Prabhat (LBNL), David Pugmire (ORNL)

www.biostat.harvard.edu/˜paciorek

Research supported by DOE DE-AC02-05CH11231

March 2012

Climate extremes

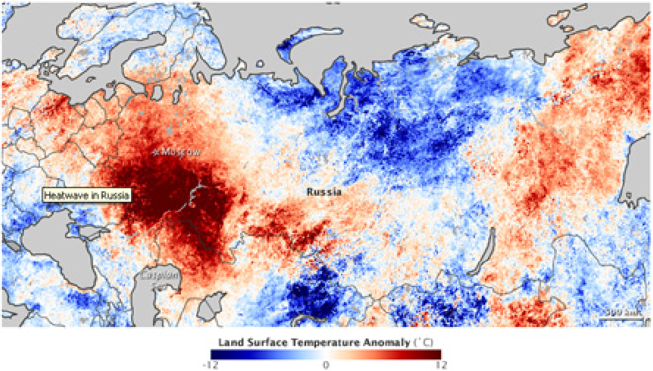


Figure: July 20-27 2010 temperature departures relative to 2000-2008 baseline, from the Washington Post

Analysis of climate extremes

- Other examples: 2003 European heat wave, 2011 Texas drought/heat wave, 2011 Mississippi River flooding, 2000 English floods, 2011 Thailand floods
- Climate scientists are interested in detecting, attributing, and projecting changes in extremes.
    - This involves analyzing the observational record and climate model output (both hindcasts and forecasts).
- My focus here is on continuous outcomes, but changes in event frequency and intensity (hurricanes, tornados, storm surges, etc.) are of great general interest.

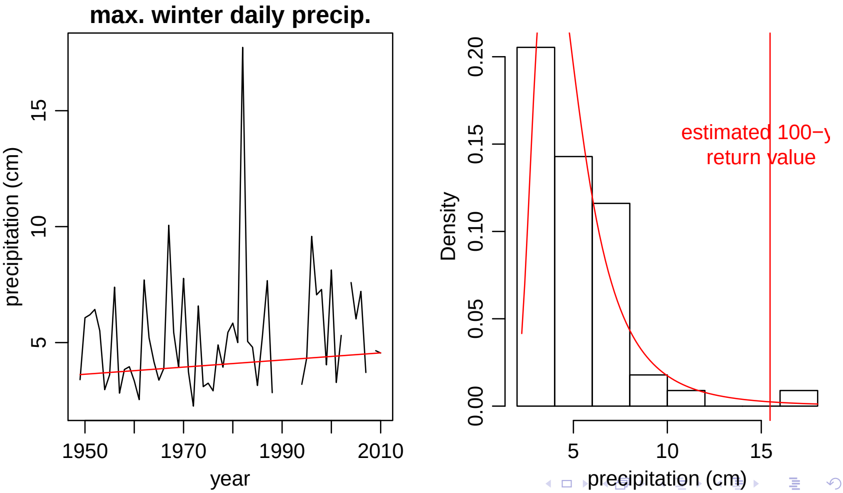
Statistical extreme value theory

- The Generalized Extreme Value (GEV) distribution:

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

- Location parameter μ, scale parameter σ, shape parameter ξ, generally fit via ML.
- Unites three distributions:
  1. $\xi < 0$: Weibull distribution; bounded to the right
  2. $\xi = 0$: Gumbell distribution; exponential tail
  3. $\xi > 0$: Frechet distribution; heavy tail
- Asymptotic theory says that the distribution of block maxima (or minima) converges to the GEV distribution as the block size goes to infinity.
- This provides a statistically rigorous way to analyze extremes and estimate probabilities of extreme events.

Example: Berkeley, CA winter precipitation



Return values (levels)

- A 100-year flood is the size of flood expected to occur once every 100 years on average, also called the 100-year return value. I.e., a tail probability of $p = 0.01$.
- By the quantiles of the GEV distribution, the MLE for the $1/p$-year return value is:

$$\hat{z}_p = \hat{\mu} - \frac{\hat{\sigma}}{\hat{\xi}}\left(1 - (-\log(1-p))^{-\hat{\xi}}\right)$$

- Uncertainty can be estimated based on the delta method.
- Analysis of extremes is necessarily based on limited data and involves extrapolation, though the asymptotic theory provides some justification.

Nonstationary extreme value analysis

Climate change concerns lead us to investigate whether extremes are changing over time. Extremes may also vary by season and with covariates (in particular teleconnections such as ENSO).

- A basic strategy:

$$F(x) = \exp\left(-\left[1+\xi_t\left(\frac{x_t-\mu_t}{\sigma_t}\right)\right]^{-1/\xi_t}\right)$$

- One might have all three parameters vary with time (linearly, polynomially, or based on splines).
- Analyses often find little evidence (based on likelihood ratio tests) that ξ (and even σ) are varying with time, though ξ in particular is hard to estimate even in a stationary model.
- A basic model is linear in time in μ only, as a first-order estimate of the trend over time.

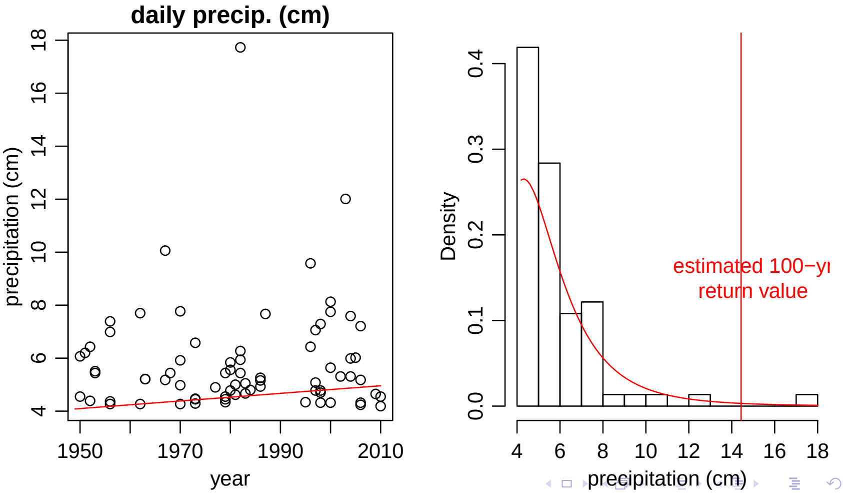
Point process modeling

- Using block maxima seems wasteful of data. Instead model all the exceedances over a high threshold, c (e.g., the 95%ile or 99%ile of all rainy days in the data).
- The point process model specifies the probability of the number of exceedances (the intensity measure) and the likelihood of the actual exceedances (the intensity function):

$$L(\mu, \sigma, \xi; x_1, \ldots x_n) \propto \exp\left(-n_y \left[1 + \xi \left(\frac{c - \mu}{\sigma}\right)\right]^{-1/\xi}\right) \cdot \prod_{i=1}^{N(A)} \frac{1}{\sigma} \left[1 + \xi \left(\frac{x_i - \mu}{\sigma}\right)\right]^{-1/\xi - 1}$$

- The parameters are equivalent to the GEV parameters and can be used to compute return values.
- Asymptotics are with respect to the threshold getting larger.

Example: Berkeley, CA winter precipitation



Spatial extreme value analysis

- Given the sparsity of data and the spatial structure of weather, an obvious goal is to do a spatial analysis of multiple locations.
    - Borrow strength
    - Acknowledge joint uncertainty
- Standard spatial analyses have assumed spatially-correlated parameters, but conditionally IID observations.
    - Hierarchical Bayesian approaches have been a common approach: Cooley, Gelfand, Sang, Shaby, and others
    - Computation is a big hurdle and MCMC performance can be poor
- Analysts often remove consecutive exceedances to reduce temporal autocorrelation
- Some recent work on models that allow for spatially-correlated observations.

Our perspective

- Given the size of observation and climate model output datasets and the increasing spatial resolution of models, a hierarchical modeling strategy fit by MCMC is not practical for most large-scale and production-mode climate analysis.
- Our focus:
    - Location-specific analysis (embarrassingly parallel)
    - Basic models for temporal change (linear)
    - Stratify by season rather than modeling seasonality
    - Work with return values (reduce dimensionality from parameter space)
    - Joint uncertainty via bootstrapping
    - Parallel software development

Bootstrapping

- By resampling the same blocks at each location, one preserves the spatial structure and can estimate joint uncertainty
- Embarrassingly parallel
- Atmospheric oscillations (ENSO, NAO, etc.) operate with multi-year periodicity and induce within-year and across-year autocorrelation
  - Basic approach is to bootstrap in year-long blocks
  - Open question: should one account for longer-term dependence?

Spatial strategy 1: Multi-stage analysis

- Given location-specific fits, one can consider a second-stage analysis of return values estimated at each location, $\hat{v}_i$:

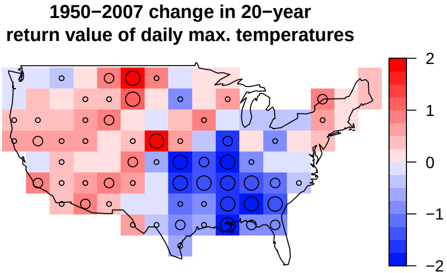
  $$\hat{v}_i = g(s_i) + \epsilon_i$$

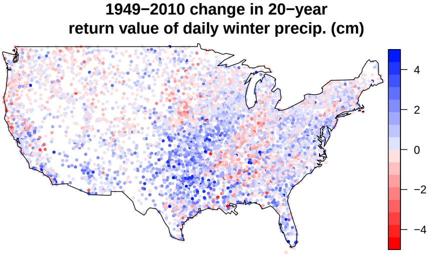
  with $g(\cdot)$ specified as a Gaussian spatial process and $\widehat{\mathrm{Cov}}(\epsilon)$ based on bootstrap.
- Similar to meta analyses and spatial smoothing of air pollution time series results
- Some unpublished work by Richard Smith on this.

Spatial strategy 2: Local likelihood

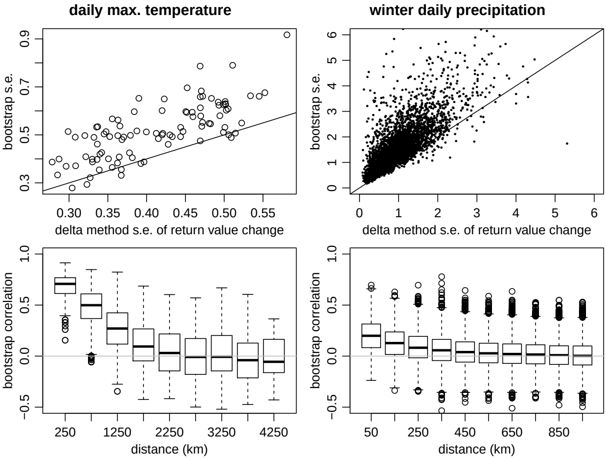
- Ramesh & Davison (2000; JRSSB) suggest to use local likelihood to smooth in time
- Here we propose to use local likelihood to smooth in space, using cross-validation of the log-predictive density to choose the bandwidth
    - Fit for each location, borrowing strength in a neighborhood (one might also consider accounting for elevation in the neighborhood definition)
    - Normal density smoothing kernel, truncated at 3σ
    - Common threshold based on quantile of focal location
    - Common parameter values, but one might consider locally linear parameters
- Bootstrap can again provide uncertainty estimates

Location-specific fits


1950−2007 change in 20−year
return value of daily max. temperatures
2
1
0
−1
−2

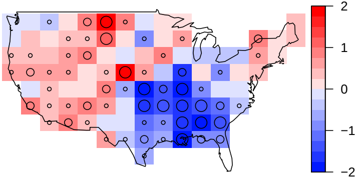

1949−2010 change in 20−year
return value of daily winter precip. (cm)
4
2
0
−2
−4

Bootstrap-based uncertainty and spatial dependence

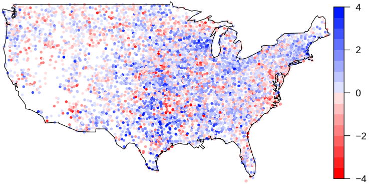


Multi-stage analysis

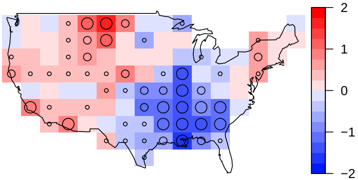
Unsmoothed return value change for temperature



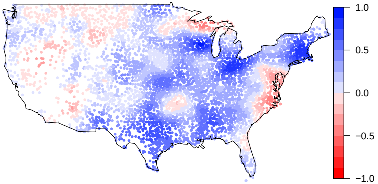
Unsmoothed return value change in summer precip. (cm)



Kriged return value change for temperature



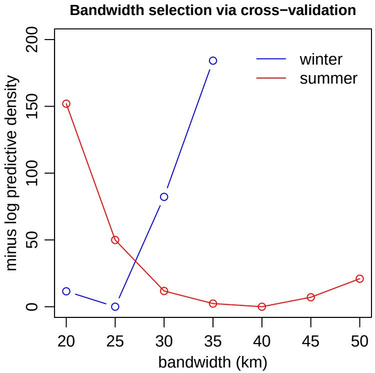
Kriged return value change in summer precip. (cm)



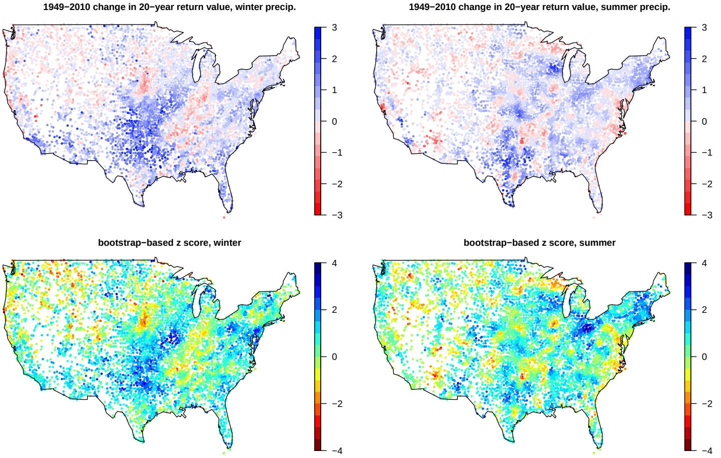
Comments: Temperature kriging has little effect.
Precipitation kriging based on diagonal of bootstrap covariance.

Local likelihood bandwidth selection

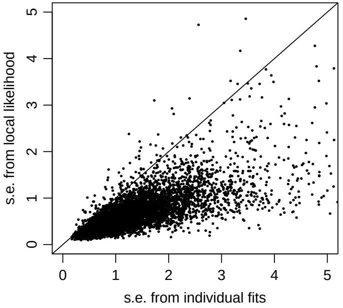
Hold out 10% of stations (approximately 550), estimating parameters from neighboring stations.

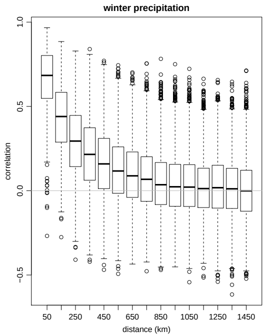


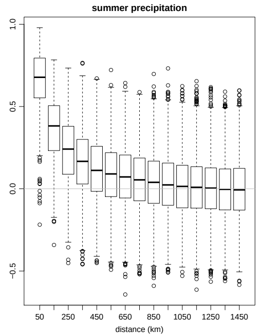
Local likelihood fits


1949−2010 change in 20−year return value, winter precip.
1949−2010 change in 20−year return value, summer precip.
3
2
1
0
−1
−2
−3
bootstrap−based z score, winter
bootstrap−based z score, summer
4
2
0
−2
−4

Local likelihood uncertainty







Open questions

- How use bootstrap-based error covariance in a second-stage smoothing context?
- How should we display and assess joint uncertainty?
- Would the False Discovery Rate approach be helpful for assessing the collection of z-scores, particularly given the larger bootstrap standard errors?

R software development

- The *ismev* package (accompanying the Coles books "Introduction to Statistical Modeling of Extreme Values") fits GEV and Point Process models, with a general $X\beta$ form for all three parameters
- I am building the following capabilities on top of the *ismev* functionality:
    - Handling missing values in point process modeling (common in observational data), assuming MAR missingness
    - Fitting point process models given only the exceedances; this greatly speeds computation
    - Calculating return values
    - Including delta-method-based uncertainty for return values and differences in return values
    - Including block bootstrap capability
    - Allowing local likelihood fitting

Parallel software deployment

- An LBNL/LLNL/ORNL/LANL team is developing climate data analysis tools within the context of the VisIt parallel visualization software (developed at the national labs).
- The core idea is to allow for extremes analysis in VisIt (and also in the new UV-CDAT software) by calling R functionality.
- VisIt will handle (in parallel) data input/output, calendaring, data reduction to block maxima or exceedances, mobilizing multiple R instances, collection of location-specific results, and visualization.
  - VisIt's VtK data structures passed to R
  - R code called by VtK
- R will handle location-specific model specification, likelihood maximization, calculation of return values, and bootstrapping uncertainty.