Introduce Myself - Elizabeth Shamseldin

- PhD Candidate under Richard Smith at UNC-Chapel Hill
- Working with Bayesian methods in spatial statistics
- Opportunity to work on extreme value project at NCAR

Extreme Precipitation: An Application Modeling N-Year Return Levels at the Station Level

Presented by: Elizabeth Shamseldin
Joint work with: Richard Smith, Doug Nychka, Steve Sain, Dan Cooley

Statistics Group, IMAGe

National Center for Atmospheric Research

April 27, 2006

Motivating Question

Extreme precipitation under various climate scenarios

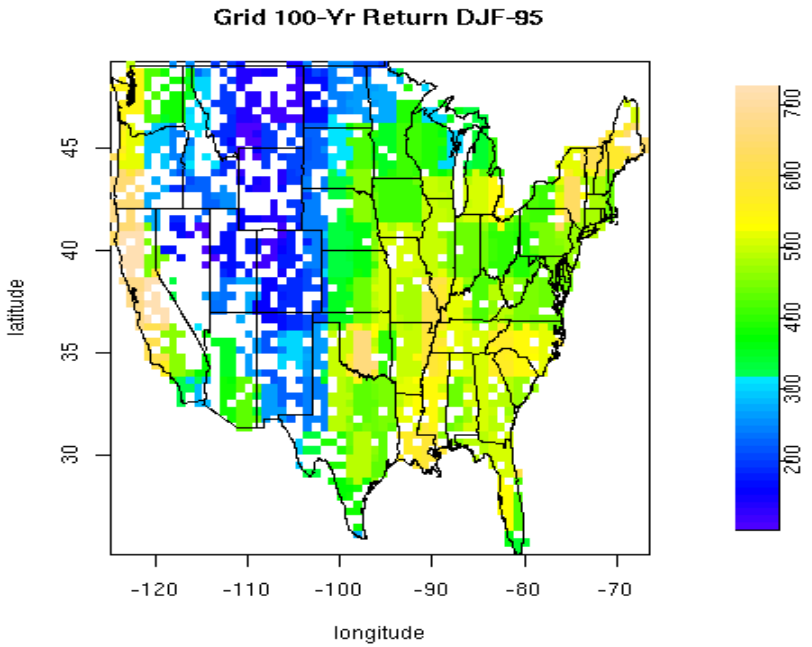
- Can regional climate model return level estimations be used to obtain return level predictions at the station level?
- Relationship between extremes of grid-cell data (re-analysis NCEP) and weather station data (NCDC)
- Leads to relationship between climate model data and point source data

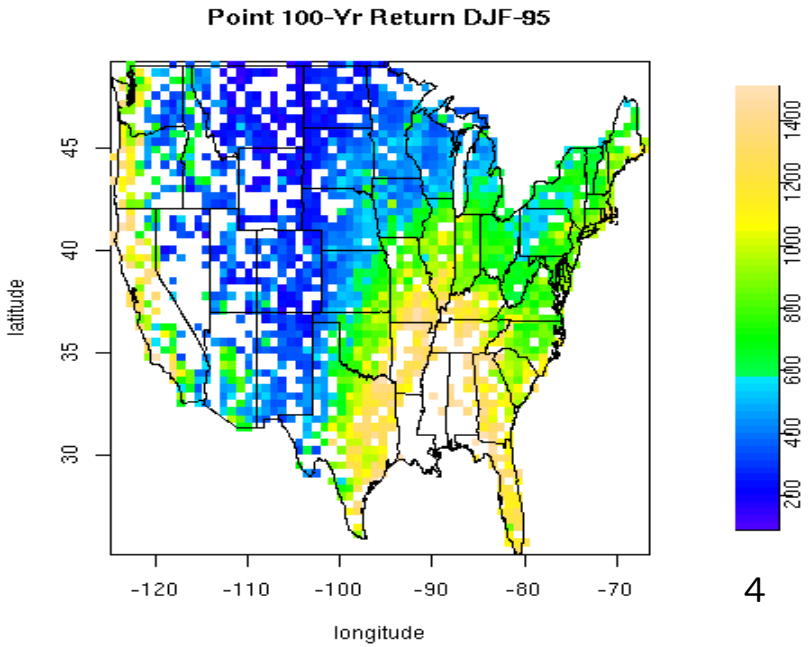
Data

- Point-source observational data from NCDC, originally obtained from Dr. Pavel Groisman
- Covers period 1950-1999 over 5873 stations.
- The data are daily rainfall values; units are tenths of a millimeter.
- Grid-cell data are from NCEP
- Covering period 1948–2003 with no missing data on 288 2.5^{o} grid cells, converted to the same units as the NCDC data.
- Rainfall values are considered over the four seasons
- Threshold values are determined by the 95^{th} and 97^{th} percentiles
- Clustering method is used to define peaks

Outline of Techniques

- Relationship of grid cell n-year return levels to the n-year return levels at point locations is explored
- Patterns are similar, scales differ
- Need model to translate between grid level and point level returns





Outline of Techniques

Tail of the Generalized Extreme Value distribution (GEV) is fit to the grid cell data above a given threshold

$$\Pr\{Y \leq y\} \;=\; \exp\left\{-\left(1+\xi\frac{y-\mu}{\psi}\right)_{+}^{-1/\xi}\right\} \qquad (1)$$

where Y is a random variable, μ is a location parameter, ψ a scale parameter and ξ is the extreme-value shape parameter; μ and ξ can take any value in $(-\infty, \infty)$ but ψ has to be > 0.

Technique

- Peaks Over Threshold model: considers $\Pr\{X \leq u + x \mid X > u\}$ for a given threshold u; the parameters are directly tied to the threshold value

- Point process approach is similar to the peaks over threshold method: all observations over a given threshold are considered. Tail of GEV is estimated.

- However, in the PP approach, the parameters are not tied to the threshold value.

Theory **Provided the model fits the data:**

- PP approach should produce equivalent parameter values as the POT approach

- Parameters are independent of the threshold (adjusting for estimation error)

- Ideal threshold determined by considering where the parameter values stabilize

Theory

Choose to model n-year return due to interpretability. *Event so extreme it is expected to happen once every n years:* ($\frac{1}{n}$ probability/yr)

The n-year return values can be directly obtained using the estimated GEV parameters. Define y_n by the equation:

$$\left(1+\xi\frac{y_n-\mu}{\psi}\right)^{-1/\xi} = \frac{1}{n}$$

which leads to the formula:

$$y_n = \begin{cases}\mu+\psi\frac{n^{\xi}-1}{\xi} & \text{if } \xi\neq 0,\\ \mu+\psi\log n & \text{if } \xi = 0.\end{cases} \tag{2}$$

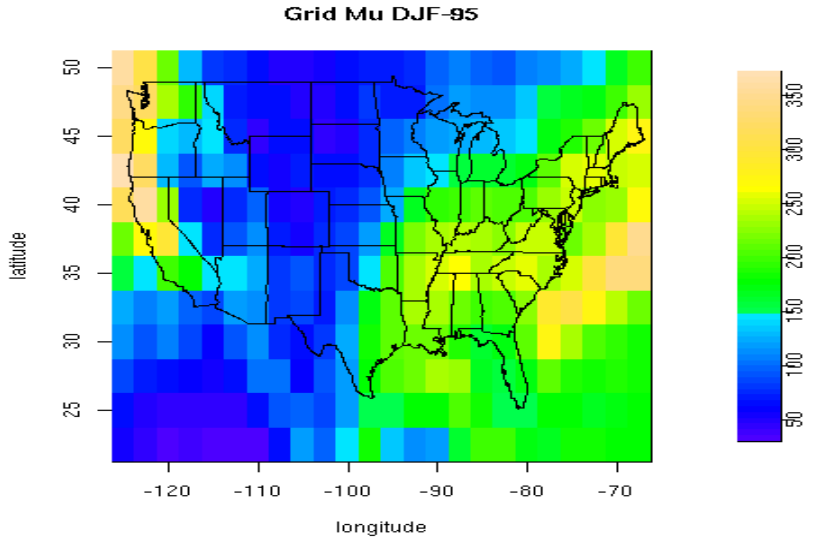
Recap

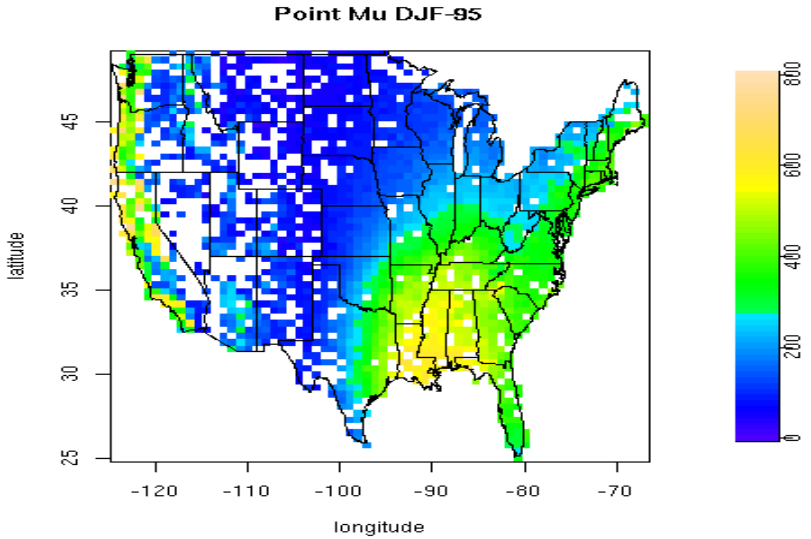
- GEV parameter estimates are obtained and checked through various diagnostics

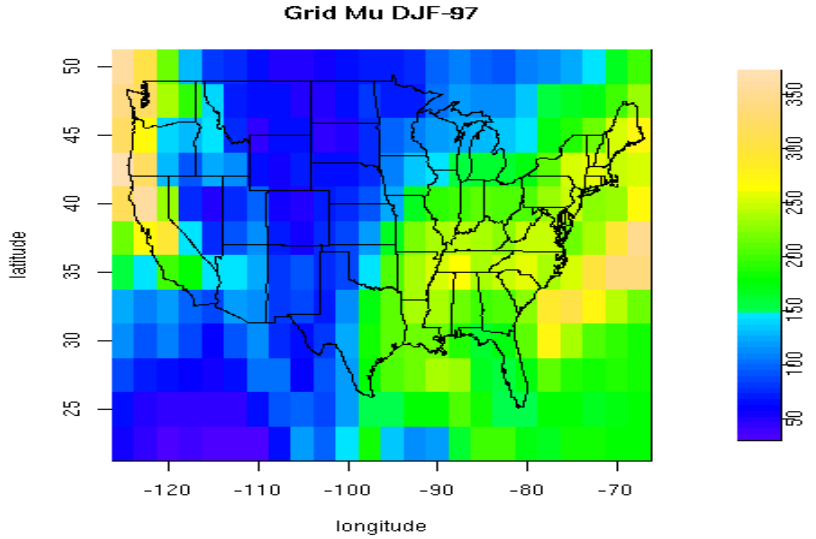
- GEV parameter estimates are used to generate n-year return levels at grid and station levels

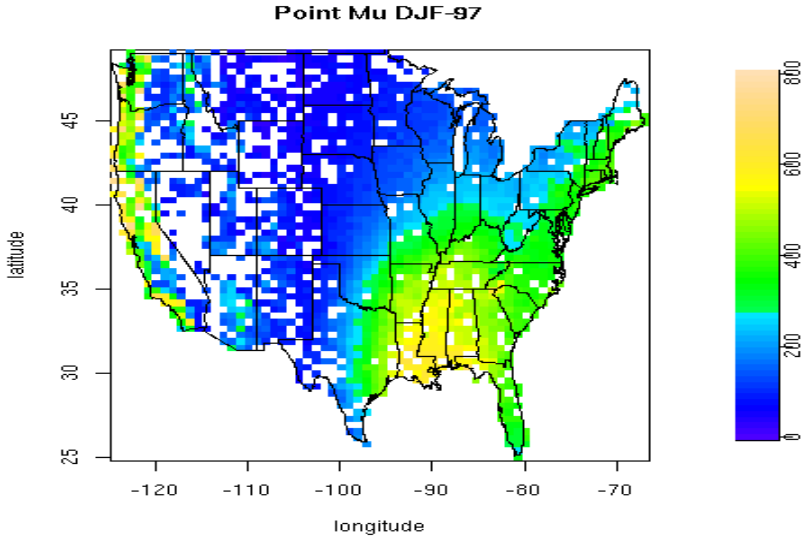
- Various models are explored to predict point location n-year return from grid cell n-year return

GEV Model Fit - Mu Parameter: 95 Grid - Point vs 97 Grid - Point

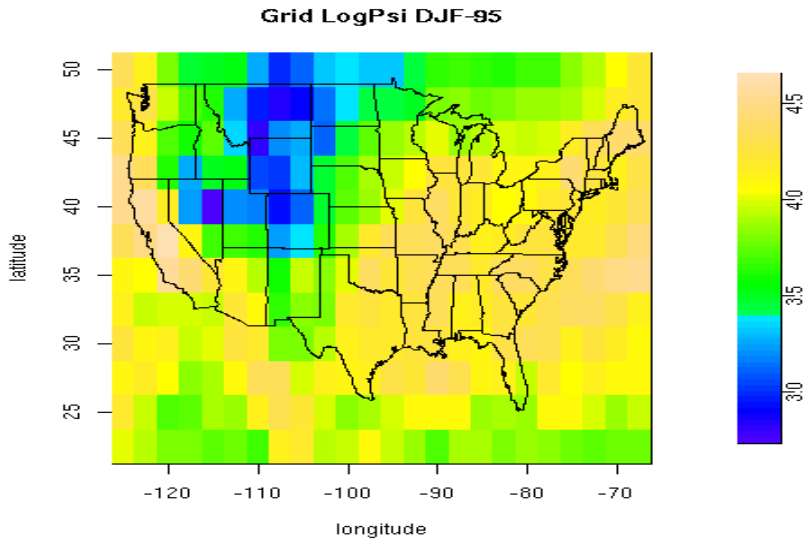


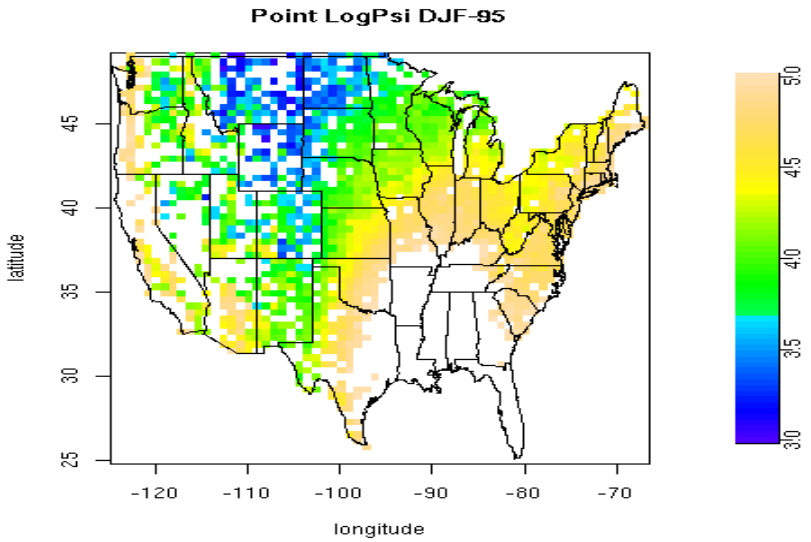


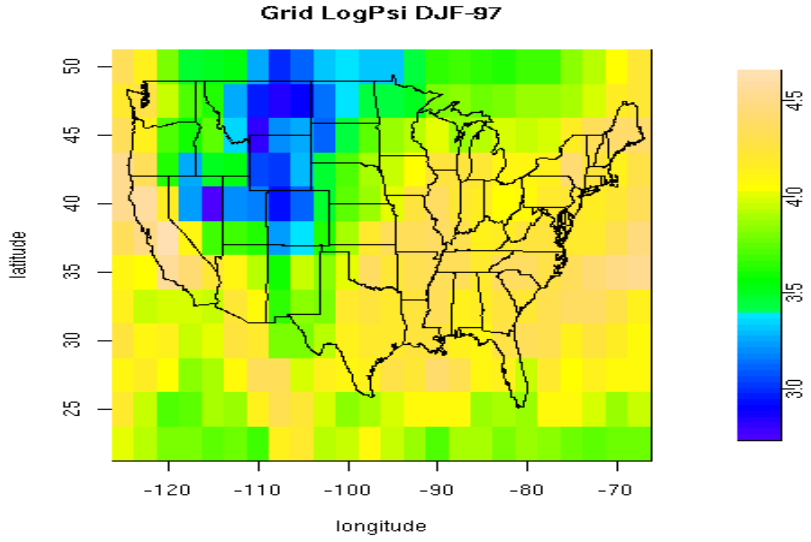


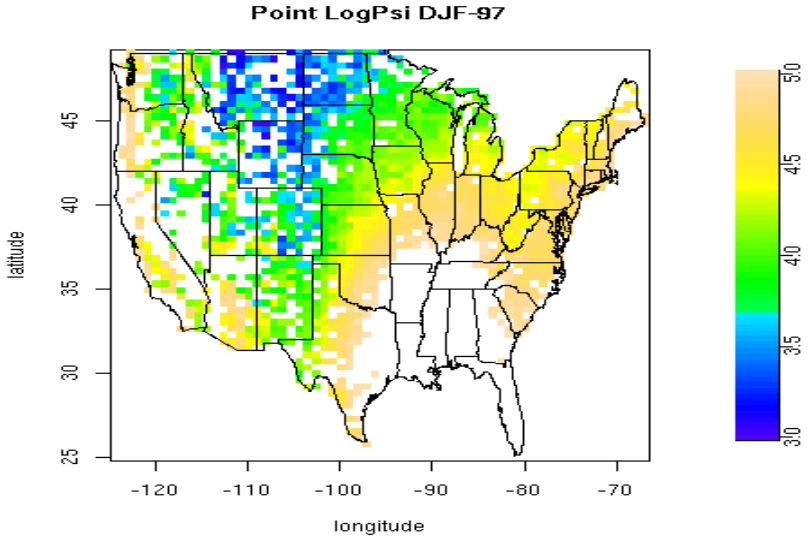


GEV Model Fit - LogPsi Parameter: 95 Grid - Point vs 97 Grid - Point

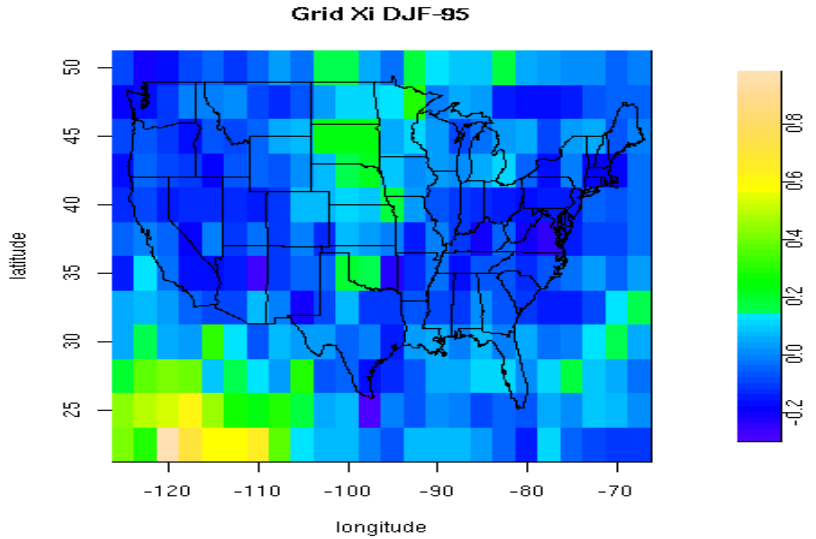


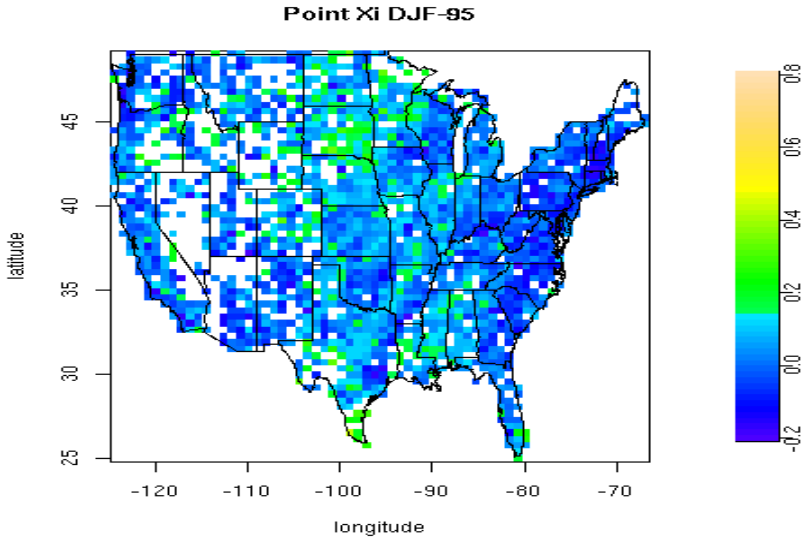


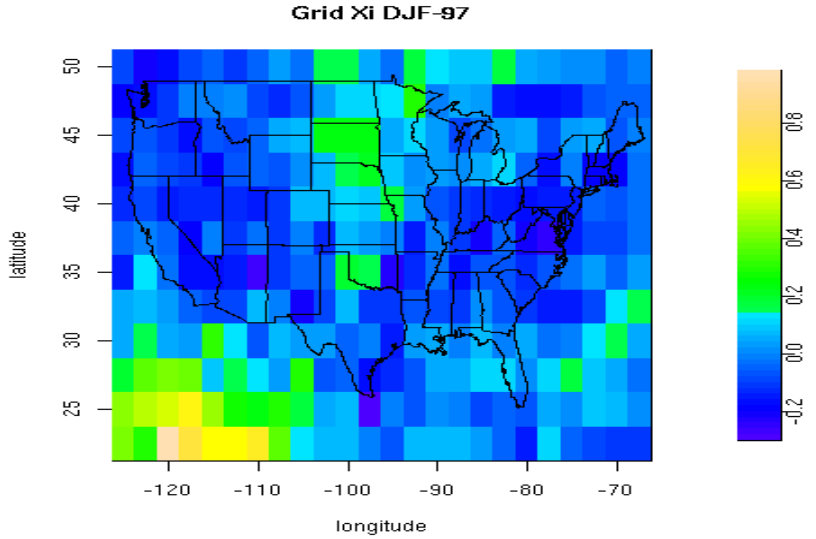


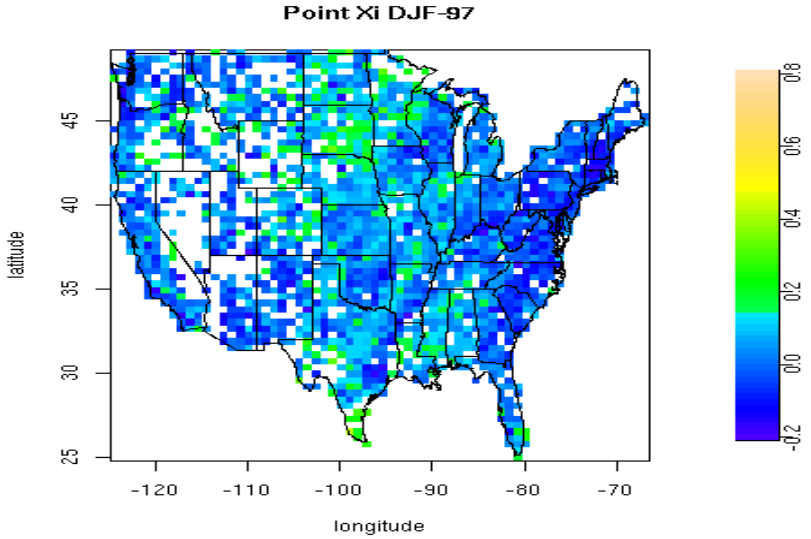


GEV Model Fit - Xi Parameter: 95 Grid - Point vs 97 Grid - Point

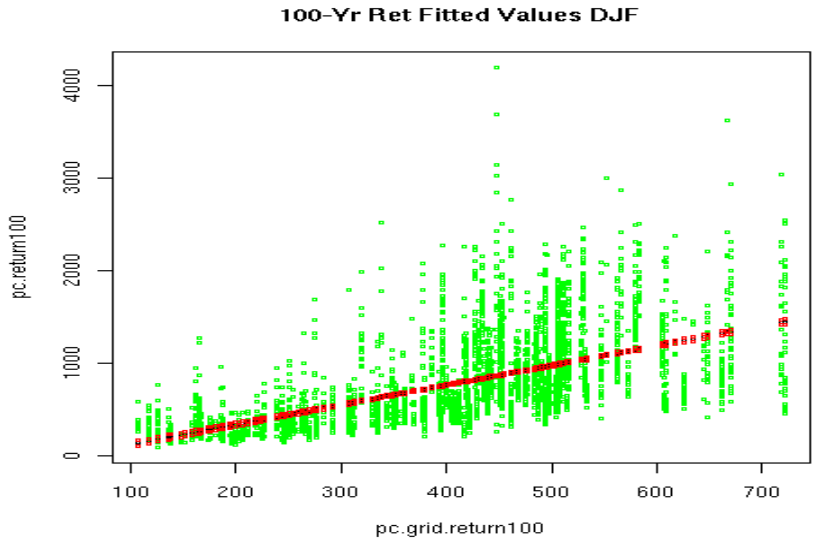


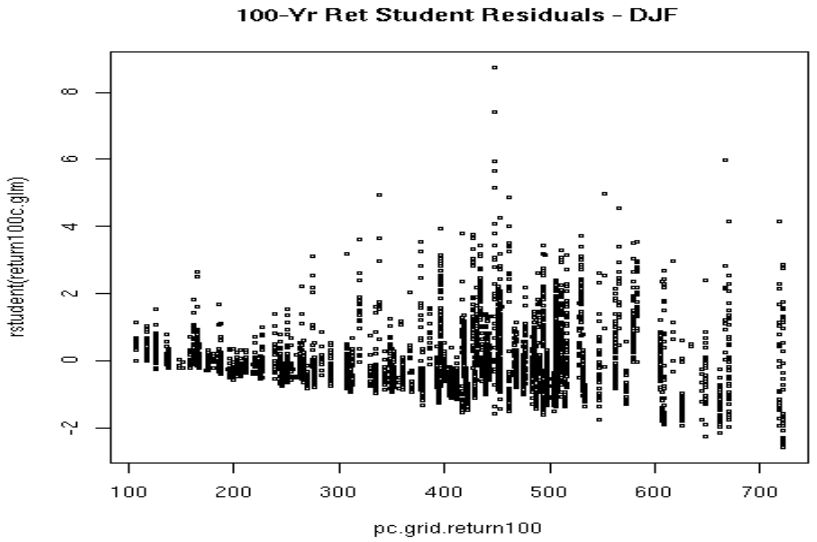


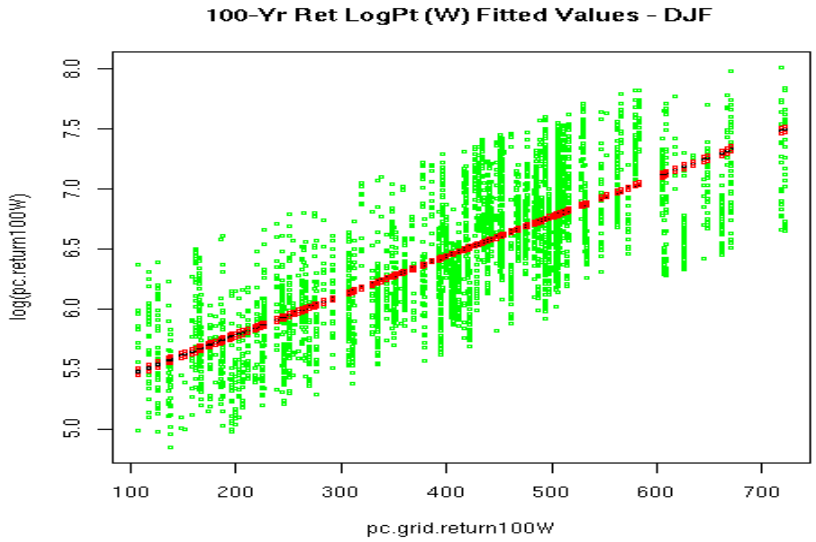


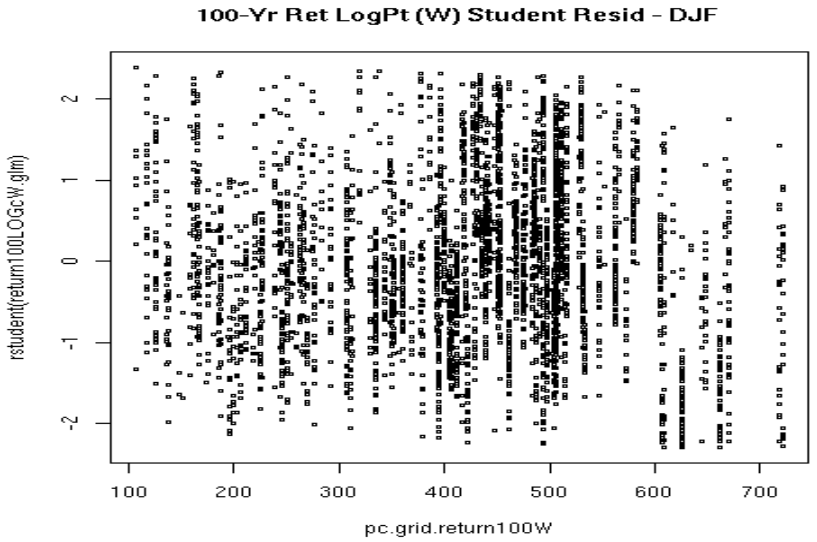


100-Year Return: Point vs Grid - LogPoint vs Grid

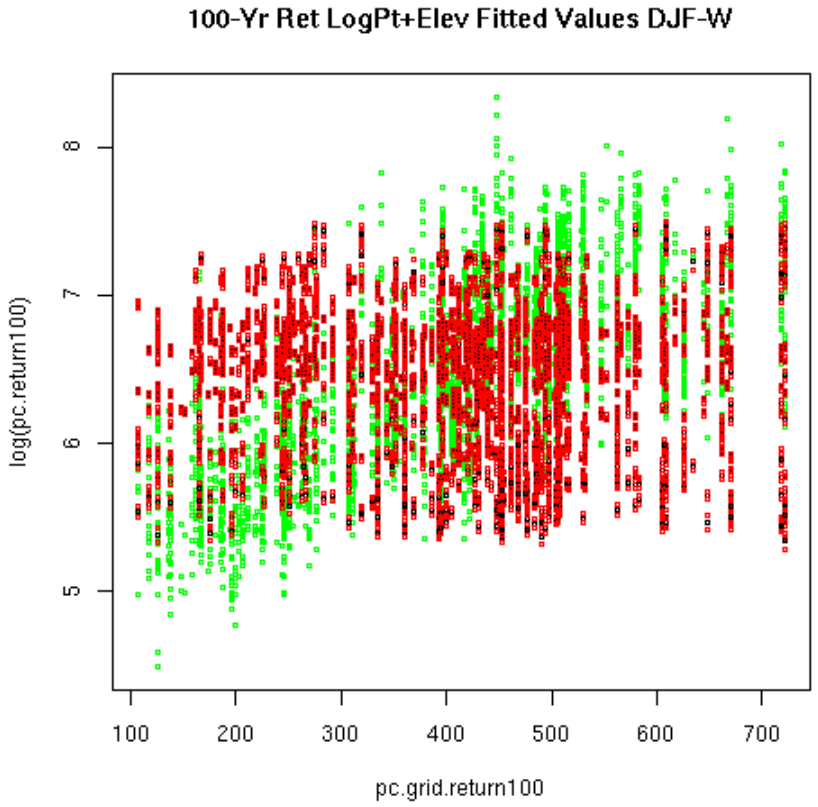


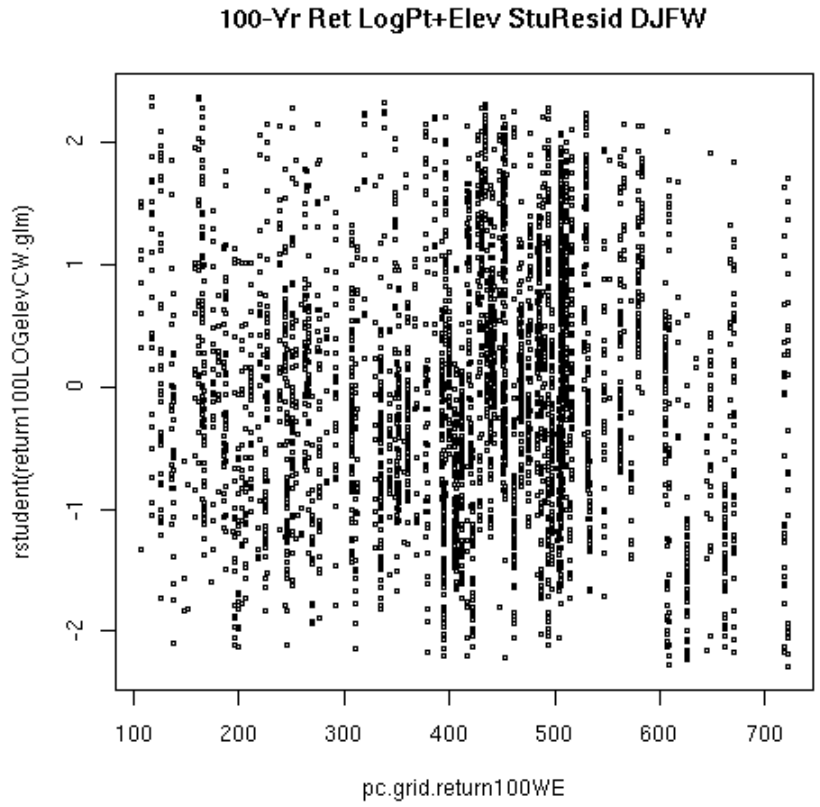






100-Year Return: LogPoint + Elevation vs Grid



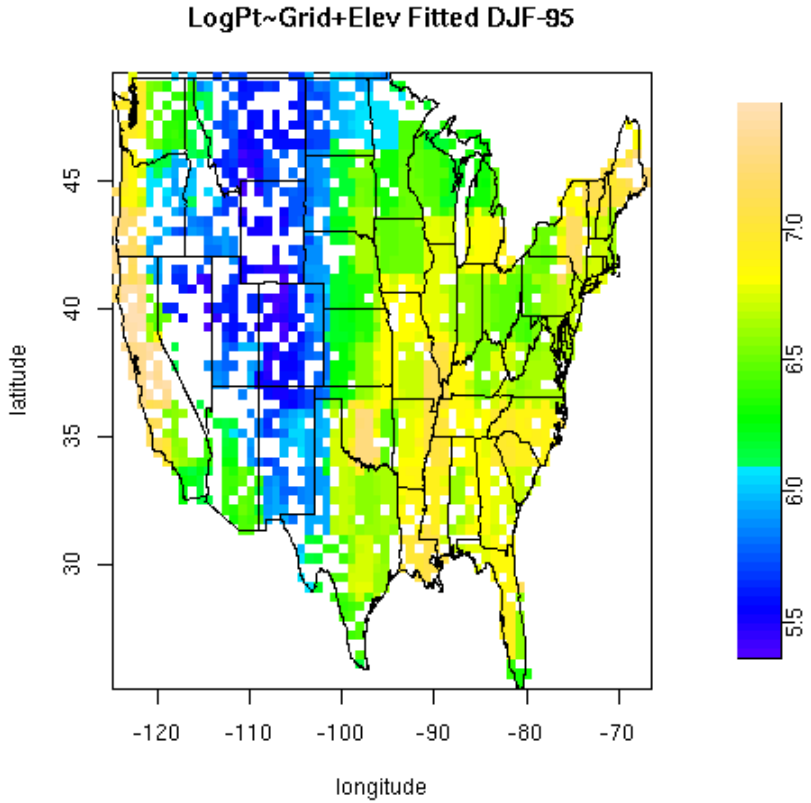


Model Results and Comparison Univariate Regression Predicting Point Locations 100-Year Return Level Using Grid Cell 100-Year Return

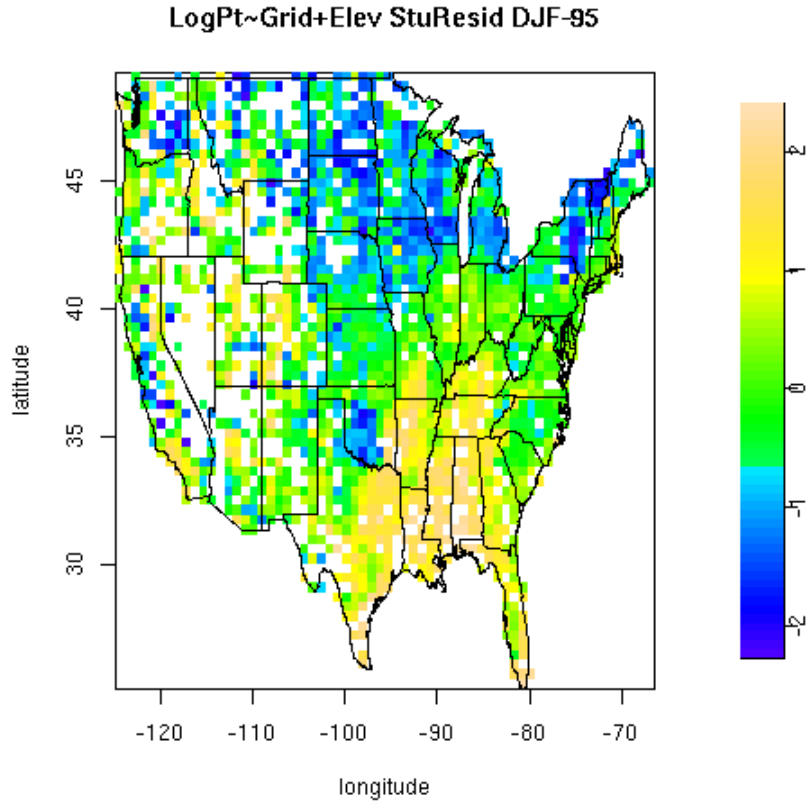
WINTER - 100	OBS	AIC	Intercept	X1	X2
log(Pt) Grid+Elev	4025	3425	5.3186	0.0030	-0.000124
97 log(Pt) Grid+Elev	4013	3356	5.3074	0.0030	-0.000113
SPRING - 100	OBS	AIC	Intercept	X1	X2
log(Pt) Grid+Elev	4138	1924	5.8960	0.0023	-0.000258
97 log(Pt) Grid+Elev	4145	2282	5.9736	0.0021	-0.000281

100-Year Return: LogPoint ~ Grid + Elevation Spatial Trend

Fitted 100-Yr Return Values

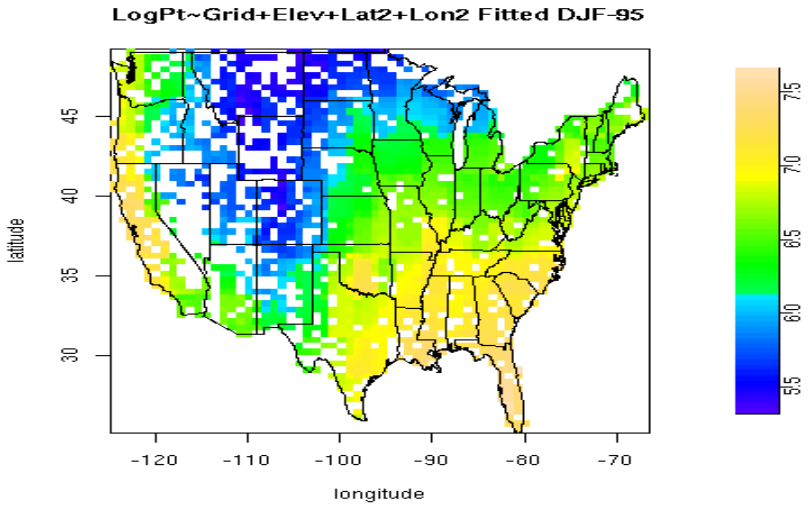


Residuals

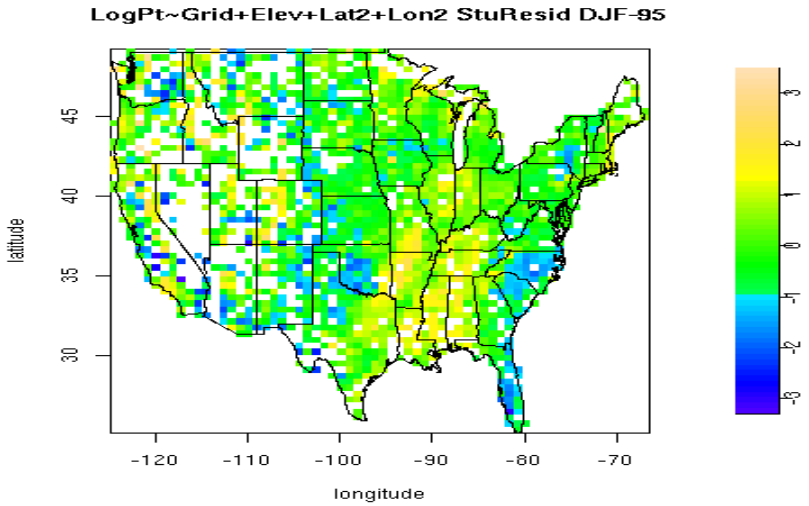


100-Year Return: LogPt $\sim$ Grid + Elevation Including Lat and Lon

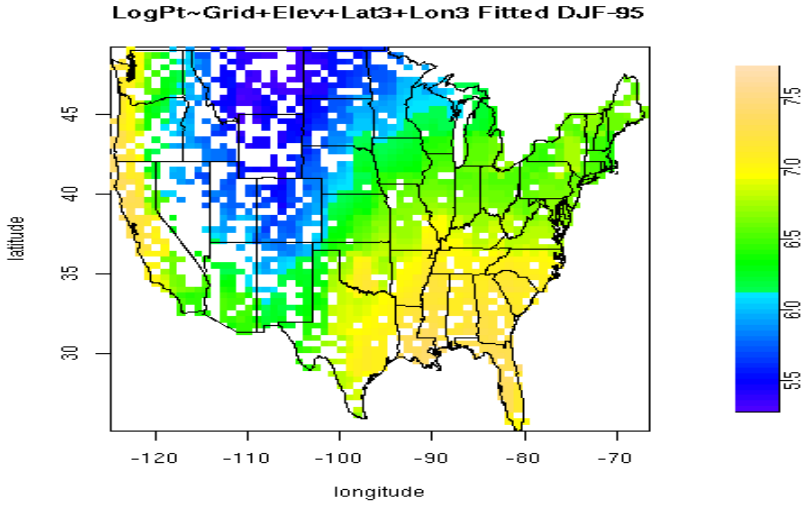
Quadratic Model



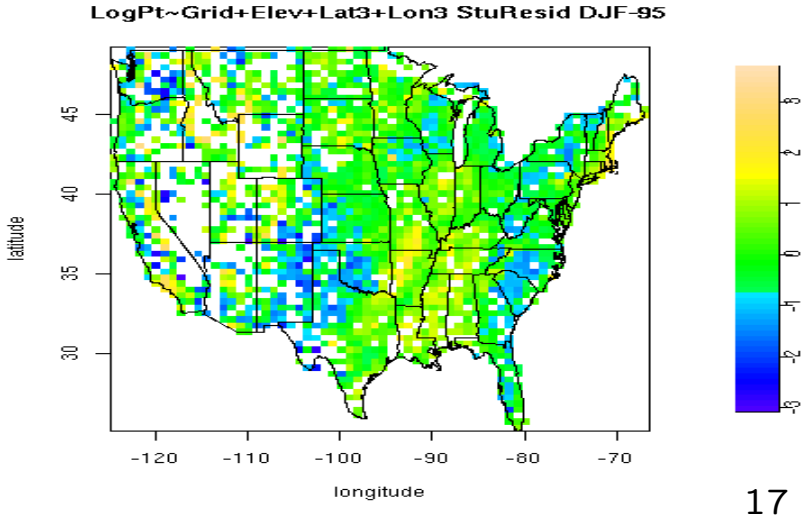
Residuals



Cubic Model

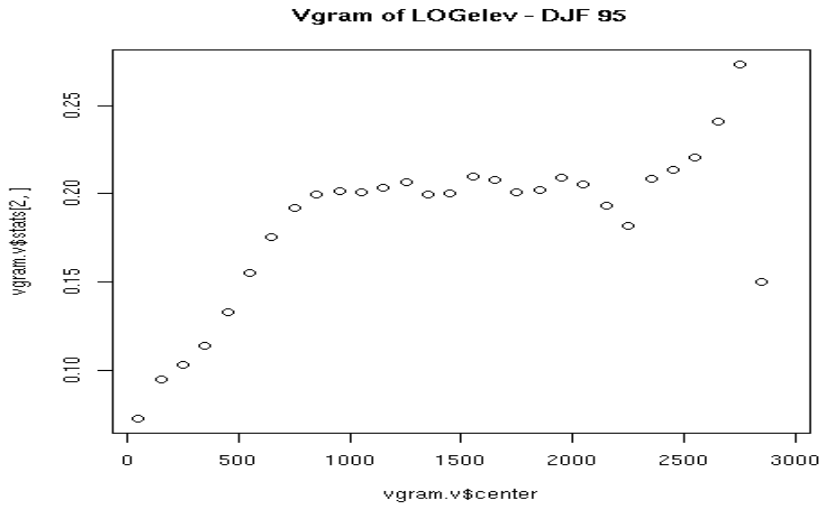


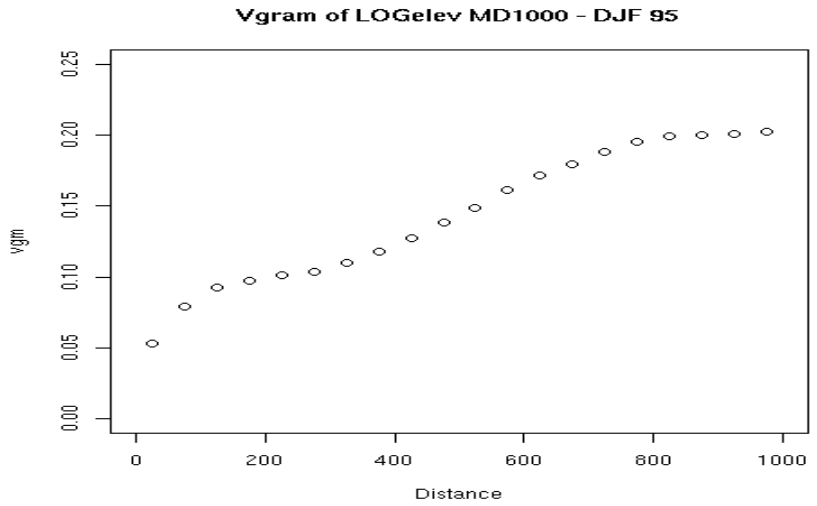
Residuals



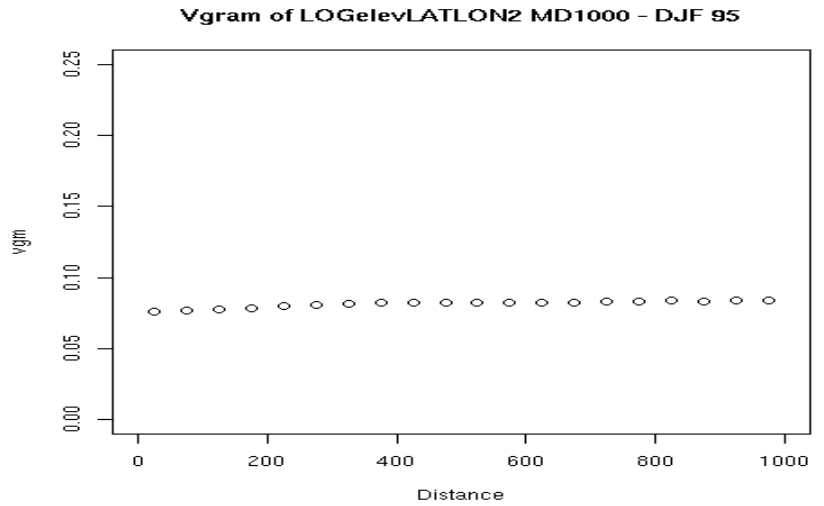
100-Year Return: LogPt ∼ Grid + Elevation Including Lat and Lon

Variograms of Log(Pt)+Elevation

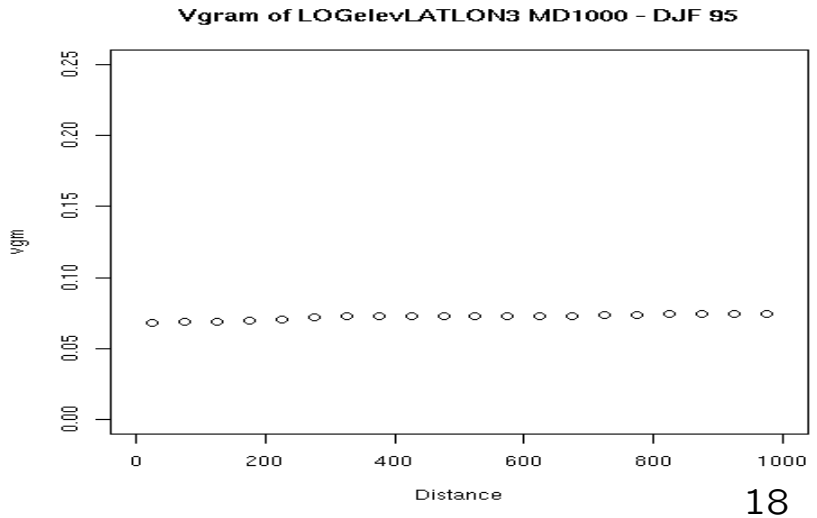




Quadratic Variogram



Cubic Variogram



Other Models Considered

Cubic vs Quartic

- Cubic and quartic models show little difference in the remaining spatial trend
- Cubic model: all terms and interaction terms are significant in the model
- Quartic model: some terms are no longer significant, including elevation
- AIC vs BIC

Elevation

- Quadratic models in elevation: did not resolve the spatial correlation among station return levels as well as the higher order models in latitude and longitude

Conclusion and Future Work

- Modeling the tail of the GEV Distribution appears to produce stable GEV parameter estimates and model coefficients within seasons and across 95% and 97% thresholds

- 100 year return levels are successfully modeled by season at the point (station) level using grid-level return values, station elevation, and station latitude and longitude coordinates

- The regression relationship between climate model predictions and future modeled extremes is best expressed through a cubic model in latitude and longitude to account for the spatial correlation, with $R^2 = 0.77$ and standard errors within 1.008 and 1.056 tenths of a millimeter.

- Future work includes plans to test grid-point models on RCM and CCSM data

References

Coles, S.G. (2001) An Introduction to Statistical Modeling of Extreme Values. Springer Verlag, New York.

Davison, A.C. and Smith, R.L. (1990), Models for exceedances over high thresholds (with discussion). J.R. Statist. Soc., 52, 393-442.

Fisher, R.A. and Tippett, L.H.C. (1928), Limiting forms of the frequency distributions of the largest or smallest member of a sample. Proc. Camb. Phil. Soc. 24, 180190.

Gumbel, E.J. (1958), Statistics of Extremes. Columbia University Press.

Katz, R.W., Parlange, M.B. and Naveau, P. (2002), Statistics of extremes in hydrology. Advances in Water Resources (fill in full details of reference)

Kharin, V.V. and Zwiers, F.W. (2000), Changes in the extremes in an ensemble of transient climate simulations with a coupled atmosphere-ocean GCM. Journal of Climate 13, 37603788.

Leadbetter, M.R., Lindgren, G. and Rootz?en, H. (1983), Extremes and Related Properties of Random Sequences and Series. Springer Verlag, New York.

Pickands, J. (1975), Statistical inference using extreme order statistics. Ann. Statist. 3, 119131.

Smith, R.L. (1989), Extreme value analysis of environmental time series: An application to trend detection in ground-level ozone (with discussion). Statistical Science 4, 367-393.

Smith, R.L. (1990), Extreme value theory. In Handbook of Applicable Mathematics 7, ed. W. Ledermann, John Wiley, Chichester. Chapter 14, pp. 437-471.

Smith, R.L. (2003), Statistics of extremes, with applications in environment, insurance and finance. Chapter 1 of Extreme Values in Finance, Telecommunications and the Environment, edited by B. Finkenstadt and H. Rootzen, Chapman and Hall/CRC Press, London, pp. 178. http://www.unc.edu/depts/statistics/postscript/rs/semsta

Smith, R.L. and Weissman, I. (1994), Estimating the extremal index. J.R. Statist. Soc. B 56, 515528.

Zwiers, F.W. and Kharin, V.V. (1998), Changes in the extremes of the climate simulated by CCC GCM2 under CO2 doubling. Journal of Climate 11, 22002222.

Theory

To take account of these deficiencies, an alternative class of methods has been developed, often known as the *Peaks Over Thresholds* (POT) approach. For the distribution of the excess values, a common family of probability density functions is the *Generalized Pareto Distribution* (GPD), introduced by Pickands (1975), and given by

$$\Pr\{X \leq u + x \mid X > u\} \;=\; 1 - \left(1 + \xi\frac{x}{\sigma}\right)_{+}^{-1/\xi}. \qquad (3)$$

One drawback in the POT model is that the parameters are directly tied to the threshold value, u.

A third approach, the *point process approach* (Smith 1989, 2003, Coles 2001), although operationally very similar to the POT approach, uses a representation of the probability distribution that leads directly to the GEV parameters (μ, ψ, ξ).

Theory

The Point Process method considers N peaks, $Y_1 ... Y_N$, observed at times $T_1 ... T_N$. Pairs are viewed as points in the space $[0, T] \times (u, \infty)$ (u=threshold), which form a nonhomogeneous Poisson process with intensity measure:

$$\lambda(t, y) = \frac{1}{\psi}\left(1 + \xi\frac{(y-\mu)}{\psi}\right)_+^{\frac{-1}{\xi}-1}$$

Theory

By standard formulae for a Poisson Process, the likelihood is of the form:

$$\begin{aligned} L(\mu,\psi,\xi) &= \prod_{i=1}^{N} \lambda(T_i,Y_i)\cdot \exp\left\{-\int_0^T\int_u^\infty \lambda(t,y)dtdy\right\} \\ &= \prod_{i=1}^{N} \lambda(T_i,X_i)\cdot \exp\left\{-T\left(1+\xi\frac{u-\mu}{\psi}\right)_+^{-1/\xi}\right\}. \end{aligned} \quad (4)$$

Theory

In practice we work with the negative log likelihood, $\ell = -\log L$, which leads to

$$\begin{aligned} \ell(\mu, \psi, \xi) &= N \log \psi + \left(\frac{1}{\xi} + 1\right) \sum_{i=1}^{N} \log \left(1 + \xi \frac{Y_i - \mu}{\psi}\right)_+ \\ &+ T \left(1 + \xi \frac{u - \mu}{\psi}\right)_+^{-1/\xi} \end{aligned} \tag{5}$$

where T is the length of the observation period in years and the $(...)_+$ symbols essentially mean that the expression are only evaluated if $1 + \xi \frac{u-\mu}{\psi} > 0$ and $1 + \xi \frac{Y_i - \mu}{\psi} > 0$ for each i (if these constraints are violated, L is automatically set to 0).

Theory

- The basic method of estimation is therefore to choose the parameters (μ, ψ, ξ) to maximize (**??**) or equivalently to minimize (5).

  This is performed by numerical nonlinear optimization.

- In practice it is convenient to replace (μ, ψ, ξ) by $(\theta_1, \theta_2, \theta_3)$ where $\theta_1 = \mu,\ \theta_2 = \log \psi,\ \theta_3 = \xi$ (defining θ_2 to be $\log \psi$ rather than ψ itself makes the algorithm more numerically stable, and has the advantage that we don't have to build the constraint $\psi > 0$ explicitly into the optimization procedure).