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

March 22, 2006

Outline

- Motivating Question
- Techniques
- Theory
- Results
- Conclusion and Future Work

Motivating Question

The question under investigation is whether regional climate model return level estimations can be used to obtain return level predictions at the station level.

Outline of Techniques

- Relationship of grid cell data to the n-year return levels at point locations is explored

- Tail of the Generalized Extreme Value distribution (GEV) is fit to the grid cell data above a given threshold
  - Similar to the Peaks Over Threshold method
  - However, here the method used for parameter estimation is a Point Process approach
  - Leads directly to the GEV parameters

Techniques

- GEV Parameter estimates are used to generate n-year return levels
- n-year returns at the point (station) locations are estimated the same way
- Various models are explored to predict point location n-year return from grid cell n-year return

Theory

The Generalized Extreme Value or GEV distribution is defined by the equation

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

In practice, however, fitting (1) directly to sample annual maxima has a number of disadvantages.

- Many of the observational series are short (less than 25 years)
- Most observational series contain missing values - it is unclear how to define the annual maximum when a significant fraction of the daily values are missing.

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

$$\Pr\{X \leq u + x \mid X > u\} \;=\; 1 - \left(1 + \xi\frac{x}{\sigma}\right)_{+}^{-1/\xi}. \qquad (2)$$

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 approach is similar to the peaks over threshold method in that all observations over a given threshold are considered.

- However, in the POT model the parameters are directly tied to the threshold value whereas in the PP approach, the parameters are not tied to the threshold value.

- Here we estimate the tail of the GEV. Instead of conditionally modeling the tail as the GPD does, we are directly modeling the tail values over a given threshold.

- The parameters for the GEV obtained through the point process approach lead directly to the GEV parameters.

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:

$$\begin{aligned} L(\mu,\psi,\xi) &= \prod_{i=1}^{N} \lambda(T_i,Y_i)\cdot \exp\left\{-\int_0^T\int_u^\infty \lambda(t,y)dtdy\right\} \\ &= \prod_{i=1}^{N} \lambda(T_i,X_i)\cdot \exp\left\{-T\left(1+\xi\frac{u-\mu}{\psi}\right)_+^{-1/\xi}\right\}. \end{aligned} \quad (3)$$

Theory

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

$$\begin{aligned} \ell(\mu, \psi, \xi) &= N \log \psi + \left(\frac{1}{\xi} + 1\right) \sum_{i=1}^{N} \log \left(1 + \xi \frac{Y_i - \mu}{\psi}\right)_+ \\ &+ T \left(1 + \xi \frac{u - \mu}{\psi}\right)_+^{-1/\xi} \end{aligned} \tag{4}$$

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 (3) or equivalently to minimize (4).

  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).

Theory

- The PP approach should produce equivalent parameter values as the POT approach, provided the threshold is high enough for the model to fit the data.

- Provided the model fits the data, the parameters are independent of the threshold (adjusting for estimation error). The ideal threshold can be determined by considering where the parameter values stabilize.

- In the GPD approach, the scale parameter still varies with threshold, and is not necessarily the same as the scale parameter of the PP approach.

Theory

The n-year return values can be directly obtained using the estimated GEV parameters obtained from the PP approach. 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:

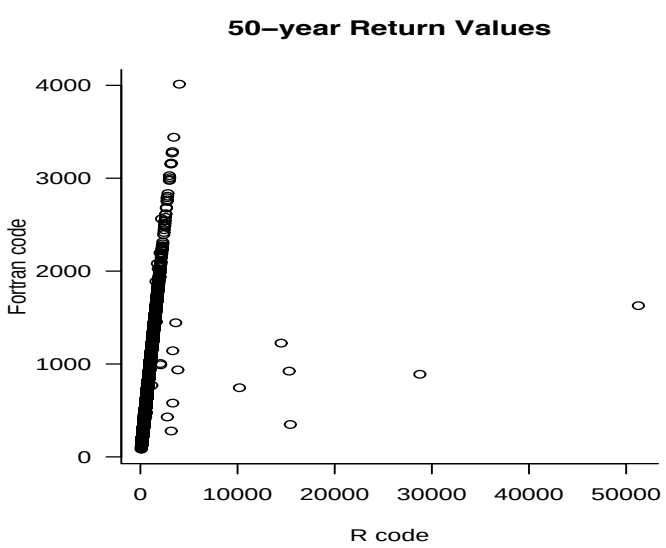
$$y_n = \begin{cases} \mu + \psi \frac{n^{\xi} - 1}{\xi} & \text{if } \xi \neq 0, \\ \mu + \psi \log n & \text{if } \xi = 0. \end{cases} \tag{5}$$

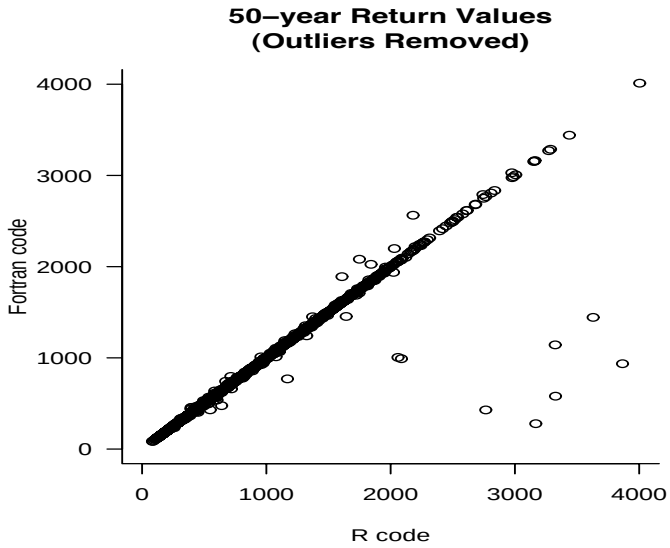
Data

- 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

Results

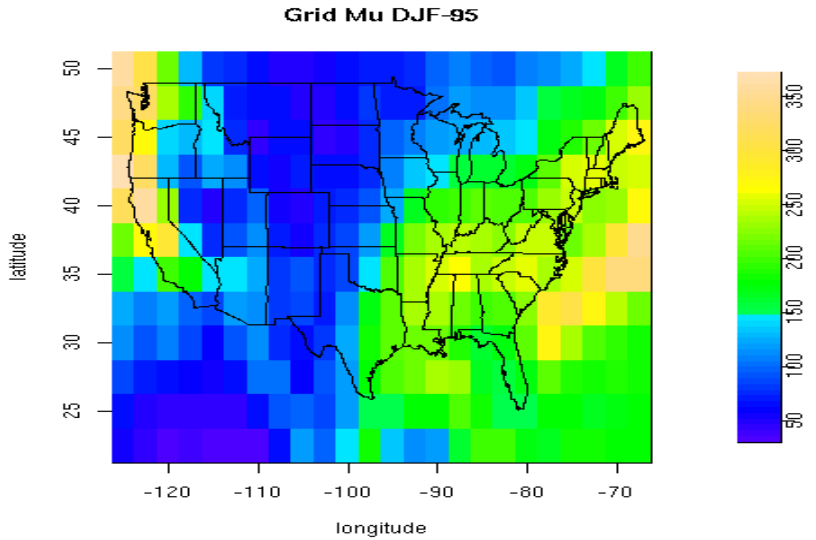
Comparisons of R and Fortran Codes for NCDC Rainfall Data

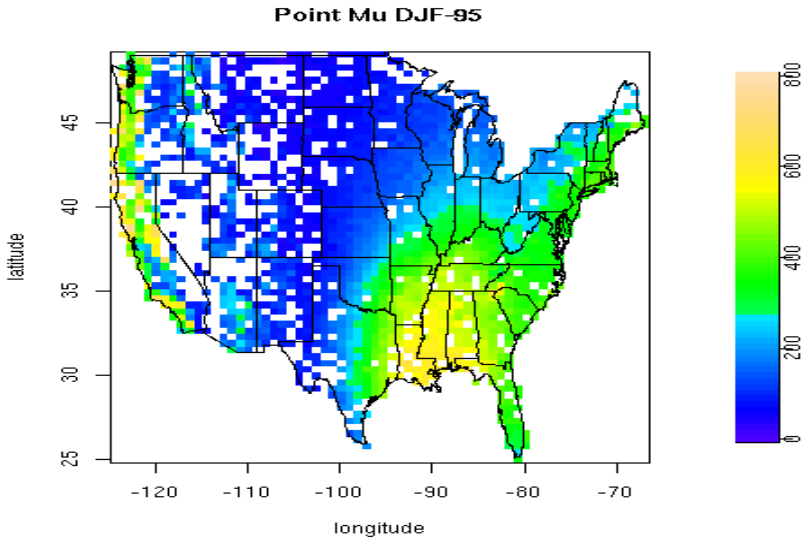


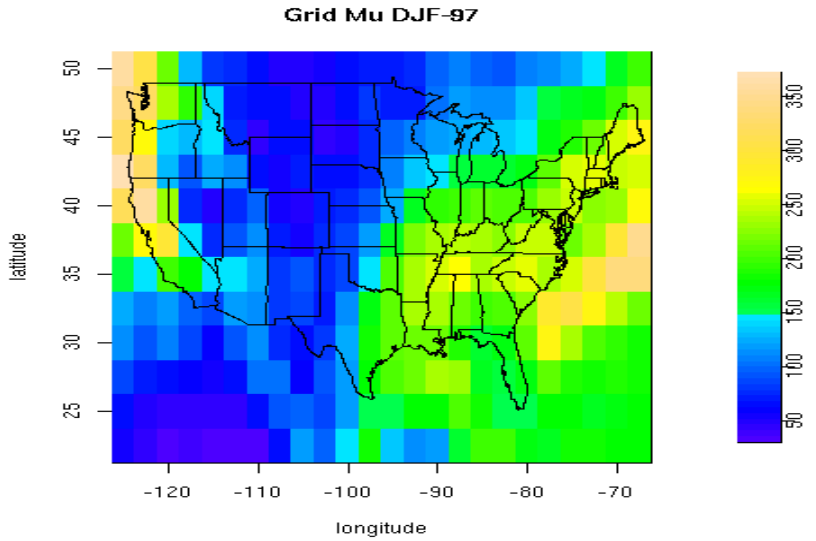


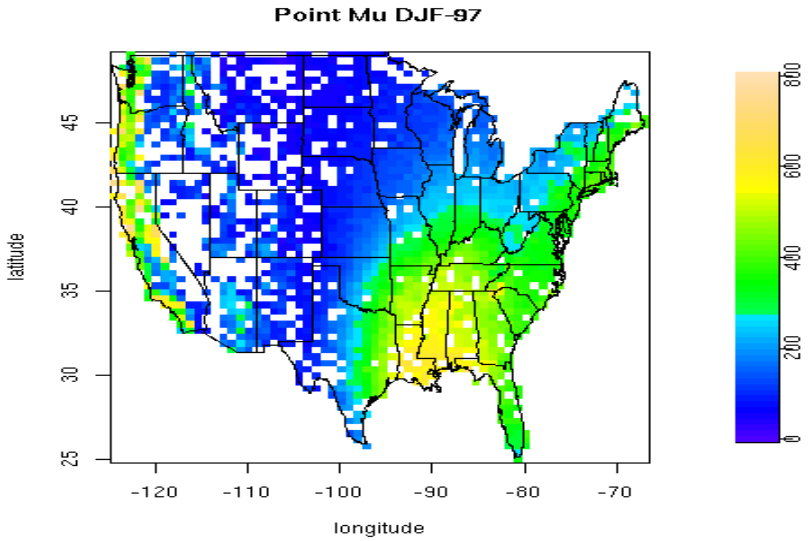
The parameter estimates obtained through Richard Smith's point process algorithm using the Quasi-Newton method are directly comparable to estimates obtained through the $pp.fit$ function, using the Nelder-Mead method, in the R software package $ismev$ for the point process approach.

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

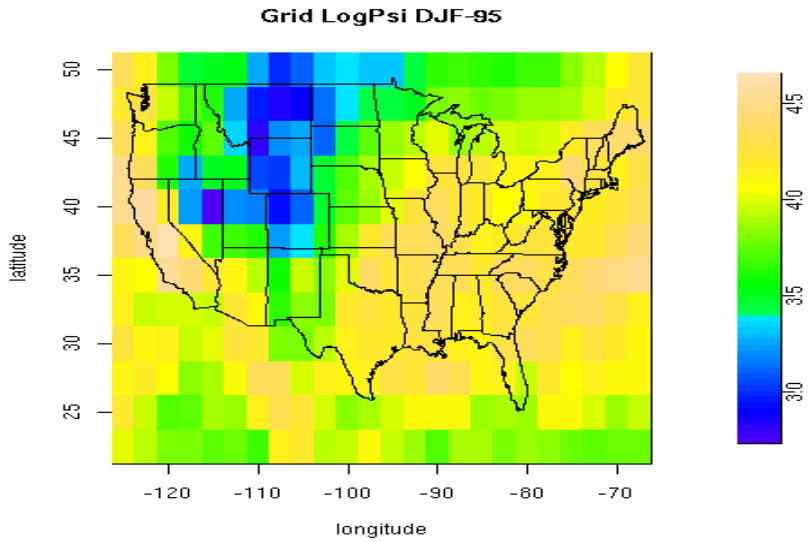


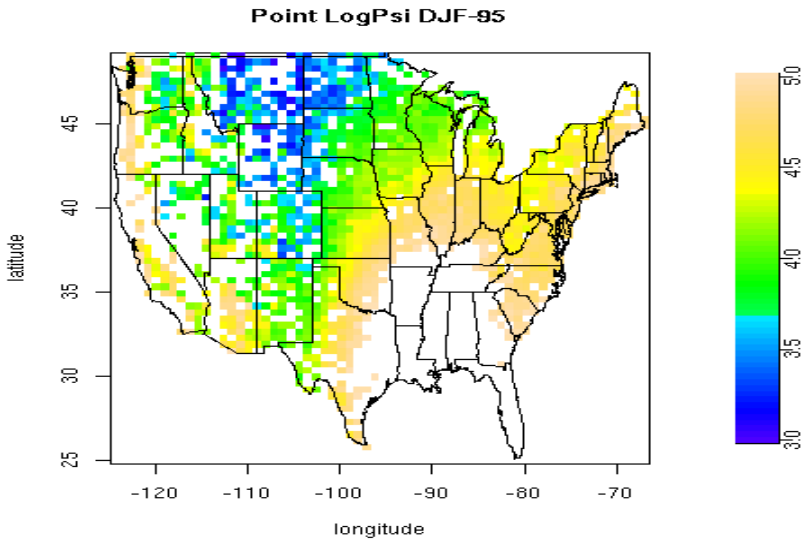


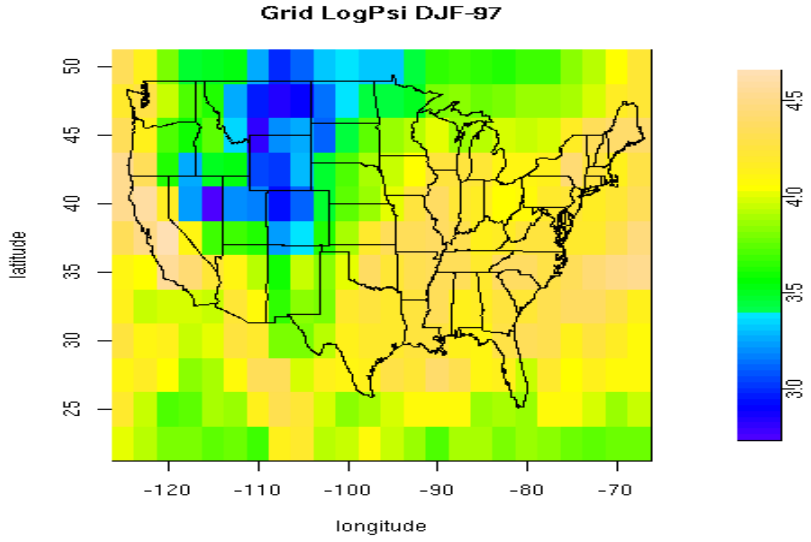


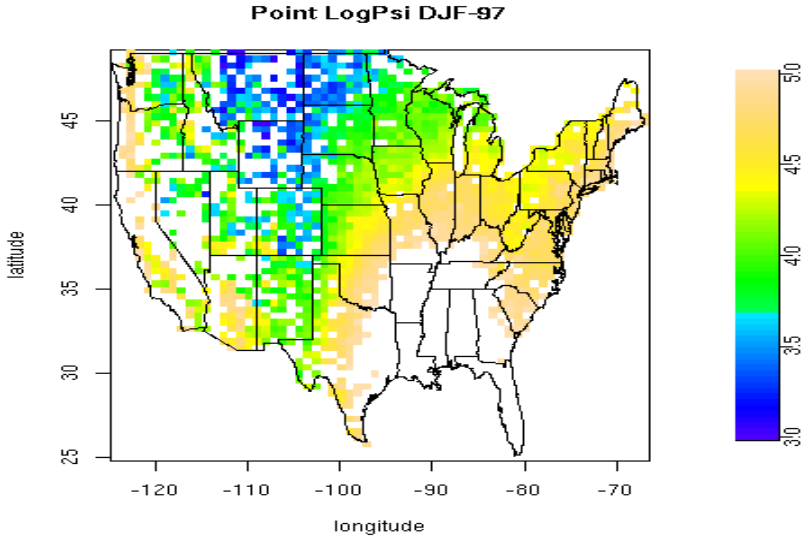


GEV Model Fit - LogPsi Parameter: 95 Grid - Point vs 97 Grid - Point

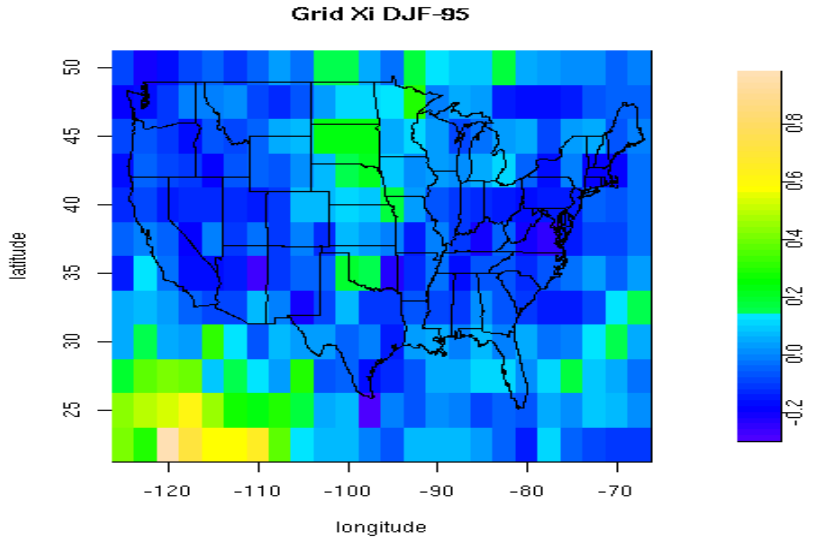


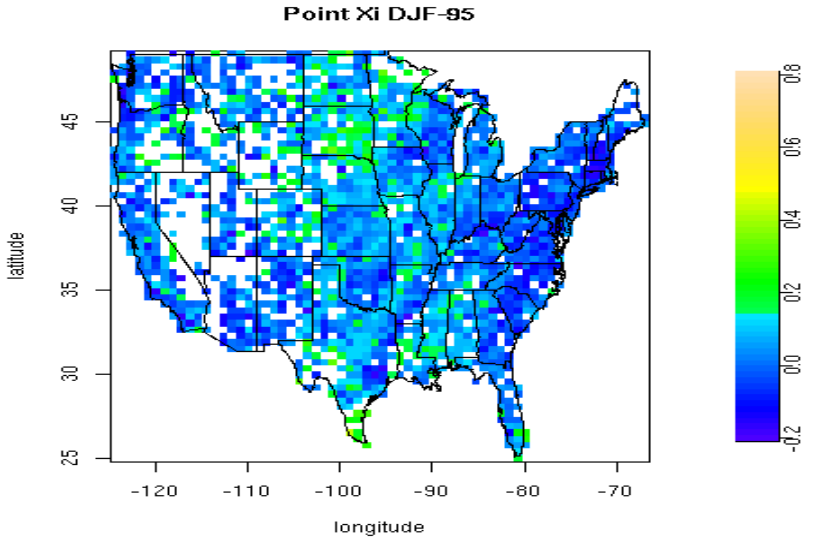


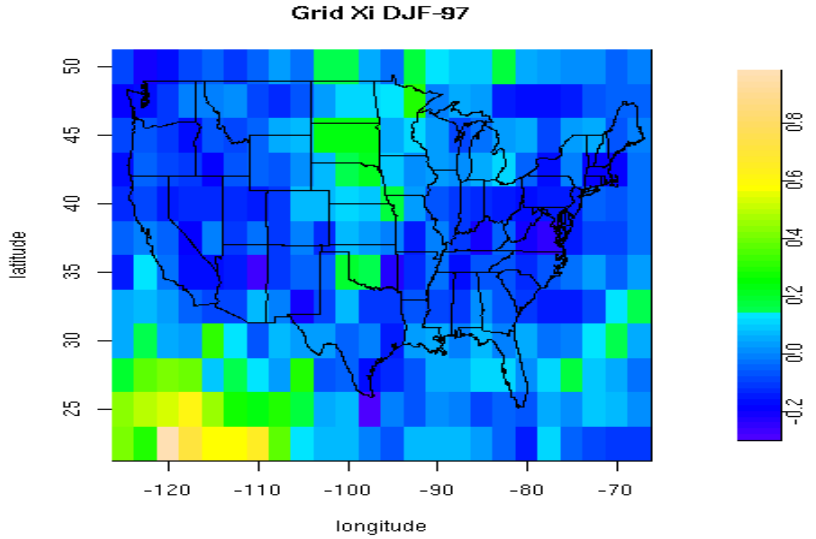


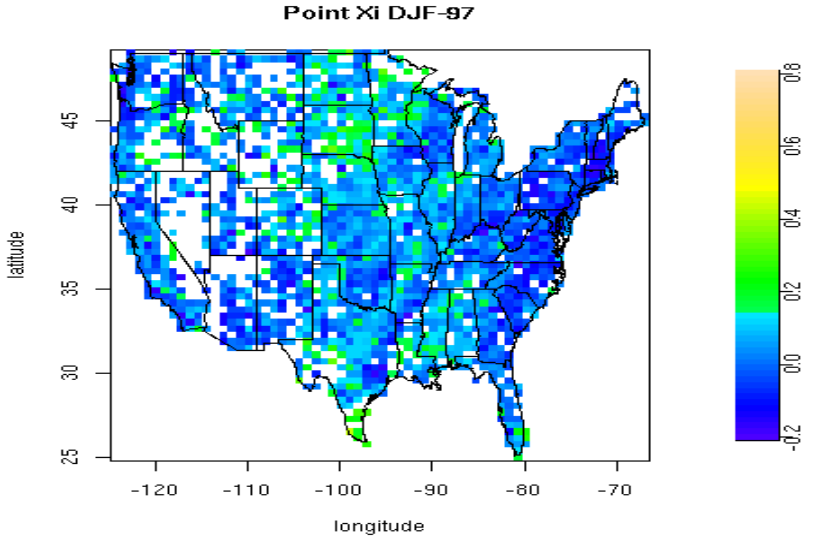


GEV Model Fit - Xi Parameter: 95 Grid - Point vs 97 Grid - Point

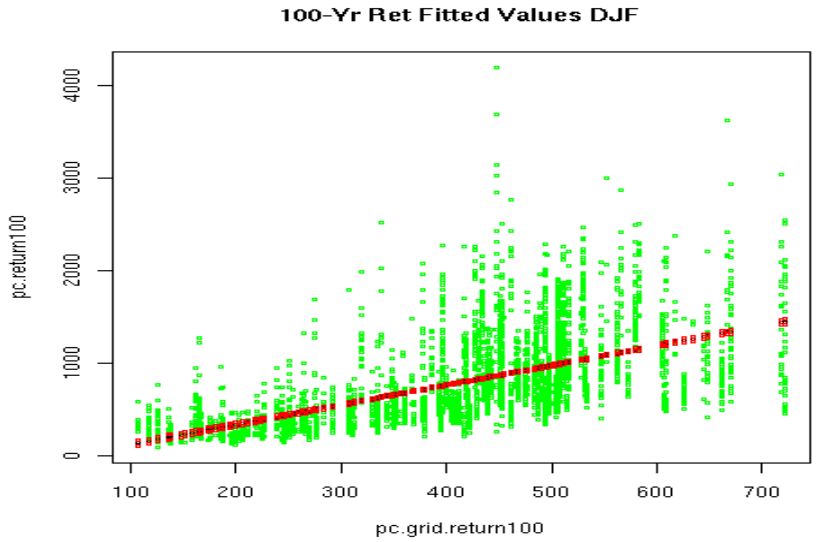


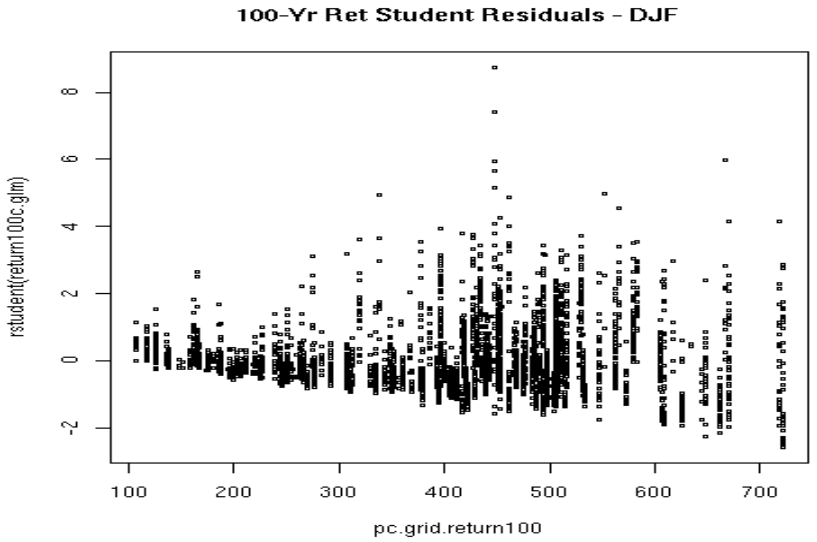


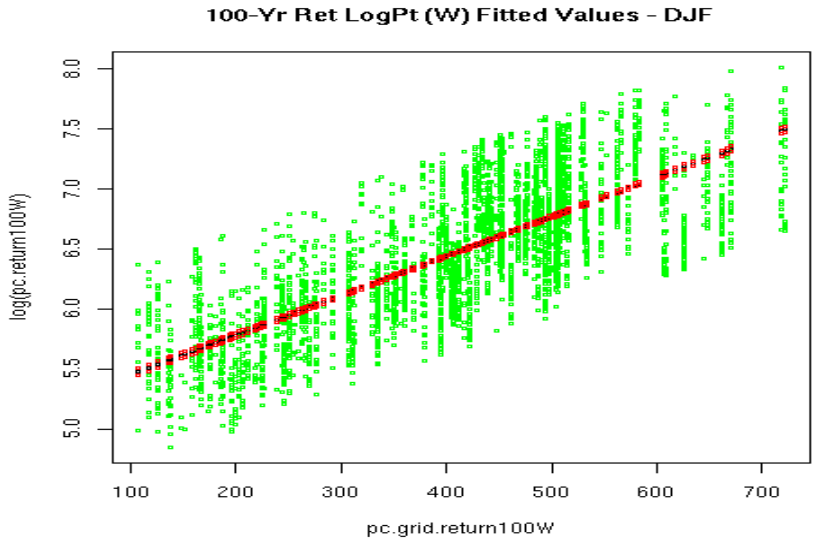


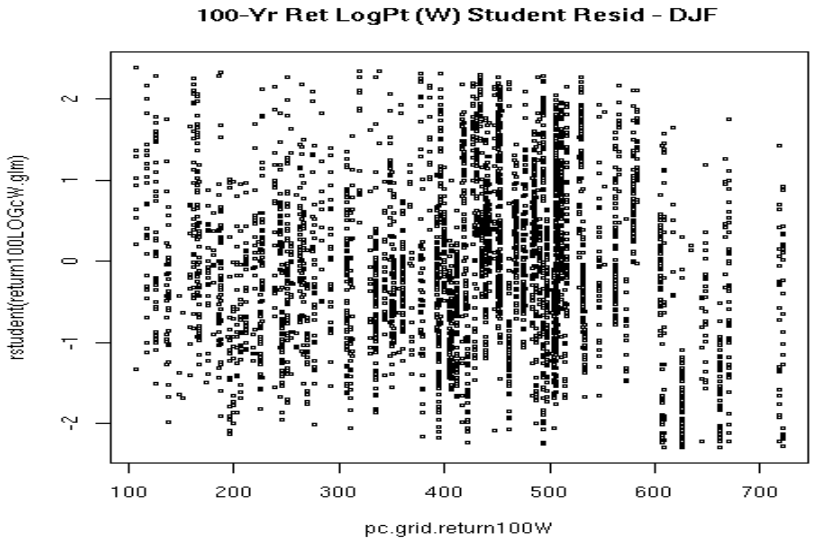


100-Year Return: Point vs Grid - LogPoint vs Grid

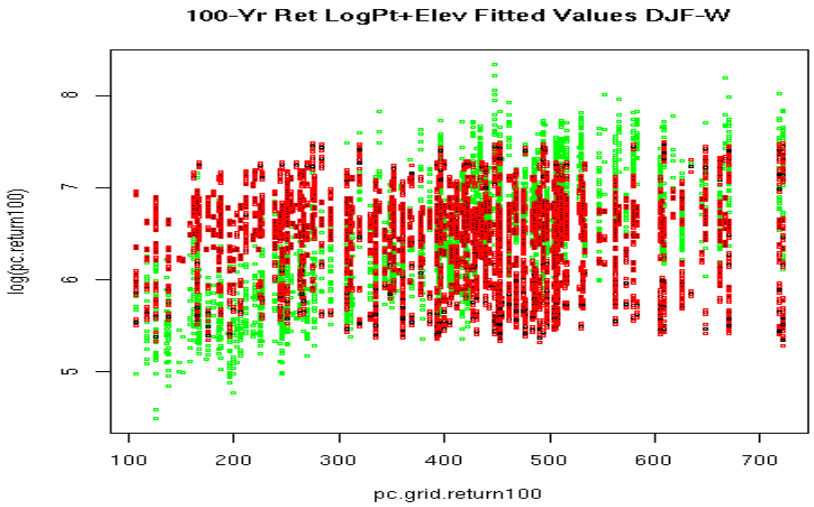


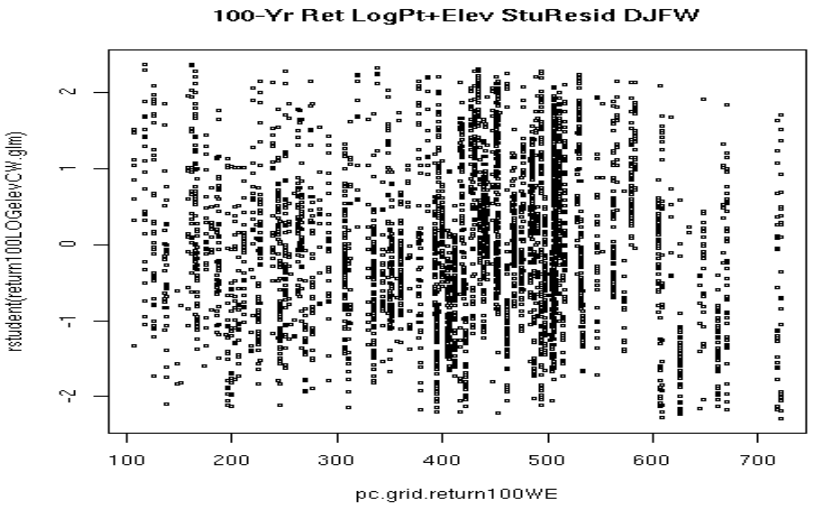




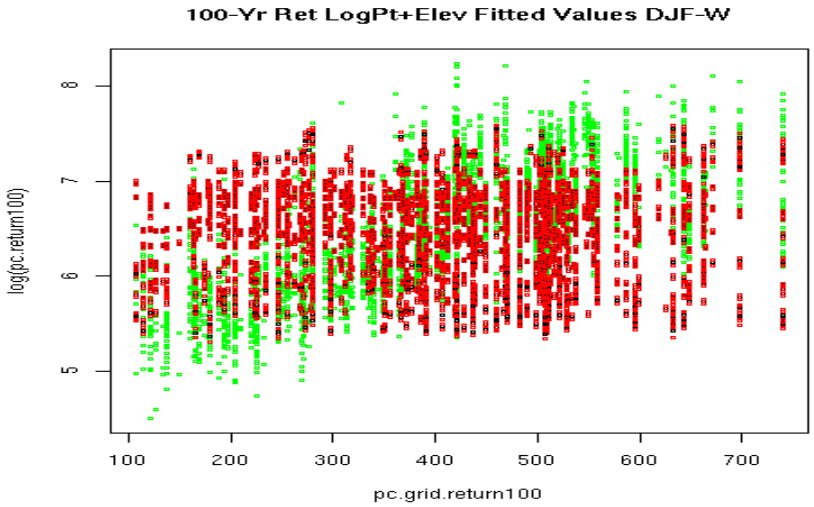


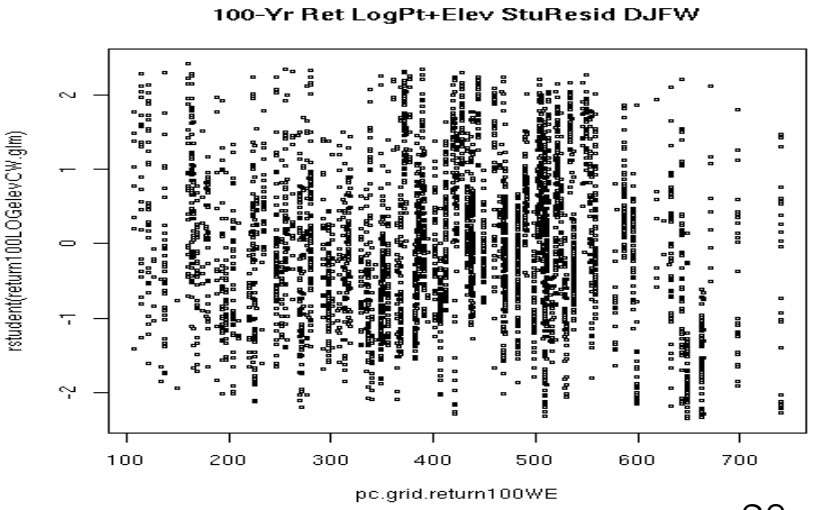
100-Year Return: LogPoint + Elevation vs Grid



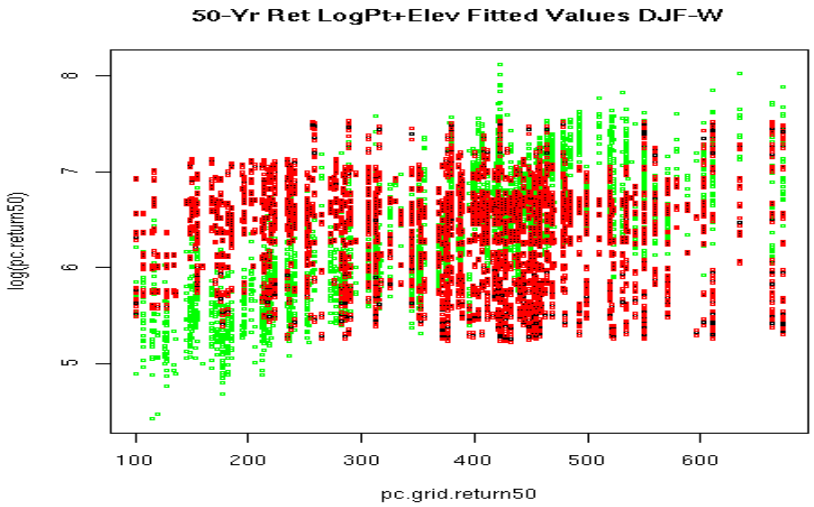


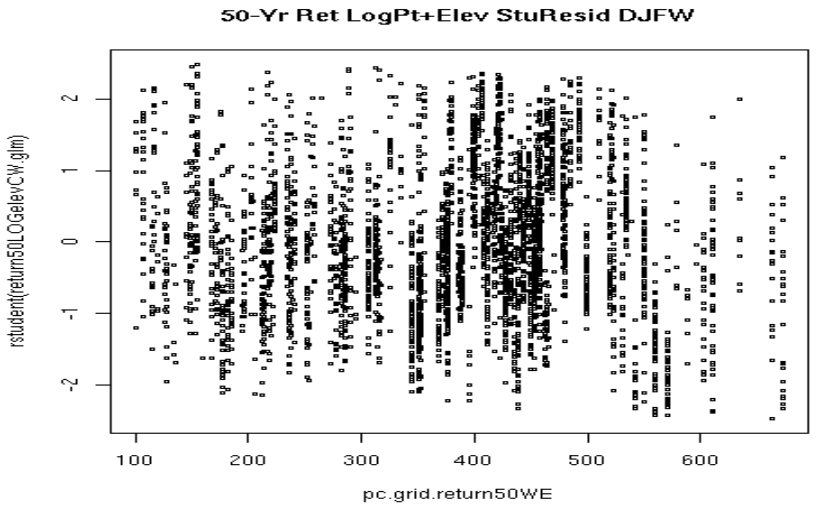
100-Year Return w/ 97% Threshold: LogPoint + Elevation vs Grid



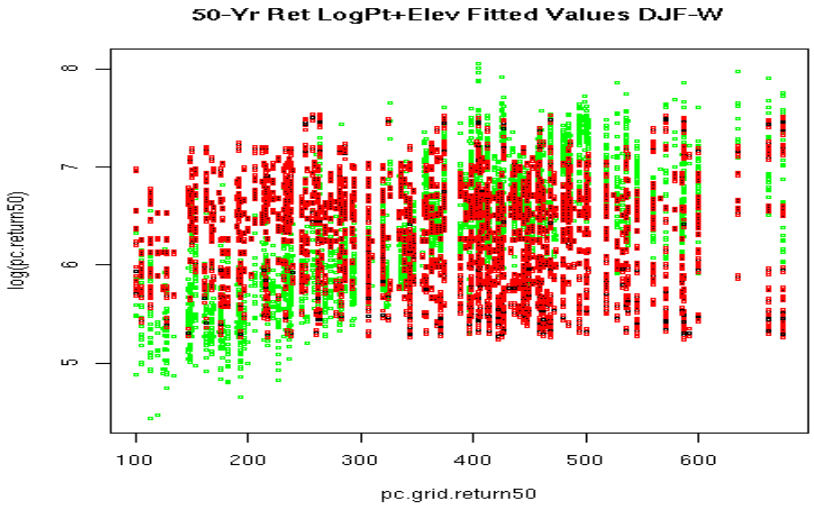


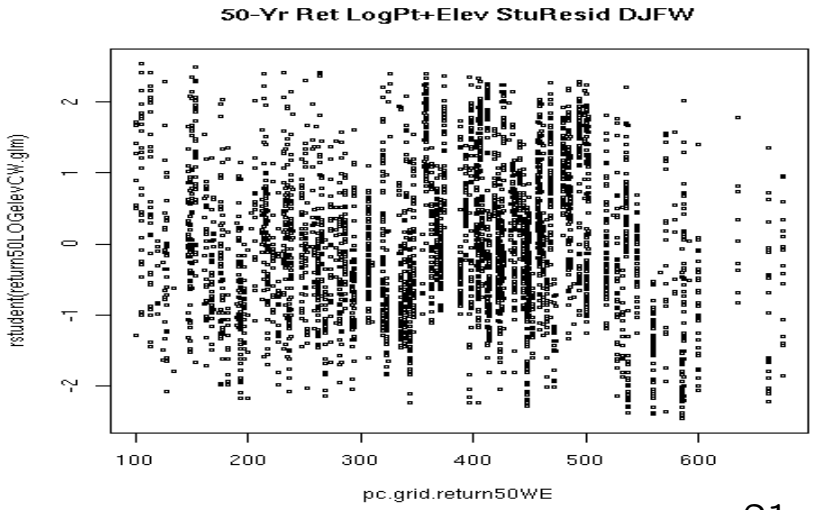
50-Year Return: LogPoint + Elevation vs Grid



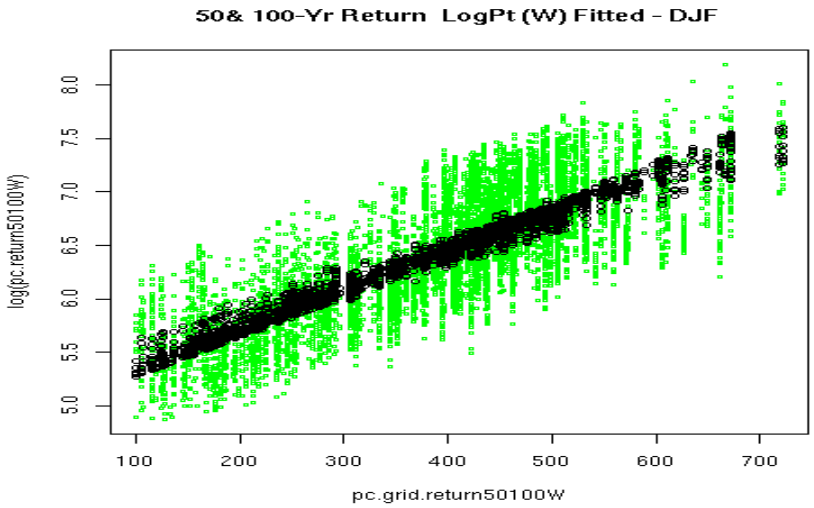


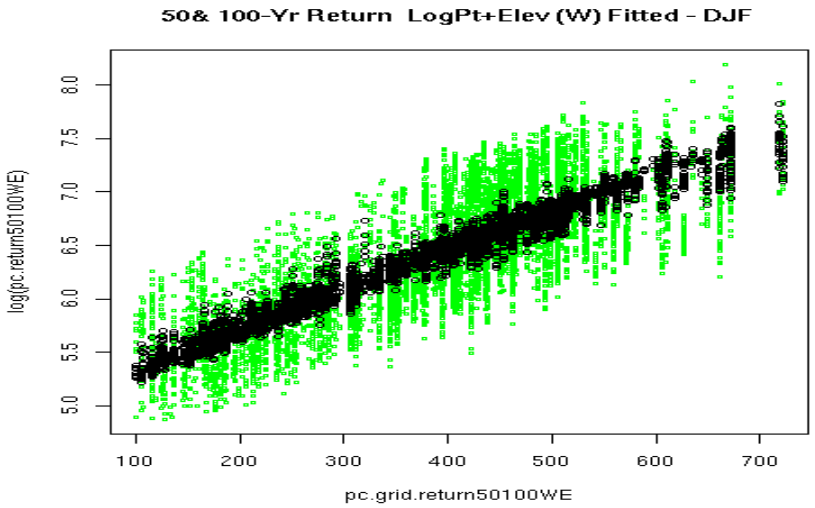
50-Year Return w/ 97% Threshold: LogPoint + Elevation vs Grid



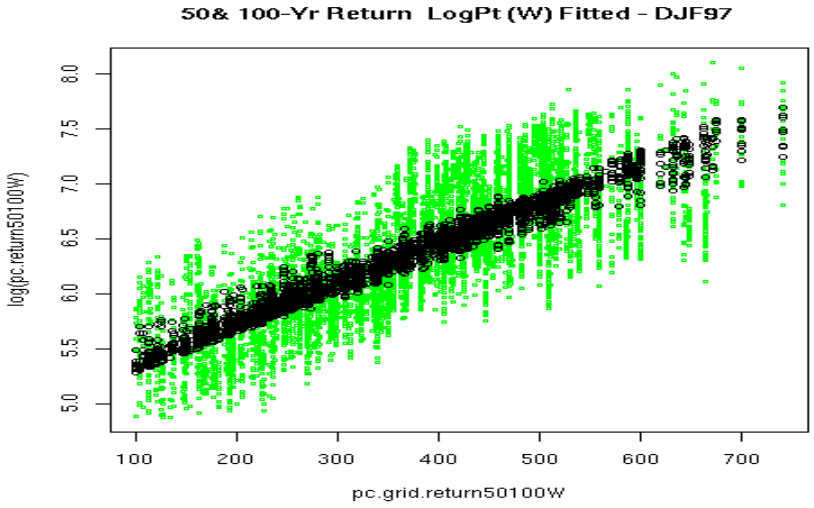


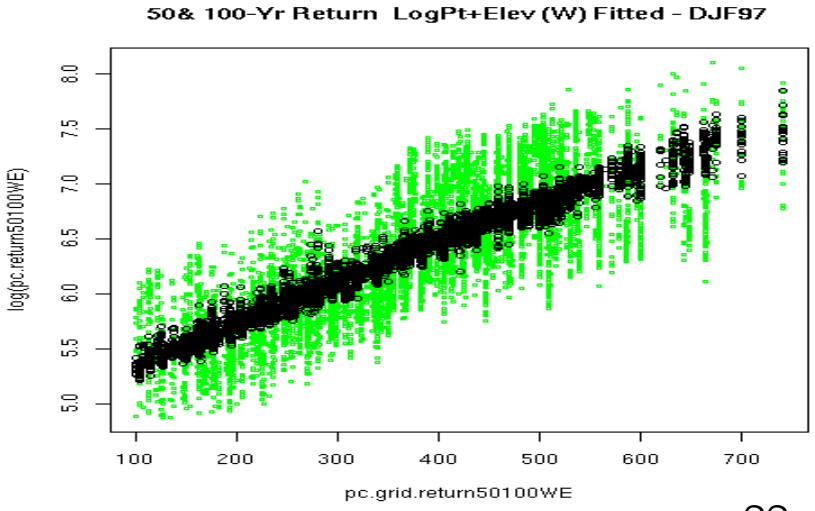
Joint Modeling of 50 and 100-Year Returns: 95% Threshold





Joint Modeling of 50 and 100-Year Returns: 97% Threshold





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
Point Grid	4207	62011	-85.7749	2.1267	
log(Point) Grid	4207	4724	5.1957	0.0032	
log(Pt) Grid+Elev	4207	4681	5.3775	0.0029	-0.000110
W: log(Pt) Grid	4026	3479	5.1250	0.0033	
W: log(Pt) Grid+Elev	4025	3425	5.3186	0.0030	-0.000124
97 W: log(Pt) Grid	4003	3343	5.1253	0.0033	
97 W: log(Pt) Grid+Elev	4013	3356	5.3074	0.0030	-0.000113
SPRING - 100	OBS	AIC	Intercept	X1	X2
Point Grid	4339	64296	-96.0847	2.4299	
log(Point) Grid	4339	3776	5.4950	0.0029	
log(Pt) Grid+Elev	4339	3332	5.9362	0.0022	-0.000253
W: log(Pt) Grid	4132	2324	5.4042	0.0031	
W: log(Pt) Grid+Elev	4138	1924	5.8960	0.0023	-0.000258
97 W: log(Pt) Grid	4150	2761	5.4572	0.0030	
97 W: log(Pt) Grid+Elev	4145	2282	5.9736	0.0021	-0.000281

Conclusion and Future Work

- Modeling the tail of the GEV Distribution appears to produce stable estimates as indicated across seasons and 95% and 97% thresholds

- 100 and 50-year return levels are successfully modeled by season at the point (station) level using grid-level return values and station elevation

- Model coefficients are consistent within seasons across 95% and 97% thresholds

- Future work includes possible consideration of quadratic models

- Future work includes plans to test grid-point models on 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.