Using Extreme Value Models for Analyzing River Flow

by

Lanier Senter

(Under the direction of Lynne Seymour)

Abstract

This thesis studies several statistical issues for extreme value models. Gumbel, Frechet, and Weibull distributions are used to find a generalized extreme value model using different block maxima are investiaged along with point process models and threshold models. An instructive example of water flow is presented and analysis is included to demonstrate these models.

Index words: Gumbel, Frechet, Weibull, Generalized Extreme Value Model, Point Process, Threshold

Using Extreme Value Models for Analyzing River Flow

by

Lanier Senter

B.S., Meredith College, 1998

A Thesis Submitted to the Graduate Faculty
of The University of Georgia in Partial Fulfillment
of the
Requirements for the Degree

Master of Science

Athens, Georgia

2004

Using Extreme Value Models for Analyzing River Flow

by

Lanier Senter

Approved:

Major Professor: Lynne Seymour

Committee: William P. McCormick
Jaxk Reeves

Electronic Version Approved:

Maureen Grasso
Dean of the Graduate School
The University of Georgia
August 2004

Acknowledgements

I would like to convey my full appreciation toward my major professors, Dr. Lynne Seymour and Dr. Bill McCormick, for all the help, knowledge, and guidance throughout the work on my thesis. I would also like to thank Dr. Jaxk Reeves for being on my committee. I would like to thank them all for being wonderful teachers and friends during my time here.

I would like to thank the department and my parents and friends for making my home at the University of Georgia a most enjoyable experience in my life. Without them it would have not been the same.

I would like to also thank all of the students I have met here at Georgia, without the assistance and motivation from all of you ... I could not have done it.

GO DAWGS!!!

Table of Contents

Page

List of Figures

List of Tables

Chapter 1

Introduction

Extreme Value Models are used in a number of different significant areas. Examples taken from Coles, 2001 are as follows.

- Climate related Extreme Values: There are many factors that deal with certain climate conditions, such as rainfall, temperatures, and wind speeds. These can be observed at the extreme maximum or minimum levels, or studied for seasonality. These types of models can be measured and manipulated to predict future estimates, protect the environment, and make risk assessments.

- Financial Series: Financial extremes include closing prices of the Dow Jones Index and comparing dollar exchange rate (ex. UK sterling/US and UK sterling/Canadian). Extreme value models can help provide key market information and make risk assessments on financial markets.

- Environmental Extreme Values: This type of extreme value modeling includes areas such as Annual sea-levels, pollution levels, traffic patterns, and ocean wave modeling. Many of these areas are studied and analyzed for the variations in the extremes to improve the estimation of variation within a certain period or time.

- Some other areas of applications include alloy strength prediction, portfolios used for insurance, memory cell failure, and food science.

The objective of Extreme Value Models is to describe the stochastic behavior of observations that have very large or small levels. For example, the main goal of this work is to discover the presence or absence of trend in the maximum water flow of Peachtree Creek.

Goals of Modeling Extreme Values

The foremost goal of Extreme Value Modeling is to estimate parameters through maximum likelihood, while explaining the measure of uncertainty from sampling variability. After estimating parameters it is useful to assess the model diagnostics and check goodness of fit, by comparing models that include all possibilities of covariates, maximum values, or any other useful information in the analysis.

1.1 OUTLINE

The rest of Chapter 1 will illustrate and provide information about the Peachtree Creek dataset.

The second chapter will concentrate on some basic definitions to assist in explaining and comprehending general remarks about extremes. Straightforward concepts for estimating parameters will be introduced. Estimating model parameters, checking goodness of fit, and introducing covariates to specific extreme value models will also be covered. Also included in this chapter are modeling extremes for stationary and non-stationary processes and modeling extremes via the point process of high level exceedences.

The third chapter will apply the theory discussed in the earlier chapters to the Peachtree Creek data, described below.

1.2 DESCRIPTION OF DATA

The data to be studied were maximum daily water flow recorded from Peachtree Creek, near downtown Atlanta, GA. Observations were recorded starting on June 21, 1958 and ending on September 30, 1998, giving 14,713 observations. The data are shown in Figure 1.

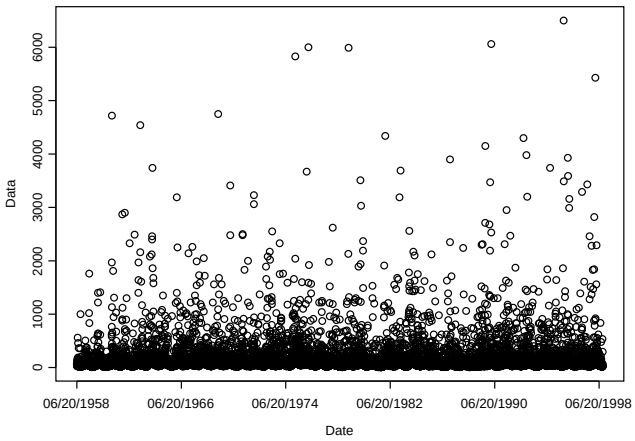

Figure 1.1: Peachtree Creek Waterflow Dataset

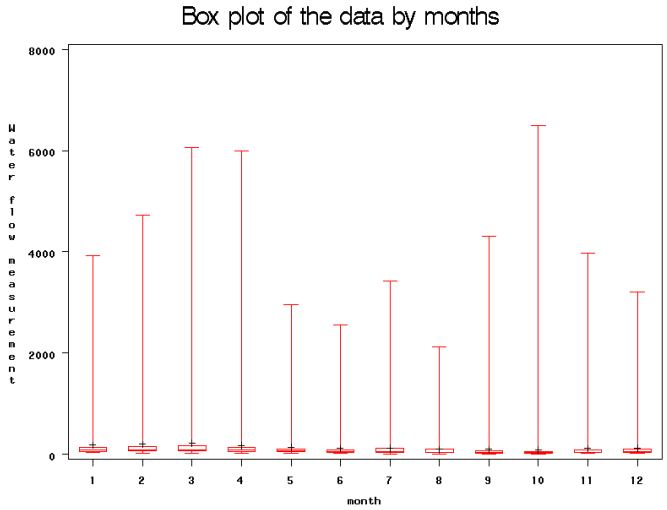

Figure 1.2: Boxplot of Daily Maxima by Month

The huge variability in the data is clear. There is no clear minimum value, but a clear maximum value. A more helpful graph is shown in Figure 1.2, this shows the maxima, minima, median, quartile 1, and quartile 3 for each month. The maximum observation is 6500, which occurred on October 4, 1995. Along with further research, Hurricane Opal's remains came through Atlanta, Georgia on this day.

Chapter 2

EXTREME VALUE MODELS

2.1 BASIC DEFINITONS AND CONCEPTS

The following and hereafter are all definitions, theorems, and ideas taken from Coles (2001).

Extreme value models use a set of observations, x_i, to estimate the probability structure of large or small values of the random variable, X_i, which represents a quantity whose outcome is uncertain.

The foundation and main goal of this analysis is to focus on the statistical behavior of

$$M_n = \max\{X_1, \ldots, X_n\},$$

where $X_1, \ldots, X_n$, is a sequence of independent random variables having a common distribution function F. X_i will usually correspond to values measured on a time-scale, such as hourly, daily, or yearly maxima. In the Peachtree Creek dataset daily maxima are observed, while monthly and annual maximum observations can also be found.

The distribution of M_n can be derived for all values of n:

$$\begin{aligned} Pr\{M_n \leq z\} &= Pr\{X_1 \leq z, \ldots, X_n \leq z\} \\ &= Pr\{X_1 \leq z\} * \ldots * \{Xn \leq z\} \\ &= \{F(z)\}^n \end{aligned}$$

The above formula requires i.i.d. (independent and identically distributed) observations, which is reasonable, for example, when the data represent block maxima such as annual maxima, this is done to make sure there is no correlation between observations, hence giving independence. Usually the distribution function, F, is unknown. Common approaches are

to estimate F by using general statistical techniques or accepting that F is unknown, and examining families of models for F^n, which can only be estimated on the basis of extremes. In the case of M_n based on i.i.d observations $x_1, x_2, \ldots$ with $x_i \sim F$, the behavior of M_n is quite simple. If ω also $= \exp\{x : F(x) < 1\}$, the right endpoint of the distribution function F, then $M_n \to \omega$. The convergence can be in probability or almost sure. To prevent this type of degenerate asymptotic behavior, one introduces the linear normalization. For appropriate sequences $\{a_n\}$ and $\{b_n\}$, one has that $M_n^* = (M_n - b_n)/a_n$ has a nondegenerate limiting distribution. Suitable choices of the sequences a_n and b_n will make the location and scale parameter of M_n^* become stable as n increases. The next theorem will cover any possible limit of the distribution.

Theorem 2.1.1: Let $X_1, \ldots, X_n$ be i.i.d.. If there exist sequences of constants $\{a_n > 0\}$ and $\{b_n\}$ such that

$$Pr\{(M_n - b_n)/a_n \leq x\} \to G(x) \text{ as } n \to \infty$$

where G is a non-degenerate distribution function, then G belongs to one of the following families:

$$\text{Gumbel: } G(x) = \exp\{-\exp[-(\frac{x-b}{a})]\} \quad -\infty < x < \infty \tag{2.1}$$

$$\text{Frechet: } G(x) = \begin{cases} 0, & x \leq b \\ \exp\{-(\frac{x-b}{a})^{-\alpha}\}, & x > b \end{cases} \tag{2.2}$$

$$\text{Weibull: } G(x) = \begin{cases} \exp\{-[-(\frac{x-b}{a})^{\alpha}]\}, & x < b \\ 1, & x \geq b \end{cases} \tag{2.3}$$

Each family has a location parameter, b, and scale parameter, a. The Frechet and Weibull family have a shape parameter α.

The three types of limits that occur in this theorem have distinct behaviors, related to the different tail behaviors of the distribution function F of the X_i. The conditions which ensure that a distribution belongs to one limit law for the maxima depends on the type of tail behavior. For example, the Pareto distribution has a Frechet limit, a uniform distribution has

a Weibull limit, and the Normal distribution has a Gumbel limit. It is possible to combine all three families into one family of models with the next distribution function.

$$G(x) = \exp\{-[1 + \xi(\frac{x-\mu}{\sigma})]^{-1/\xi}\}, \quad 1 + \xi\frac{(x-\mu)}{\sigma} > 0. \tag{2.4}$$

This is the **generalized extreme value** (GEV) family of distributions. This family has three parameters: a location parameter, μ, a scale parameter, σ, and a shape parameter, ξ. Note that all three limit distributions are contained in the GEV family where the Gumbel limit occurs as the limit of the GEV as $\xi \to 0$. This is expressed in the next result.

Theorem 2.1.2: If there exist a sequence of constants $\{a_n > 0\}$ and $\{b_n\}$ such that

$$Pr\{(M_n - b_n)/a_n \leq x\} \to G(x) \quad \text{as} \quad n \to \infty,$$

for a non-degenerate distribution function G, then G is a member to the GEV family

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

defined on $\{z : 1 + \xi(z-\mu)/\sigma > 0\}$, where $-\infty < \mu < \infty, \sigma > 0$ and $-\infty < \xi < \infty$.

The limit in this theorem is used as an approximation for large values of n, which suggests the use of the GEV family for modeling the distribution of maxima of long sequences. For large n,

$$\begin{aligned} Pr\{M_n \leq z\} &\approx G\{(x - b_n)/a_n\} \\ &= G^*(x), \end{aligned} \tag{2.5}$$

where G^* is another member of the GEV family.

We can now use the method of maximum likelihood to estimate the parameters of G^*. The **likelihood function** is given by

$$L(\theta) = \prod_{i=1}^{n} f(X_i, \theta) \tag{2.6}$$

and the **log-likelihood function** is

$$l(\theta) = \log L(\theta) = \sum_{i=1}^{n} \log f(X_i, \theta) \tag{2.7}$$

where $x_1, \ldots, x_n$ are independent realizations of a random variable with probability density function $f(x;\theta)$.

The **maximum likelihood estimator** $\hat{\theta}$ of θ is defined as the value of θ that maximizes $L(\theta)$ or $l(\theta)$.

It is also necessary to look at the approximations of standard errors and confidence intervals to obtain the best estimates.

The matrix $I_E(\theta)$, which measures the expected curvature of the log-likelihood surface, is the **expected information matrix**.

$$I_E(\theta) = \begin{bmatrix} e_{1,1}(\theta) & \cdots & e_{1,d}(\theta) & \\ \vdots & \ddots & e_{i,j}(\theta) & \vdots \\ \vdots & e_{i,j}(\theta) & \ddots & \vdots \\ e_{d,1}(\theta) & \cdots & e_{d,d}(\theta) & \end{bmatrix} e_{i,j}(\theta) = E\{\frac{-\partial^2}{\partial\theta_i \partial\theta_j} l(\theta)\} \tag{2.8}$$

Theorem 2.1.3: Let $x_1, \ldots, x_n$, be independent realizations from a distribution within a parametric family F, and let $l(\cdot)$ and $\hat{\theta}$ denote the log-likelihood function and the MLE of the d-dimensional parameter θ, respectively. Under suitable regularity conditions, the following asymptotic limit law holds. For large n,

$$\hat{\theta} \sim MVN_d(\theta, I_E(\theta)^{-1})$$

where θ denotes the true value of the parameter and $I_E(\theta)^{-1}$ is the inverse matrix. MVN (μ, Σ) denotes a multivariate normal distribution with mean μ, and variance-covariance matrix Σ, where $\mu = E(X)$ and $\Sigma = E(X-\mu)(X-\mu)^T$, and X is a d×1 column vector.

Information from Theorem 2.1.3 can be used to obtain approximate confidence intervals for individual components of $\theta = (\theta_1, \ldots, \theta_d)'$. Let $I_E(\theta)^{-1} = (\phi_{i,j})$. It then follows that

$$\hat{\theta}_i \sim N(\theta_i, \phi_{i,i}),$$

using this and $Z_{\alpha/2}$, which is the $1-(\alpha/2)$ quantile of the standard normal distribution, it can be shown that the approximate $100(1-\alpha)\%$ asymptotic confidence interval for θ_i is

$$\hat{\theta}_i \pm Z_{\alpha/2}\sqrt{\phi_{i,i}}$$

Generally the true value of θ is unknown. It is common to approximate I_E with the **observed information matrix**, $I_o(\theta)$, defined by

$$I_0(\theta) = \begin{bmatrix} \frac{-\partial^2}{\partial\theta_1^2} l(\theta) & \cdots & \frac{-\partial^2}{\partial\theta_1\theta_d} l(\theta) \\ \vdots & \ddots & \vdots \\ \frac{-\partial^2}{\partial\theta_d\theta_1} l(\theta) & \cdots & \frac{-\partial^2}{\partial\theta_d^2} l(\theta) \end{bmatrix} \tag{2.9}$$

evaluated at $\theta = \hat{\theta}_0$, and denoting the elements of the inverse of the matrix by $\phi^*_{i,j}$. It follows from above that an approximate $100(1-\alpha)\%$ confidence interval for θ_i is

$$\hat{\theta}_i \pm Z_{\alpha/2}\sqrt{\phi^*_{i,i}},$$

which is typically more accurate than the previous confidence interval.

A method for quantifying the uncertainty in the MLE is based on the **deviance function**, defined by

$$D(\theta) = 2\{l(\hat{\theta}) - l(\theta)\}, \tag{2.10}$$

where $\hat{\theta}$ is the MLE and $l(\theta)$ is the log-likelihood.

Theorem 2.1.4: Let $x_1, \ldots, x_n$ be independent realizations from a distribution within a parametric family F, and let $\hat{\theta}_0$ denote the MLE for the d-dimensional model parameter θ_0. Then for large n, under suitable regularity conditions, the deviance function $D(\theta) = 2\{l(\hat{\theta}_0) - l(\theta)\}$ satisfies

$$D(\theta_0) \sim \chi^2_d$$

Making important model selection decisions is based on a Chi-square distribution, with d degrees of freedom.

Another intention is to select a model efficiently. Suppose that $M1$ is a model with parameter vector θ, and a model $M0$ is the subset of model $M1$ obtained by constraining k of the components of θ. Let $l_1(M1)$ be the maximized log-likelihood for the model $M1$, and let $l_0(M0)$ be the maximized log-likelihood for the model $M0$, and define

$$D = 2\{l_1(M1) - l_0(M0)\} \tag{2.11}$$

to be the **deviance statistic**.

Theorem 2.1.5: Let $l_0(M0)$ and $l_1(M1)$ be the maximized values of the log-likelihood for models $M0$ and $M1$ respectively. A test of validity of model $M0$ relative to $M1$ at the α level of significance is to reject $M0$ in favor of $M1$ if $D = 2\{l_1(M1) - l_0(M0)\} > c_\alpha$, where c_α is the $(1-\alpha)$ quantile of the χ^2_k distribution.

2.2 DIAGNOSTIC PLOTS

Before assessing if a Generalized Extreme Value distribution should be used, the goodness-of-fit graphical techniques are used.

Probability and Quantile Plots

Given an ordered sample of independent observations

$$x_{(1)} \leq x_{(2)} \leq \ldots \leq x_{(n)}$$

from a population with estimated distribution function $\hat{F}$, a **probability plot** graphs the pairs

$$\{(\hat{F}(x_i), i/(n+1)) : i = 1, \ldots, n\},$$

If $\hat{F}$ is a reasonable model for the population distribution, the points of the probability plot should lie close to the line with slope of 1.

A **quantile plot** graphs

$$\{(\hat{F}^{-1}(i/(n+1)), x_i) : i = 1, \ldots, n\}$$

If $\hat{F}$ is a reasonable estimate of F, then the quantile plot should also have points that lie close to the unit diagonal. The quantities x_i and $\hat{F}$ inverse $(i/(n+1))$ provide estimates of the $i/(n+1)$ quantile of the distribution of F.

Return levels

When data is blocked into sequences of length n (where n is large), it will generate a series of block maxima $M_{n,1}, \ldots, M_{n,m}$, with the GEV distribution fitted. Most often the

case is annual maxima, in which case n is the number of observations in a year. Estimates of extreme quantiles of the annual maxima distribution are obtained by inverting the GEV distribution. The **return level** Z_p associated with the **return period** $1/p$, is given by

$$Z_p = \begin{cases} \mu - \sigma/\xi[1 - \{-\log(1-p)\}^{-\xi}], & \xi \neq 0 \\ \mu - \sigma \log\{-\log(1-p)\}, & \xi = 0 \end{cases} \tag{2.12}$$

where $G(z_p) = (1-p)$.

The level Z_p is expected to be exceeded on average once every $1/p$ years; Z_p is exceeded by the annual maximum in any particular year with probability p.

The relationship of the GEV model to its parameters is interpreted in terms of the quantile expressions from above. By defining $y_p = -\log(1-p)$, it can be shown that

$$Z_p = \begin{cases} \mu - \sigma/\xi[1 - Y_p^{-\xi}], & \xi \neq 0 \\ \mu - \sigma \log Y_p, & \xi = 0 \end{cases} \tag{2.13}$$

Now plotting Z_p against Y_p on a logarithmic scale, or Z_p plotted against $\log Y_p$, will produce 3 different cases. If $\xi = 0$, the plot will be linear; if $\xi < 0$, the plot is convex with asymptotic limit as $p \to 0$ at $\mu - \sigma/\xi$; and if $\xi > 0$ the plot is concave with no finite bound.

This graph is a **return level plot**, which is used for model presentation and validation.

2.3 MODELING EXTREMES FOR STATIONARY DATA

A random sequence $X_1, \ldots, X_n$ is said to be **stationary** if given a set of integers $\{i_1, \ldots, i_k\}$ and any integer m, the joint distributions of $\{X_{i1}, \ldots, X_{ik}\}$ and $\{X_{i1+m}, \ldots, X_{ik+m}\}$ are identical.

Some typical ways of analyzing stationary data are to study the extremal index, cluster sizes, block maxima, and threshold models.

Theorem 2.3.1: Let X_i be a stationary process and X_i^* be a sequence of independent variables with the same marginal distribution. Define $M_n = \max\{X_1, \ldots, X_n\}$ and $M_n^* =$

$\max\{X_1^*, \ldots, X_n^*\}$. Under suitable regularity conditions,

$$Pr\{(M_n^* - b_n)/a_n \leq z\} \to G1(z), \quad n \to \infty$$

$\{a_n > 0\}$ and $\{b_n\}$ are normalizing sequences, where G1 is a non-degenerate distribution function if and only if

$$Pr\{(M_n - b_n)/a_n \leq z\} \to G2(z), \quad n \to \infty,$$

where

$$G2(z) = G1^{\theta}(z)$$

for a constant θ such that $0 < \theta \leq 1$.

The constant θ is called the **extremal index**, which can be used to explain the propensity of the process to cluster at extreme levels. An example of when this might happen is daily temperatures: it would be obvious that streaks of very hot days, or very cold days would have a tendency to occur close to one another. The way to get around this complexity is to decluster the data.

When modeling stationary series it is helpful to look at the model of block maxima or by suggesting a threshold.

The block method is used to reduce the data to maxima over successive blocks.

$$M_{n,1} = \max_{(1 \leq i \leq n)} x_i; \ldots; M_{n,k} = \max_{(1+(k-1)n \leq i \leq kn)} x_i$$

Modeling block maxima, the stationary dependence in the data can be ignored, so it can be treated as if the data were independent. This result is also included in the Generalized Extreme Value family. Maximum likelihood estimators are used to estimate the parameters.

Declustering corresponds to a filtering of the dependent observations to obtain a set of threshold excesses that are approximately independent.

2.4 MODELING EXTREMES IN NON-STATIONARY SEQUENCES

Non-stationary data change systematically over time, such as data containing seasons or trends. When dealing with non-stationary data it is useful to incorporate a non-identically

distributed assumption to the model of independent data. This concept is suitable for blocked data, such as annual maximum, as an alternative to the possibly dependent original data.

Let $X_1, X_2, \ldots$ be annual maxima where the X_i's are independent but not identically distributed and $X_i \sim GEV(\mu_i, \sigma_i, \xi_i)$, $i = 1, 2, \ldots$ where GEV(μ, σ, ξ) denotes the generalized extreme value distribution with parameters (μ, σ, ξ). Estimating μ and σ is useful; the shape parameter, ξ, is usually unrealistic to try to model, and is normally assumed to be $\xi(t) = \xi_o$.

$$G(x) = \exp\{-[1 - \xi(\frac{X - \mu}{\sigma})]^{-1/\xi}\}. \tag{2.14}$$

It is often helpful to look for trends by introducing covariates in the μ and σ parameters. There are many ways this could be modeled; some models can have a trend over time, such as

$$\mu(t) = \begin{cases} \beta_0 + \beta_1 t & \text{linear} \\ \beta_0 + \beta_1 t + \beta_2 t^2 & \text{quadratic} \\ \beta_0 + \beta_1 e^{\beta_2 t} & \text{exponential} \end{cases} \tag{2.15}$$

To ensure that the variance parameter, σ, is always positive for all values of t, the exponential equation is often used

$$\sigma(t) = \exp(\beta_0 + \beta_1(t)). \tag{2.16}$$

Extreme Value parameters can be written in the form

$$\theta(t) = h(X^T \beta)$$

where θ may be μ, σ, or ξ, h is a specified function, β is the vector of unknown parameters that are being estimated, and X is a design matrix.

Using the example models above it follows that linear and quadratic trends in μ are

$$\mu(t) = [1, t] \begin{bmatrix} \beta_0 \\ \beta_1 \end{bmatrix} \quad \mu(t) = [1, t, t^2] \begin{bmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \end{bmatrix} \tag{2.17}$$

or, using a covariate,

$$\mu(t) = [1, \text{ covariate}(t)] \begin{bmatrix} \beta_0 \\ \beta_1 \end{bmatrix} \tag{2.18}$$

Estimating Parameters

Maximum likelihood parameter estimation is well adapted to changes in the model structure.

From above, letting β be the complete vector of parameters, the likelihood is

$$L(\beta) = \prod_{t=1}^{m} g(Z_t; \mu(t); \sigma(t); \xi) \tag{2.19}$$

where g denotes the density of the GEV distribution and $x_1 = z_1$, $x_2 = z_2, \ldots, x_n = z_n$ are the observed data values.

If the ξ parameter never equals zero then the log likelihood is

$$\begin{aligned} l(\mu, \sigma, \xi) &= -\sum_{i=j}^{m} \{\log \sigma(t) + (1 + \frac{1}{\xi(t)}) \log[1 + \xi(t)(\frac{Z_t - \mu(t)}{\sigma(t)})] \\ &\quad + [1 + \xi(t)(\frac{Z_t - \mu(t)}{\sigma(t)})]^{-1/\xi(t)}\} \end{aligned} \tag{2.20}$$

such that for $t = 1, \ldots, m$

$$1 + \xi(t)(\frac{Z_t - \mu(t)}{\sigma(t)}) > 0. \tag{2.21}$$

Model Selection

Selecting the appropriate model is very important; it is possible to have numerous different models. The main concept is to choose a simple model that explains a great deal of the variation in the data. The deviance statistic uses Maximum likelihood of nested models to check model sufficiency. The deviance statistic, as noted before, is

$$D = 2\{l_1(M1) - l_0(M0)\}, \tag{2.22}$$

where $l_1(M1)$ and $l_0(M0)$ are the maximized log-likelihoods under models $M1$ and $M0$. This tests the hypothesis

$$H0 : M0$$

$$H1 : M1$$

D is asymptotically χ^2_k where k is the number of extra parameters in $M1$ than in $M0$. Large values of D indicate that $M1$ explains the variance in the data significantly better than the

reduced model $M0$. Small values of D indicate that it is permissible to reduce from the more complicated model $M1$ to the simpler model $M0$.

Model Diagnostics

As discussed previously it is valuable to look at the probability and quantile plots. With a range of possible models, suppose we accept the model $X_t \sim GEV(\mu(t), \sigma(t), \xi) = G(t)$. Then Y_t is defined as

$$\tilde{Y}_t = \frac{1}{\xi} \log[1 + \xi(\frac{X_t - \mu(t)}{\sigma(t)})] \tag{2.23}$$

with probability distribution function

$$PR(\tilde{Y}_t \leq Z) = e^{-e^{-Z}} \tag{2.24}$$

which is the equation used to make the pairs of the probability and quantile plots, by using y_t as the observed values, and denoting $y_t(1), \ldots, y_t(m)$ as the ordered values. The probability and quantile pairs, respectively, are

$$\left\{ \frac{i}{(m+1)},\ e^{-e^{-\tilde{y}(i)}} \right\} \quad i = 1, \ldots, m \tag{2.25}$$

$$\{\tilde{y}(i),\ \ -\log(-\log(\frac{i}{m+1}))\} \quad i = 1, \ldots, m. \tag{2.26}$$

2.5 EXTREME VALUE MODELING WITH POINT PROCESSES

Point processes are useful for modeling because they are based on exceedances above high levels and consequently incorporate the entire data set.

Theorem 2.5.1: If there exist a sequence of constants $\{a_n > 0\}$ and $\{b_n\}$ such that

$$Pr\{(M_n - b_n)/a_n \leq x\} \rightarrow G(x) \ \text{ as } \ n \rightarrow \infty$$

for a non-degenerate distribution function G, then G is a member of the GEV family

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

where $[1 + \xi(z - \mu)/\sigma] > 0$, $-\infty < \mu < \infty$, $\sigma > 0$ and $-\infty < \xi < \infty$.

Now assuming that n is large, an approximate maximum is:

$$Pr(M_n \leq u) \approx G^*(u) = G((u - b_n)/a_n) \tag{2.27}$$

for large u, G^* is also a member of the GEV family.

Using the mechanics of the above theorem we can now let $G^* = GEV(\mu, \sigma, \xi)$ where the parameters correspond to the parameters for the distribution of annual maxima. Letting $n = n_o$ and $G = G^*$, where n_o corresponds to the number of observations in one year, typically 365 or 366, the following can be calculated.

$$\begin{aligned} Pr\{M_{n_o} \leq u\} &= P^{n_o}\{X_1 > u\} \\ &= \exp\{n_o \ln(1 - P\{X_1 > u\})\}, \end{aligned} \tag{2.28}$$

where M_{n_o} is the annual max.

By Taylor series, $\ln(1 - x) = -x$, approximately, for small x, so

$$\begin{aligned} &= \exp\{n_o \ln(1 - P\{X_1 > u\})\}, \\ &\approx exp\{-n_o P(X_1 > u)\} \end{aligned}$$

but,

$$\begin{aligned} Pr\{M_{n_o} = u\} &\approx G(u) \\ G(u) &= \exp\{-[1 + \xi(\frac{u - \mu}{\sigma})]^{-1/\xi}\}. \end{aligned}$$

We now have two approximations, which implies that the two exponents in the above equations are roughly equal, so

$$p = Pr\{X_1 > u\} \approx^* \frac{1}{n_0}[1 + x_i(\frac{u - \mu}{\sigma})]^{-1/\xi} \tag{2.29}$$

where $X_1 > u$ is the daily observation above a certain level u, which must be substantially large for the approximation to be of good quality.

Also notice that while p is independent of n_o, the right hand side must also be approximately independent of n_o, because the parameters μ and σ, and how large μ is, depends on n_o.

Now, for large u

$$n_o p = n_o P\{X_1 > u\} \approx [1 + \xi(\frac{u-\mu}{\sigma})]^{-1/\xi} \tag{2.30}$$

and,

$$N_{n_0}(u) = \sum_{K=1}^{n_0} I[X_k > u] \tag{2.31}$$

which is the number of exceedances of level u in 1 year. Then by the Poisson approximation to the Binomial distribution,

$$N_{n_0}(u) \approx \text{ Poisson } (\lambda) \tag{2.32}$$

where $\lambda = [1 + \xi(\frac{u-\mu}{\sigma})]^{-1/\xi}$.

Another way of obtaining the Poisson distribution is by looking at the number of exceedances of u over an interval $[t_{i-1}n_o, t_i n_o]$.

Looking at the interval $[sn_o, tn_o]$ we find the mean of the Poisson distribution by the Poisson approximation to the Binomial.

$$\begin{aligned}
\text{Mean} &= np \\
&= [(t-s)n_o + 1]p \\
&\approx \frac{(t-s)n_0 + 1}{n_0}[1 + \xi(\frac{u-\mu}{\sigma})]^{-1/\xi} \\
&\approx (t-s)[1 + \xi(\frac{u-\mu}{\sigma})]^{-1/\xi} \\
&= (t-s)\lambda \\
\lambda &= [1 + \xi(\frac{u-\mu}{\sigma})]^{-1/\xi}.
\end{aligned} \tag{2.33}$$

If the intervals do not overlap, then the variables $N_u([s_1, t_1])$ and $N_u([s_2, t_2])$ are independent. We now have

$$N_u([s,t]) = \sum_{K=Sn_0}^{tn_0} I[X_k > u] \sim \text{ Poisson } [(t-s)\lambda]. \tag{2.34}$$

The above discussion describes the parts of a **Poisson point process**.

A **Poisson point process,** N, on an interval $[0, T]$ of real numbers is a stochastic process with index sets A which is a subset of $[0, T]$ such that

1. $N(A) \sim$ Poisson $(|A|\lambda)$, where $|A|$ = length of A

2. $N(A)$ and $N(B)$ are independent random variables if $A \cap B = \acute{\oslash}$

3. $N(A \cup B) = N(A) + N(B)$ if $A \cap B = \acute{\oslash}$

Such a process is called a **homogeneous Poisson process** with **intensity parameter**, λ, where λ is the rate an event occurs.

A **non-homogeneous Poisson Process,** N, satisfies the above definition, but

$$N(A) \sim \text{Poisson } (\int_A \lambda(t)dt).$$

The function $\lambda(t)$ is called the **intensity function** and $\Lambda(A) = (\int_A \lambda(t)dt) = E[N(A)]$ is called the **mean measure** or the **intensity measure**.

Let

$$N_{n_0}(A) = \sum_{K:\frac{K}{n_0} \in A} I[X_k > u], \quad A \subset [0,1]. \tag{2.35}$$

Then in the case of i.i.d. random variables $\{x_k, 1 \leq k \leq n_o\}$, N_{n_o} is approximately a homogeneous Poisson process with intensity

$$\lambda = \{1 + \xi[\frac{u-\mu}{\sigma}]\}^{-1/\xi} \tag{2.36}$$

where (μ, σ, ξ) correspond to the GEV parameters for the annual maxima.

Statistical Inference

In an interval $[0, T]$ which will be our observation interval for a Poisson point process, N, suppose that points are observed at $x_1, \ldots, x_m$.

To obtain a likelihood, consider small intervals at the locations x_i say $I_i = [x_i, x_i + \delta_i]$ where δ_i is very small, and consider one event in each interval and no points outside interval.

$$P(N(I_1) = 1, N(I_2) = 1, \ldots, N(I_m) = 1, N([0,T] : UI_i) = 0)$$

by independence

$$\prod_{j=1}^{m} P(N(I_j) = 1) P(N([0,T] \backslash \bigcup_{j=1}^{m} I_j = 0) \tag{2.37}$$

since the $|I_j| = \delta_j$ are small, it leads to

$$\approx \prod_{j=1}^{m} P(N(I_j) = 1) P(N[0,T] = 0) \tag{2.38}$$

by the Poisson distribution

$$\begin{aligned} &\approx \prod_{j=1}^{m} (\int_{I_j} \lambda(t)dt) \exp(-\int_0^T \lambda(t)dt) \\ &\times \prod_{j=1}^{m} \lambda(X_j)\delta_j \exp(-\int_0^T \lambda(t)dt). \end{aligned} \tag{2.39}$$

Dividing by $\prod_{j=1}^{m} \delta_j$ and letting δ_j go to zero, we obtain

$$L(\theta; X_1, \ldots, X_m) = \prod_{j=1}^{m} \lambda(X_i; \theta) \exp(-\int_0^t \lambda(t; \theta)dt) \tag{2.40}$$

where the intensity function $\lambda(t) = \lambda(t; \theta)$ depends on an unknown parameter θ.

Finding the MLE in the homogeneous case is as follows by substituting in

$$\begin{aligned} \lambda(t;\theta) &\equiv \lambda \text{ in } L(\theta; X_1, \ldots, X_m) \\ &= L(\lambda; X_1, \ldots, X_m) = \lambda^m e^{-t\lambda} \\ &\Rightarrow l(\lambda) = m \ln \lambda - T\lambda \\ 0 &= \frac{\partial l(\theta)}{\partial \lambda} = \frac{m}{\lambda} - T \Rightarrow \hat{\lambda} = \frac{m}{T}. \end{aligned} \tag{2.41}$$

As a result, the MLE is the rate of points occurring in the window $[0, T]$.

Handling m-year data

Consider the point process as before where (μ, σ, ξ) are GEV parameters for annual maximum. We now consider this for the maxima over m-years, by the following

$$N_{n_0}(m)([s,t] \times [Z, \infty]) = \sum_{K=sm_{n_0}}^{tm_{n_0}} I[X_k \geq Z] \tag{2.42}$$

with $0 \leq s < t \leq 1$, and z large. Also note

$$\begin{aligned} &\Lambda([s,t] \times [Z, \infty]) \\ \approx\ &-\ln P(N_{n_0}^{(m)}([s,t] \times [Z, \infty]) = 0) \end{aligned}$$

$$
\begin{aligned}
&= -\ln P\{\max_{sm_{n_0}\le k\le tm_{n_0}} X_k < Z\} \\
&= -\ln P^{(t-s)mn_0}\{X_1 \le Z\} = -(t-s)m\ln P\{Mn_0 < Z\} \\
&\approx m(t-s)[1+\xi(\frac{Z-\mu}{\sigma})]^{-1/\xi}. \qquad (2.43)
\end{aligned}
$$

Chapter 3

APPLICATION OF THEORY

This chapter is concerned with the Peachtree Creek flow data, and will be analyzed using S-PLUS functions for extreme value modeling by Stuart Coles. The first section presented the entire dataset; as mentioned it is very useful to look at the annual maxima. The annual maxima are shown in Figure 3.1.

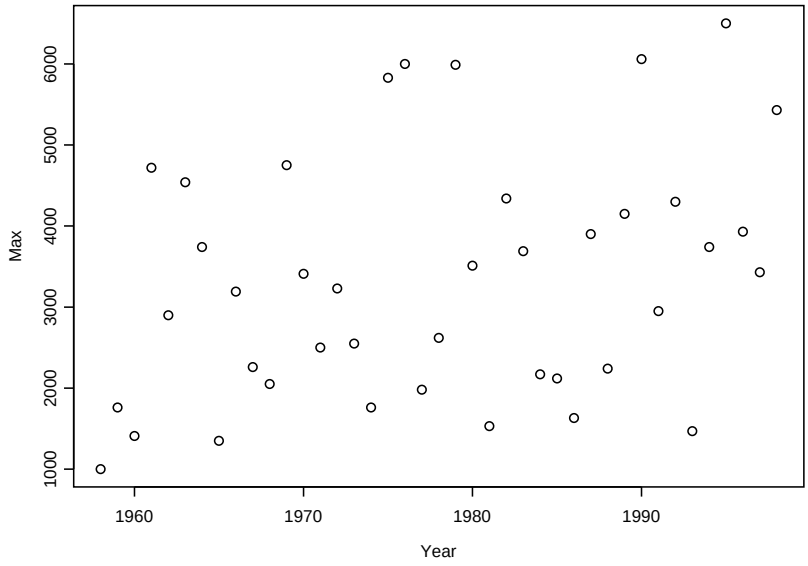

Figure 3.1: Graph of Annual Maxima Plotted Against Time

Sometimes it is also helpful to look at the annual minimas, which are plotted in Figure 3.2 below.

3.1 Generalized Extreme Value Models for Annual Maxima

The next step is to try and fit a GEV model and find maximum likelihood estimators for the parameters.

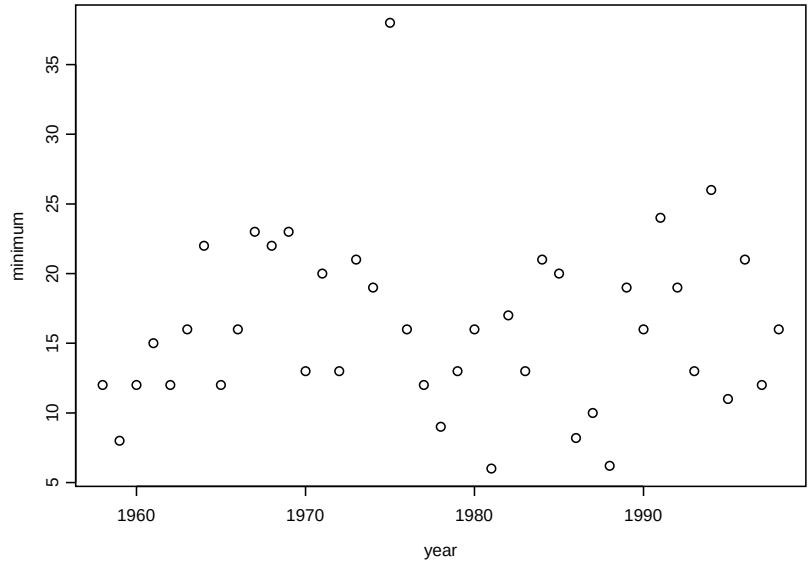

Figure 3.2: Graph of Annual Minima Plotted Against Time

Fitting the GEV model to the annual maxima the negative log-likelihood is **355.9351**. The MLEs and the standard errors for the parameters are given in Table 3.1.

Table 3.1: Results of GEV Model with Annual Maximas

Parameter	MLE	Standard Error
μ	2670.8326	235.9649
σ	1249.1856	179.5865
ξ	-0.0640	0.1731

These estimates give the GEV model of:

$$G(x) = \exp\{-[1 + (-.0640)(\frac{X - 2670.83}{1249.19})]^{-1/(-.0640)}\}. \tag{3.1}$$

Looking at the diagnostic plots in Figure 3.3, one is able to check the quality of model fit.

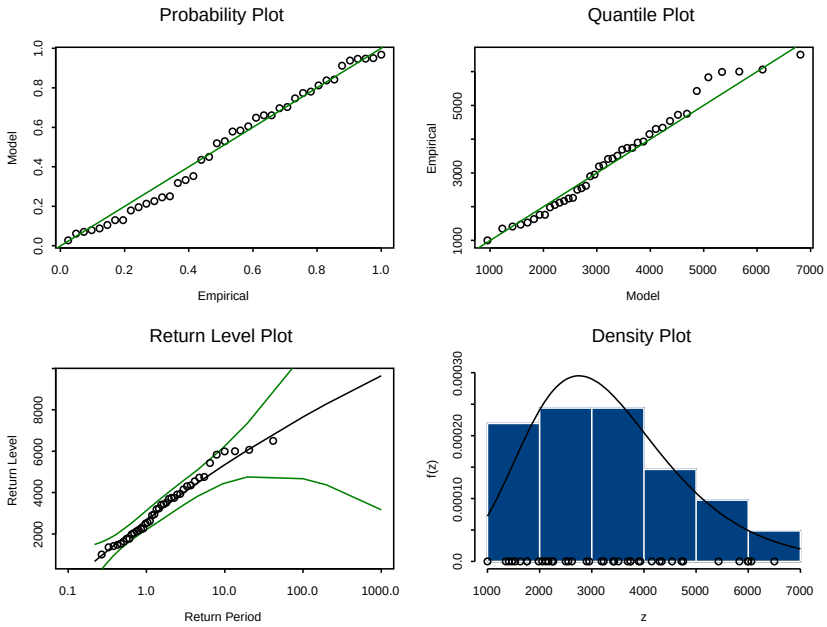

Figure 3.3: Probability, Quantile, Return Level, and Density Plots of Annual Maximum

The Probability and Quantile plots appear to follow the unit diagonal rather well. The return level plot satisfies the representation of empirical estimates.

It is also practical to introduce a covariate to test for a trend. Looking at the data it seems to follow a linear trend. The years in the data vary from 1958 to 1998, so it is useful to rescale the year variable so that it can be centered around a simple range.

$$\text{Year}^* = (\text{Year} - 1978)/20, \tag{3.2}$$

which indexes the year on a scale of $[-1, 1]$. To allow for linear time trend in the location parameter μ, the parameter must be modeled as a linear function of time

$$\mu(t) = \beta_0 + \beta_1(t). \tag{3.3}$$

Fitting this GEV model the negative log-likelihood is **353.9169**. The MLEs and standard errors for the parameters are given in Table 3.2.

Table 3.2: Results of GEV Model with Annual Maximums and Linear Time Trend

Parameter	MLE	Standard Error
$\mu : \beta_0$	2707.2251	224.7376
β_1	659.1208	337.6346
σ	1193.8939	170.1023
ξ	-0.0707	0.1706

These estimates give the GEV model of :

$$G(x) = \exp\{-[1 + (-.0707)(\frac{X - [2707.23 + 659.12(\text{Year})^*]}{1193.894})^{\frac{-1}{.0707}}. \tag{3.4}$$

The diagnostic plots for GEV with covariate year* are the following.

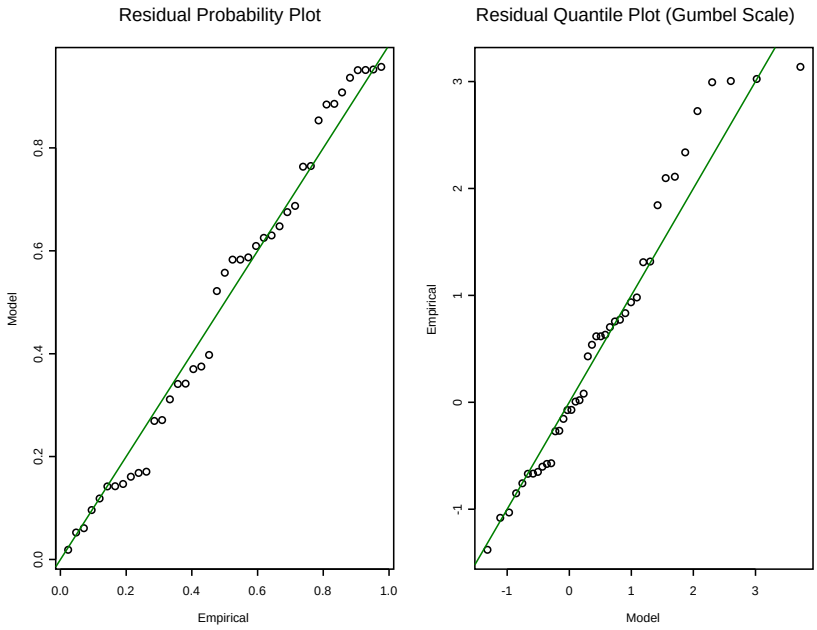

Figure 3.4: Residual Probability and Residual Quantile Plots with Covariate Year*

The probability and quantile plots still seem to follow the unit diagonal moderately well.

To check which model is better, we find the deviance statistic $D = 2\{l_1(M1) - l_0(M0)\}$.

$$\begin{aligned} D &= 2\{355.925 - 353.9169\} \\ &= 4.01 \end{aligned} \tag{3.5}$$

Since the equation is only adding one variable to help explain the model this Deviance statistic follows a Chi-square distribution with 1 degree of freedom.

$$P(\chi_1^2 > 4.01)P - \text{value} = .045231. \tag{3.6}$$

This indicates that there is a linear time trend in the data. In both of the above GEV models, the shape parameter ξ shows evidence of being zero. Using this information it is useful to look at the Gumbel model, which does not include ξ.

Fitting the Gumbel model to the annual maxima the negative log-likelihood is **355.9904**. The MLEs and standard errors for the parameters are given in Table 3.3:

Table 3.3: Results of Gumbel model with Annual Maxima

Parameter	MLE	Standard Error
μ	2628.479	200.5442
σ	1217.317	151.3739

This leads to the Gumbel model:

$$G(x) = \exp\left\{-\exp\left[-\left(\frac{X - 2628.479}{1217.317}\right)\right]\right\}.$$

Using the covariate Year*, as above, to test for a linear trend along with the annual maxima, the Gumbel model gives a negative log-likelihood is **353.9989**. The MLEs and standard errors of the parameters are given in Table 3.4.

This leads to the Gumbel model:

$$G(x) = \exp\left\{-\exp\left[-\left(\frac{X - (2662.404 + 618.316(\text{Year}^*))}{1160.900}\right)\right]\right\}.$$

Table 3.4: Results of Gumbel model with Annual Maxima and covariate Year*

Parameter	MLE	Standard Error
$\mu : \beta_0$	2662.4074	191.2631
β_1	618.316	307.1478
σ	1160.900	144.0741

To check which model is better, we find the deviance statistic $-D = 2\{l_1(M1) - l_0(M0)\}$.

$$\begin{aligned} D &= 2\{355.9904 - 353.9989\} \\ &= 3.983 \end{aligned}$$

Since the equation is only adding one variable to help explain the model this Deviance statistic follows a Chi-square distribution with 1 degree of freedom.

$$P(\chi_1^2 > 3.983) P - \text{value} = 0.45962.$$

This indicates there is a linear time trend in the data, which agrees with the previous results.

3.2 Generalized Extreme Value Model for Annual Minima

Doing the same tests for the annual minimums, the results are as follows.

GEV model negative log-likelihood is **129.8342**. The MLEs and the standard errors for the parameters are given in Table 3.5:

Table 3.5: Results of GEV Model with Annual Minimas

Parameter	MLE	Standard Error
μ	13.4589	0.8772
σ	5.0538	0.6238
ξ	-0.0475	0.0997

This estimates give the GEV model of:

$$G(x) = \exp\{-[1 + (-.0475)(\frac{X - 13.46}{5.054})]^{-1/-.0475}\}. \tag{3.7}$$

Looking at the diagnostic plots in Figure 3.5 one can check the quality of model fit.

In the diagnostic plots, the points seem to follow somewhat of a step function around the unit diagonals.

When introducing the covariate Year* (same as above) to check for a linear trend, the negative log-likelihood is **129.8326**. The MLEs and standard errors for the parameters are given in Tabler 3.6:

Table 3.6: Results of GEV Model with Annual Minima and Linear Time Trend

Parameter	MLE	Standard Error
$\mu : \beta 0$	13.4564	0.8783
$\beta 1$	-0.07599	1.359
σ	5.0507	0.6261
ξ	-0.0468	0.1010

These estimates give the GEV model of:

$$G(x) = \exp\{-[1 + (-.0468)(X - (\frac{[13.456 - .07599(\text{Year}^*)]}{5.0507})]^{\frac{1}{0.0468}}. \tag{3.8}$$

The diagnostic plots are in Figure 3.5.

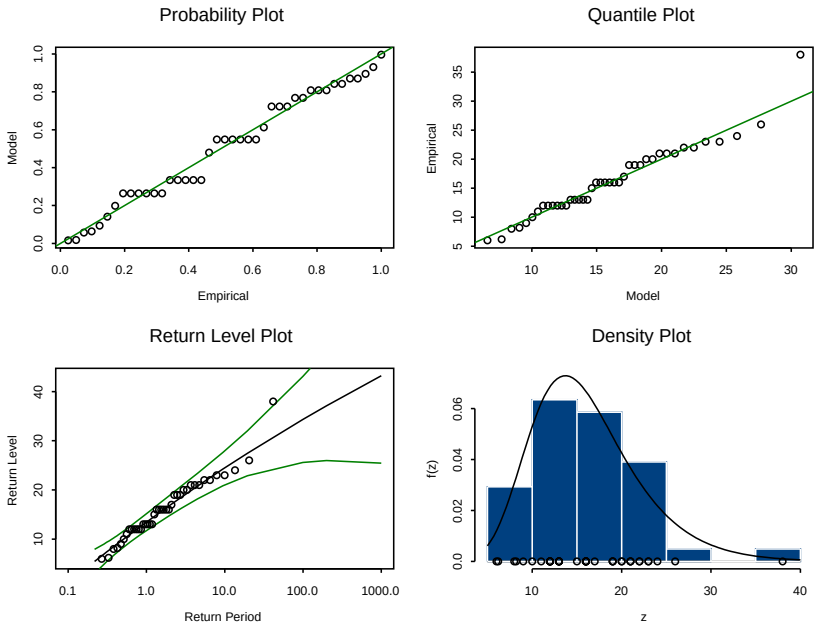

Figure 3.5: Probability, Quantile, Return Level and Density Plots for Annual Minimums

Selecting the better model by doing a deviance statistic leads to the conclusion that the original GEV is better, because the Chi-square test gives a P-value of .96809. You can also look at the parameter β_1 and see that there is not much significance in the new model.

3.3 Point Process Model with Daily Maxima

Now, looking at the entire dataset we can find MLEs and standard errors for the point process model. We can use model parameters together with variable thresholds.

Below is a graph that plots the entire dataset along with a threshold u, which equals 1470. Also plotted on this graph is another threshold which contains a periodic function of period one year, which introduces seasonality. This has the equation:

$$a + b\sin(\frac{2\pi(x-d)}{365.25}). \tag{3.9}$$

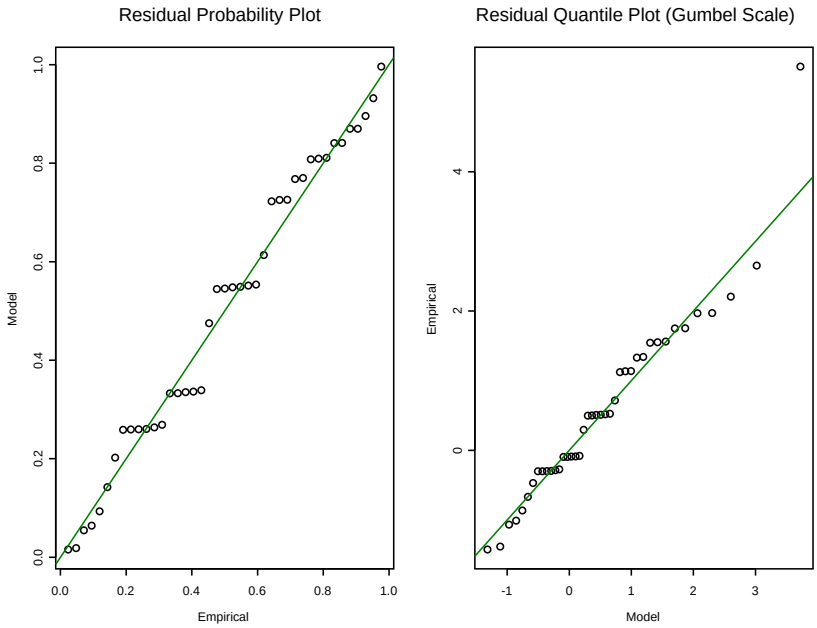

Figure 3.6: Probability and Quantile Plots for Annual Minima w/ Covariate Year*

The goal of this example is to strive for approximately 10% of the data to fall above the threshold, which is done by simple trial and error of the variables a, b, d, and u. We find $a = 2000$, $b = 1200$, $d = 50$. The period is over 365.25 days to ensure that leap years are accounted for.

The following results for the point process model are as follows.

The threshold of **u=1470** gives a negative log-likelihood of **1092.215**, the number of exceedances is **143**. This data set has approximately 14713 observations.

The parameters and standard errors are given in Table 3.5:

These estimates give the following model:

$$G(x) = \{\exp[1 + (.10045)(\frac{x - 2686.99}{1023.46})]^{-1/.10045}\}. \tag{3.10}$$

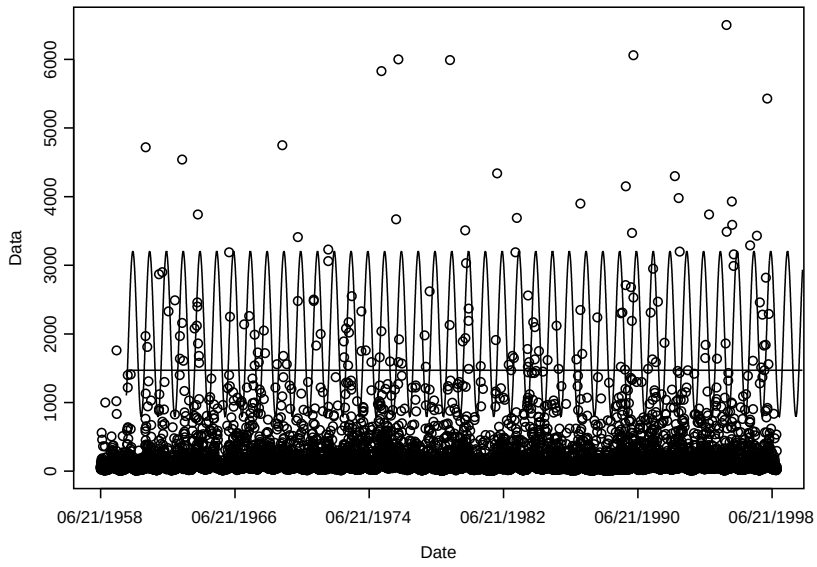

Figure 3.7: Graph of Entire Dataset with a Threshold u, and a Seasonality Threshold

Two periodic covariates are introduced to help explain the model and show seasonality effects in the location parameter μ. The two covariates are

$$\sin(\frac{2\pi x}{365.25}), \quad \cos(\frac{2\pi x}{365.25}).$$

The negative log-likelihood is **1090.764** and the number of exceedances is **143**. The parameters and standard errors are given in Table 3.8:

The parameters are of the form:

$$\begin{aligned} \mu(t) &= 2596.983 + (-256.573)\sin(\frac{2\pi x}{12}) + 14.525\cos(\frac{2\pi x}{12}) \\ \sigma(t) &= 6.8514 \\ \xi(t) &= -.2259. \end{aligned}$$

Using a deviance statistic to pick the best model

$$D = 2\{1092.215 - 1090.764\}$$

Table 3.7: Results of Point Process Model with Threshold $u = 1470$

Parameter	MLE	Standard Error
μ	2686.9912	145.6682
σ	1023.4605	101.2105
ξ	0.10045	0.1081

Table 3.8: Results of Point Process Model with Threshold u = 1470 and Periodic Covariates.

Parameter	MLE	Standard Error
$\mu : \beta_0$	2596.9829	144.9903
β_1	-256.5731	180.5731
β_3	14.5247	109.7353
σ	6.8514	0.1073
ξ	-0.2259	0.1116

$$= \quad 2.902$$

The number of parameters in the model first is 3 and the number of parameters in the second model is 5, which gives the deviance statistic a Chi-Square distribution with 2 degrees of freedom.

$$P(\chi_2^2 > 2.902) P - \text{value} = .23434.$$

This indicates that the original model is better and there is no effect of seasonality.

The threshold of $u = a + b\sin(\frac{2\pi(x-a)}{365.25})$ gives a negative log-likelihood of **1108.197**, and the number of exceedances is **144**, which is approximately 10% of the data.

The parameters and standard errors are given in Table 3.9.

This leads to the GEV model:

$$G(x) = \{\exp[1 + (.0443)(\frac{x - 2583.7061}{1160.986})^{-1/.0443}\}.$$

Table 3.9: Results of Point Process model with threshold $u = a + b\sin(\frac{2\pi(x-d)}{365.25})$

Parameter	MLE	Standard Error
μ	2583.7061	163.2399
σ	1160.9859	107.9677
ξ	0.0443	0.09836

Two periodic covariates are introduced to help explain the model and show seasonality effects in the location parameter μ. The two covariates are

$$\sin(\frac{2\pi x}{365.25}), \quad \cos(\frac{2\pi x}{365.25}).$$

The negative log-likelihood is **1091.815** and the number of exceedances is **144**. The parameters and standard errors are given in Table 3.10.

Table 3.10: Results of point process with threshold $u = a + b\sin(\frac{2\pi(x-d)}{365.25})$ and periodic covariates.

Parameter	MLE	Standard Error
$\mu : \beta_0$	3044.5101	179.6485
β_1	558.5669	211.5935
β_3	-890.0917	133.1975
σ	6.7521	0.1333
ξ	-0.2012	0.1758

The parameters are of the form:

$$\begin{aligned}
\mu(t) &= 3044.5101 + 558.567\sin(\frac{2\pi x}{365.25}) - 890.09\cos(\frac{2\pi x}{365.25}) \\
\sigma(t) &= 6.752 \\
\xi(t) &= .2012.
\end{aligned}$$

Using a deviance statistic to pick the best model

$$D = 21108.197 - 1091.815$$

$$= \quad 32.764$$

The number of parameters in the first model is 3 and the number of parameters in the second model is 5, which gives the deviance statistic a Chi-Square distribution with 2 degrees of freedom.

$$P(\chi^2_2 > 32.764) P - \text{value} < .0001.$$

The second model is the better choice, which states that seasonality does have an effect.

The better of each pair of these threshold models can be compared since both models have approximately the same number of exceedances. Looking at Table 3.11, both models seem to explain the data set equally well.

Table 3.11: Comparison of Point process models.

PProc Threshold daily (constant μ, constant σ, constant ξ)	-LL = 1092.215
PProc Threshold daily (periodic μ, constant σ, constant ξ)	-LL = 1091.815

3.4 Point Process Model with Monthly Maxima

Another interesting test is to do a point process among the monthly maxima. A graph of the monthly maxima is below in Figure 3.8.

Also plotted is a threshold u, which equals 2300, along with another threshold which contains a period of one-month, which introduces seasonality. This has the equation:

$$a + b \sin(\frac{2\pi(x-d)}{12}). \tag{3.11}$$

Also plotted is in Figure 3 is a frequency histogram of the previous 144 maximum observations by month. Looking at this graph, there seems to be a periodic trend by month. The goal of this example is to strive for approximately 10% of the data to fall above the threshold, as stated above. Thus $a = 2300$, $b = 550$, $d = 8$. The period is over 12 months and the variable x is in the interval $[1, 476]$.

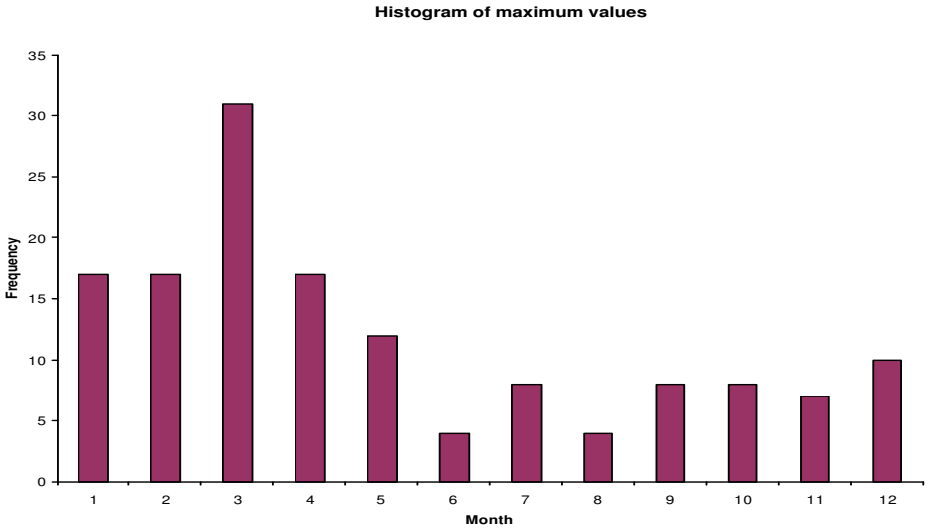

Figure 3.8: Frequency Histogram of top 144 Maxima By Month

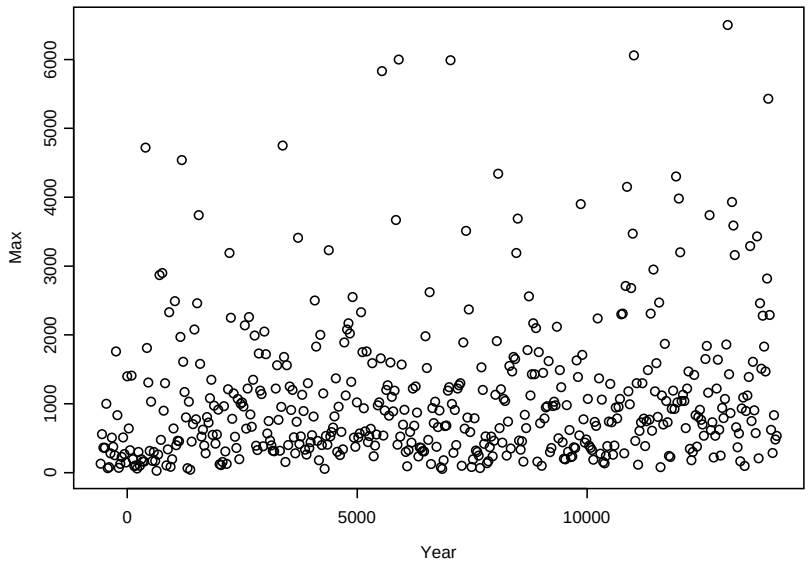

Figure 3.9: Graph of Monthly Maxima Plotted Against Time

The following results for the point process model are given in Table 3.7.

The threshold of **u=2300** gives a negative log-likelihood of **270.4264**, the number of exceedances is **49**, this data set has approximately 476 observations.

The parameters and standard errors are given in Table 3.12:

Table 3.12: Results of Point Process Model with Monthly Maximum with threshold $u = 2300$

Parameter	MLE	Standard Error
μ	6165.3514	532.4311
σ	697.7749	317.43278
ξ	-0.2225	0.17645

These estimates give the following model:

$$G(x) = \{\exp[1 + (-.2225)(\frac{X - 6165.35}{697.7749})^{-1/(-.2225)}\}. \tag{3.12}$$

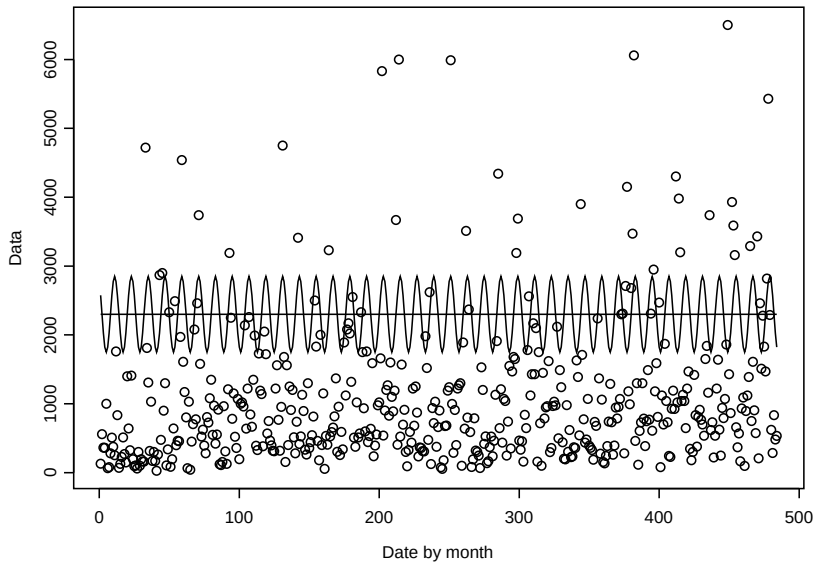

Figure 3.10: Graph of Montly Maxima Plotted with Threshold u, and a Seasonality Threshold

For this threshold with the periodic function, two covariates were introduced to help explain the model and show seasonality effects in the location parameter. The two covariates are

$$\sin(\frac{2\pi x}{12}), \quad \cos(\frac{2\pi x}{12}).$$

The negative log-likelihood is **270.385** and the number of exceedances is **49**. The parameters and standard errors are given in Table 3.13:

The parameters are of the form:

$$\begin{aligned} \mu(t) &= 6123.8341 - 52.94482\sin(\frac{2\pi x}{12}) + 70.7181\cos(\frac{2\pi x}{12}) \\ \sigma &= 683.363 \\ \xi &= -0.2304. \end{aligned}$$

Using a deviance statistic to pick the best model

$$D = 2270.426 - 270.385$$

Table 3.13: Results of Point Process Model with Threshold $u = 2300$ and Periodic Covariates.
Parameter MLE Standard Error

Parameter	MLE	Standard Error
$\mu : \beta_0$	6123.8341	529.2874
β_1	-52.94482	329.2575
β_3	70.7181	353.375
σ	683.363	297.8925
ξ	-0.2304	0.1697

$$= \quad 0.082$$

The number of parameters in the first model is 3 and the number of parameters in the second model is 5, which gives the deviance statistic a Chi-Square distribution with 2 degrees of freedom with a p-value of .95983.

Looking at this, it is apparent that the seasonality covariate is not significant.

The threshold of $u = a + b\sin(\frac{2\pi(X-d)}{12})$ gives a negative log-likelihood of **265.8351**, the number of exceedances is **49**, which is approximately 10% of the data.

The parameters and standard errors are given in Table 3.14.

Table 3.14: Results of Point Process model with threshold $u = a + b\sin(\frac{2\pi(x-d)}{12})$

Parameter	MLE	Standard Error
μ	6165.3514	532.4311
σ	697.7749	317.4328
ξ	-0.2225	0.1765

This leads to the GEV model:

$$G(x) = \exp\{1 + (-.255)(\frac{X - 6165.3514}{697.7749})^{1/.2225}.$$

Two periodic covariates are introduced to help explain the model and show seasonality effects in the location parameter μ. The two covariates are

$$\sin(\frac{2\pi x}{12}), \quad \cos(\frac{2\pi x}{12}).$$

The negative log-likelihood is **264.0142** and the number of exceedances is **48**. The parameters and standard errors are given in Table 3.15.

Table 3.15: Results of point process with threshold $u = a + b\sin(\frac{2\pi(x-d)}{2})$ and periodic covariates.

Parameter	MLE	Standard Error
$\mu : \beta_0$	6078.7429	506.8731
β_1	-278.9827	295.5058
β_3	483.2911	339.4034
σ	738.8544	262.0202
ξ	-0.187	0.138

The parameters are of the form:

$$\mu(t) = 6078.7429 - 278.9827(\sin(\frac{2\pi x}{12})) + 483.2911\cos(\frac{2\pi x}{12}).$$

Using a deviance statistic to pick the best model

$$\begin{aligned} D &= 2\{265.8381 - 264.0142\} \\ &= 3.6478 \end{aligned}$$

The number of parameters in model 1 is 3 and the number of parameters in the second model is 5, which gives the deviance statistic a Chi-Square distribution with 2 degrees of freedom.

$$P(\chi_2^2 > 3.6478)p - \text{value} = 0.16140.$$

Again, when dealing with the monthly maximums, seasonality does not have an effect.

The better of each threshold model can be compared since both models have approximately the same number of exceedances. Looking at Table (3.16), the second model seem to explain the data set better.

Table 3.16: Comparison of Point process models.

PProc Threshold monthly (constant μ, constant σ, constant ξ)	-LL = 270.4264
PProc Threshold monthly (constant μ, constant σ, constant ξ)	-LL = 265.8381

3.5 Summary

In summary, the Peachtree Creek data was investigated for a trend in the maximum. A linear trend was found in the annual maximum in both the GEV and Gumbel models. The minimum seemed to have no trend at all. Using a seasonality threshold in a point process model it does have a significant effect with the seasonality covariates. When dealing with the monthly maximum the seasonality threshold with no covariates seems to explain the most, indicating that there is also seasonality present. In conclusion there does seem to be a linear time trend among the data along with some seasonality, which states that the Peachtree Creek's water flow has increased over time. The best fit model is the Gumbel model with a linear time trend. This information is useful to explain and provide information about flooding and erosion.

Bibliography

Coles, Stuart (2001). An Introduction to Statistical Modeling of Extreme Values. Great Brittain: Springer Verlag London Limited.

Coles, Stuart (2001). S-Plus Functions for Extreme Value Modeling. http://www.maths.bris.ac.uk/ masgc/ismev/summary.html

Appendix A

S-Plus

```
Script 1

-Takes cleandata and fixes the datethen plots all the data points,
also plots every 5 years of data.

cleandata[,3]=seq.dates("06/20/1958","09/30/1998",by="days")
plot(cleandata[,3],cleandata[,2],xlab="Date",ylab="Data")
a<-cleandata$F2[1:1656]
a
b<-cleandata$F3[1:1656]
b
plot(b,a,xlab="DATE",ylab="riverflow")
c<-cleandata$F2[1657:3482]
c
d<-cleandata$F3[1657:3482]
d
plot(d,c,xlab="DATE",ylab="riverflow")
e<-cleandata$F2[3483:5309]
e
f<-cleandata$F3[3483:5309]
f
plot(f,e,xlab="DATE",ylab="riverflow")
g<-cleandata$F2[5310:7135]
g
h<-cleandata$F3[5310:7135]
h
plot(h,g,xlab="DATE",ylab="riverflow")
i<-cleandata$F2[7136:8961]
i
j<-cleandata$F3[7136:8961]
j
plot(j,i,xlab="DATE",ylab="riverflow")
k<-cleandata$F2[8962:10787]
```

```
k
l<-cleandata$F3[8962:10787]
l
plot(l,k,xlab="DATE",ylab="riverflow")
m<-cleandata$F2[10788:12614]
m
n<-cleandata$F3[10788:12614]
n
plot(n,m,xlab="DATE",ylab="riverflow")
o<-cleandata$F2[12615:14713]
o
p<-cleandata$F3[12615:14713]
p
plot(p,o,xlab="DATE",ylab="riverflow")
```

Script 2

- Plots all data points by lines

```
cleandata2[,3]=seq.dates("06/20/1958","09/30/1998",by="days")
plot(cleandata[,2],cleandata[,1],type='l',xlab="Date",ylab="Data")
```

Script 3

- Plots the annual max's (from OUT1). Plots the monthly max's (from OUT2). Plots the monthly max's in intervals of 10 years.

```
out1
out2
plot(out1$year,out1$MaxYear,xlab="Year", ylab="Max")

a<-out2$YearMonth[1:127]
a
b<-out2$MaxMonth[1:127]
b
plot(a,b,xlab="Dates in SAS by 10 years", ylab="Max")
c<-out2$YearMonth[128:235]
c
d<-out2$MaxMonth[128:235]
d
plot(c,d,xlab="Dates in SAS by 10 years", ylab="Max")
e<-out2$YearMonth[236:335]
e
f<-out2$MaxMonth[236:335]
f
```

```
plot(e,f,xlab="Dates in SAS by 10 years", ylab="Max")
g<-out2$YearMonth[336:484]
g
h<-out2$MaxMonth[336:484]
h
plot(g,h,xlab="Dates in SAS by 10 years", ylab="Max")

plot(out2$YearMonth,out2$MaxMonth,xlab="Year", ylab="Max")
```

Script 4

- Fits annual maxs into GEV function. Also, fits annual max's with the covariate year. Gives Probability, Quantile, and Return level plots for each. The covariate year is rescaled to values of (-1,1) to fit into the exponential better.

```
source("ismev2.dat")
source("ismev2.fns")

gev.fit(out1$MaxYear)
fm.gev <- gev.fit(out1$MaxYear)
gev.diag(fm.gev)

cov <- cbind ((out1$year-1978)/20)
cov
gev.fit(out1$MaxYear,ydat=cov, mul=1)
fm.gev <- gev.fit(out1$MaxYear,ydat=cov, mul=1)
gev.diag(fm.gev)
```

Script 5

- Tries to fit annual max's with the covariate year (not rescaled year, WRONG!)

```
source("ismev2.dat")
source("ismev2.fns")

cov <- cbind (out1$year)
cov
gev.fit(out1$MaxYear,ydat=cov, mul=1)
fm.gev <- gev.fit(out1$MaxYear,ydat=cov, mul=1)
gev.diag(fm.gev)
```

Script 7

- Fits the annual max's with the covariate year (0, 40)gives the same thing as (-1,1).

```
source("ismev2.dat")
 source("ismev2.fns")

cov <- cbind (out1$year-1958)
cov
gev.fit(out1$MaxYear,ydat=cov, mul=1)
fm.gev <- gev.fit(out1$MaxYear,ydat=cov, mul=1)
gev.diag(fm.gev)
```

Script 8

- This takes all the data points and performs a threshold test at u=1500, and performs a test that includes a covariate of the sin and cos curve, and also the x - which is a covariate for timeit is a vector from 1 to the length of the data values.

```
source("ismev2.dat")
source("ismev2.fns")
cleandata[,3]=seq.dates("06/21/1958","09/30/1998",by="days")
x<-1:length(cleandata$F2)
x
u1 <-rep(1500,length(cleandata$F2))
u1
usin <- function(x,a,b,d) { a + b*sin(((x-d)*2*pi)/365.25)}
wu <- usin(x,2000,1200,50)
wu
plot(cleandata[,3],cleandata[,2],xlab="Date",ylab="Data")
lines(u1)
lines(wu)
x

ydat <- cbind(sin(x*2*pi/365.25),cos(x*2*pi/365.25),x)
ydat
mydata.pp <- pp.fit(cleandata$F2,u=wu,ydat=ydat,mul=1:2,sigl=1:2,siglink=exp)
mydata.pp <- pp.fit(cleandata$F2,u=wu,ydat=ydat,mul=1:3,sigl=1:2,siglink=exp)
pp.fit(cleandata$F2,1500)
```

Script 9

```
- This program is trying to model threshold with the monthly max's.

source("ismev2.dat")
source("ismev2.fns")

x<-1:length(out2$MaxMonth)
x
out2[,7]=x
u1 <-rep(2300,length(out2$MaxMonth))
u1
usin <- function(x,a,b,d) { a + b*sin(((x-d)*2*pi)/12)}
wu <- usin(x,1700,850,500)
wu
plot(out2[,7],out2[,5],xlab="Date by month",ylab="Data")
lines(u1)
lines(wu)

ydat <- cbind(sin(x*2*pi/12),cos(x*2*pi/12),x)
ydat
mydata.pp <- pp.fit(out2$MaxMonth,u=wu,ydat=ydat,mul=1:3,sigl=1:2,siglink=exp)
pp.fit(out2$MaxMonth,2300)
```

Scriptmin

```
-This program fits the minimums into a GEV.  Using OUT3.

source("ismev2.dat")
 source("ismev2.fns")

gev.fit(out3$MinYear)
fm.gev <- gev.fit(out3$MinYear)
gev.diag(fm.gev)
```

```
Scriptmin2

- This program fits the minimums into a GEV with a covariate of year.  Using OUT3.

source("ismev2.dat")
source("ismev2.fns")

cov <- cbind ((out3$year-1978)/20)
cov
gev.fit(out3$MinYear,ydat=cov, mul=1)
fm.gev <- gev.fit(out3$MinYear,ydat=cov, mul=1)
gev.diag(fm.gev)
```