The Generalized Pareto Distribution and Threshold Analysis of Normalized Hurricane Damage in the United States Gulf Coast

Anthony Daspit[1], Kumer Pial Das[2]

[1]JP Morgan Chase, 1111 Fannin Street, Houston TX 77002
[2]Department of Mathematics, Lamar University, Beaumont, TX 77710

Abstract

This study concerns the probability distribution of the most damaging hurricanes to strike the United States. Economic damage is normalized to adjust for temporal shifts in societal vulnerability. Consistent with the extreme value theory, a generalized Pareto (GP) distribution is fitted to the excess in damage over a high threshold. The focus of the statistical analysis is primarily on diagnostics to determine an appropriate threshold. Conclusive evidence is provided that such data have a heavy tail (i.e., a GP distribution with a positive shape parameter)

Key Words: Normalized hurricane data, generalized Pareto distribution, mean excess plot

1. Introduction

As extreme value theory has developed in the past few decades it has found application in climatology, actuarial science, and engineering among others. Of special interest in insurance and reinsurance has been modeling the loss behavior in the tails of a distribution.

In this study, we will analyze the tail behavior of hurricane losses in the United States Gulf of Mexico shoreline and the Atlantic shore of Florida. Pertinent data will span the years 1926-2009. To derive the most accurate portrayal of hurricane events and the historical economic metrics, data was gathered from the U.S. Census Bureau, the Bureau of Economic Analysis (BEA), and the National Oceanic and Atmospheric Association's National Hurricane Center (NOAA). In order to accurately compare historic loss amounts, a proper normalization method must be used to account for not only inflation, but also changes in population and wealth along the affected areas. At the writing of this paper the Bureau of Economic Analysis' estimate of 2010 inflation-adjusted current cost net stock of fixed assets and consumer goods has not been made available, therefore the analysis is done in 2009 dollars. After storm losses were normalized, Mathwave's EasyFit© software was used to estimate the parameters of the exceedences at various threshold levels.

The paper is organized as follows. Formulating the objective of the study in section 1, a short discussion of Pielke-Landsea normalization procedure is given in section 2. We also normalize hurricane damage data in section 2. Fundamental to the GP distribution and parameter estimation procedure is described in section 3. The paper is concluded in section 4.

2. Pielke-Landsea Normalization Procedure

When modeling infrequent historical catastrophic events a large time span must be utilized. This causes difficulty in comparison across time and changing socio-economic conditions. The economic data from a hurricane in the 1930's is very different from the data from an equally intense storm that occurs in the 21st century. Only adjusting for inflation is inadequate as it does not take into account the change in the density of the affected population nor does it consider the change in per capita wealth. The Pielke-Landsea (PL) procedure takes these factors into consideration [7]. In order to normalize past storm damages to present values, the assumption is made that the losses are proportional to inflation, wealth, and population. Pielke-Landsea leave open that other factors could be added that represent changes in the insurance industry itself like deductibles and policy types. They also point out that since storm damage is generated by buildings, rather than people, and the ratio of population to housing has changed, adjustments would have to be made to accommodate for the rise in buildings per capita for more accurate results.

Pielke and Landsea proposed the normalized formula as:

$$\mathbf{NL_{Present\ Year}} = \mathbf{L_y} * \mathbf{i_y} * \mathbf{W_y} * \mathbf{P_{y,c}}$$

where:

$\mathbf{NL_{Present\ Year}}$ = a storm's losses normalized to present value.

$\mathbf{y}$ = year of storm's impact.

$\boldsymbol{c}$ = counties of storms maximum intensity.

$\mathbf{L_y}$ = a storm's damages in year $\mathbf{y}$, in year $\mathbf{y}$ dollars (not adjusted for inflation).

$\mathbf{i_y}$ = inflation factor, the ratio of the present implicit price deflator for Gross Domestic Product to that of year $\mathbf{y}$.

$\mathbf{W_y}$ = wealth factor, the ratio of the inflation adjusted current cost net stock of fixed assets and consumer goods as per capita to that year $\mathbf{y}$.

$P_{y,c}$ = affected population factor, the ratio of the change in the population of the coastal counties most affected by the storm from year **y** to present.

2.1 Calculating the PL Factors

Inflation (i_y)

To adjust for the change in the value of the dollar, the implicit price deflator for gross domestic product (IPDGDP) for the years 1929-2009 from the BEA are used.

$$i_y = \frac{IPDGDP_{2009}}{IPDGDP_y}$$

Wealth (W_y)

The national wealth is captured by the BEA in there estimate of current-cost net sock of fixed assets and consumer durable goods (CCNS). Because wealth is reported in billions of current-year dollars for the entire nation, is has to be adjusted for inflation and population. It must be disaggregated to a real (noninflated) per capita value as inflation, wealth, and population are independently distinguished in the normalization procedure. United States population estimates are available from the U.S. Census Bureau. Since censuses are done every ten years, linear interpolation was used for intermediate years.

$$W_y = \frac{\text{Real Wealth Factor}}{\text{US Population Factor}} = \frac{\left(\frac{\left(\frac{CCNS_{2009}}{CCNS_y}\right)}{i_y}\right)}{\left(\frac{\text{US population}_{2009}}{\text{US population}_y}\right)}$$

Affected Population ($P_{y,c}$)

The NOAA National Hurricane Center provides data on affected counties by storm. For each hurricane, the affected counties were documented and there current and historical populations were recorded from the US Census Bureau. Since censuses are done every ten years, linear interpolation was used for intermediate years. With the county-level population data, an affected population factors was calculated by the summing of the historical and 2009 populations of the counties affected the year (y) of the storm then calculating a ratio of those figures.

$$P_{y,c} = \frac{\text{Population of Affected Counties}_{2009}}{\text{Population of Affected Counties}_y}$$

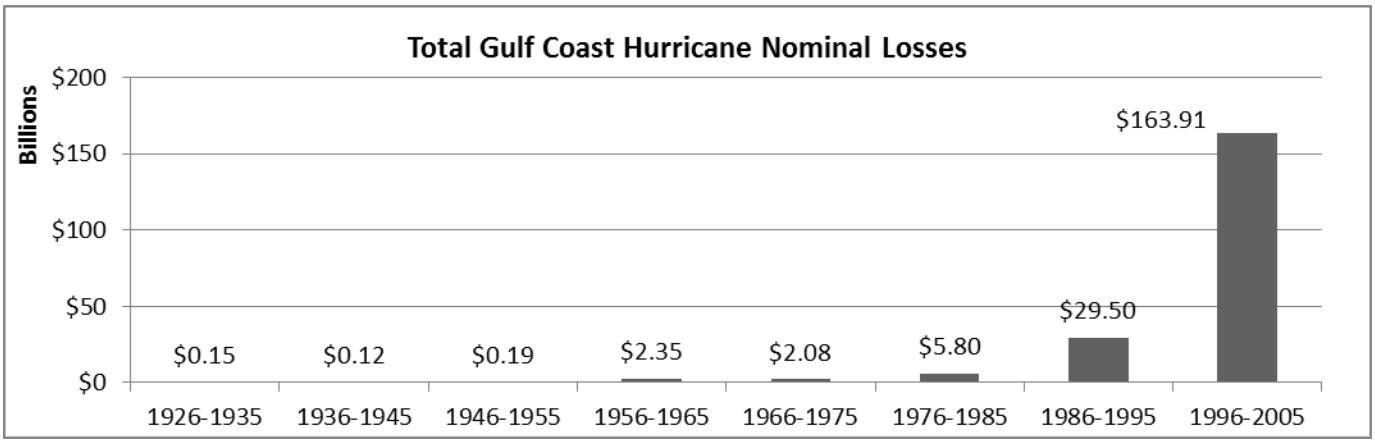

Figure 1: A chart of aggregate nominal storm losses. The obvious upward trend is deceptive as it doesn't take into account change in inflation, population, or wealth.

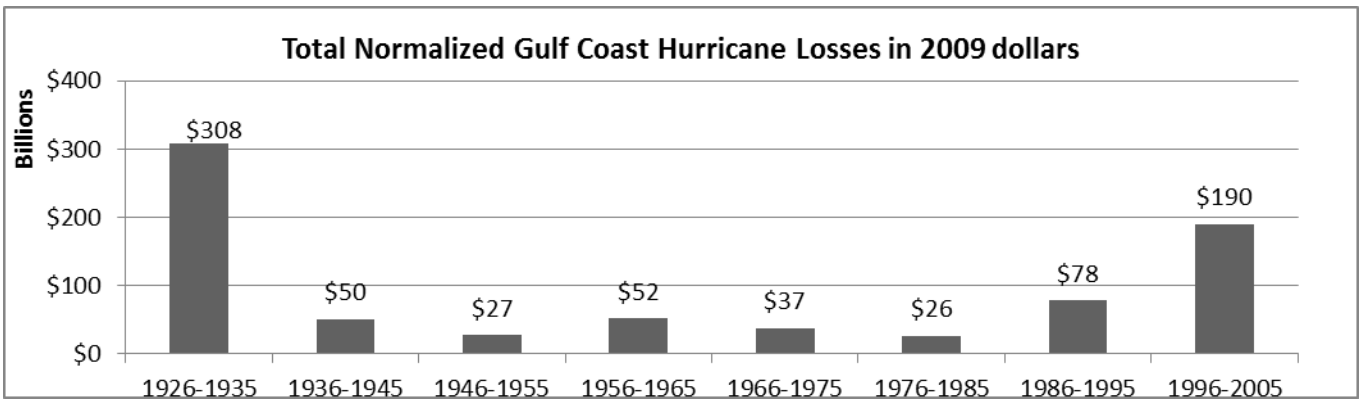

Figure 2 : Normalized values show that the losses from 1926 to 1935 were actually significantly worse than any other decade.

After normalizing the losses, the results were a mean storm damage of $34.79 billion with a characteristic long right-tail.

3. Extreme Value Theory

Pickands [6] defined that if X is a random quantity with a distribution $F(x)$ and μ is a threshold amount then the Generalized Pareto Distribution (GPD) can approximate $F(x|\mu) = P(X \leq \mu + x | X > \mu)$. The probability density function of the GPD $f(x)$ can be written as follows:

$$f(x) = \begin{cases} \frac{1}{\sigma}\left(1 + \kappa\frac{(x-\mu)}{\sigma}\right)^{-1-\frac{1}{\kappa}} & \kappa \neq 0 \\ \frac{1}{\sigma}e^{\left(-\frac{(x-\mu)}{\sigma}\right)} & \kappa = 0 \end{cases} \qquad \begin{array}{ll} x \geq 0, & \kappa > 0 \\ \mu < x < \mu - \frac{\sigma}{\kappa}, & \kappa < 0 \\ x \geq \mu, & \kappa = 0 \end{array}$$

Where κ is defined as the shape parameter, the scale parameter is defined by σ, and μ is defined as the location parameter.

The main drawback in extreme value theory analysis has been that when given a dataset we only use the values that exceed the threshold in estimating parameters. This leads to a limited data set which to estimate the parameters. Setting the threshold itself is not an easy task, as it is increased model bias is reduced but fewer events are available in parameter estimation.

3.1 Maximum Likelihood Estimations of Parameters

The method of Maximum Likelihood Estimation (MLE) was chosen for parameter estimations because of its well documented use in determining the parameter values for GPD distributed data. Estimates from using this method are usually unbiased, therefore for all sample sizes the parameter of interest is calculated correctly. The first step in MLE process is the likelihood function L.

$$L(X|\mu,\sigma,\kappa) = \prod_{i=1}^{n} f(x_i|\mu,\sigma,\kappa)$$

$$L(X|\mu,\sigma,\kappa) = \begin{cases} \frac{1}{\sigma^n}\left[1+\kappa\frac{x_n-\mu}{\sigma}\right]^{\frac{-1}{\kappa}} \prod_{1}^{n}\left[1+\kappa\frac{x_i-\mu}{\sigma}\right]^{-1}, \kappa \neq 0 \\ \frac{1}{\sigma^n} e^{\frac{-(x_n-\mu)}{\sigma}}, \kappa = 0 \end{cases}$$

After taking the logarithm of the likelihood function we maximize by setting the derivatives $\frac{\partial lnL}{\partial \sigma}, \frac{\partial lnL}{\partial \kappa}$ equal to zero. A maximum likelihood estimator cannot be obtained for μ, because the likelihood function is unbounded with respect to μ. Since μ is the lower bound of the random variable X, we may use the lowest sample value as a constraint then the likelihood function is maximized with respect to μ when $\mu = x_1$.

$$\hat{\mu} = x_1$$

$$\hat{\sigma} = \frac{\hat{\kappa}}{e^{n\hat{\kappa}}-1} - (x_n - \hat{\mu})$$

$$\sum_{i=1}^{n}\left[e^{n\hat{\kappa}} + \frac{x_n - x_i}{x_i - \hat{\mu}}\right]^{-1} = \frac{n}{e^{n\hat{\kappa}}-1} - \frac{1}{\hat{\kappa}e^{n\hat{\kappa}}}$$

3.2 Estimations of Parameters Results

Mathwave's EasyFit© was used to calculate the parameter estimates. Plots of the parameter values against the changing threshold levels were created to check for parameter stability. A long "flat" spot appears in between $13 billion and $17.5 billion because there were no storms with normalized losses in this range; therefore no change in the data set occurs as the threshold was moved through these levels. Parameter stability cannot be assessed accurately in these long flat spots, but they were relatively stable up to 12.5 billion. For all threshold levels between 7.5 and 26 billion, by both the Kolmogrov-Smirnov and Anderson-Darling test, the GPD is an acceptable fit at α=.01.

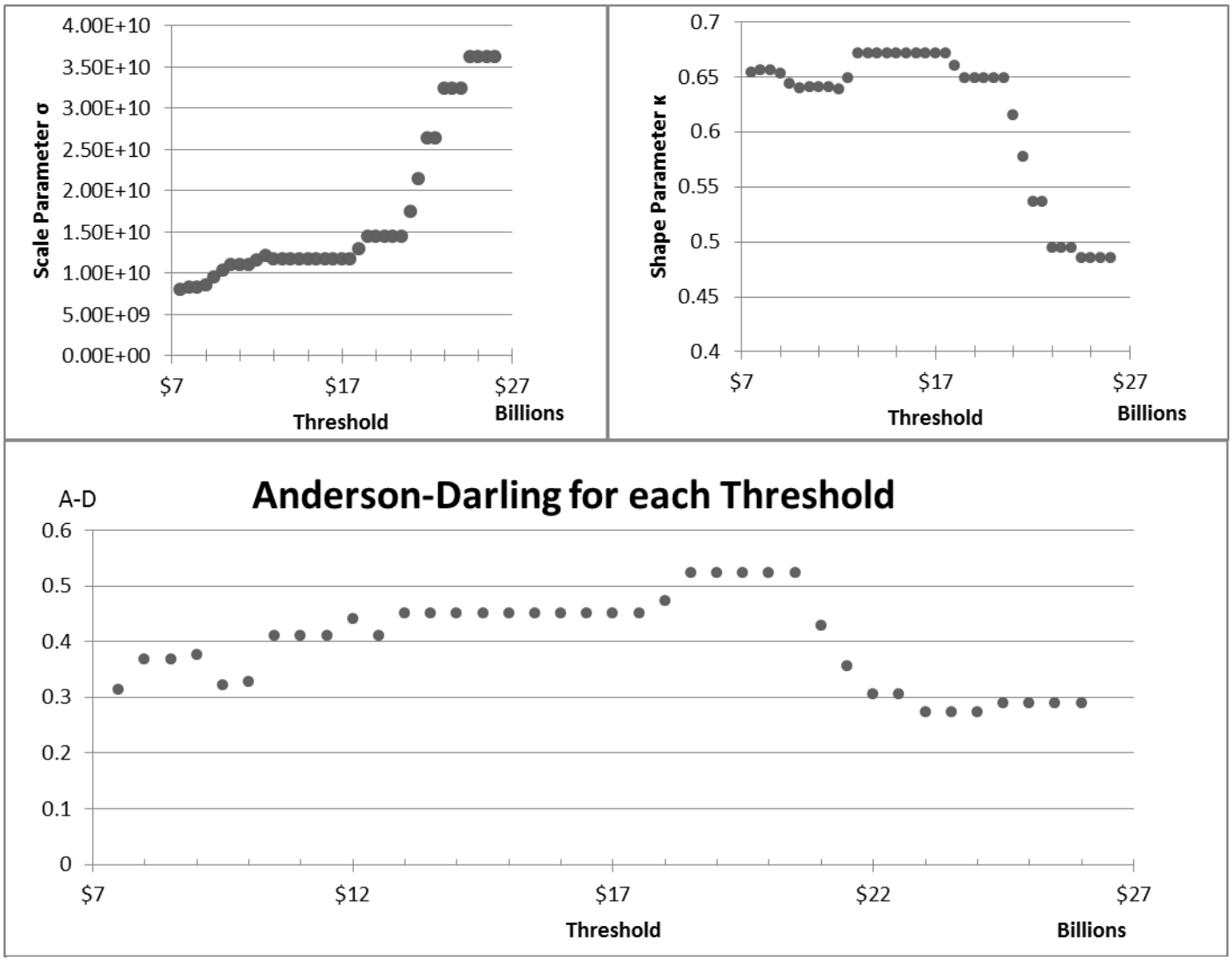

Figure 3: Plots of the parameters show strong stability up to $12.5 billion. Anderson –Darling scores reveal that the GPD is an extremely good fit for all thresholds.

3.3 Means Excess Plot

A graphical approach is used to define the possible range for the threshold choice and to assess the soundness of the threshold decision using the Mean Excess Plot.

Let X be a random variable with the cumulative distribution function F. Then the distribution of the excesses over the threshold t has the c.d.f.:

$$F_t(x) = P(X \leq x + t | X > t) = \frac{F(x+t) - F(t)}{1 - F(t)}$$

$$0 \leq x < x_F - t \qquad x_F < \infty \qquad x_F \text{ is the right endpoint of } F$$

Now take the expected value of F_t

$$E(F_t(x)) = e(t) = E(X - t | X > t)$$

The function e(t) describes the mean excess function of the random variable X. If F ~$GPD_{\kappa,\sigma}(x)$ then the excesses have the c.d.f.:

$$F_t(x) = GPD_{\kappa,\sigma(t)} \quad \sigma(t) = \sigma + \kappa t$$

$$0 \leq x < \infty \text{ if } \kappa \geq 0 \text{ and } 0 \leq x \leq \frac{-\sigma}{\kappa} - t \text{ if } \kappa < 0$$

The excess distribution over higher thresholds remains a GPD with the same shape parameter but with a scale parameter that is linearly increasing with the threshold level. From the equations above the mean excess function can be calculated by:

$$e(t) = \frac{\sigma(t)}{1 - \kappa} = \frac{\sigma + \kappa t}{1 - \kappa}$$

$$0 \leq t \leq \infty \text{ if } 0 \leq \kappa < 1 \text{ and } 0 \leq t \leq \frac{-\sigma}{\kappa} \text{ if } \kappa < 0$$

4. Concluding Remarks

Overall the mean excess e(t) graph is positively sloped which is consistent with $0 < \kappa < 1$ for all threshold levels. In addition, the relationship between the mean excess value and the threshold level is strongly linear for the threshold levels between 7.5 billion and 26 billion,

which according to equation e(t) is to be expected if the data come from a GPD. Had the data ceased to be properly distributed at certain threshold levels, then the mean excess plot would become erratic or curved and threshold selection at those unstable levels would violate the assumption that the data was from a GPD. Therefore, from the mean excess graph alone, limitations cannot be placed on threshold selection. After reviewing the Anderson-Darling statistics (Figure 3, bottom), the minimum is between 23 and 24 billion but the GPD is an extremely good fit throughout the threshold levels at $\alpha = .01$. The only limitations on threshold selection come from the parameter stability charts (Figure 3, top). As discussed earlier, after a threshold of $12.5 billion, the parameter stability cannot be in violation of the parameter stability assumption, the mean excess plot is linear, and the GPD is a good fit according to the Anderson-Darling test.

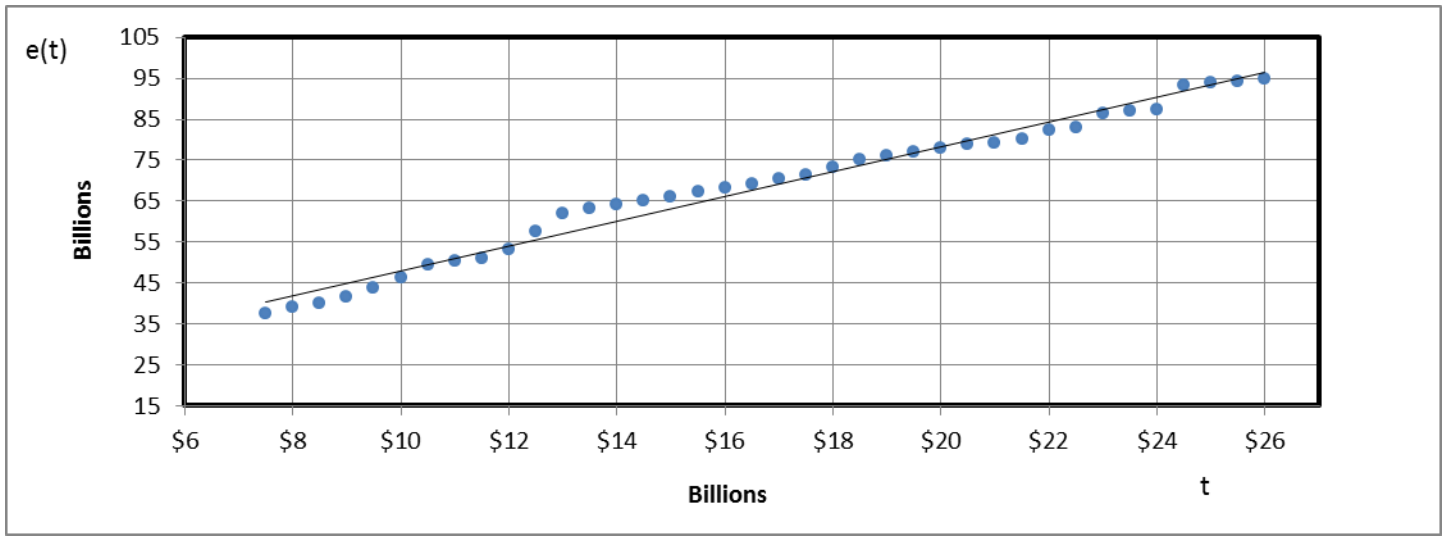

Figure 4: The Mean Excess Plot shows strong linearity through all threshold levels.

References

1. Bureau of Economic Analysis BEA. 2009a. "Table 1.1: Current-cost net stock of fixed assets and consumer durable goods." U.S. Dept. of Commerce, Washington, D.C.

2. Bureau of Economic Analysis BEA. 2009b. "Table 1.1.9: Implicit price deflators for gross domestic product." U.S. Dept. of Commerce, Washington, D.C.

3. Davison, A. C. 1984. Modelling excesses over high thresholds, with an application. In: *Statistical Extremes and Applications*, ed. J. Tiago de Oliveira, 461-482. Reidel, Dordrecht, The Netherlands.

4. Davison, A. R. L. Smith, Models for exceedances over high thresholds, Journal of the Royal Statistical Society. SeriesB 52 (3) (1990) 393–442.

5. National Hurricane Center NHC. 2011. "NHC/TPC archive of hurricane seasons." *NOAA, National Weather Service*, National Centers for Environmental Prediction, Miami, Fla.

6. Pickands, J. 1975. Statistical inference using extreme order statistics, *Ann. Statist.*, 3, 119–131.

7. Pielke, R. A., Jr., Landsea, C. W., Downton, M., and Musulin, R. 1999. "Evaluation of catastrophe models using a normalized historical record: Why it is needed and how to do it." *J. Insur. Reg.*, 18_2_, 177–194.

8. Pielke, R. A., Jr., J. Gratz, C. W. Landsea, D. Collins, M. A. Saunders, and R. Muslin, 2008:

Normalized hurricane damage in the United States: 1900-2005. Natural Hazard Review, 9, p.29-42.

9. Smith, R. L. 1987. Estimating tail of probability distributions, *Ann. Statist.*, 15, 1174–1204.

10. U.S. Census Bureau. 2010. "2010 census of population and housing: Population and housing units: 1940 to 2010." U.S. Dept. of Commerce, Economics and Statistics Administration, Washington, D.C.,

11. U.S. Census Bureau. 2000. "County population census counts 1900–90." U.S. Census Bureau, Population Division, Washington, D.C.,

12. U.S. Census Bureau. 2010. "Ranking tables for counties: Population in 2000 and population change from 1990 to 2010 _PHC-T-4_." U.S. Census Bureau, Population Division, Washington, D.C.,

13. U.S. Census Bureau. 2011. "National population datasets: Entire dataset." U.S. Census Bureau, Population Division, Washington, D.C.,