Predicting Crude Oil Price Using the Non-Stationary Extreme Value Modeling

Audrene Tiakor* Asim Kumer Dey† Kumer Pial Das *

Abstract
Extreme value analysis is an area of statistical analysis that can be used in many disciplines. These disciplines include engineering, science, actuarial science and statistics. Extreme Value Theory (EVT) deals with the extreme deviations from the median probability distribution and is used to study rare but extreme events. When considering the use of EVT to model data where extremes exist, one must consider whether extreme events are stationary or non-stationary. There are two methods that can be used within EVT for effective modeling of data; the Block Maxima (BM) method, which follows a generalized extreme value (GEV) distribution, and the Peaks Over Threshold Method which follows a generalized Pareto distribution (GPD). For this study, EVT will be used to model spot prices for the West Texas Intermediate (WTI) crude oil data from January 1986 to December 2016. With the spot prices for crude oil data, descriptive statistics will be used to model and interpret the characteristics of the data set, while determining whether the data contain extreme data. Next, hypotheses testing will take place to explore the applicable concepts such as the assumption or normality and independence. Considering that there are many factors that cause fluctuation in crude oil prices such as supply and demand, natural disasters and various world crises, hypothesis testing will also be used to determine whether the data is stationary or nonstationary. With the conclusion that the data are nonstationary, last, the BM method for non-stationary extreme events will be used to analyze return levels. The return levels provide insight about the cost of future crude oil prices.
2017 Mathematics Subject Classification: 60G70

Keywords: Extreme Value Theory, Block Maxima Method, Normality Testing, Stationarity, Non-stationary Generalized Extreme Value, Return Levels.

1 Introduction

Crude oil is an important commodity in today's world. Crude oil is used for many necessities and luxuries such as transportation, plastic, clothing, food,

*Lamar University, TX.
†University of Texas at Dallas, TX.

and insulation. Depending on the price of crude oil per barrel, the prices for our necessities and luxuries fluctuate. When the cost of crude oil per barrel is high, prices for necessities and luxuries increase. When the cost of crude oil per barrel is low, then the price for which we pay for our necessities and luxuries decreases. Understanding how crude oil prices affect the way we spend money. Data collected crude oil price over time can be analyzed to predict what the prices of crude oil will look like in future years. This gives us insight to how the fluctuation in prices and the fluctuation in the amount of money spent in the future as crude oil prices continue to fluctuate. EVT is developed to model and assess the risks caused by the extreme events [7]. For this application, the extreme events analyzed are the fluctuations in crude oil prices. EVT deals with the extreme deviations from the median of probability distributions and seeks to assess the probability of events that are more extreme than a certain large value [5]. Using the data of the WTI spot prices for crude oil from the U.S Energy Information Administration website from the years of 1986 to 2016, the extreme data values (extremely high prices for crude oil per barrel) can be found and the probability of the prices that exceed a certain large value can be analyzed. The purpose for this study is many fold: First, descriptive statistics are used to analyze the characteristics of the data, Second, hypothesis testing is used to ensure that important assumptions are met, which is very crucial to how effective the data can be modeled for further analysis. Third, after choosing which method in EVT is most effective, parameters are estimated using the maximum likelihood estimation (MLE) approach. Return levels are analyzed to assess the probability of future crude oil prices and then a brief discussion about the construction of confidence intervals for the return levels has been presented.

2 Extreme Value Theory

Usually, statistics are used to understand and describe where the bulk of the data lies in a distribution. EVT deals with the data that falls in the tail of the distribution, which may contain outliers. Let $X_1, X_2, ..., X_n$ be a sequence of independent random variables, with a common distribution function G. Define $M_n = max\{X_1, X_2, ..., X_n\}$. These X_i's represent values of a process that is measured on a regular time scale such as hourly measurements of sea level or daily mean temperature. M_n represents the maximum of the process over n time units of the observation [4]. M_n can be found for all values of n by the following:

$$\begin{aligned} Pr\{M_n \leq z\} &= Pr\{X_1 \leq z, \cdots, X_n \leq z\} \\ &= Pr\{X_1 \leq z\} \cdots Pr\{X_n \leq z\} \\ &= \{G(z)\}^n \end{aligned}$$

The distribution function G is unknown. To go into further details about how G can be found, there are two main limit theorems that constitute the statistical basis for applications in EVT. One of the main limit theorem leads to the GEV, and the other theorem by Gnedenko-Pickands-Balkema-Haan leads to the Generalized Pareto Distribution (GPD) [12].

2.1 Block Maxima Method

Since the distribution function G is unknown, by the extremal types theorem presented by Fisher and Tippett in 1928, if there exist sequences of constants $\{a_n > 0\}$ and $\{b_n\}$ such that

$$Pr\{(M_n - b_n)/a_n \leq z\} \rightarrow G(z)$$

as $n \rightarrow \infty$, where G is a non-degenerate function, then G could possibly belong to a family of models known as Gumbel (Type I), Frechet (Type II) and Wiebull (Type III). The family of distributions are as follows:

Gumbel:

$$G(z) = exp\Big\{-exp\Big[-\Big(\frac{z-b}{a}\Big)\Big]\Big\}$$

Frechét:

$$G(z) = \begin{cases} 0 & z \leq b \\ exp\Big\{-\Big(\frac{z-b}{a}\Big)^{-\alpha}\Big\}, & z > b \end{cases}$$

Wiebull:

$$G(z) = \begin{cases} exp\Big\{-\Big[-\Big(\frac{z-b}{a}\Big)^{\alpha}\Big]\Big\}, & z < b \\ 1 & \text{if } z \geq b \end{cases}$$

Where in Type I, $a > 0$ and $a > b$ and in Types II and III, $\alpha > 0$.

The Gumbel, Frechét and Weibull distributions can be combined into a family of models having the distribution of the following form [4]:

$$G(z) = exp\Big\{-\Big[1+\xi\Big(\frac{z-\mu}{\sigma}\Big)^{-\frac{1}{\xi}}\Big]\Big\} \tag{1}$$

which is defined on the set $\{z : 1+\xi z-\mu)|\sigma > 0\}$, where the location parameter (μ), scale parameter (σ) and shape parameter (ξ) satisfy respectively, $-\infty < \mu < \infty$, $\sigma > 0$ and $-\infty < \xi < \infty$. Equation 1 is known as the GEV family of distributions, where Gumbel, Frechét and Weibull are sub families, that are defined by the following respectively: $\xi \longrightarrow 0$, $\xi > 0$ and $\xi < 0$. This method in which the distribution function G can be found is known as the BM method.

2.2 Peaks Over Threshold (POT) Method

Let $Y_1, Y_2, ..., Y_n$ be a sequence of independent and identically distributed (i.i.d) random variables with common function $F(y)$. Since EVT deals with the data within the tail of the distribution, to model the upper tail of $F(y)$, consider exceedances over a threshold u and let $X_1, X_2, ..., X_k$ denote the excess (or peaks). Peaks Over Threshold (POT) is used when taking these peak values occurrences during any period of time from a continuous record [6]. POT depends on the threshold u which is defined by X_i:

Predicting Crude Oil Price Using the Non-Stationary EVT Approach

$$X_i = Y_i - u|Y_i > 0 \quad (2)$$

where the exceedances over u for $i = 1, 2, ..., k$ are asymptotically distributed and follow a generalized Pareto distribution (GPD) [11]. The two limit theorems associated with GEV and GPD show that information on the distribution of extremes can be gathered in two ways: one, by measuring the maximum of a sample whose size n goes to infinity, or two, by recording the excess function of that sample when increasing the threshold u to its upper limit [12]. For this study, the BM method is used. The BM method is used for the spot prices for crude oil per barrel data because the data can be be divided into non-overlapping periods of equal size. Before using the BM method, descriptive statistics are used to model the data with a purpose of finding the characteristics of the spot prices for crude oil per barrel data.

3 Crude Oil Data: Preliminary Analysis

West Texas Intermediate (WTI) spot prices for crude oil are collected from the U.S Energy Information Administration website from the years of 1986 to 2016. Fig. 1 yields the time plot of the data which reveal some high picks as well as a possible trend. Histogram of the price Fig. 2 indicates that the data tend to follow a long tailed distribution.

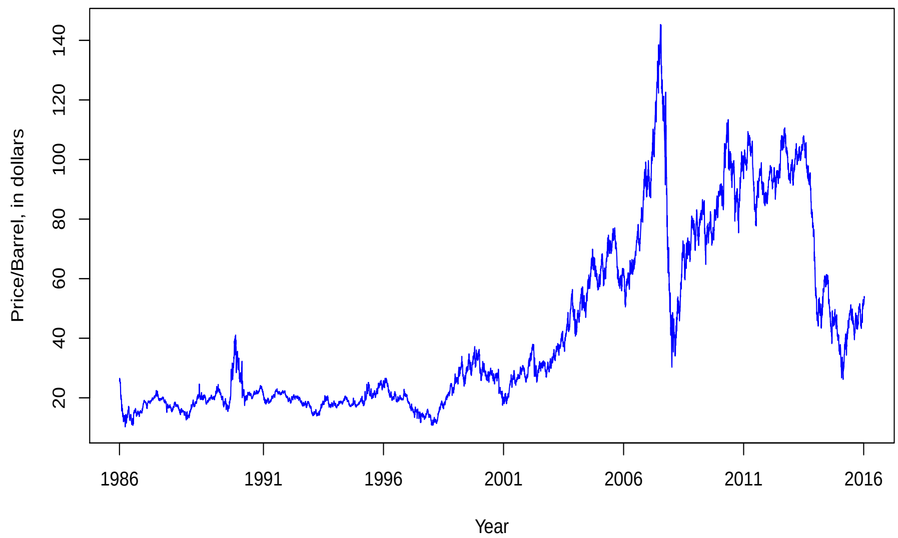

Figure 1: WTI Crude Oil Price, 1986-2016

Predicting Crude Oil Price Using the Non-Stationary EVT Approach

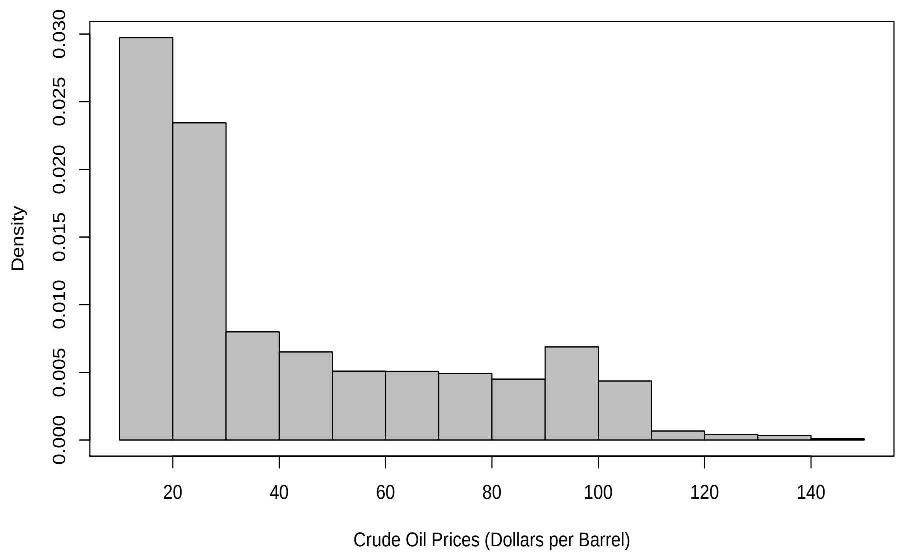

Figure 2: Histogram of WTI Crude Oil Price, 1986-2016

Table 1 lists the summary statistics of the data. Note that there is a big difference between third quartile (Q3) and Maximum value which indicates presence of outliers in the data. From the boxplot Fig. 3 it is clear that there exist some outliers in the data.

Table 1: Summary Statistics of Crude oil price

Min.	1st Qu. (Q1)	Median	Mean	3rd Qu. (Q3)	Max.
10.25	19.38	28.01	42.87	63.47	145.31

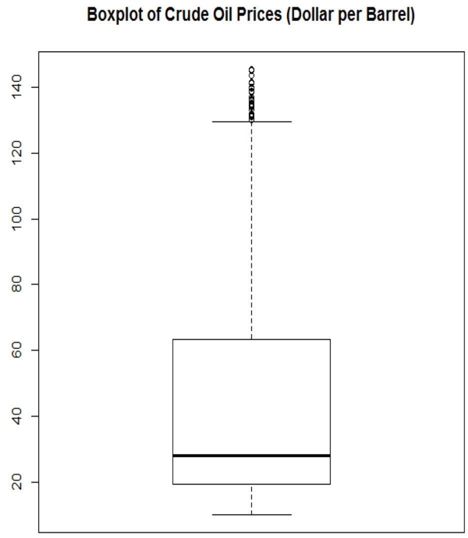

Figure 3: Boxplot of Crude Oil Prices

Again if an observation is farther than $1.5f_s$, where $f_s = IQR = Q_3 - Q_1$, from the closest fourth, it is classified as a *mild outlier* and if $3f_s$ from the nearest fourth, then it is called an *extreme outlier*. For crude oil data $f_s = Q_3 - Q_1 = 44.09$ and therefore, $1.5f_s = 66.14$ and $3f_s = 132.27$. We observe 1868 out of 7820 observations in the data are mild outliers and 27 of them are extreme outliers.

We can also construct some hypothesis test to check whether the data deviate from normal density. Table 2 summarizes the results of the Kolmogorov-Smirnov test, where the *null hypothesis* can be written as

H_o: The data follow a normal distribution.

Table 2: Results of Kolmogorov-Smirnov Test

Test	Test Statistic (D)	p-value
Kolmogorov-Smirnov	1	< 0.01

Here the $p-$value is very small which implies the deviation of the data from classical normal distribution.

4 Modeling Crude Oil Price

In sections 3 we conduct a primary analysis of the data set and we have found that the data set deviates from normal distribution and it also has a long tail.

Predicting Crude Oil Price Using the Non-Stationary EVT Approach

We use the generalized extreme value (GEV), the *block maxima approach* to model our long tail crude oil price data, i.e., the annual maximum crude oil price (X_t) are considered to follow $X_t \sim GEV(\mu, \sigma, \xi)$. The fitted models are summarized in Table 3.

Table 3: Estimated parameters of the (stationary) GEV model, standard errors are in parenthesis.

$\hat{\mu}$	$\hat{\sigma}$	$\hat{\xi}$
29.40 (3.20)	14.27 (3.74)	0.849 (0.28)

However, Fig. 4 shows that both the probability plot and QQ-plot have clear deviation from 45−degree line. We can also observe that the return level plot shows high variability and the density plot does not yield a good fit. Therefore it is clear that the stationary GEV Model does not capture the variability of the data very well.

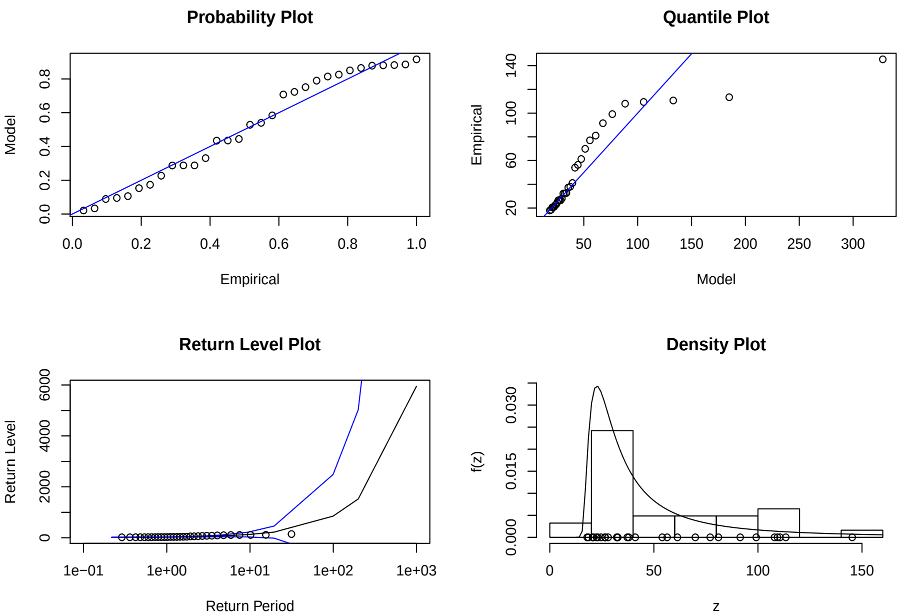

Figure 4: Diagnostic plots, stationary GEV model

The time plot Fig. 1 in section 3 indicates a non-stationarity trend in the data. A non-stationary processes have characteristics that change systematically thorough time (Non-homogenous) due to many reasons, e.g., time trend, seasonal trend, covariate relationship, etc. Variations through time are commonly modeled as a polynomial trend in the location parameter. Here we considere the annual maximum crude oil price in year t follow the following GEV distribution:

Predicting Crude Oil Price Using the Non-Stationary EVT Approach

$$X_t \sim GEV(\mu(t), \sigma, \xi), \quad (3)$$
$$\mu(t) = \beta_0 + \beta_1 t,$$

where, scale parameter, σ, and shape parameter, ξ are constant while the location parameter $\mu(t)$ remains a function of time. The corresponding distribution function for non-stationary generalized extreme value (GEV) [1, 2] can be written as

$$F(x; \mu(t), \sigma, \xi) = exp\{-\Big[1 + \xi\Big(\frac{y - \mu(t)}{\sigma}\Big)^{-\frac{1}{\xi}}\Big]\}. \quad (4)$$

Table 4: Estimates parameters of the non-stationary GEV model, standard errors are in parenthesis.

$\hat{\beta}_0$	$\hat{\beta}_1$	$\hat{\sigma}$	$\hat{\xi}$
11.58 (4.80)	1.73 (0.289)	14.01 (2.673)	0.382 (0.196)

Table 4 gives the estimates for $\hat{\beta}_o$ and $\hat{\beta}_1$ as 11.58 and 1.73 respectively, which yields the estimated linear trend as Eq. 5. The estimates for the scale and shape parameters are 14.01 and 0.382, respectively. Fig. 5 combines the linear trend of the non-stationary GEV model with observed data.

$$\begin{aligned} \hat{\mu}(t) &= \hat{\beta}_o + \hat{\beta}_1 t \\ &= 11.58 + 1.731t \end{aligned} \quad (5)$$

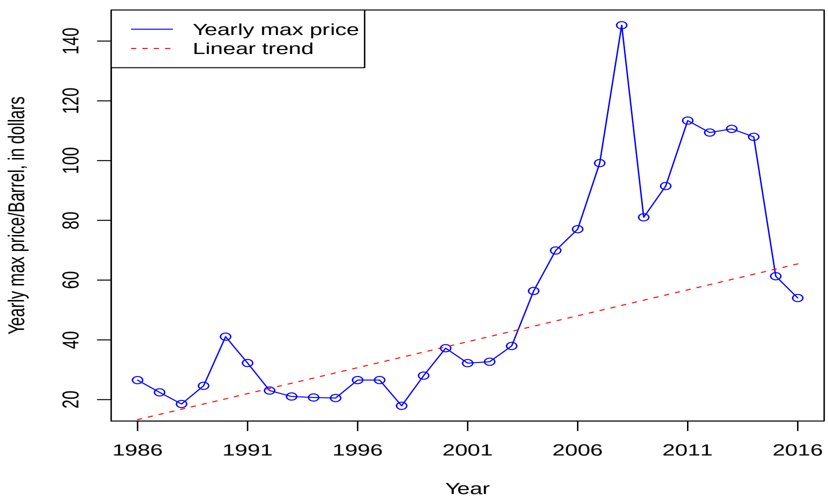

Figure 5: Linear trend of the non-stationary GEV model

Predicting Crude Oil Price Using the Non-Stationary EVT Approach

Diagnostic plots Fig. 6 of Non-stationary GEV models shows that both the probability plot and QQ-plot are approximately $45-$degree line. This implies that the Non-stationary GEV Model with linear trend in location parameter is a reasonable fit for modeling this data set.

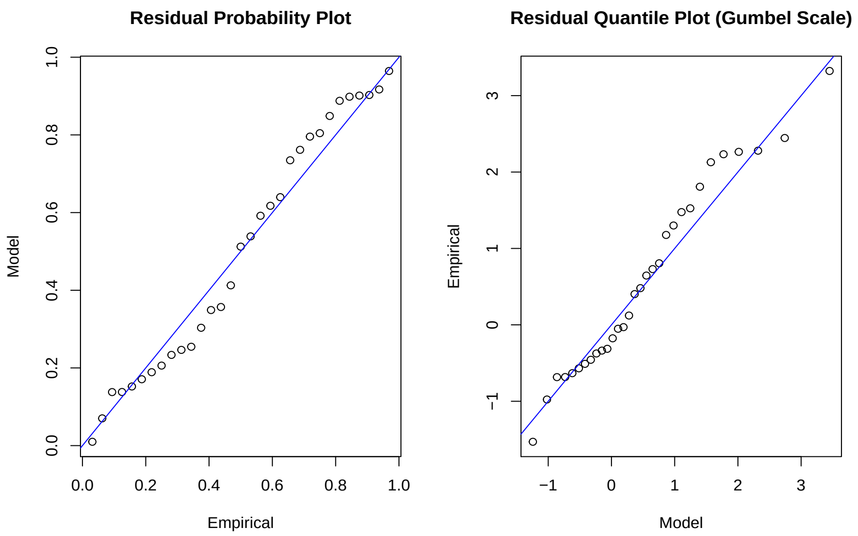

Figure 6: Diagnostic plots of non-stationary GEV model

We can also check whether non-stationary GEV Model significant over the stationary case. Here we need to calculate deviation statistic D as

$$\begin{aligned} D &= 2\{\ell_1(\mathcal{M}_1) - \ell_0(\mathcal{M}_0)\} \\ &= 15.742, \end{aligned} \tag{6}$$

where, $\ell_1\{\mathcal{M}_1)\}$ and $\ell_0\{\mathcal{M}_0)\}$ are the maximised log-likelihood under the non-stationary statinary models, respectively. The asymptotic distribution of D is given by the χ_1^2 distribution. Here D is much greater than $\chi_1^2(0.05) = 3.841$. Therefore, allowing for a linear dependence in time improve on our model which allows for a linear trend through time.

The assumption of independence can be investigated by the Von-Neumann test [14] where the null hypothesis is that the data series consists of independent elements. Table 5 gives the results of the Von Neumann test where the p-value is less than α=0.05, which leads to the conclusion that the data is not independent. However block maxima method is more 'relaxed' in regards to independence where the Peaks Over Threshold method strictly relies on independence of the data [10]. Therefore the block maxima method is utilized here.

Table 5: Von Neumann test of independence.

Test	Test Statistic (C)	p-value
Von Neumann	-88.31988	< 0.01

5 Return Level Estimation

The extreme quantile of crude oil price per barrel which is expected to exceed on average once every r years can be obtained by

$$z_r = (\beta_0 + \beta_1 t) + \frac{\sigma}{\xi}\big[\big(- \log(1 - r^{-1})\big)^{-\xi} - 1\big]. \tag{7}$$

Here we have time-varying return levels $z_r(t)$ [13]. For example, an estimate of the price level we might expect to see for crude oil once every 100 years is given by

$$z_{100}(t) = (11.58 + 1.73t) - \frac{14.01}{0.38}\big[\big(- \log(1 - 100^{-1})\big)^{-0.382} - 1\big].$$

Note that $z_{100}(t)$ is a function of t and therefore, the 100-year return levels estimate vary for t $= 32, 34, \cdots$, i.e. for the years 2017, 2018,$\cdots$, respectively. We calculate the crude oil price return level for return period $r = 10$, 20, 50, and 100 years. Fig. 7 shows the return levels as a function of time. We can consider the return levels as forecasts of crude oil price as we move through time.

Predicting Crude Oil Price Using the Non-Stationary EVT Approach

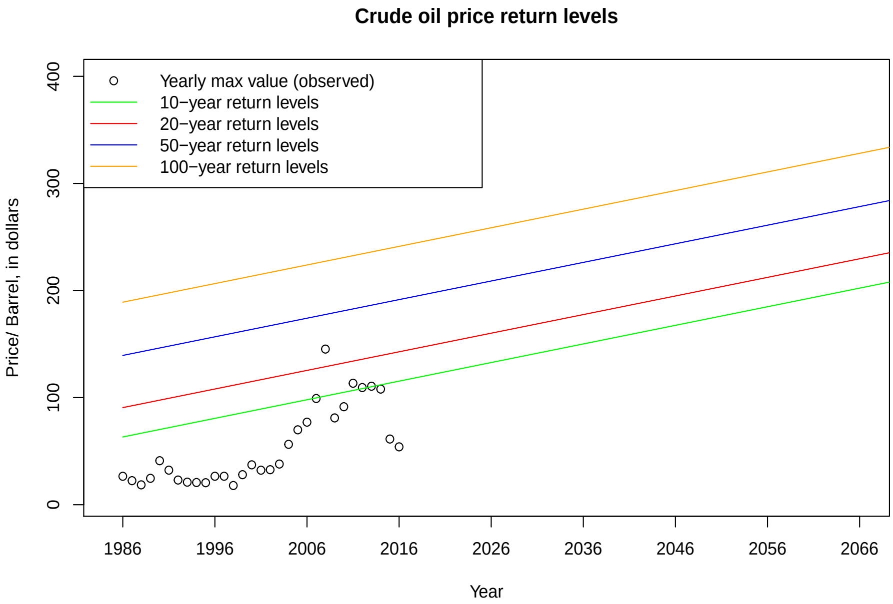

Figure 7: Crude oil price return levels for different return periods

6 Summary and Concluding Remarks

For this study, precautionary measures were considered in effectively modeling crude oil data to yield accurate return levels. These results predict the future cost of West Texas Intermediate (WTI) crude oil per barrel. The measures taken to ensure accuracy include: descriptive statistics to analyze the data and understand the data's unique characteristics and hypothesis testing to assure that certain assumptions are met and to check whether the conditions for GEV were stationary or non-stationary. After careful considerations to make sure that the data could be analyzed and modeled effectively, the model for GEV of non-stationary extremes was used to model the data, estimate the location, shape and scale parameters and find the return levels for different return periods.

References

[1] Caroni, C., Panagoulia, D. (2016, April). Non-Stationary Modelling of Extreme Temperatures In A Mountain Area of Greece. *Springer Science+Business Media Dordrecht*, 14(2), 217-228.

[2] Cheng, L., AghaKouchak, A., Gilleland, E., Katz, R. W. (2014). Non-Stationary Extreme Value Analysis In A Changing Climate. *REVSTAT-Statistical Journal*, 127, 353-369.

[3] Chiou, P., Miao, W. (2016, January). Predicted Return Level of Annual Maximum Earthquake in Northern California. *Research Gate.*

[4] Coles, S. (2001). *An Introduction to Statistical Modeling of Extreme Values* Classical Extreme Value Theory an Models (pp. 45-73).

[5] Dey, A, and Das, K., (2016). Quantifying the risk of extreme aviation accidents. *Elsevier*, 463, 345-355.

[6] Das, K., Edwards, A. S. (2016). Using Statistical Approaches to Model Natural Disasters. *American Journal of Undergraduate Research*, 13(2), 87-104.

[7] Liu, G. (2013). Using Statistical Approaches to Model Natural Disasters. *Fitting Extreme Values Into Generalized Pareto Distribution*, . Beaumont, TX: Lamar University.

[8] Obeysekera, J., Salas, J. D. (2014, July). Quantifying the Uncertainty of Design Floods under Nonstationary Conditions. *Journal of Hydrologic Engineering*, 1438-1445.

[9] Obeysekera, J., Salas, J. D. (2016). Frequency of Recurrent Extremes under Nonstationarity. *Journal of Hydrologic Engineering*, 1-9.

[10] OECD. (2016). OECD Studies on Water Mitigating Droughts and Floods in Agriculture: Policy Lessons and Approaches on Water, *OECD Publishing*, Paris. http://dx.doi.org/10.1787/9789264246744-en.

[11] Pickands, J. (1975). Statistical inference using order statistics. *Ananls of Statistics*, 3, 119-131.

[12] Pisarenko, V. F., Sornette, A., Sornette, D., Rodkin, M. V. (2014). Characterization of the Tail of the Distribution of Earthquake Magnitudes by Combining the GEV and GPD Descriptions of Extreme Value Theory. *Pure and Applied Geophysics*, 171, 1599-1624.

[13] Author unknown. Chapter 5: Non-stationary extremes.(n.d.). Retrieved September 4, 2016, from http://www.mas.ncl.ac.uk/ nlf8/teaching/mas8306/notes/chapter5.pdf.

[14] von Neumann, John. *Distribution of the Ratio of the Mean Square Successive Difference to the Variance.*,Ann. Math. Statist. 12(4) (1941), no. 4, p 367–395.