A Statistical Analysis of Extreme Wave Heights in the North Sea

by
Poorwa Singh

MSc. Thesis
July 2006

Delft University of Technology
Faculty of Applied Mathematics
Section: Probability, Risk and Statistics

Poorwa Singh
1242326

Poorwa Singh
1242326

Table of Contents

List of Figures

Poorwa Singh
1242326

List of Tables

Acknowledgements

I would like to thank my supervisors Prof. Roger Cooke and Dr. Pieter van Gelder for all the support and encouragement – both of whom have given me the guidance I needed during the work of this thesis. I would also like to express my gratitude to Prof. Jolanta Misiewicz for the patience and support to help me over both the big and small hurdles.

Also, thanks also go to Prof. Jan van Noortwijk and Dr. Rik Lopuhaa as the members of this thesis committee. I'd like to express my appreciation to Sebastian Kuniewski for all the last minute help, especially with Matlab. Last but not least, I'd like to thank Lech (Leszek) Grzelak for all the help and patience over the two years of the MSc. program.

Poorwa Singh
1242326

Overview

This aim of this master's thesis is to calculate the height of a wave that only occurs once in 10,000 years. Chapter 1, the introduction, briefly describes the stations where the data comes from. Chapter 2 gives a bit of background. It explains what a wave is and how it is measured. The next chapter describes a bit of the history about our environment of interest – the North Sea.

After all the background information, the methods used are described. This begins in chapter 4, methods. This starts with a bit of background on the distribution used. Then it continues out onto the two different methods used. First the regression method, which does not go into much detail and secondly, the Bayesian method. This method begins with an equation which is needed to reach the desired wave height. This equation in then transformed using two conditional independence assumptions. Solving this equation in described in detail in chapter 4. Then, a few fits and the corresponding wave heights are shown. Finally, chapter 5 concludes the paper.

Poorwa Singh
1242326

Poorwa Singh
1242326

Introduction

The Netherlands is one of the lowest lying countries in the world. Bordered by the North Sea, the Netherlands needs a thorough flood protection system to protect its people and its land. To get a better idea of the degree of protection required, the frequency of extreme heights of waves that occur during storms should be known. The data comprises of twenty four years of wave heights that have been measured at nine different stations in the North Sea. From this, one can extrapolate to get the height of the wave that occurs only once in 10,000 years.

Locations' of Stations

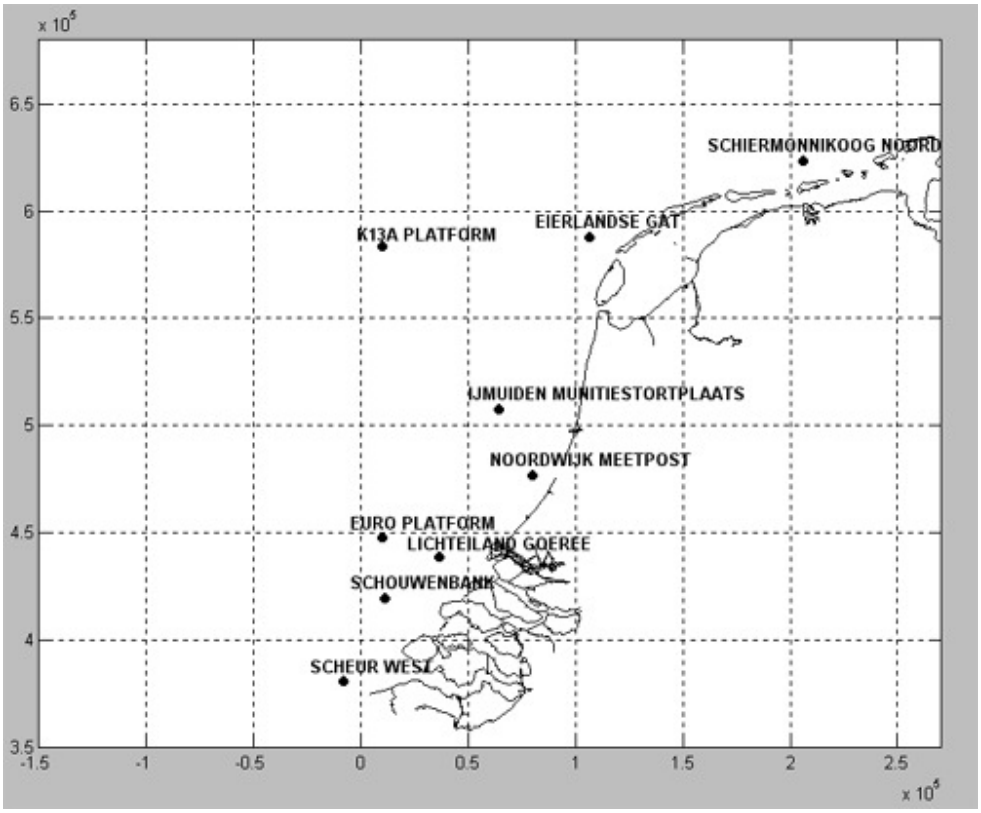


Figure 1

The data for the stations in figure 1 is obtained from www.golfklimaat.nl. The exact locations are in appendix A.1. The data used are wave heights that have been recorded at each of these stations, every three hours from January 1st, 1979 to December 31st, 2002. This means that there are 2,920 measurements for each station each year, making a total of 70,128 data points for each of the nine stations. More data exists, but it is not used as the measurements for the other times and stations, are much more sporadic.

Poorwa Singh
1242326

I have used the following abbreviations for the stations:

Station Abbreviations

Station	Abbreviation
1. Eierlandse Gat	ELD
2. Euro Platform	EUR
3. K13A Platform	K13
4. Lichteleiland Goeree	LEG
5. Noordwijk Meetpost	MPN
6. Scheur West	SCW
7. Schiermonnikoog Noord	SON
8. Schouwenbank	SWB
9. Ijmuiden Munitie Stortplaats	YM6

Table 1

Poorwa Singh
1242326

Background

To better comprehend the objective of this project, understanding the definition and the causes of waves are helpful. Waves are undulations in the surface of the ocean, which are caused by the wind. The faster the wind, the longer the wind blows, and the bigger the area over which the wind blows, the bigger the waves. The height of a wave is the vertical distance between the bottom of a trough and the top of a nearby crest. The trough is the part of the ocean wave that is displaced below the still water line and the crest is the portion that is displaced above the still water line. This is often used to refer to the highest point of the wave. See Figure 2 from "On Tides and Weather"

How the Height of a Wave is Measured

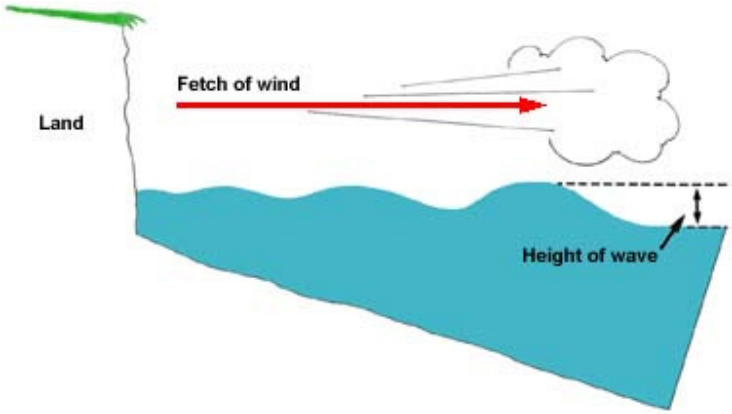


Figure 2

The data used is from www.golfklimaat.nl. There are many techniques for measuring wave heights. Wave height data from golfklimaat "are measured in the North Sea using three different types of measuring instruments, namely:

- rods (step-gauges)
- buoys (waverider, wavec, directional waverider)
- radar (radar)

Step gauges are large tubes on which electrodes are placed at regular intervals. These gauges are mounted on platforms or measuring poles. Using electronics, a continuous record is kept of the highest electrode that is just under water. In this way, it is possible to establish the changes in the surface of the sea during a certain period of time and thereby draw conclusions about the characteristics of the wave movements. It is only wave heights and periods that are measured with a step gauge, not the direction of the waves.

Of the three buoys listed above, the waverider is the oldest, and it does not measure direction. The buoy is convex in shape, a little under a meter in diameter. The buoy measures accelerations in a vertical direction caused by the force of waves against the buoy. From this it is possible to calculate changes in height of the surface of the sea and thereby the characteristics of the wave movement. The wavec buoy is the oldest buoy that can measure wave direction. This buoy, with a diameter of 2.5 m, is much bigger than the waverider. In addition to vertical accelerations, the buoy also measures its own inclinations caused by the movements of the waves. This makes it possible, not only to measure wave heights and periods, but also to gain information on the direction to which the waves move.
The directional waverider is the modern version of the wavec, but is the size of a normal waverider and it and basically works in the same way.

Wave radar is a modern version of the step gauge. The radar is mounted on a platform or on a measuring pole. The radar bundle is pointed vertically downwards. The distance between the radar and the surface of the sea is measured by reflection and, in this way, the state of the sea is recorded.” Golfklimaat.nl .

Poorwa Singh
1242326

History

The area of interest, in our case, is the North Sea, specifically, the southern part of the North Sea. In this environment, the wave heights are no more than one meter in calm conditions. The wave period (the time between one wave and the next) is 3 to 4 seconds in calm conditions, and increases to between 10 and 15 seconds during storms.

Sometimes, even though there is neither storm nor strong wind, waves that occur can be quite large. These waves that generated elsewhere, in a distant wind field, and have subsequently moved on, outside this field. As the distance grows from the source of its energy, the wave height gradually decreases and the period of the wave lengthens. This is known as the swell.

On the open sea, the swell can move forward for days on end, and may come from all directions. In the considered region, though, swell can only come from the North. It can only be from the northern reaches of the North Sea or the Atlantic Ocean, hence it is rarely more than a day old.

The waves that are measured in the southern part of the North Sea are always a mixture of wind-generated waves and swell. Under calm conditions, the influence of the swell is often visible, especially in the wave period. During storms, however, it is always the wind-generated waves that are dominant. (golfklimaat).

Poorwa Singh
1242326

Poorwa Singh
1242326

Methods

The goal of this paper is to find a method that best predicts the one in 10,000 year wave in the North Sea. The data used are from nine stations located on the North Sea, collected over 24 years. (golfklimaat.nl) Two different methods are tried to get a good estimate of this wave. The first of the two methods is based on extreme value theory. Extreme value theory deals with the maximum and minimum of independent, identically distributed (i.i.d.) random variables. The properties of the distribution of extremes (maximum or minimum), extreme order statistics, and exceedances over or below thresholds are determined by the tails of the distribution. Focusing on the tails is advantageous as there are certain distributions that are designed specifically for the end of the distribution.

Of the possible extreme value distributions, here we will use the Generalized Pareto distribution (GPD) to estimate the tail of our wave height data. GPD is used because it allows a continuous range of possible shapes that includes both the exponential and Pareto distributions as special cases, both of which are used to model exceedances. The probability density function for the generalized Pareto distribution used in Matlab is from the Mathworks site. It has shape parameter $k \neq 0$, scale parameter σ, and threshold parameter θ, is

$$y = f(x|k,\sigma,\theta) = \left(\frac{1}{\sigma}\right)\left(1 + k\frac{(x-\theta)}{\sigma}\right)^{-1-\frac{1}{k}} \quad y = f(x|k,\sigma,\theta) = \left(\frac{1}{\sigma}\right)\left(1 + k\frac{(x-\theta)}{\sigma}\right)^{-1-\frac{1}{k}}$$

for $\theta < x$, when $k > 0$, or for $\theta < x < \theta - \frac{\sigma}{k}$ when $k < 0$.

In the limit for $k = 0$, the density is

$$y = f(x|0,\sigma,\theta) = \left(\frac{1}{\sigma}\right)e^{-\frac{(x-\theta)}{\sigma}}$$

for $\theta < x$.

If the shape parameter, k is greater than 0, equal to 0, or less than 0, then the cases "correspond respectively to the extreme value type II (Frechet), extreme value type I (Gumbel), and reverse Weibull domains of attraction." Also, "for $k < 0$, the distribution has zero probability density for $x > \theta - \frac{\sigma}{k}$, while for $k \geq 0$, there is no upper bound."

This first method is regression analysis. This method is often used in analyzing extreme data. The aim of regression analysis is to construct mathematical models that describe or explain relationships that may exist between the waves

of the stations. Hence, a given advantage with this procedure is that it includes the information from all the stations. To see the relationship between the stations, first, 1000, one in 10,000 year waves are generated. The intent is to see the effect of the 1000 waves, from stations two to nine, on the first station. Method 1 is located in Appendix, A.2.

The second method is based on Bayesian theory. This is not often used in analyzing extreme data. It is advantageous in this case because, it incorporates all available data, allowing us to use the wave heights from all the stations, over a threshold, instead of focusing on only one station. The Bayesian method then uses this information to get an idea about the prior distributions of the parameters. Then updates the prior and gets the posterior, which in turn will help predict the wanted wave.

Poorwa Singh
1242326

Method 2

Background

The second method tried is one that is not often seen in extreme value theory – Hierarchical Bayes. The advantage of the Bayesian model is that it assimilates data from difference sources. "the Bayesian [model] requires a sampling model (the likelihood) and, in addition, a prior distribution on parameters. Unknown parameters are considered random and all inferences are based on their distribution conditional on observed data (the posterior distribution)." (Carlin and Louis, p. 6).[K] Also, prediction is naturally incorporated when using Bayes. "The concept of posterior prediction matches with the fact that the principal inferential objective of an extreme values analysis is of predictive nature."(Beirlant et al., p. 429). But a disadvantage of the Bayes approach is that the problem of prior elicitation leads to subjectiveness.

As an example, take the case of a one stage model. Let **y** be the observed data of a random variable, Y, where the density function of Y is f(y | **θ**). **θ** represents the vector of parameters. Let π(**θ**) denote the density of the prior distribution of **θ**. The likelihood of θ is f(**y** | **θ**) , which equals $\prod_{i=1}^{m} f(y_i \mid \theta)$ if independent. According to Bayes' theorem,

$$\pi(\theta \mid y) = \frac{f(y \mid \theta)\pi(\theta)}{\int_{\Omega} f(y \mid \theta)\pi(\theta)d\theta} \propto f(y \mid \theta)\pi(\theta)$$

where Ω is the parameter space. This allows us to update our initial beliefs about **θ**, represented by the prior π(**θ**), to be converted into the posterior distribution, π(**θ**|**y**).(Beirlant et al., p 430).

Used Model

In this case, a two stage Bayesian model is used. A diagram for this model is shown below in Figure 3:

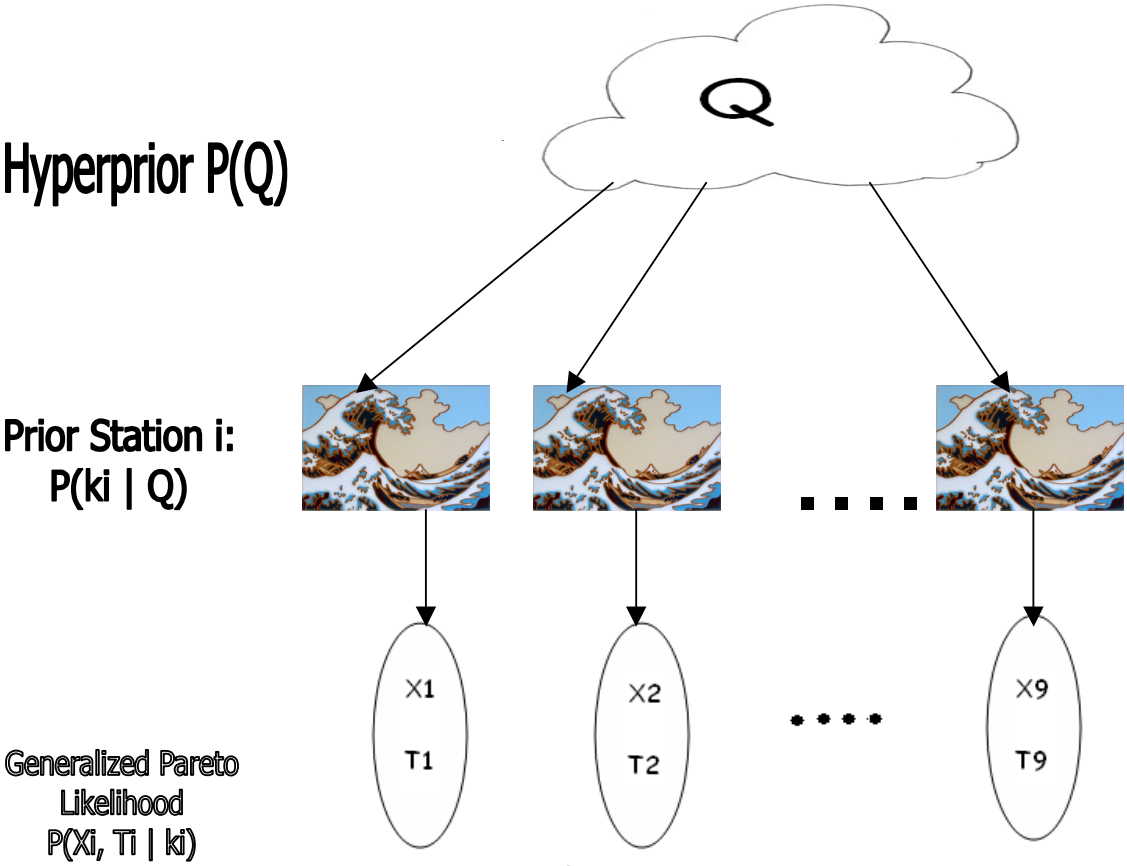


Figure 3: Bayesian Two Stage Hierarchical Model

Where the information from station *i* is characterized by an exposure T_i and the events X_i. (Cooke et al.) In this case, the T_i is always the same at 24 years, or 70,128 time measurements. The X_i's follow a Generalized Pareto distribution, the parameters of which are uncertain, and drawn from a prior distribution. The parameters of the prior distribution are also uncertain. "This uncertainty is characterized by a hyperprior distribution over the parameters of the prior…the hyperprior is a distribution P(Q) over the parameters Q of the prior distribution from which the [generalized Pareto] intensities [k_1,…,k_9] are drawn"(Cooke et al., p. 4) The advantage of using this model is that the information from the other stations is also taken into account when calculating the parameters for one station, even though the parameters and the prior distributions can be calculated separately. The model is characterized by:

$$f(X_1,\ldots,X_9,k_1,\ldots,k_9,Q) \qquad \textbf{Equation 1}$$

This model is simplified by the following conditional independence assumptions:

CI.1 Given Q, k_i is independent of $\{X_j, k_j\}_{j\neq i}$
CI.2 Given k_i, X_i is independent of $\{Q, k_j, X_j\}_{j\neq I}$ (Cooke et al.)
Expression CI.1 means that if we know the hyperprior Q, the parameter for station i is independent of the events, X_j, and the parameters, k_j, of the other stations. Expression CI.2 means that if we know the parameter, k, for station i, then the events for station i, X_i, is independent of the hyperprior, Q, of the parameters of the other stations, k_j, and of the number of events of the other stations, X_j.

We need:
$f(k_1|X_{11},\ldots,X_{1n(1)},\ldots,X_{91,\ldots},X_{9,n(9)})$ **Equation 2**

Using Bayes' Theorem:

$$= \frac{f(k_1, X_{11},\ldots,X_{1n(1)},\ldots,X_{91,\ldots},X_{9,n(9)})}{f(X_{11},\ldots,X_{1n(1)},\ldots,X_{91,\ldots},X_{9,n(9)})}$$

$$= \frac{f(X_{11},\ldots,X_{1n(1)}|k_1, X_{21},\ldots,X_{2n(2)},\ldots,X_{91,\ldots},X_{9,n(9)})\, f(k_1, X_{21},\ldots,X_{2n(2)},\ldots,X_{91,\ldots},X_{9,n(9)})}{f(X_{11},\ldots,X_{1n(1)},\ldots,X_{91,\ldots},X_{9,n(9)})}$$

- $f(X_{11},\ldots,X_{1n(1)}|k_1,X_{21},\ldots,X_{2n(2)},\ldots,X_{91,\ldots},X_{9,n(9)})$
- $f(k_1|\, X_{21},\ldots,X_{2n(2)},\ldots,X_{91,\ldots},X_{9,n(9)})$

using from CI.2 this becomes:

$f(X_{11},\ldots,X_{1n(1)}|k_1)\, f(k_1|\, X_{21},\ldots,X_{2n(2)},\ldots,X_{91,\ldots},X_{9,n(9)})$

Now, a threshold that is high enough must be chosen. That is, a height, in cm, must be chosen, such that the peaks above it are far enough apart that they do not influence each other—the peaks are independent. This is needed so that CI.2 the X_i's can be taken as independent, given k_i, is realistic. Then, the equation becomes:

$\propto \Pi_{j=1}^{n(1)} \mathbf{f(X_{1j}\,|\,k_1)}\, f(k_1|\, X_{21},\ldots,X_{2n(2)},\ldots,X_{91,\ldots},X_{9,n(9)})$ **Equation 3**

Poorwa Singh
1242326

Part 1, Equation 3

Now, the two parts of Equation 3 need to be calculated separately. First, the bold part of Equation 3, $f(X_{1j} \mid k_1)$, can be calculated using the following:

To calculate $f(X_{1j} \mid k_1)$, first a k_1 is needed. k_1 is drawn from some prior, q. The k's are needed to help interpreting and choosing this prior, q.

The k's for the data can be calculated using $P(X > x\theta \mid X > \theta)$ where θ is the predetermined threshold. This is because using $P(X > x\theta \mid X > \theta)$, $x > 1$, the scale parameter, k, can be calculated, and then using that, $f(X_{1j} \mid k_1)$ can be calculated.

$$P(X > x\theta \mid X > \theta) = \frac{P(X > x\theta)}{P(X > \theta)} = \frac{1 - F(x\theta)}{1 - F(\theta)} \qquad \textbf{Equation 4}$$

Here, F is a function such that the second moment is infinite, $\int x^2 dF(x) = \infty$ but there exists an $\varepsilon > 0$ such that $\int |x|^{\varepsilon} dF < \infty$. Then there exists $\alpha \in (0,2)$ such that $1 - F \approx x^{-\alpha} K(x)$ where K is a slowly varying function at ∞, that is, $\frac{K(\theta x)}{K(\theta)} \to c$, some constant. Choose θ such that for $x \in (1, M)$, the observed interval area being (θ, Mθ), $\left|\frac{k(\theta x)}{k(\theta)}\right| < \delta$ (small).

So from $\frac{1 - F(x\theta)}{1 - F(\theta)} \approx \frac{(x\theta)^{-\alpha} k(x\theta)}{\theta^{-\alpha} k(\theta)} = x^{-\alpha} \frac{k(x\theta)}{k(\theta)} \approx x^{-\alpha} c$ **Equation 5**

$X_i \sim F$
$Y_i = 1(X_i > \theta)$
$Z_i = 1(X_i > x\theta)$

To calculate out k:

$$\frac{1/n \sum Z_i}{1/n \sum Y_i} = \frac{E(Z_i)}{E(Y_i)} = \frac{P(X_1 > x\theta)}{P(X_1 > \theta)} = \frac{\left(1 + \frac{k\theta}{\sigma}\right)^{\alpha}}{\left(1 + \frac{kx\theta}{\sigma}\right)^{\alpha}} = \frac{\left(\frac{1}{\theta} + \frac{k}{\sigma}\right)^{\alpha}}{\left(\frac{1}{\theta} + \frac{kx}{\sigma}\right)^{\alpha}}.$$

Then, as θ → ∞, this becomes:

$$\frac{\left(\frac{k}{\sigma}\right)^{\alpha}}{\left(\frac{k}{\sigma} x\right)^{\alpha}} = x^{-\alpha}.$$

For α:

$$x^{-\alpha}c = \frac{E(Z_i)}{E(Y_i)}$$

$$-\alpha \log x + \log c = \log\left(E\left(Z_i\right)\right) - \log\left(E\left(Y_i\right)\right)$$

In this case, -α = -1/k, so α = 1/k and c = 1.

Part 2, Equation 3

The second part of Equation 3, $f(k_1 | X_{21},\ldots,X_{2n(2)},\ldots,X_{91,\ldots},X_{9,n(9)})$, is:

$$f(k_1 \mid X_{2,1},...,X_{9,n(9)}) = \int_q f(k_1 \mid q, X_{2,1},...,X_{9,n(9)})dP(q) =$$

$$\int_q \int_{k_2,...,k_n} f(k_1 \mid q, k_2,...,k_n, X_{2,1},...,X_{9,n(9)}) f(q, \overbrace{X_2,...,X_{9,n(9)}}^{\tilde{X}})d\tilde{k}\,dq$$

CI.2

$$\propto \int_q \int_{\tilde{k}} f(k_1 \mid q) f(\tilde{X} \mid q, \tilde{k}) f(q, \tilde{k}) d\tilde{k}\,dq \text{, where } \tilde{k} = (k_2,...,k_9)$$

$$\propto \int_q \int_{\tilde{k}} f(k_1 \mid q) f(\tilde{X} \mid q, \tilde{k}) \overbrace{f(\tilde{k} \mid q)d\tilde{k}\,dF(q)}^{\prod_{i=2}^{9} f(k_i \mid q)dk_i\,dF(q)}$$

$$\int_q \int_{\tilde{k}} f(k_1 \mid q) \prod_{j=2}^{9} \prod_{i=1}^{n(q)} f(X_{i,j} \mid k_j) \prod_{j=2}^{9} f(k_j \mid q) d\tilde{k}\,df(q)$$

Combining both the parts of Equation 3, $f(k_1 | X_{11},\ldots,X_{1n(1)},\ldots,X_{91,\ldots},X_{9,n(9)})$ is equal to

$$\prod_{j=1}^{n(1)} \underbrace{f(X_{1j} \mid k_1)}_{part1} \int_q \int_{\tilde{k}} \underbrace{f(k_1 \mid q)}_{part2} \prod_{j=2}^{9} \prod_{i=1}^{n(9)} \underbrace{f(X_{i,j} \mid k_j)}_{part3} \prod_{j=2}^{9} \underbrace{f(k_j \mid q)}_{part4} d\tilde{k}\,dF(q) \qquad \textbf{Equation 6}$$

Equation 6 can now be solved in parts. Parts 1 and 3 are likelihood functions distributed according to the GPD, with shape parameter k and scale parameter σ. Parts 2 and 4 are also similar to each other. Let $f(k_i \mid q)$ be distributed according to some distribution, which will be discussed later. The distributions of the wave heights can be used to get a heuristic prior, q, from estimates of $k_{1,\ldots,}k_9$. Once we have the distribution of the wave heights we need to get the height at the appropriate quantile. This quantile is :

Start with the inverse function of the GPD:

$$F^{-1}(x \mid k, \sigma, \theta) = \frac{\sigma(1-(1-x)^k)}{k} + \theta$$

$X_1,\ldots X_n$ separated maxima

$P(X > c) = 1 - F(c)$,

Or $N_t \sim$ Poisson(λ), where N_t is the total number of observations up to time t. In our case, t = 24 years.

$E(N_t) = \lambda t \Rightarrow \lambda \approx E(N_t)/t$

For t = 24, let $n_t = m$.

$E(n_t) = n_t = \lambda t \Rightarrow n_t/t = m/24$

We know that $X_n = F^{-1}\left(1-\frac{1}{n}\right)$

In our case:

$$X_{E(N_{10,000})} = F^{-1}\left(1-\frac{1}{E\left(N_{10,000}\right)}\right) = F^{-1}\left(1-\frac{1}{10,000\cdot\frac{m}{24}}\right)$$

$$= F^{-1}\left(1-\frac{24}{10,000\cdot m}\right)$$

N is Poisson because, as the number of observations is bigger than 50, the binomial distribution is approximated by Poisson. it simply counting the number of X_i's above the value, c. Here, m is the number of events, that is, it is the number of peaks above the threshold.

Before applying hierarchical Bayes, appropriate thresholds must be chosen. We would like a threshold that results in an average of around two storms per year, but are still independent. The peaks are a minimum of 24 hours apart for independence. Below are the thresholds and corresponding peaks for each station.

Threshold and Peaks for Each Station

Station	ELD	EUR	K13	LEG	MPN	SCW	SON	SWB	YM6
Number of Peaks(m)	57	64	61	57	60	56	58	60	59
Threshold, θ [cm]	539	459	529	459	439	349	504	414	499

Table 2

Using the Data to get a Heuristic k

The first and third parts of Equation 6 need an estimate for k.
From Equation 3, we know that to calculate $P(X_{1j} \mid k_1)$

$$P(X > x\theta \mid X > \theta) = \frac{P(X > x\theta)}{P(X > \theta)} = \frac{1 - F(x\theta)}{1 - F(\theta)}$$

Let $\frac{E(Z_i)}{E(Y_i)}$ where $\begin{matrix} Y_i = 1(X_i > \theta) \\ Z_i = 1(X_i > x\theta) \end{matrix}$, so $\frac{P(X > x\theta)}{P(X > x)} = \frac{E(Z_i)}{E(Y_i)}$, and from Equation 6 this equals

$$x^{-\alpha} c = \frac{E(Z_i)}{E(Y_i)} \qquad \textbf{Equation 7}$$

$$-\alpha \log x + \log c = \log(E(Z_i)) - \log E(Y_i)$$

Where c = 1 and k = 1/α.

Approach 1: Direct Fitting

From above, we will try and fit an α to $x^{-\alpha} = \frac{E(Z_i)}{E(Y_i)}$ directly thereby getting a solution for k. Yi, the total number of X_i that is above the threshold is 57 out of a total of 4,965 peaks over threshold. This makes $E(Y_i) = 57/4{,}965 = 0.0115$. $E(Z_i)$ is not so simple as it depends on x. Since, for each x there is a different answer, $E(Z_i)$ is a function that is dependent on x, shown below in Figure 4:

Poorwa Singh
1242326

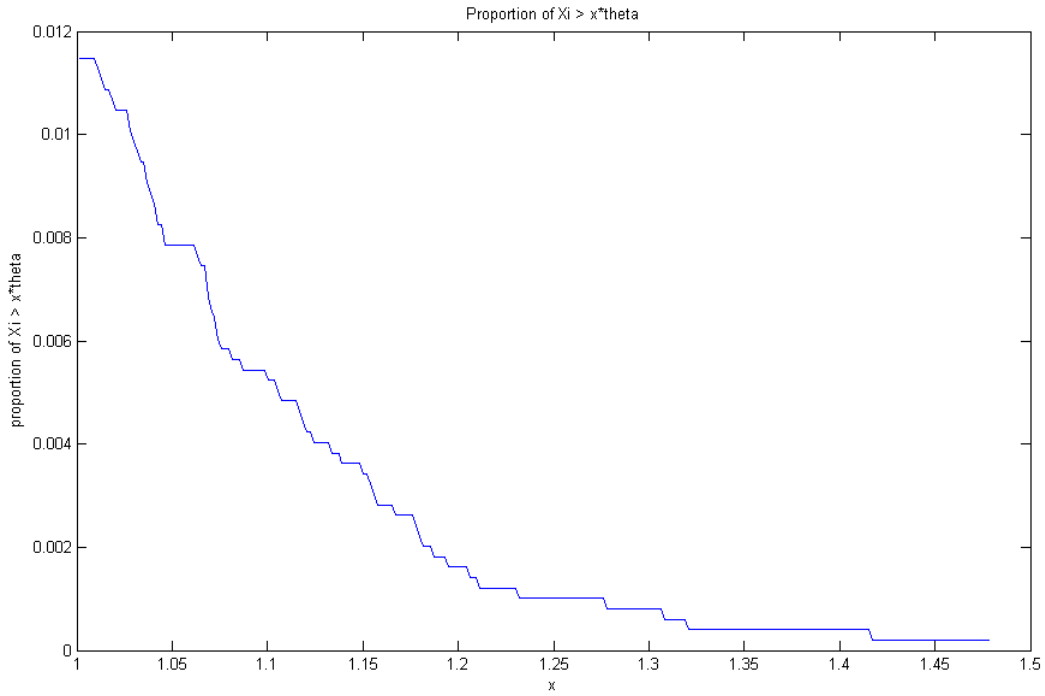


Figure 4

From above we have the following equations:

$$x^{-\alpha}c = \frac{E(Z_i)}{E(Y_i)}$$

$$-\alpha \log x + \log c = \log(E(Z_i)) - \log E(Y_i)$$

Since c = 1, the points can be linearly fitted directly for an α.

$-\alpha \log x = \log E(Z_i) - \log E(Y_i) = \log E(Z_i) + 4.4671$.

Below is the plot of log x against $\log E(Z_i) - \log E(Y_i)$.

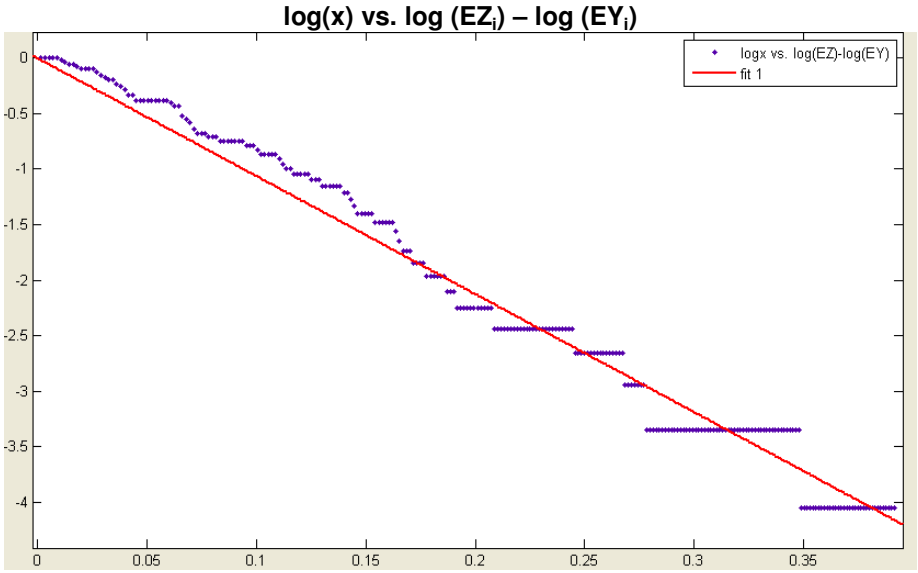


Figure 5

To find the best α, a line is fitted to the above to Figure 5. For a linear polynomial, the following formula is used: f(x) = p1*x + p2. In this case, x is our log(x) and p2 is forced to be zero. The equation we used is: f(log(x)) = -α log(x). Using the best fit, the p1 is -10.63 = -α, hence α is 10.63. See appendix A.6.
To see how well this fits, a QQ-plot is drawn. See Figure 6 below:

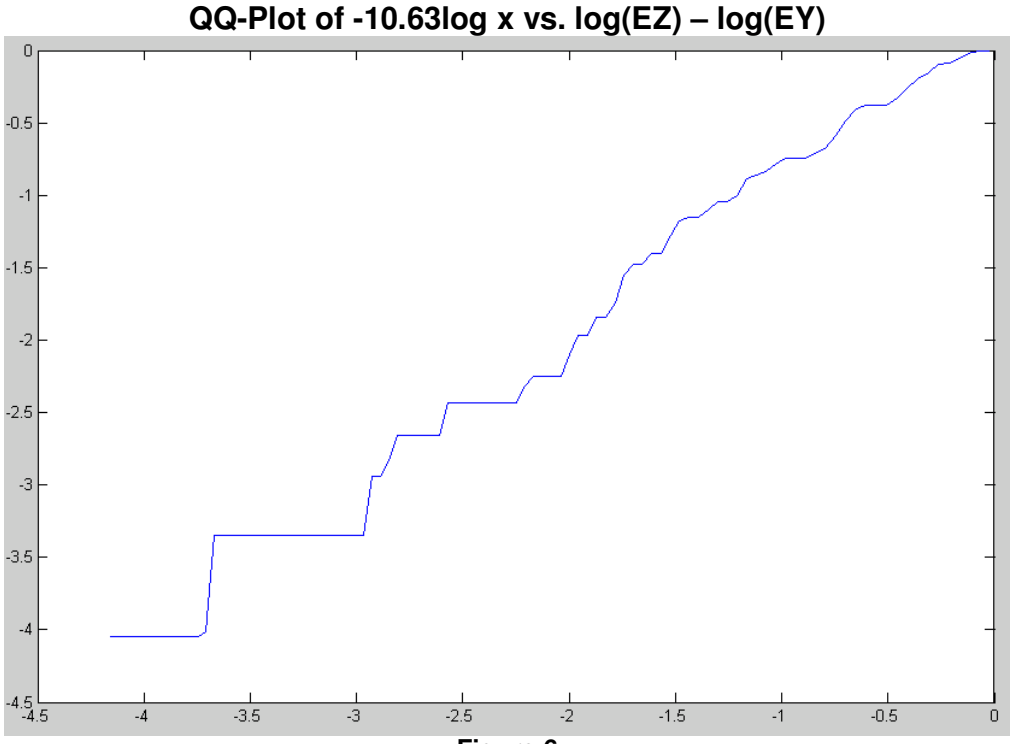


Figure 6

This plot shows that the quantiles are very similar, especially after the first part, of plot Figure 5. One can see this, as the QQ-plot quickly goes towards y = x.

Appendix A.7 shows fits for the other stations. The α and k values are listed below in Table 3.

Estimates of k and their R-squared values

Station	-α log x = log(EZ) – log(EY)		
	α	k = 1/α	R-square
ELD	10.63	0.09	0.98
EUR	11.27	0.09	0.98
K13	9.68	0.10	0.92
LEG	10.05	0.10	0.98
MPN	11.08	0.09	0.90
SCW	11.53	0.09	0.88
SON	8.227	0.12	0.96
SWB	12.15	0.08	0.92
YM6	9.57	0.10	0.94

Table 3

With these k's, calculating the σ's of the GPD is now possible. Where W = X – θ, for X > θ.

The expected value of W is:

$$EW = \int_0^\infty w \cdot \frac{1}{\sigma}\left(1 + k\frac{w}{\sigma}\right)^{-1-1/k} dw = \frac{\sigma}{1-k}$$

We also know that $E(W) = \frac{\sum W_i}{n}$, which can be approximated using the data.

Estimates of k and their R-squared values

Station	μ ≈ 1/n∑Wi	θ	-α log x = log(EZ) – log(EY)	
			k = 1/α	σ
ELD	61.96	539	0.09	56.13
EUR	45.38	459	0.09	41.35
K13	67.75	529	0.10	60.75
LEG	56.74	459	0.10	51.09
MPN	53.78	439	0.09	48.93
SCW	37	349	0.09	33.79
SON	74.53	504	0.12	65.47
SWB	40.07	414	0.08	36.77
YM6	61.59	499	0.10	55.15

Table 4

To check to see if these k's and σ's are close to our data, the empirical cdf, using the peaks, is plotted against the theoretical cdf, using the above k's and σ's. In Figure 7 there is an example of such a plot. Notice that these parameters are a pretty good fit. The other stations are in appendix A.8

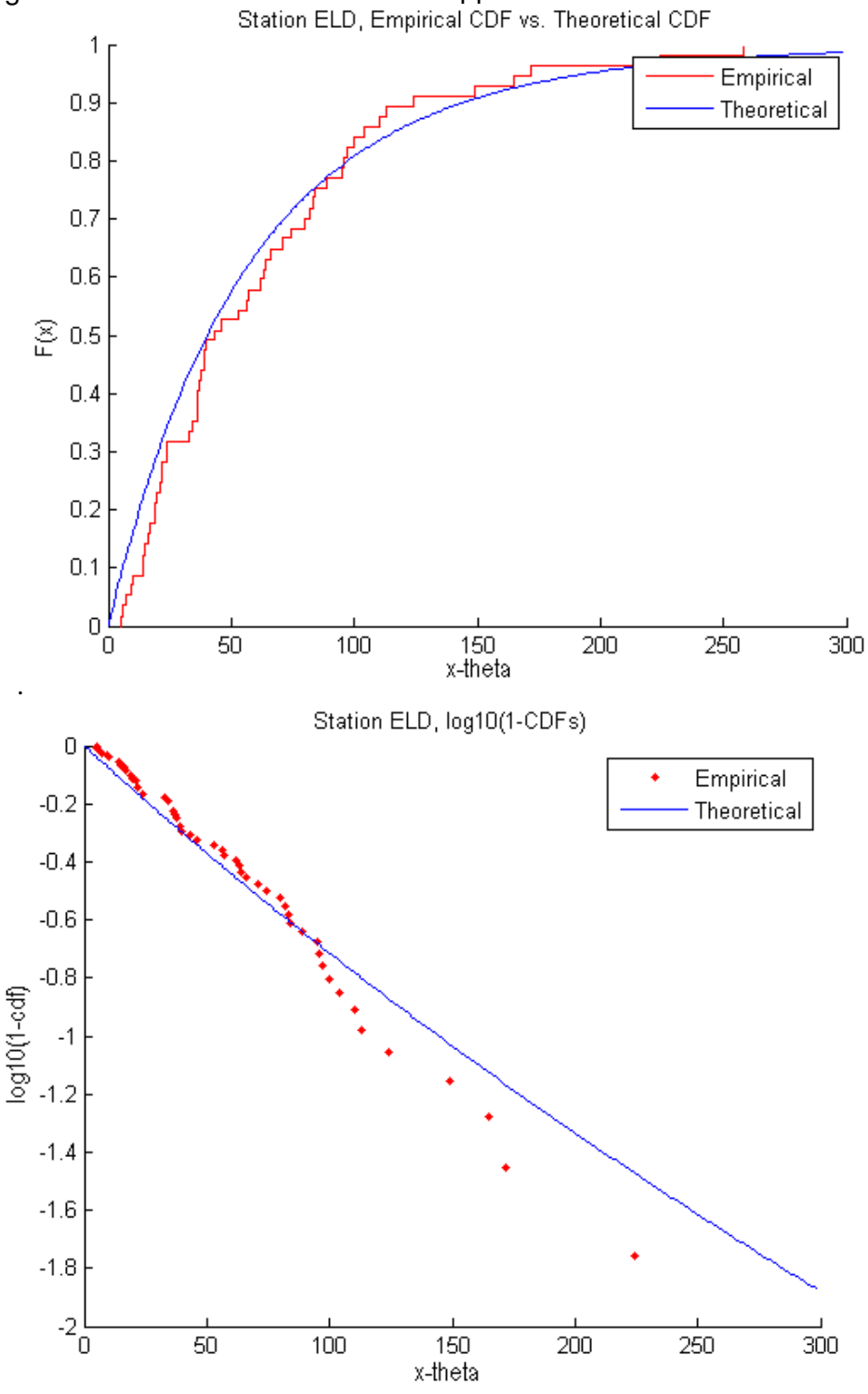


Figure 7

The above figures show that the GPDs using the fitted k's and σ's look pretty good. The other stations are in A.8. To make sure, the Kolmogorov-Smirnov (KS) goodness of fit test is performed for each station. See Table 5.

Goodness of Fit

	ELD	EUR	K13	LEG	MPN	SCW	SON	SWB	YM6
h	0	0	0	0	0	0	0	0	0
p-value	0.26	0.94	0.57	0.31	0.36	0.39	0.26	0.25	0.53
ks-stat	0.13	0.06	0.10	0.13	0.12	0.12	0.13	0.13	0.10
crit.val	0.18	0.17	0.17	0.18	0.17	0.18	0.18	0.17	0.17

Table 5

Table 5 shows the goodness of fit values for these k's and σ's. Most of the fits are good. h tells us whether or not to reject the null hypothesis that the empirical distribution is drawn from the corresponding theoretical GP distribution. If h = 0, then we do not reject the null hypothesis at significance level alpha, and if h = 1, then we reject the null hypothesis at significance level alpha. All but one of the p-values is above the 5% level. This means that the null hypothesis, the empirical distribution is drawn from the theoretical distribution with the respective k's and σ's, is true all the time, except for station SCW.

Below are the predicted one in 10,000 year waves for each of the stations using only the respective GPDs.

Predicted Wave Heights for Each Station

	ELD	EUR	K13	LEG	MPN	SCW	SON	SWB	YM6
Height, above θ [cm]	942.79	685.16	1088.81	885.86	809.99	542.50	1298.22	581.39	990.21
Complete Height[cm]	1481.79	1144.16	1617.81	1344.86	1248.99	891.50	1802.22	995.39	1489.21

Table 6

Approach 2: Maximum Likelihood Approach

As the direct fitting approach did not give a good parameter estimates all the stations, another method to get better k's and σ's should be tried. A different way to calculate k and σ is using the maximum likelihood method:

The density of the Generalized Pareto distribution is:

$$f(x) = f(x,k,\sigma) = \frac{1}{\sigma}\left(1 + k\frac{x-\theta}{\sigma}\right)^{-1-\frac{1}{k}}$$

The likelihood function is:

$$L = \prod_{i=1}^{n} f(x_i,k,\sigma) = \prod_{i=1}^{n} \frac{1}{\sigma}\left(1 + k\frac{x_i-\theta}{\sigma}\right)^{-1-\frac{1}{k}} = \frac{1}{\sigma^n}\prod_{i=1}^{n}\left(1 + k\frac{x_i-\theta}{\sigma}\right)^{-1-\frac{1}{k}}$$

$$\ln L = -n\ln\sigma - \left(1+\frac{1}{k}\right)\sum_{i=1}^{n}\ln\left(1 + k\frac{x_i-\theta}{\sigma}\right)$$

$$\frac{\partial \ln L}{\partial \sigma} = \frac{-n}{\sigma} - \left(1+\frac{1}{k}\right)\sum_{i=1}^{n}\frac{1}{1+k\frac{x_i-\theta}{\sigma}}\left(-k\frac{x_i-\theta}{\sigma^2}\right) = 0$$

$$\frac{\partial \ln L}{\partial k} = \frac{1}{k^2}\sum_{i=1}^{n}\ln\left(1 + k\frac{x_i-\theta}{\sigma}\right) - \left(1+\frac{1}{k}\right)\sum_{i=1}^{n}\frac{1}{1+k\frac{x_i-\theta}{\sigma}}\frac{x_i-\theta}{\sigma}$$

Let $w_i = x_i - \theta$. After multiplying $\frac{\partial \ln L}{\partial \sigma}$ by σ and $\frac{\partial \ln L}{\partial k}$ by k:

$$-n + (k+1)\sum_{i=1}^{n}\frac{w_i}{\sigma + kw_i} = 0$$

$$\frac{1}{k}\sum_{i=1}^{n}\ln\left(1 + k\frac{w_i}{\sigma}\right) - (k+1)\sum_{i=1}^{n}\frac{w_i}{\sigma + kw_i} = 0 \qquad \textbf{Equation 8}$$

From the first of the above two equations 9:

$$(k+1)\sum_{i=1}^{n}\frac{w_i}{\sigma + kw_i} = n$$

Which is the same as:

$$\frac{(k+1)}{k}\sum_{i=1}^{n}\frac{(kw_i+\sigma)-\sigma}{\sigma + kw_i} = n \qquad \textbf{Equation 9}$$

From both of the above equations 9 and equation 10:

$$\frac{1}{k}\sum_{i=1}^{n}\ln\left(1 + \frac{kw_i}{\sigma}\right) = n \qquad \textbf{Equation 10}$$

From equation 10

$$\frac{k+1}{k}\left(n-\sum_{i=1}^{n}\frac{\sigma}{\sigma+kw_i}\right)=n$$

$$n\left(\frac{k+1}{k}-1\right)=\frac{k+1}{k}\sum_{i=1}^{n}\frac{\sigma}{\sigma+kw_i} \qquad \textbf{Equation 11}$$

$$n=(k+1)\sum_{i=1}^{n}\frac{1}{1+\frac{k}{\sigma}w_i}$$

Let a = k/σ and b = k, then equation 12 becomes

$$n=(b+1)\sum_{i=1}^{n}\frac{1}{1+aw_i} \qquad \textbf{Equation 12}$$

and equation 11 is:

$$\frac{1}{b}\sum_{i=1}^{n}\ln\left(1+aw_i\right)=n \qquad \textbf{Equation 13}$$

From equations 13 and 14:

$$b=n\left(\sum_{i=1}^{n}(1+aw_i)\right)-1=G(a,w_1,...,w_n)$$

$$b=\frac{1}{n}\sum_{i=1}^{n}\ln(1+aw_i)=H(a,w_1,...,w_n) \qquad \textbf{Equation 14}$$

Since $w_1,\ldots,w_n$ are known, we must find an a such that G – H = 0. Using that, k and σ can be calculated. The results are below in Table 7

.

Estimates of k and σ using MLE

	ELD	EUR	K13	LEG	MPN	SCW	SON	SWB	YM6
θ [cm]	539	459	529	459	439	349	504	414	499
$\hat{k}$	-0.15	-0.06	-0.29	-0.17	-0.33	-0.27	-0.09	-0.18	-0.11
$\hat{\sigma}$	71.40	47.98	88.26	66.25	72.15	47.34	81.32	47.12	68.45
θ-σ/k	1003.23	1301.35	836.71	854.16	654.93	525.65	1404.94	679.62	1124.24

Table 7

Table 8 shows the results of the Kolmogorov-Smirnov test to see how well these parameters actually fit.

Goodness of fit of MLE

	ELD	EUR	K13	LEG	MPN	SCW	SON	SWB	YM6
p-val	0.69	0.99	0.56	0.91	0.71	0.42	0.72	0.76	0.76
KS-stat	0.09	0.05	0.10	0.07	0.09	0.12	0.09	0.08	0.09

Table 8

Notice the p-values in table 8, are all well above the 5 % level, meaning that the empirical cdf is likely from the theoretical cdf.

Predicted Wave Heights [cm] for Each Station

	ELD	EUR	K13	LEG	MPN	SCW	SON	SWB	YM6
Height, above θ [cm]	365.66	370.95	290.94	322.19	208.61	164.73	538.64	221.56	418.54
Complete Height [cm]	904.66	829.95	819.94	781.19	647.61	513.73	1042.64	635.56	917.54

Table 9

Below, in Figure 8, is an example of the empirical and theoretical distributions. See appendix A.9 for all the figures from all the stations.

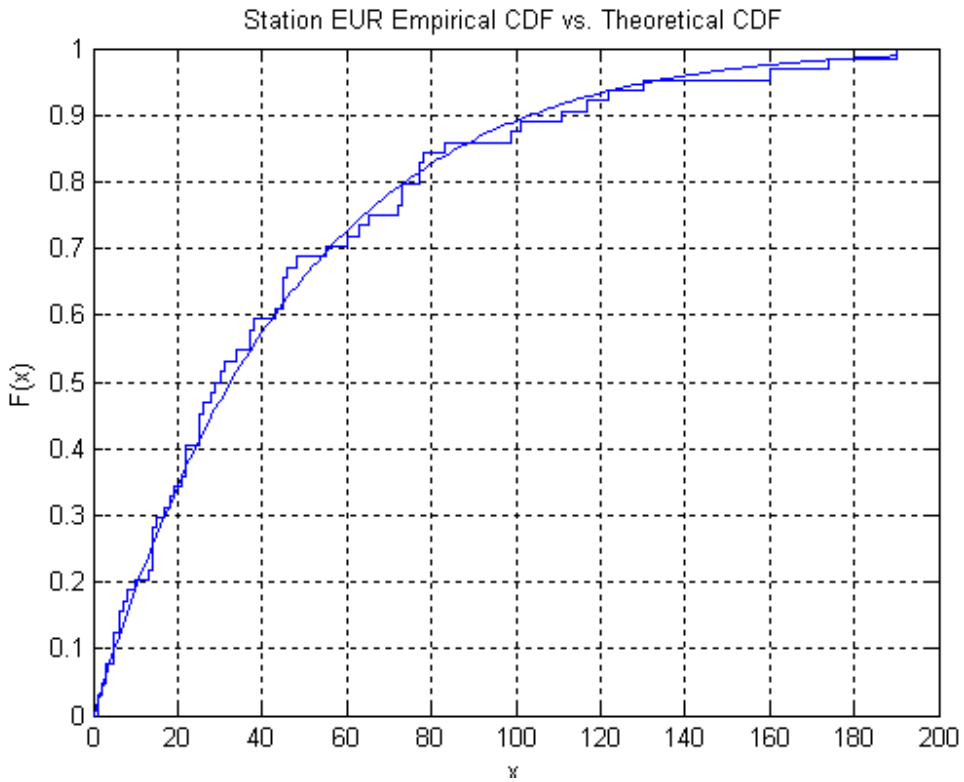


Figure 8

Approach 3: Method of Moments

The MLE method works well, but one more method is tried to see if a better fit can be achieved. This is the Method of Moments(MOM).The method of moments (MOM) for the GPD were introduced by Hosking and Wallis(1987). This method's basic idea is that estimators for unknown parameters can be derived from the expressions for the population moments. The r-th moments of the GPD exists if k < 1/r. Provided that they exist, the mean and variance of the GPD are given by: (Beirlant et al., p. 150).

$$E(Y) = \frac{\sigma}{(1-k)},$$

$$\text{var}(Y) = \frac{\sigma^2}{(1-k)^2(1-2k)}$$

A sample of $Y_1,\ldots Y_{Nt}$ i.i.d. GP random variables is available. Where $Y_i = (X_i - \theta)$ = (original height of peak minus the threshold) The order statistics associated with $Y_1,\ldots Y_{Nt}$ are denoted by $Y_{1,Nt} \leq \ldots \leq Y_{Nt,Nt}$. Replace E(Y) with $\overline{Y} = \sum_{i=1}^{N_t} Y_i / N_t$ and var(Y) by $S_Y^2 = \sum_{i=1}^{N_t} (Y_i - \overline{Y})^2 / (N_t - 1)$. Using the above equations and the replacements, yields the following MOM estimators:

$$\hat{k}_{MOM} = \frac{1}{2}\left(1 - \frac{\overline{Y}^2}{S_Y^2}\right)$$

$$\hat{\sigma}_{MOM} = \frac{\overline{Y}}{2}\left(1 + \frac{\overline{Y}^2}{S_Y^2}\right)$$

Using these equations gets the following k's and σ's:

Estimates of k and σ using MOM

	ELD	EUR	K13	LEG	MPN	SCW	SON	SWB	YM6
$\hat{k}$	-0.18	-0.03	-0.16	-0.16	-0.28	-0.18	-0.07	-0.19	-0.07
$\hat{\sigma}$	72.81	46.88	78.75	66.01	68.77	43.80	79.41	47.80	42.74
θ-σ/k	943.5	2021.67	1021.19	871.56	684.61	592.33	1638.43	665.58	1109.57

Table 10

Then, using these k's and σ's as the parameters in the theoretical equation, the empirical vs. theoretical distributions are plotted. Figure 9 is a plot of station K13's empirical and theoretical distributions. The other stations are in appendix A.10.

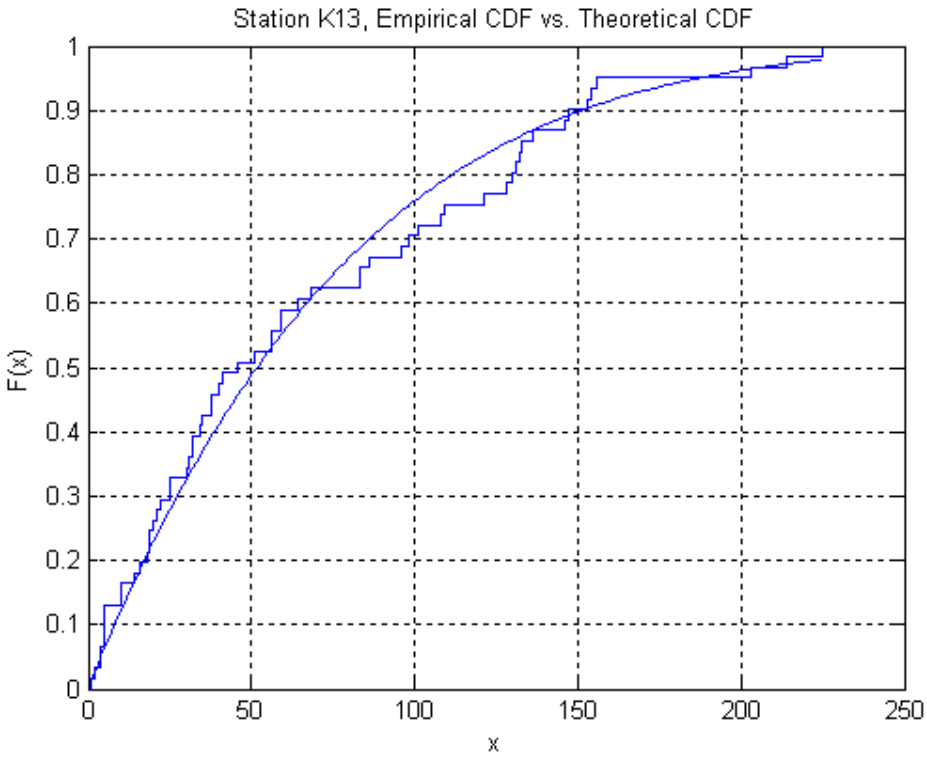


Figure 9

Again, this seems to fit pretty well. Since we cannot tell with simply the plot, the KS-test is done. Results are shown below in table 11.

Goodness of Fit of MOM

	ELD	EUR	K13	LEG	MPN	SCW	SON	SWB	YM6
h	0	0	0	0	0	0	0	0	0
p-value	0.7261	0.99	0.84	0.90	0.84	0.61	0.67	0.76	0.86
ks-stat	0.0899	0.44	0.08	0.07	0.08	0.10	0.09	0.09	0.08
Crit.val	0.18	0.17	0.17	0.18	0.17	0.18	0.18	0.17	0.17

Table 11

Notice that the parameters do not fit as well as the MLE ones. So of the three methods, MLE is the best estimator of the k's and σ's. Also, the heights of the desired one in 10,000 year wave, predicted simply using one station, are below in Table 12.

Predicted Wave Heights for Each Station

	ELD	EUR	K13	LEG	MPN	SCW	SON	SWB	YM6
Height, above θ [cm]	344.62	405.40	391.76	326.02	232.11	200.79	586.56	212.54	483.25
Complete Height [cm]	883.62	864.4	920.76	785.02	671.11	549.79	1090.56	626.54	982.25

Table 12

Poorwa Singh
1242326

The hyperprior, q

After calculating a good estimate for k, the next part of equation 6 must be calculated.

$$\prod_{j=1}^{n(1)} \underbrace{f(X_{1j} \mid k_1)}_{part1} \int_q \int_{\tilde{k}} \underbrace{f(k_1 \mid q)}_{part2} \prod_{j=2}^{9} \prod_{i=1}^{n(9)} \underbrace{f(X_{i,j} \mid k_j)}_{part3} \prod_{j=2}^{9} \underbrace{f(k_j \mid q)}_{part4} d\tilde{k}\, dF(q)$$

To do this, a hyperprior distribution, q, must first be determined. Usually, a hyperprior is estimated by a conjugate class, unfortunately, there is no conjugate class for GPD's.

Using the few data points we have, we check to see which distribution fits k the best.

As all k's are negative, they must be modified slightly to be able to use certain distributions.

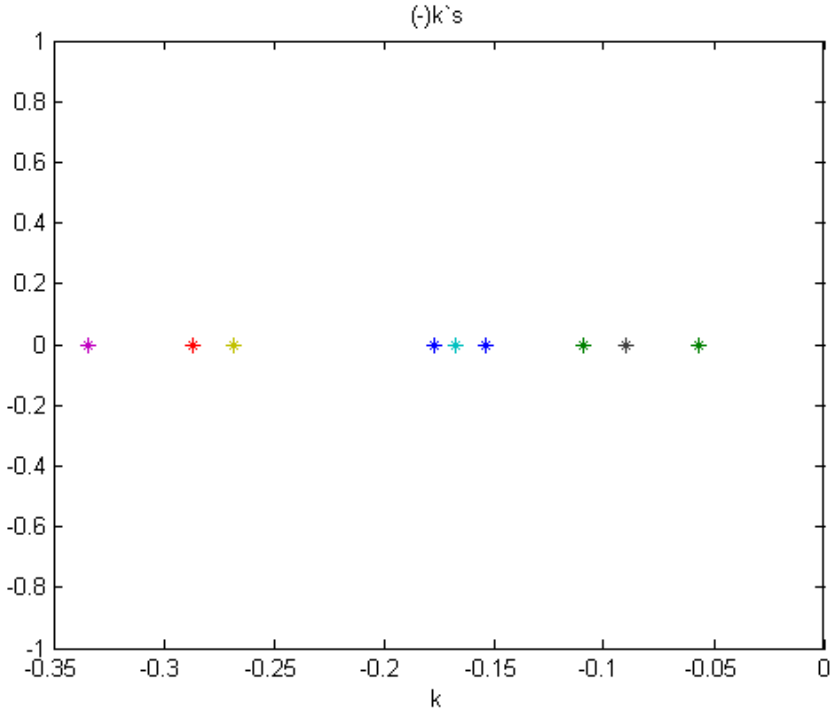


Figure 10

Results from KS-Test for different distribution fits

Distribution fits for k

		h	p-value	ks-statistic	critical value
k	**Normal**	1	0.01	0.51	0.43
	Ext. value	1	0.00	0.55	0.43
Negative k	**Exponential**	0	0.42	0.28	0.43
	Gamma	1	0.00	0.99	0.43
	Normal	1	0.01	0.51	0.43
	Ext. value	0	0.22	0.33	0.43
k minus minimum	**Exponential**	0	0.29	0.31	0.43
	Normal	1	0.01	0.51	0.43
	Ext. value	1	0.00	0.55	0.43

Table 13

The data with the best fit is the negative of the original k's. Let G = -K, the best fit to G is the exponential distribution. To give the exponential distribution more flexibility, the gamma distribution will be used. Recall, when α of the gamma distribution is equal to 1, it is the exponential distribution. Let $\alpha = 1/v^2$ and $\beta = v^2q$. Let G be distributed according to a gamma distribution, with parameters α and β.

$$f(g \mid \alpha, \beta) = \frac{1}{\beta^{\alpha}\Gamma(\alpha)} g^{\alpha-1} e^{-g/\beta}, \ \alpha > 0,$$

Where $E(G) = \alpha\beta = q$
And $Var(G) = \alpha\beta^2 = qv^2$
We need to calculate α and β. We know that the expected value of G, E(G) = -E(K) and the average of the MLE fits of $k_1,\ldots,k_9$ is $\frac{\sum_{i=1}^{9} k_i}{9} = -0.1827$. This value gives us an idea of a range for the q's to be in—that is if G ~ exponential, then E(G) = q, where q has a wide range, namely 0 to 10.

Now, Equation 7 is calculated for q ~ U[0,3], q ~ U[0,10] and q ~ U[3,7]. These q's are used to randomly generate many values of g's, for v = 0.1, 0.3, 0.5, 0.7, 0.9, 1, and 2. For each range of q and each v, Equation 7 is calculated.
Recall that:

$$f(k_1 \mid X_{11},\ldots,X_{1n(1)},\ldots,X_{91,\ldots},X_{9,n(9)}) =$$
$$\prod_{j=1}^{n(1)} f(X_{1j} \mid k_1) \int_{q}\int_{\tilde{k}} f(k_1 \mid q) \prod_{j=2}^{9}\prod_{i=1}^{n(9)} f(X_{i,j} \mid k_j) \prod_{j=2}^{9} f(k_j \mid q) d\tilde{k}\, dF(q)$$

The distribution of the (-)k_1's derived is then used to calculate the predictive distribution

$$F(X) = E_{k_1}\left(GPD\left(X, k_1, \sigma\right)\right) \qquad \textbf{Equation 15}$$

The fact that $\sigma = E\left(X-\theta\right)*\left(1-k\right)$ is used to get the appropriate σ's. Let $E\left(X-\theta\right) = \frac{\sigma}{1-k} = \mu - \theta$, so $\sigma = E\left(X-\theta\right)*\left(1-k\right) = E\left(\mu-\theta\right)*\left(1-k\right)$.

Here we run into a numerical problem, there are so many wave heights that product of f($X_{i,j}$ | k_j) is almost zero. One modification that can be made is when calculating the product, $\prod_{i=1}^{n(j)} f(X_{i,j} \mid k_j)$ directly does not work, take the log of it, sum the f($X_{i,j}$ | k_j) and then convert it back using the exponential function. This gives better results. Below, in figures 11, 12, and 13, are the outcomes of this approach.

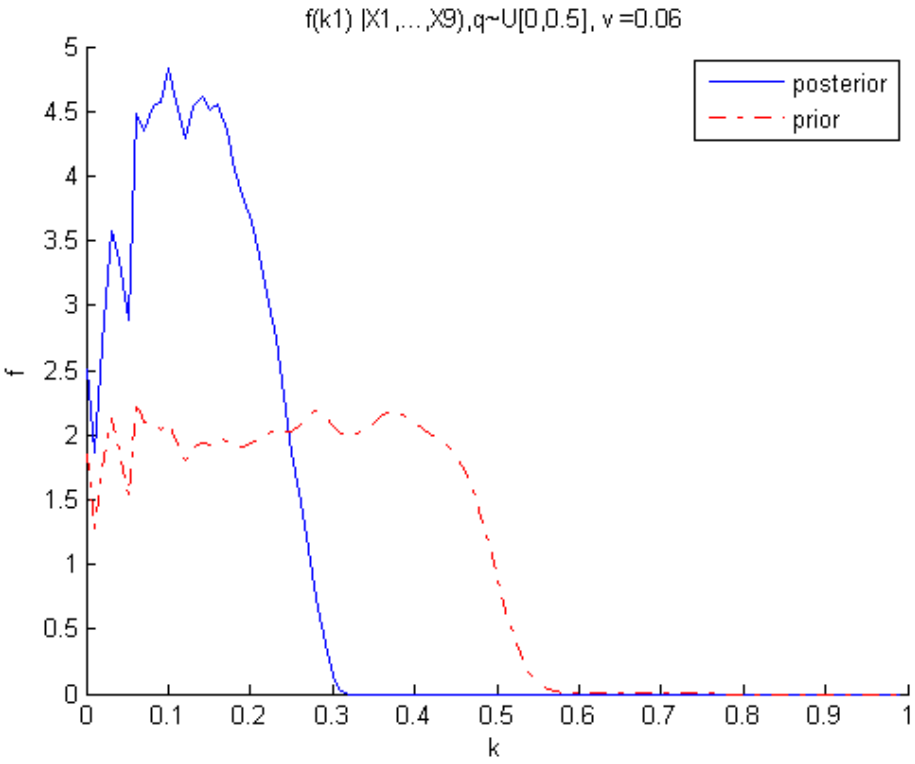


Figure 11

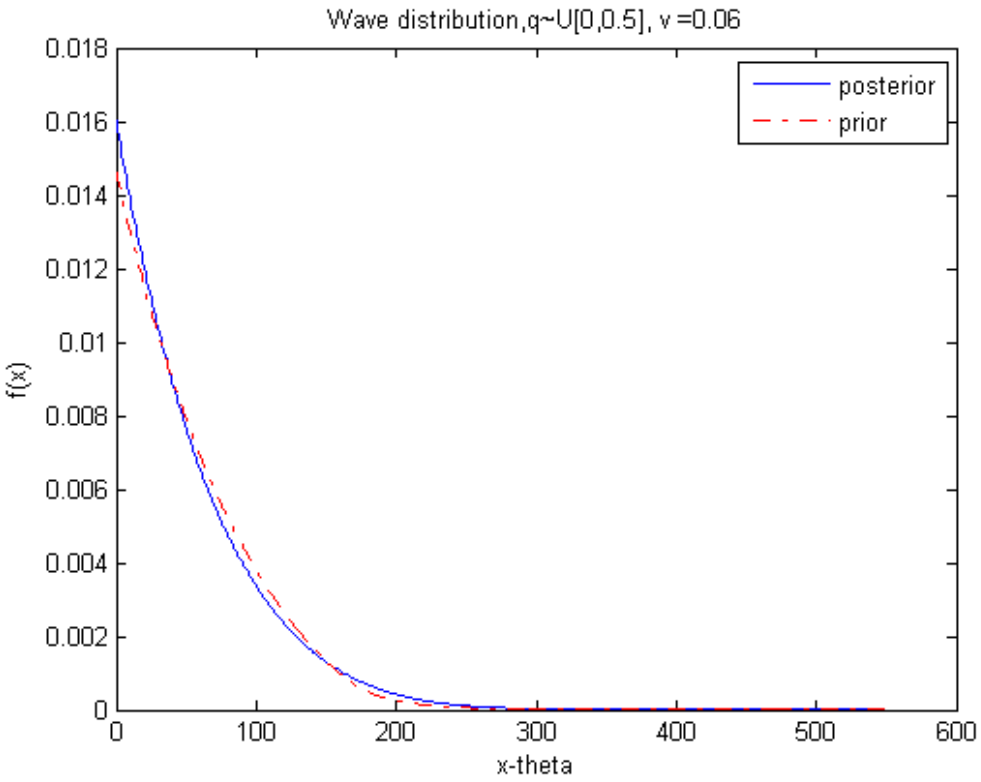


Figure 12

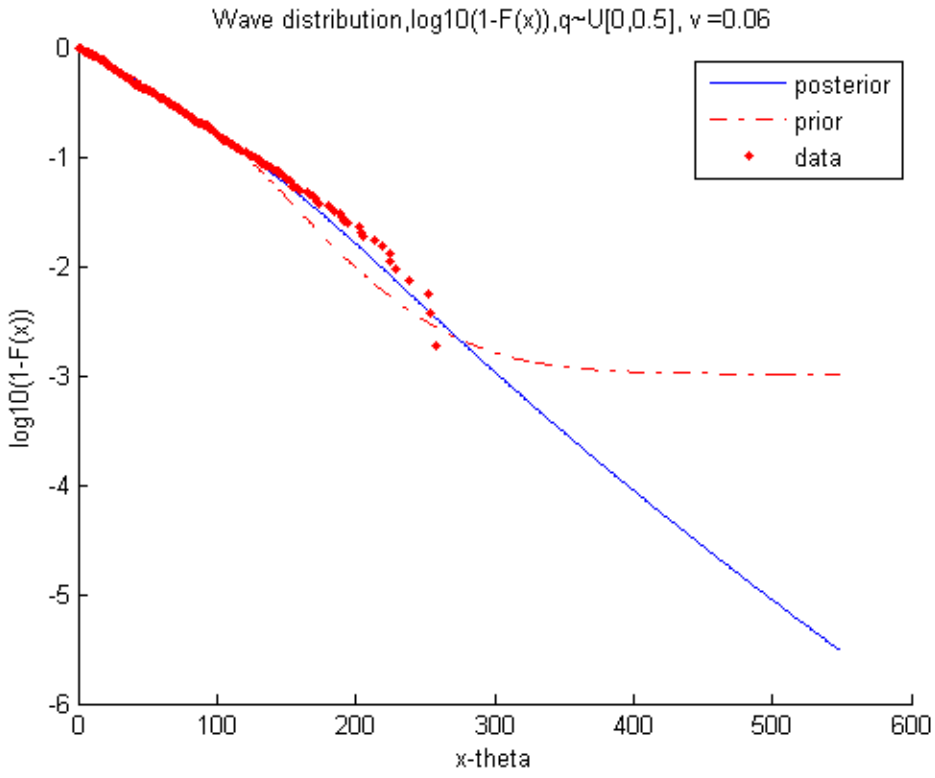


Figure 13

Figure 11 shows that the prior k's are not very good, but the posterior k's are between 0 and 0.3, which is where the most of the (absolute) k's from the data are. The wave heights of the prior and posterior in Figure 12 ,do not look much different from each other, but the difference can be seen in the log ccdf, Figure 13. Notice the prior is smaller than the posterior at the beginning, until x-θ is around 300, then it abruptly becomes bigger and stays that way. The data, on the other hand, does not increase with the prior, showing us that the posterior is indeed an improvement of the prior. Other stations are in A.12

1 in 10,000 Year Wave Heights [cm], using all the Data from each Station

	q~U[0,0.5]		q~U[0,3]	
v =	**prior**	**posterior**	**prior**	**posterior**
0.0005	1008.93	1149.65	601.29	1051.42
0.001	1040.49	1158.27	618.32	1086.03
0.02	960.13	1126.55	700.65	1266.39
0.04	989.58	1141.76	627.71	1120.58
0.06	975.42	1133.8	635.2	1128.68
0.08	988.34	1139.05	638.12	1139.27
0.1	985.12	1134.87	632.61	1128.55
0.2	977.58	1134.88	637.23	1135.74
0.5	1028.87	1169.93	669.41	1159.29
0.8	1081.65	1213.18	718.81	1186.43

Table 14

Table 14 shows the wanted wave heights, using all the data from each station. Notice that posteriors are much closer to each other than the priors, which are completely different. For q ~ U[0,0.5], this can be seen in Figure 13, where towards the end, the posterior is much lower than the prior. These wave heights are higher than the wave heights from just one station, using the MLE fit for the GPD, Table 9. This seems logical as this method uses the information from the other stations as well. The desired wave that would occur only once in 10,000 years is between eleven and twelve meters in height.

Conclusion

The Hierarchical Bayes model is an interesting approach to extreme value theory problems. Basically, first a threshold is decided for each station. The threshold should be high enough that it only allows for around two storms a year and makes the point independent of each other by disregarding all close points below the threshold. This results in our data of nine stations. Then, a GPD is fitted to each station by the MLE. This gives us the parameters, k and σ. Now, we have an idea about k (prior) and also an idea about the hyperprior q. Using these little

bits of information, we try solving equation 6 for several k's and q's. This procedure updates our initial inference, giving us a better estimate of the actual distribution of X. A few distributions of X are calculated and from these, the 1 in 10,000 year waves are determined.

Another Trial

Recall: $\prod_{j=1}^{n(1)} \underbrace{f(X_{1j} \mid k_1)}_{part1} \int_q \int_{\tilde{k}} \underbrace{f(k_1 \mid q)}_{part2} \prod_{j=2}^{9} \prod_{i=1}^{n(9)} \underbrace{f(X_{i,j} \mid k_j)}_{part3} \prod_{j=2}^{9} \underbrace{f(k_j \mid q)}_{part4} d\tilde{k}\, dF(q)$

Using Hierarchical Bayes, try a Beta(α,β) distribution for k instead of the Gamma. An advantage of the beta is that it is defined on a bounded interval. One adjustment has to be made to the Beta distribution. Since its range is [0,1] and we want it to be from [-0.5,0.5], let k ~ Beta(α,β) – 0.5. Also, α and β have to be greater than one, α > 1 and β > 1, because this makes the beta density concave, meaning that the k's are spread more in the middle. If α and β are less than one, then the beta density would be convex, implying the k's are accumulated at the ends.

α and β are determined by the fact that E(k) = q = $\frac{\alpha}{\alpha+\beta}$ (expected value of a beta distribution), and let v =Var(k) = $\frac{\alpha\beta}{(\alpha+\beta)^2(\alpha+\beta+1)}$. Then,

$$\alpha = \frac{-q*(-q+q^2+v)}{v} \; and \; \beta = \frac{(-1+q)*(-q+q^2+v)}{v} \qquad \textbf{Equation 16}$$

The q is the uniformly distributed hyperprior. The v affects the range of q. Since α >1 and β >1, so using the α's and β's from Equation 16 we get the following relationship:

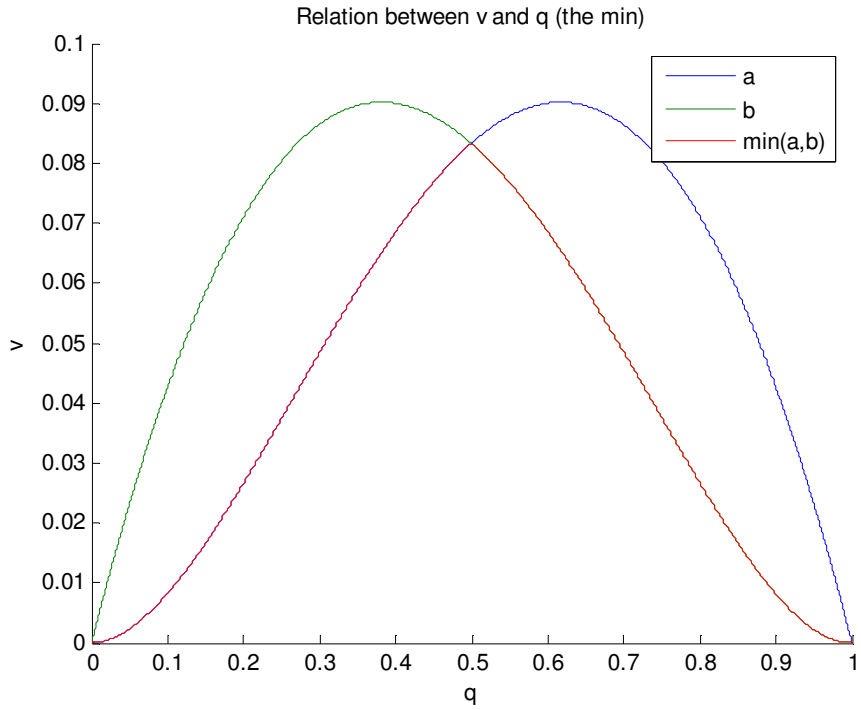


Figure 14

The relationship between v and q from α > 1, α: $v < \frac{q^2(1-q)}{1+q}$ and the relationship between v and q from β > 1, β: $v < \frac{q\left(q^2-2q+1\right)}{2-q}$. The minimum of (α,β) determines the bounds for q. v has to be less than 1/12. Hence, $v < \min\left\{\frac{q^2(1-q)}{1+q}, \frac{q\left(q^2-2q+1\right)}{2-q}\right\}$.

Then for the GPD part of the equation, we already have a distribution for the k's. We also know that $E(X-\theta \mid k) = \frac{\sigma}{1-k}$, where X | k ~GPD. Using the k's from the distribution, the σ's corresponding to them can be calculated by the following: $\sigma = (\hat{\mu}-\theta)(1-k)$.

This method uses the same hyperprior, q ~ Uniformly, and different priors, k | q ~ Beta (α, β). σ is related to the k's using $\sigma = (\hat{\mu}-\theta)(1-k)$. As a beginning, shown below are two trials of this method.

Two Examples

The examples of this method are shown below in figures 15-20 for v = 0.0005 and v = 0.001. Notice figures 15 and 16. The posterior accumulates around the MLE k's. Also, in figures 19 and 20, see how close the posteriors are to the data collected from the station. This shows the fit is indeed quite good. v = 0.0005 gives a prediction for the one in 10,000 year wave to be 410 cm over the threshold of 539 cm and v = 0.001 gives the prediction to be 420 cm above the threshold.

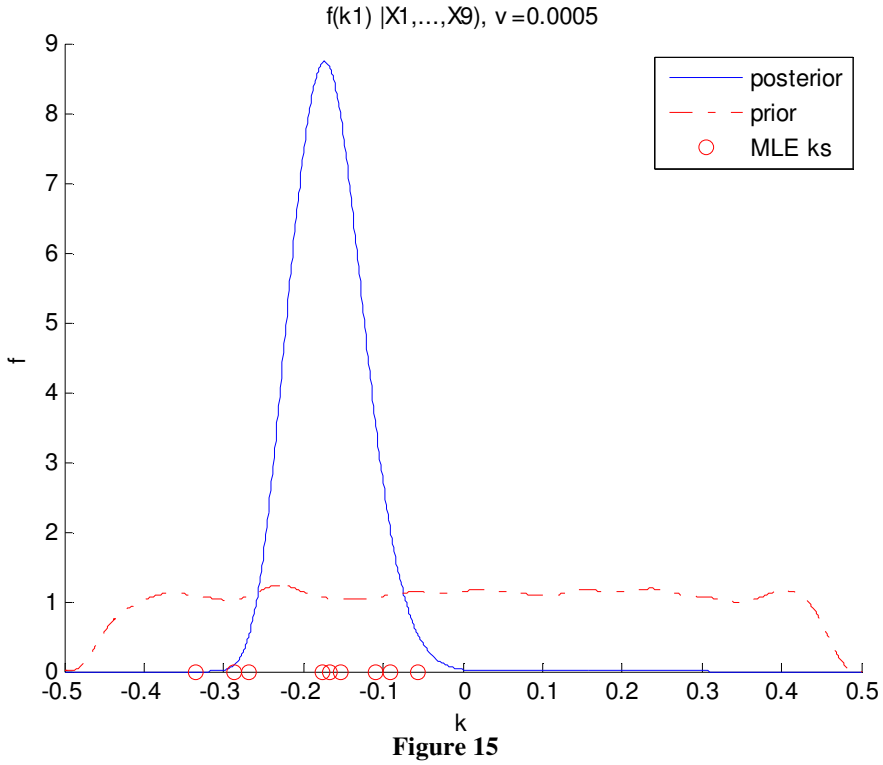


Figure 15

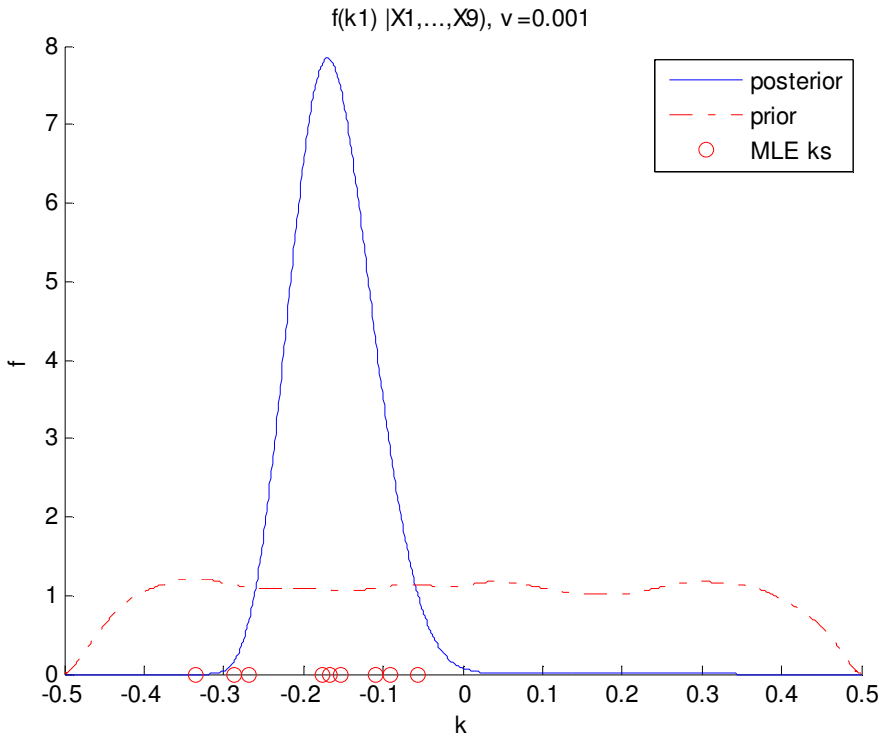


Figure 16

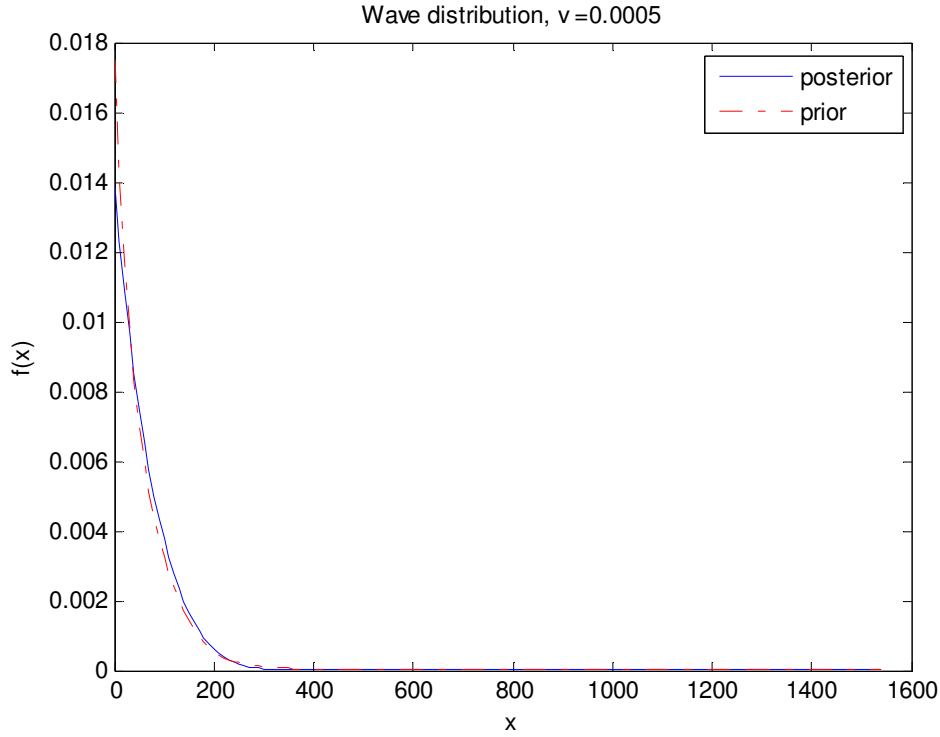


Figure 17

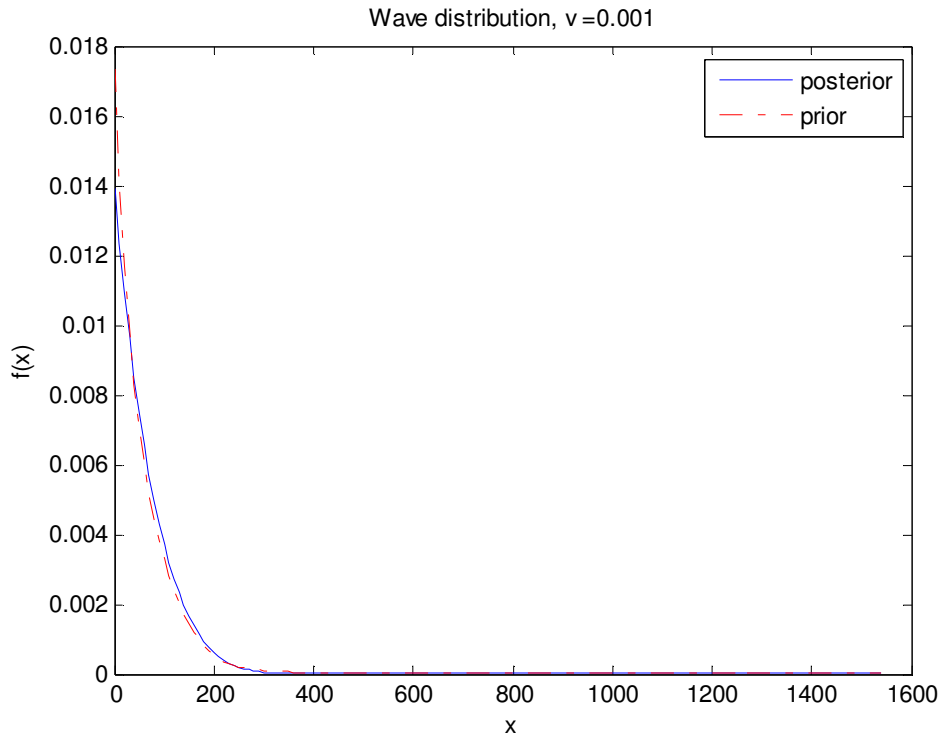


Figure 18

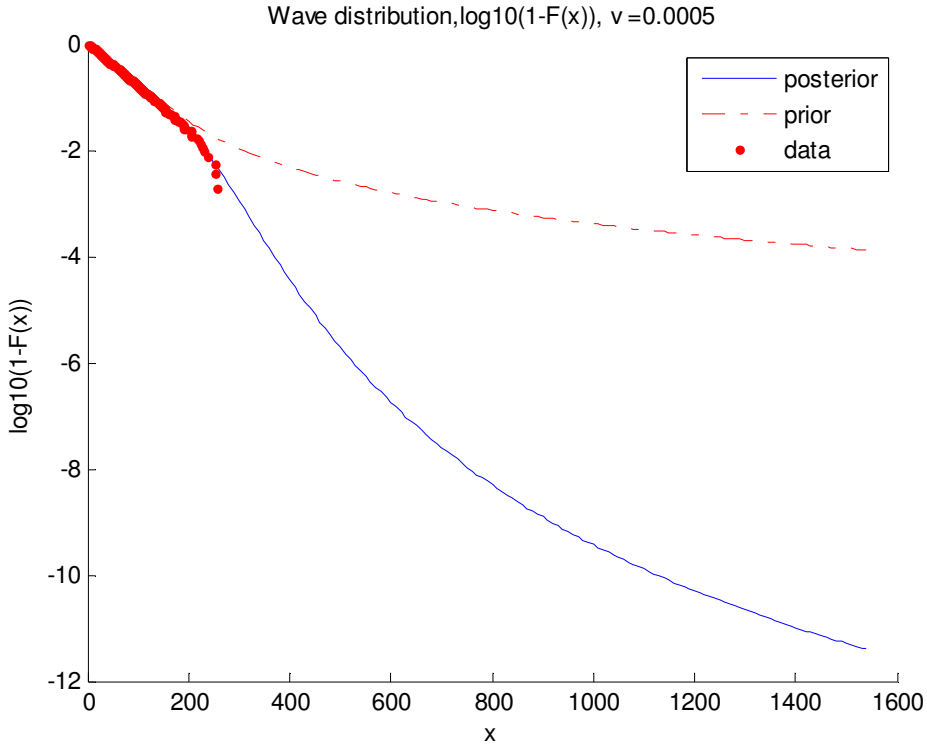


Figure 19

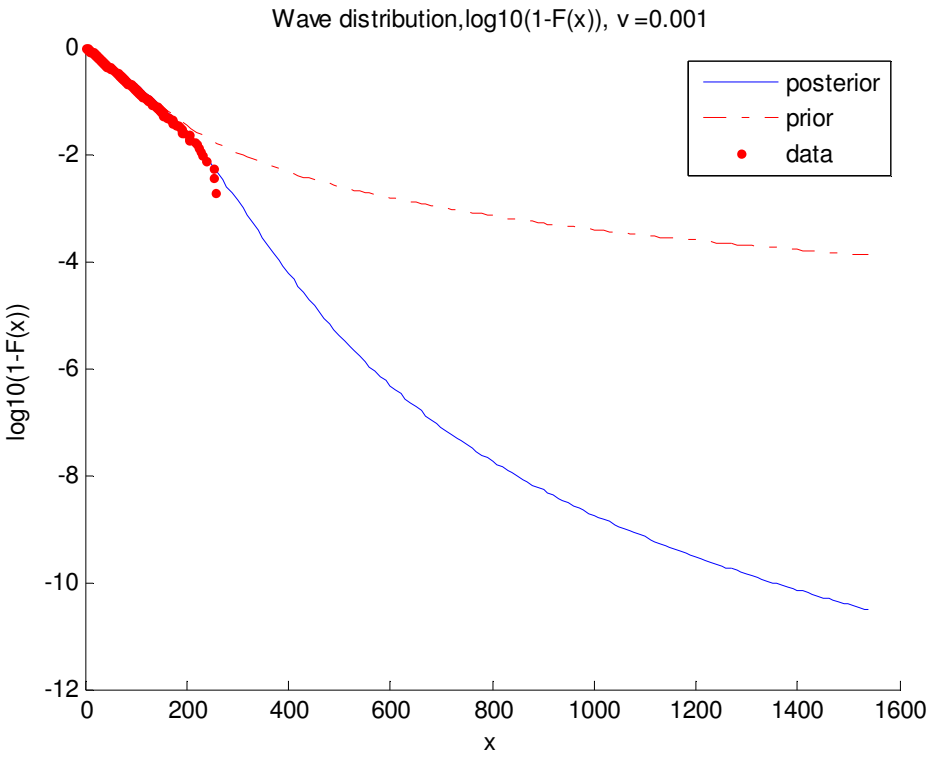

Wave distribution,log10(1-F(x)), v =0.001
posterior
prior
data
log10(1-F(x))
x
0
-2
-4
-6
-8
-10
-12
0
200
400
600
800
1000
1200
1400
1600

Figure 20

Conclusion

The objective of the thesis was to predict the height of the 1 in 10,000 year wave using twenty four years of data collected from the North Sea. Hence, extreme value statistics should be used. For this reason, normal regression was tried first. This resulted in a badly fitted model. The reasons could be that GPD only fit the data for the first station, ELD. It could also be that the 1000 predicted waves were not distributed normally, throwing off the regression analysis. This made most the β's insignificant (close to zero). Logically, this did not make sense as the correlations between the stations were relatively high.

Since the regression method did not work, a new method was tried—the two stage Bayes. This model takes all the information into account. It starts with the wave height data, which is used to estimate the k's, the shape parameters of the GPD. One step up, the priors of the k's, given q, were distributed according to the gamma distribution, and one more step up, the q's were distributed uniformly. Using all this information, the 1 in 10,000 year wave heights are predicted. The results seem reasonable because the distribution of the estimated wave heights, look like the distributions from the data of the stations.

Overall, the hierarchical Bayes method, although computationally intensive, worked better than traditional regression.

Poorwa Singh
1242326

Bibliography

1. Beirlant, Jan; Goegebeur, Yuri; Segers, Johan; and Teugels, Josef. *Statistics of Extremes, Theory and Applications.* John Wiley and Sons, Ltd. West Sussex, England. 2004.
2. Carlin, Bradley, P.; Louis, Thomas, A. *Bayes and Empirical Bayes Methods for Data Analysis.* Chapman & Hall. Suffolk, Great Britain. 1996.
3. Cooke, R.M., Bunea, C., Charitos, T., Mazzuchi, T.A. *Mathematical Review of ZEDB Two-Stage Bayesian Model*. 2002
4. Diebolt,Jean; El-Aroui, Mhamed-Ali; Garrido, Myriam; and Girard, Stephane. Institute National de Recherche en Informatique et en Automatique (INRIA). *Quasi-conjugate Byes estimates for GPD parameters and application to heavy tails modeling.* Montbonnot-St-Martin, France. 2003.
5. Documentation, Statistics Toolbox, Generalized Pareto Distribution. http://www.mathworks.com/access/helpdesk/help/toolbox/stats/bqem6vf-1.html Date accessed: 18/07/06
6. Embrechts, Paul; Kluppenberg, Claudia; and Mikosch, Thomas. *Modeling Extremal Events.* Springer. 1997.
7. http://www.golfklimaat.nl/index.cfm?page=uitleg.meetlocaties. Date accessed: 17/07/06.
8. Kotz, Samuel. Nadarajah, Saralees. *Extreme Value Distributions: Theory and Applications.* 2000. Imperial College Press. London.
9. On Tides and Weather. http://www.mikeladle.com/chapter4.html. Date accessed: 17/07/06
10. Wave Climate, Explanation. http://www.golfklimaat.nl/index.cfm?page=uitleg 05/2006. Date accessed: 17/07/06
11. Wave Climate, Explantion, Measuring. http://www.golfklimaat.nl/index.cfm?page=uitleg.meten 05/2006. Date accessed: 17/07/2006.

Poorwa Singh
1242326

Appendix

A.1

www.golfklimaat.nl

uitgebreide stationsnaam	meetnet	RD x	RD y	geografisch NB	geografisch OL	water-diepte m MSL
Aukfield platform	Noordzee	-	-	56°23'59"	02°03'56"	85
K13a platform	Noordzee	10.176	583.334	53°13'04"	03°13'13"	30
Schiermonnikoog noord	Noordzee	206.527	623.483	53°35'44"	06°10'00"	19
Eierlandse Gat	Noordzee	106.514	587.985	53°16'37"	04°39'42"	26
Ijmuiden mun.stortplaats	Noordzee	64.779	507.673	52°33'00"	04°03'30"	21
Noordwijk meetpost	Noordzee	80.443	476.683	52°16'26"	04°17'46"	18
Euro platform	Noordzee	9.963	447.601	51°59'55"	03°16'35"	32
Lichteiland Goeree	Noordzee	36.779	438.793	51°55'33"	03°40'11"	21
Schouwenbank	ZEGE	11.244	419.519	51°44'48"	03°18'24"	20
Scheur west Wandelaar	ZEGE	-7.797	380.645	51°23'32"	03°02'57"	15

A.2

Method 1

Procedure

The first method is the well-tried method of regression. The procedure for this is:

1. The peaks over threshold of one station must be selected
   a. This is done by first storing all peaks, regardless of height. A peak is defined by a point which is higher than the points immediately before and after it.
   b. A threshold is put in and all peaks below the threshold are removed.

   c. If there are two or more peaks within 24 hours of each other, the highest is taken.
2. Wave heights of the other stations at corresponding times, with time differences taken into account, must also be recorded
3. Using the peaks from the first station and the wave heights of the corresponding times from the other station, a rank correlation matrix and conditional correlation matrices are generated
4. With these correlations, generate Uniform data
5. Convert the Uniform data to the generalized Pareto (GP) distribution. x = F(U), where U is the Uniform data and F is the GP distribution.

Then, use this information to:

6. Sample for additional 24 years from each of the stations
7. Estimate the parameters of the GP
8. Compute one in 10,000 year wave for locations 1,..,k
   a. The prediction of the height of the one in 10,000 year wave: Start with the inverse function of the GPD:

$$F^{-1}(x \mid k, \sigma, \theta) = \frac{\sigma(1-(1-x)^k)}{k} + \theta$$

$X_1,\ldots X_n$ separated maxima

$P(X > c) = 1 - F(c)$,

Or $N_t \sim$ Poisson(λ), where N_t is the total number of observations up to time t. In our case, t = 24 years.

$E(N_t) = \lambda t \Rightarrow \lambda \approx E(N_t)/t$

For t = 24, let $n_t = m$.

$E(n_t) = n_t = \lambda t \Rightarrow n_t/t = m/24$

We know that $X_n = F^{-1}\left(1-\frac{1}{n}\right)$

In our case:

$$X_{E(N_{10,000})} = F^{-1}\left(1-\frac{1}{E\left(N_{10,000}\right)}\right) = F^{-1}\left(1-\frac{1}{10,000\cdot\frac{m}{24}}\right)$$

$$= F^{-1}\left(1-\frac{24}{10,000\cdot m}\right)$$

N is Poisson because, as the number of observations is bigger than 50, the binomial distribution is approximated by Poisson. Here, m is the number of events, that is, it is the number of points above the threshold. The result of Equation 1 is, given a number of events m, in 24 years, that is the 1 in 10,000 year wave.

9. Calculate β's using regression analysis: $Y_{TRUE} = \sum_i \beta_i X_{10000,i}$ (samples)

   Where Y_{TRUE} is the extrapolated height of the one in 10,000 year wave from the simulated waves, for the base station, and $X_{10000,i}$'s are the extrapolated waves from the simulated data, for all the other stations.
10. Repeat 1000 times to get a distribution for the one in 10,000 year wave, for each station.

Background: Regression Analysis

Regression analysis is "the study of the analysis of data aimed at discovering how one or more variables (called independent variables, predictor variables, or regressors) affect other variables (called dependent variables or response variables)."(golfklimaat.nl, p.1)

Basically, it is used when there is a lot of data and a model must be fitted to it to summarize the data more affectively. The regression model looks like:
$y = X\beta + \varepsilon$ where

$$y = \begin{pmatrix} y_1 \\ y_2 \\ . \\ . \\ yn \end{pmatrix}, \varepsilon = \begin{pmatrix} \varepsilon_1 \\ \varepsilon_2 \\ . \\ . \\ \varepsilon_n \end{pmatrix}, X = \begin{pmatrix} 1 & x_{11} & x_{12} & . & x_{1m} \\ 1 & x_{21} & x_{22} & . & . \\ . & . & . & . & . \\ . & . & . & . & x_{n-1,m} \\ 1 & x_{n1} & . & . & x_{nm} \end{pmatrix}, \beta = \begin{pmatrix} \beta_0 \\ \beta_1 \\ . \\ . \\ \beta_m \end{pmatrix} \text{ (p.28)}$$

Here, n=10,000 and m=8, so y, ε, and β are columns of 9 and X is a 10,000 by 9 matrix.
The y column is the response column, the x's are the regressor variables, β's are the weights of the regressor variables, and ε's are the residuals. The residuals are the differences between the actual data, y's and the model, Xβ. ε =y – Xβ.

When using regression analysis, it is not only important to fit the data, but also to see how good your model actually is. Two different such measures are the R^2-statistic and the p-value of the F-statistic. Where
$R^2 = 1 - \sum_{i=1}^{n} \varepsilon_i^2 \Big/ \sum_{i=1}^{n} (y_i - \bar{y})^2$ This value always "lies between 0 and 1 and the closer it is to 1, the better the fit."(p.14)

The second statistic is the p-value of a t-test, which tests the null hypothesis that $\beta_j = 0$. If the p-value is less than the given significance level, then it is significant and we assume $\beta_j \neq 0$.

Application

According to the above procedure, steps 1 and 2 – peaks from station 1 and the corresponding wave heights from other stations are recorded. We now need to generate more data using the data we already have. This is most accurately done if the relationship between stations is also noted. This relationship is captured by the correlations and conditional correlations of the data. Hence, these correlations are calculated to generate more accurate uniformly distributed data. See below for these correlations

Product Moment Correlation Matrix for Actual Data, Threshold 449 cm

1.00	0.52	0.62	0.63	0.61	0.49	0.42	0.50	0.69
0.52	1.00	0.66	0.83	0.76	0.81	0.06	0.91	0.72
0.62	0.66	1.00	0.61	0.59	0.55	0.14	0.61	0.62
0.63	0.83	0.61	1.00	0.74	0.80	0.21	0.83	0.76
0.61	0.76	0.59	0.74	1.00	0.69	0.26	0.77	0.71
0.49	0.81	0.55	0.80	0.69	1.00	0.23	0.90	0.66
0.42	0.06	0.14	0.21	0.26	0.23	1.00	0.11	0.27
0.50	0.91	0.61	0.83	0.77	0.90	0.11	1.00	0.68
0.69	0.72	0.62	0.76	0.71	0.66	0.27	0.68	1.00

Table 1

Rank Correlation Matrix for Actual Data, Threshold 449 cm

1.00	0.48	0.56	0.56	0.57	0.46	0.36	0.47	0.60
0.48	1.00	0.63	0.84	0.77	0.80	0.07	0.90	0.70
0.56	0.63	1.00	0.57	0.54	0.55	0.12	0.62	0.59
0.56	0.84	0.57	1.00	0.74	0.80	0.22	0.83	0.72
0.57	0.77	0.54	0.74	1.00	0.69	0.18	0.77	0.70
0.46	0.80	0.55	0.80	0.69	1.00	0.24	0.89	0.65
0.37	0.07	0.12	0.22	0.18	0.24	1.00	0.12	0.21
0.47	0.90	0.62	0.83	0.77	0.89	0.12	1.00	0.67
0.60	0.70	0.59	0.72	0.70	0.65	0.21	0.67	1.00

Table 2

Notice that the product moment and rank correlations, from tables 1 and 2 respectively, are quite high. This means that given a point from set A, one would be able to predict a relatively narrow bound for where a point from set B(or any other set) would be. The product moment correlation matrix has higher values than the rank correlation matrix. Thus, when the data is ordered, it is less correlated than when it is not. If the data is not ordered the band for the prediction of one point given another point is more accurate. This seems reasonable, as the waves at the stations are more likely to be similar to each other according to the times the wave peaks occur, and not as similar to each other at just the heights of the peaks. Conditional correlations are in A.3.

Using these correlations and conditional correlations, 1000 sets of data are generated. Then, to make sure that the generated data are in fact as similarly related to each other as the actual data is, the rank correlation of the actual data and the rank correlation of the simulated data are compared. The relationship between stations of the original data and one of the generated data sets are quite similar. They are in A.4

This new data is then transformed into the GPD for its own station. That is, say for station A, a Uniform data, U, set is created. This set is transformed into the GPD with parameters k and σ, best fitted to the original data for station A. The new data,x, is distributed according to the GPD, $x = F(U)$, where $F \sim GP(k_{A,},\sigma_A,\theta_A)$.

Then, from each set of data, the one in 10,000 year wave is calculated using equation 1. The 1000 wave heights simulated for station ELD with a threshold of 449 look like Figure 1.

1000 predictions of a 1 in 10,000 year wave for station ELD

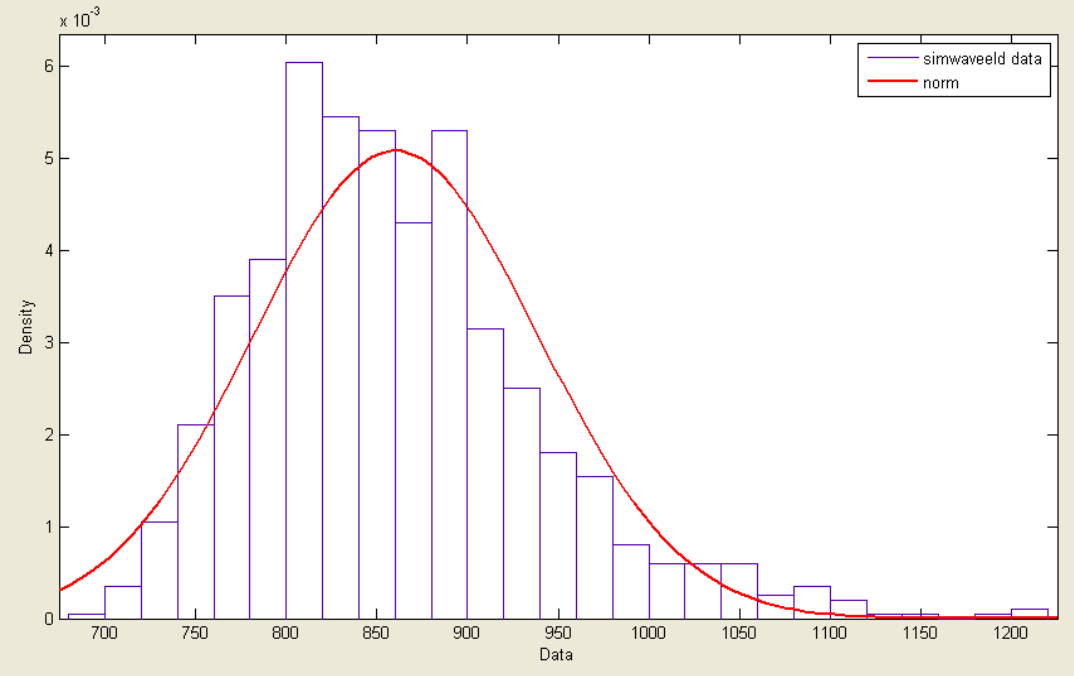


Figure 1

A Normal distribution is fitted to the 1000 simulated 1 in 10,000 year wave heights. The Normal distribution is chosen because we are no longer looking at peaks, but the height from a quantile of the distribution. This yields a mean of 860.41 cm and a standard error of 78.51cm. Unfortunately, this is not a good fit because it results in an unacceptable p-value from the Kolomogorov-Smirnov goodness of fit test of 9.4678e-004. The mean, standard deviations and p-values of the 1/10,000 year wave heights are below in Table 3. Recall, the p-value for the F-test tells us the chance one distribution comes from another. In this case, the chance the distribution of the data has a 0.09 % chance of coming from Normal(860.41, 78.51).

Simulations for a threshold of 449cm, 1000 trials, each with 160 peaks

Station	Mean	St. Error	p-value
ELD	860.41	78.51	9.47E-04
EUR	829.4	5.61	4.35E-06
K13	859.8	11.99	0.024
LEG	880.34	10.88	1.44E-05
MPN	830.24	16.44	0.14
SCW*	756.35	8.8	3.91E-05
SON*	1030.3	19.76	0
SWB*	838.03	10.83	0
YM6*	879.83	25.02	0.2

Table 3

*based on 999 trials.

Now, we must check the influence of each station on the first. This is done through the following regression analysis. In this case, y is a column of heights of the one in 10,000 year waves, from the simulations for station ELD. X is matrix where the first column is simply ones, for the β_0, and every column after that is a column of heights of the one in 10,000 year waves, from simulations, corresponding to stations EUR to YM6. Cross terms are not taken into account.

For a threshold of 449cm, the weights for stations, or β is:

Weights for Each Station for a Threshold of 449cm

1, β_0	EUR,β_1	K13,β_2	LEG,β_3	MPN,β_4	SCW,β_5	SON,β_6	SWB,β_7	YM6,β_8
-573.7	-0.68	-0.00	0.37	0.47	0.47	0.27	0.78	0.09

Table 4

This gives the regression equation:

$$y_{10,000,ELD} = -573.7582 - 0.6836 y_{10,000,EUR} - 0.0045 y_{10,000,K13} + 0.3728 y_{10,000,LEG}$$
$$+ 0.4698 y_{10,000,MPN} + 0.4394 y_{10,000,SCW} + 0.2653 y_{10,000,SON} + 0.7175 y_{10,000,SWB} + 0.0907 y_{10,000,YM6}$$

The biggest influence on station ELD is station SWB, which has a small, positive influence, and EUR which has a small negative influence. The fact that these two stations seem to influence ELD, the most seems a bit odd as neither SWB nor EUR are not close to station to ELD. The reason for this could be that the R^2 value for this analysis is 0.0379. This means that the fit is not good. R^2 always lies between 0 and 1, and the closer it is to 1 the better the fit. The p-value is 6.2228e-006. If the p-value is below 0.05, the null hypothesis can be rejected, implying a significant influence. Again, this is very small and supports the conclusion of this being a bad fit.

This analysis is performed for other thresholds to see other possible fits.

Simulations for a threshold of 509cm, 1000 trials, each with 82 peaks

Station	Mean	St. Error	p-value
ELD	865.14	103.16	5.36E-08
EUR	847.78	7.5	5.33E-18
K13	853.39	17.91	3.86E-04
LEG	889.4	13.11	4.29E-04
MPN	848.89	16.44	8.36E-04
SCW	808.41	7.86	6.08E-10
SON	1059	20.9	2.51E-04
SWB	821.48	7.87	3.10E-11
YM6	893.46	28.33	0.01

Table 5

Notice again that even with this higher threshold, the distribution of the 1 in 10,000 year wave for each station, fits quite badly. The best fit is for station YM6, with a p-value of 0.6 %.

For a threshold of 509cm, the weights for stations, or β is:

Weights for Each Station for a Threshold of 509cm

1, β_0	EUR,β_1	K13,β_2	LEG,β_3	MPN,β_4	SCW,β_5	SON,β_6	SWB,β_7	YM6,β_8
-466.22	-0.08	0.23	0.55	0.33	0.28	-0.00	0.13	0.11

Table 6

The station most influential on ELD this time is station LEG.
The R^2-statistic for this threshold is 0.0135. Again, this is not very good.
The p-value is 0.0962, which means that all the stations are insignificant, their β's can be taken as zero. The chart with all the β's, except β_0, is below in Figure 2.

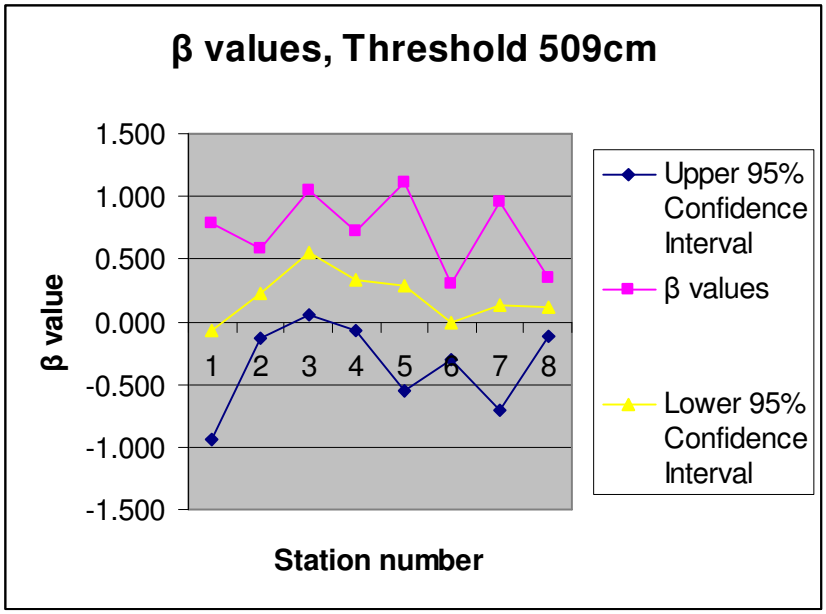


Figure 2

Notice that this time, no stations should be significant, but one is, β_3 (LEG). This implies that something is now wrong.

Now, a threshold of 529 cm is used.

Simulations for a threshold of 529cm, 1000 trials, each with 68 peaks

Station	Mean	St. Error	p-value
ELD	917	152.759	2.89E-09
EUR	852.8	7.24	9.31E-15
K13	870.8	16.38	3.41E-04
LEG	893.87	12.67	5.05E-07
MPN	867.35	14.55	1.70E-05
SCW	827.54	7.62	5.51E-12
SON	1076.9	23.23	2.50 e-004
SWB	840.4	7.06	5.85E-14
YM6	908.74	25.09	0.023

Table 7

Again, the p-values are very small, implying a bad fit.

For a threshold of 529cm, the weights for stations, or β is:

Weights for Each Station for a Threshold of 529cm

1, β_0	EUR,β_1	K13,β_2	LEG,β_3	MPN,β_4	SCW,β_5	SON,β_6	SWB,β_7	YM6,β_8
-1534.62	-0.11	0.77	0.52	0.34	0.07	0.29	0.58	0.28

Table 8

The most influential station this time is K13. This actually makes sense as K13 is quite close to ELD, in location, compared to the other stations.
The R^2-statistic for this threshold is 0.0173 which is also not very good. The p-value is 0.0262, again this means that at least one station has a significant effect on station ELD.

The chart with all the β's, except β_0, is below in Figure 3.

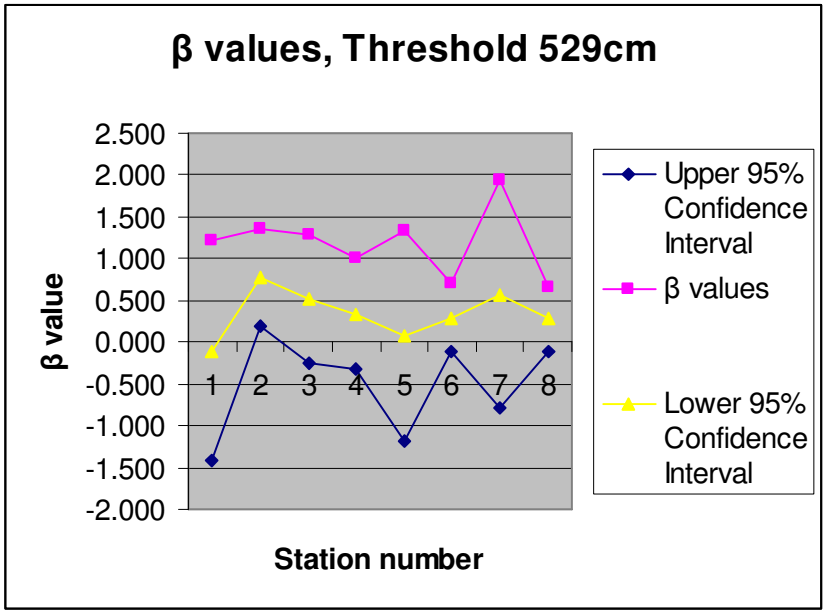


Figure 3

This time only one station is relevant, β_2(K13).

Threshold: 549 cm.

Simulations for a threshold of 549cm, 1000 trials, each with 52 peaks

Station	Mean	St. Error	p-value
ELD	1011	269.22	4.79E-14
EUR	871.02	8.25	4.93E-16
K13	858.28	20.13	0.01
LEG	911.62	11.69	3.08E-05
MPN	884.55	13.71	8.11E-05
SCW	845.66	8.27	4.67E-12
SON	1039.4	40.55	0
SWB	858.59	8.09	4.86E-15
YM6	923.24	22.37	0.02

Table 9

Notice that the p-values are increasing, but they still are not above the 5 % level. They could be increasing because the number of points is decreasing, allowing more room for error.

For a threshold of 549cm, the weights for stations, or β is:

Weights for Each Station for a Threshold of 549cm

1, β_0	EUR,β_1	K13,β_2	LEG,β_3	MPN,β_4	SCW,β_5	SON,β_6	SWB,β_7	YM6,β_8
-2824.737	-0.62	1.05	1.69	0.52	0.19	-0.10	1.18	0.44

Table 10

Same as before, the most influential station is LEG.
The R^2-statistic for this threshold is 0.0179, not such a good fit again. The p-value is 0.0216. The chart with all the β's, except β_0, is below in Figure 4.

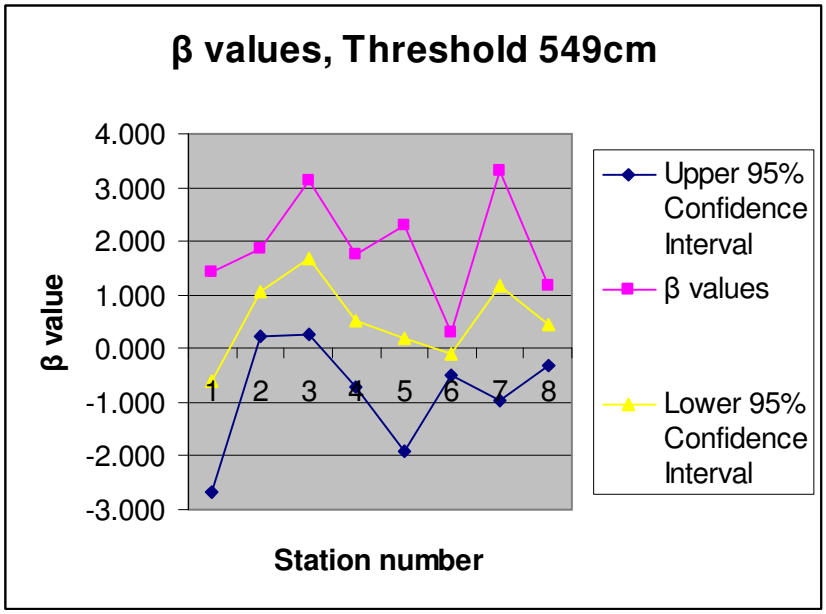


Figure 4

Notice that again only two stations are relevant, β_2(K13) and β_3(LEG).

Conclusion

For this set of data, the regression method does not work well. This could be the result of a few factors. One of these could be that although the GPD fits the first station, when used on the wave heights of the corresponding times of the other stations, it does not. This is the first cause of a bad result. Another problem could be that the normal distribution does not fit the distribution of the simulated, 1 in 10,000 year waves. As a result, when the β's are calculated, most are insignificant. As this seems a bit odd, as the stations are highly correlated, more of them should have an influence (a higher β value). Hence, this method is not a good one for estimating extreme wave heights in the North Sea.

Poorwa Singh
1242326

A.3

Conditional correlations for data based on the peaks of station ELD for a threshold of 449. condcorr1 is the conditional correlation of 2 of the stations given station 1 (ELD), condcorr12 is the conditional correlation of 2 of the stations given station 1 and 2 (ELD&EUR).

Condcorr1 =

2	3	4	5	6	7	8	9
1.0000	0.4917	0.7830	0.6824	0.7494	-0.1273	0.8711	0.5897
0.4917	1.0000	0.3693	0.3208	0.3917	-0.1069	0.4846	0.3880
0.7830	0.3693	1.0000	0.6217	0.7336	0.0330	0.7805	0.5799
0.6824	0.3208	0.6217	1.0000	0.5881	-0.0315	0.6909	0.5430
0.7494	0.3917	0.7336	0.5881	1.0000	0.0892	0.8622	0.5312
-0.1273	-0.1069	0.0330	-0.0315	0.0892	1.0000	-0.0584	-0.0124
0.8711	0.4846	0.7805	0.6909	0.8622	-0.0584	1.0000	0.5528
0.5897	0.3880	0.5799	0.5430	0.5312	-0.0124	0.5528	1.0000

ans =

3	4	5	6	7	8	9
1.0000	-0.0289	-0.0231	0.0403	-0.0513	0.1315	0.1395
0.0289	1.0000	0.1922	0.3566	0.2151	0.3224	0.2353
-0.0231	0.1922	1.0000	0.1586	0.0763	0.2687	0.2382
0.0403	0.3566	0.1586	1.0000	0.2811	0.6439	0.1671
-0.0513	0.2151	0.0763	0.2811	1.0000	0.1077	0.0782
0.1315	0.3224	0.2687	0.6439	0.1077	1.0000	0.0987
0.1395	0.2353	0.2382	0.1671	0.0782	0.0987	1.0000

ans =

4	5	6	7	8	9
1.0000	0.1917	0.3582	0.2139	0.3292	0.2418
0.1917	1.0000	0.1597	0.0753	0.2742	0.2439
0.3582	0.1597	1.0000	0.2837	0.6447	0.1632
0.2139	0.0753	0.2837	1.0000	0.1156	0.0863
0.3292	0.2742	0.6447	0.1156	1.0000	0.0818
0.2418	0.2439	0.1632	0.0863	0.0818	1.0000

ans =

5	6	7	8	9
1.0000	0.0993	0.0357	0.2277	0.2074
0.0993	1.0000	0.2271	0.5976	0.0845
0.0357	0.2271	1.0000	0.0490	0.0365
0.2277	0.5976	0.0490	1.0000	0.0024
0.2074	0.0845	0.0365	0.0024	1.0000

ans =

6	7	8	9
1.0000	0.2248	0.5934	0.0657
0.2248	1.0000	0.0420	0.0298
0.5934	0.0420	1.0000	-0.0471
0.0657	0.0298	-0.0471	1.0000

ans =

7	8	9
1.0000	0.1166	0.0154
0.1166	1.0000	0.1071
0.0154	0.1071	1.0000

ans =

8	9
1.0000	-0.1060
-0.1060	1.0000

A.4

One example of rank correlation of simulated variables:
Trial 616:

1.0000 0.5420 0.6264 0.6170 0.7158 0.5148 0.5724 0.5254 0.7557
0.5420 1.0000 0.6273 0.8201 0.7420 0.7470 0.1110 0.9051 0.7672
0.6264 0.6273 1.0000 0.5828 0.6227 0.6038 0.3709 0.6209 0.6531
0.6170 0.8201 0.5828 1.0000 0.8121 0.8087 0.3308 0.8584 0.7998
0.7158 0.7420 0.6227 0.8121 1.0000 0.7382 0.3682 0.7810 0.8310
0.5148 0.7470 0.6038 0.8087 0.7382 1.0000 0.3494 0.8821 0.6918
0.5724 0.1110 0.3709 0.3308 0.3682 0.3494 1.0000 0.1904 0.3715
0.5254 0.9051 0.6209 0.8584 0.7810 0.8821 0.1904 1.0000 0.7633
0.7557 0.7672 0.6531 0.7998 0.8310 0.6918 0.3715 0.7633 1.0000

The differences are:
diffrankcorreld(:,:,616) =

0 -0.0151 -0.0081 0.0087 -0.0076 -0.0112 0.0084 0.0067 -0.0275
-0.0151 0 0.0077 0.0193 0.0039 0.0047 -0.0072 0.0034 0.0106
-0.0081 0.0077 0 0.0250 -0.0167 0.0054 -0.0045 0.0065 0.0021
0.0087 0.0193 0.0250 0 0.0016 -0.0014 -0.0204 0.0092 0.0071
-0.0076 0.0039 -0.0167 0.0016 0 -0.0366 -0.0289 -0.0049 -0.0021
-0.0112 0.0047 0.0054 -0.0014 -0.0366 0 0.0061 -0.0047 -0.0224
0.0084 -0.0072 -0.0045 -0.0204 -0.0289 0.0061 0 0.0036 -0.0408
0.0067 0.0034 0.0065 0.0092 -0.0049 -0.0047 0.0036 0 -0.0043
-0.0275 0.0106 0.0021 0.0071 -0.0021 -0.0224 -0.0408 -0.0043 0

The max difference for this trial is 0.0408. The maximum difference for all 1000 of the trials is: 0.1632

A.5

Threshold 449cm:

Bint =

-1601.203	453.687
-1.542	0.175
-0.408	0.399
-0.076	0.821
0.174	0.765
-0.112	0.991
0.022	0.509
0.270	1.165
-0.105	0.286

Rsquared =
0.0379
Fstat =
4.8853
pval =
6.2228e-006

Threshold 469cm:

Bint =

-1169.308	1112.352
-2.048	-0.257
-0.085	0.884
0.063	1.064
0.216	0.855
-0.686	0.637
-0.126	0.354
-0.123	0.911

-0.088	0.355

Rsquared =
0.0317
Fstat =
4.0511
pval =
9.3662e-005

Threshold 489cm:

Bint =

-1278.729	739.650
-0.869	0.608
-0.069	0.701
-0.083	0.719
-0.017	0.582
-0.543	0.822
-0.316	0.202
-0.429	0.998
-0.038	0.421

Rsquared =

0.0162

Fstat =

2.0383

pval =

0.0393

Threshold 509cm:

Bint =

-1672.206	739.762
-0.936	0.782
-0.136	0.586
0.057	1.046
-0.069	0.725
-0.550	1.113
-0.307	0.306
-0.703	0.963
-0.114	0.343

Rsquared =

0.0135

Fstat =

1.6920

pval =

0.0962

Threshold 529cm:

Bint =

-3467.514	398.279
-1.427	1.211
0.186	1.352
-0.242	1.283
-0.318	1.000
-1.193	1.338
-0.113	0.700
-0.796	1.947
-0.101	0.660

Rsquared =
0.0173
Fstat =
2.1871
pval =

0.0262

Threshold 549cm:

Bint =

-5873.940	224.465
-2.671	1.431
0.213	1.877
0.248	3.136
-0.707	1.745
-1.895	2.280
-0.512	0.312
-0.954	3.314
-0.308	1.186

Rsquared =
0.0179
Fstat =
2.2570
pval =
0.0216

A.6

Fit 1:
Fit for –α log x = log (EZ) – log (EY)
Linear model Poly1:
f(x) = p1*x + p2
Coefficients (with 95% confidence bounds):
p1 = -10.63 (-10.72, -10.54)
p2 = 0 (fixed at bound)

Goodness of fit:
SSE: 8.06
R-square: 0.9811
Adjusted R-square: 0.9811
RMSE: 0.1771
Fit 2:
Fit for $x^{-\alpha} = \frac{E(Z_i)}{E(Y_i)}$
General model Power1:

f(x) = a*x^b
Coefficients (with 95% confidence bounds):
a = 1 (fixed at bound)
b = -9.213 (-9.395, -9.031)

Goodness of fit:
SSE: 0.4335
R-square: 0.9783
Adjusted R-square: 0.9783
RMSE: 0.04107

A.7

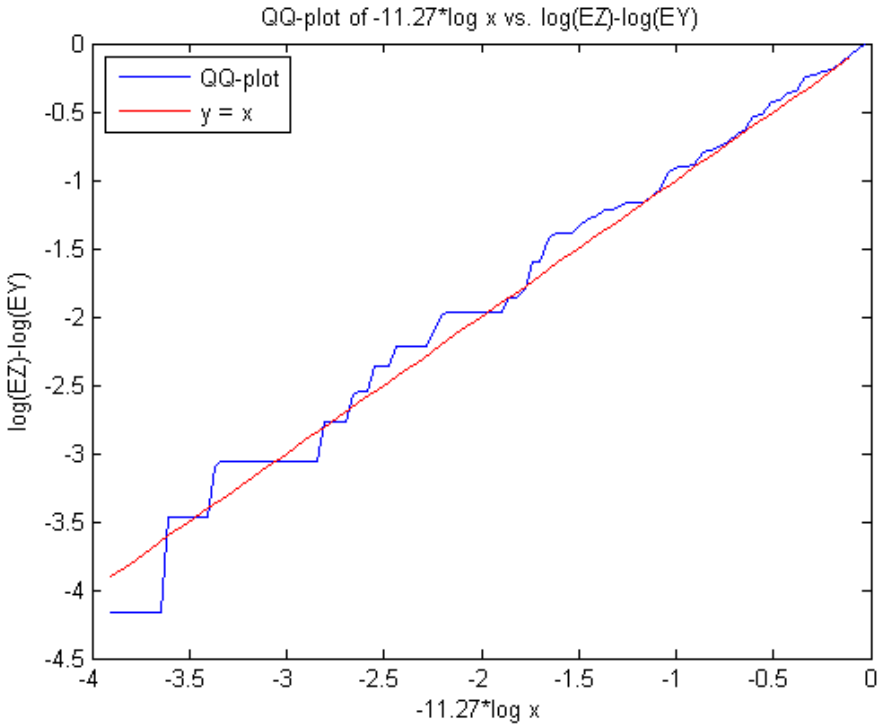
Station	Fit 1: -α log x = log(EZ) – log(EY)
ELD	f(x) = p1*x + p2 Coefficients (with 95% confidence bounds): p1 = -10.63 (-10.72, -10.54) p2 = 0 (fixed at bound) Goodness of fit: SSE: 8.06 R-square: 0.9811 Adjusted R-square: 0.9811 RMSE: 0.1771
EUR	fit1(x) = -a*x (x is log x) Coefficients (with 95% confidence bounds): a = 11.27 (11.16, 11.38) Goodness of fit: sse: 4.8924 rsquare: 0.9817 dfe: 189 adjrsquare: 0.9817 rmse: 0.1609
K13	fit1(x) = -a*x Coefficients (with 95% confidence bounds): a = 9.682 (9.476, 9.888) Goodness of fit:

	sse: 25.2691 rsquare: 0.9176 dfe: 224 adjrsquare: 0.9176 rmse: 0.3359
LEG	fit1(x) = -a*x (x is log x) Coefficients (with 95% confidence bounds): a = 10.05 (9.948, 10.16) Goodness of Fit sse: 8.9380 rsquare: 0.9767 dfe: 228 adjrsquare: 0.9767 rmse: 0.1980
MPN	fit1(x) = -a*x (x is log x) Coefficients (with 95% confidence bounds): a = 11.08 (10.78, 11.39) Goodness of fit: sse: 33.3148 rsquare: 0.9000 dfe: 179 adjrsquare: 0.9000 rmse: 0.4314
SCW	fit1(x) = -a*x (x is log x) Coefficients (with 95% confidence bounds): a = 11.53 (11.12, 11.94) Goodness of fit sse: 22.2969 rsquare: 0.8817 dfe: 124 adjrsquare: 0.8817 rmse: 0.4240
SON	fit1(x) = -a*x (x is log x) Coefficients (with 95% confidence bounds): a = 8.217 (8.123, 8.312)

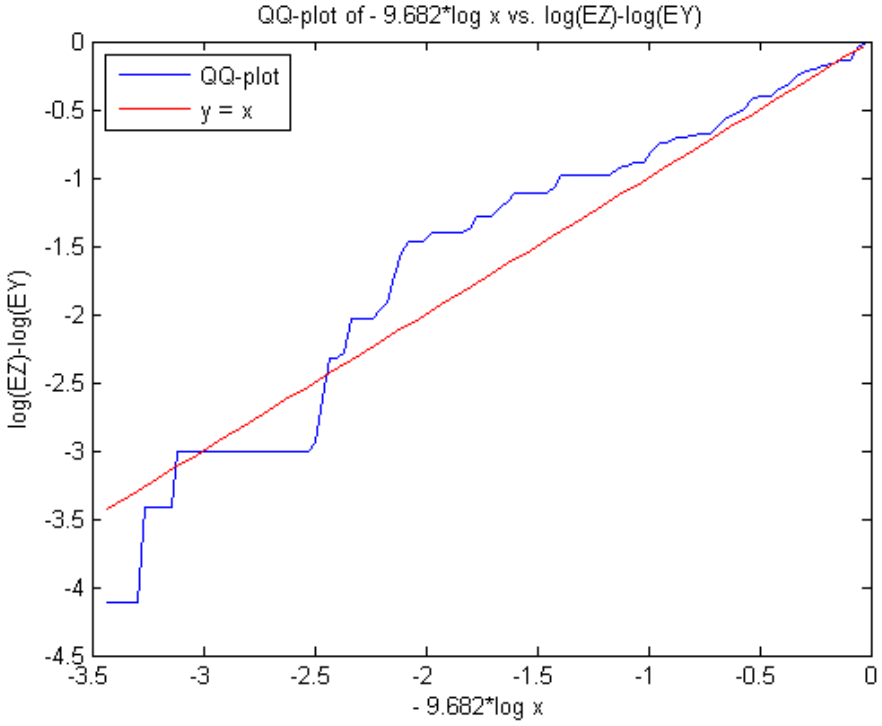
	Goodness of fit: sse: 19.0645 rsquare: 0.9603 dfe: 309 adjrsquare: 0.9603 rmse: 0.2484
SWB	fit1(x) = -a*x (x is log x) Coefficients (with 95% confidence bounds): a = 12.15 (11.86, 12.44) Goodness of fit: sse: 16.5318 rsquare: 0.9168 dfe: 149 adjrsquare: 0.9168 rmse: 0.3331
YM6	fit1(x) = -a*x (x is log x) Coefficients (with 95% confidence bounds): a = 9.565 (9.405, 9.725) Goodness of fit: sse: 21.2543 rsquare: 0.9391 dfe: 238 adjrsquare: 0.9391 rmse: 0.2988

Corresponding QQ-plots:

EUR Fit 1:

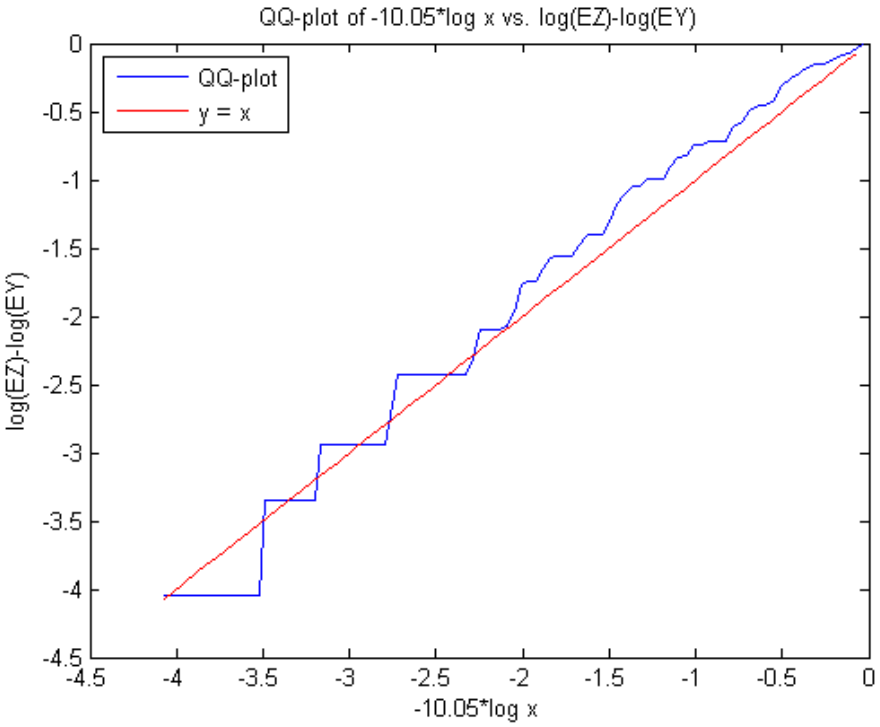

QQ-plot of -11.27*log x vs. log(EZ)-log(EY)
QQ-plot
y = x
log(EZ)-log(EY)
-11.27*log x
0
-0.5
-1
-1.5
-2
-2.5
-3
-3.5
-4
-4.5
-4
-3.5
-3
-2.5
-2
-1.5
-1
-0.5
0

K13 Fit 1:

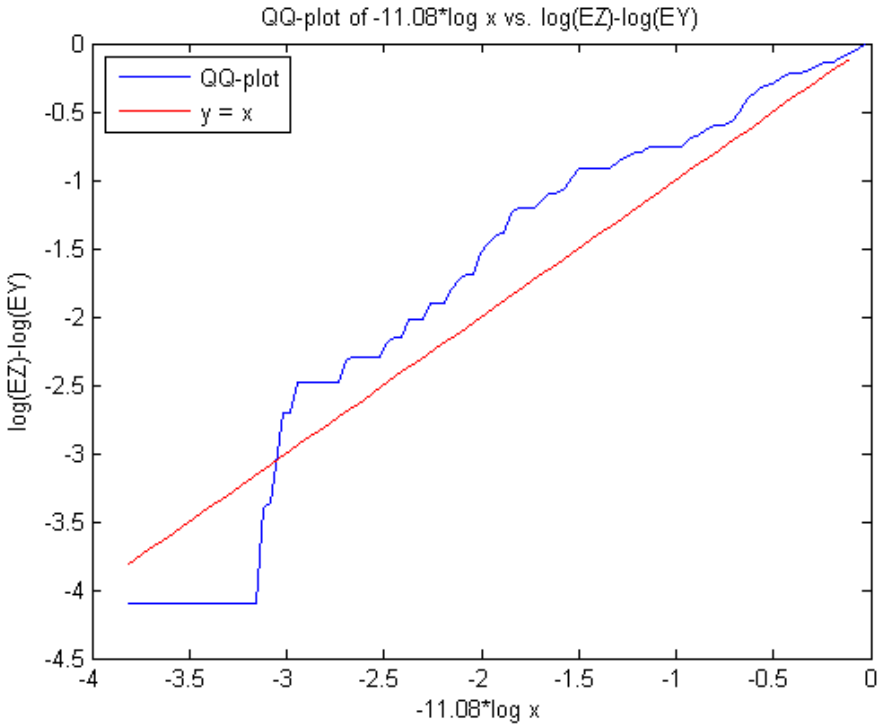

QQ-plot of - 9.682*log x vs. log(EZ)-log(EY)
QQ-plot
y = x
log(EZ)-log(EY)
- 9.682*log x
0
-0.5
-1
-1.5
-2
-2.5
-3
-3.5
-4
-4.5
-3.5
-3
-2.5
-2
-1.5
-1
-0.5
0

Poorwa Singh
1242326

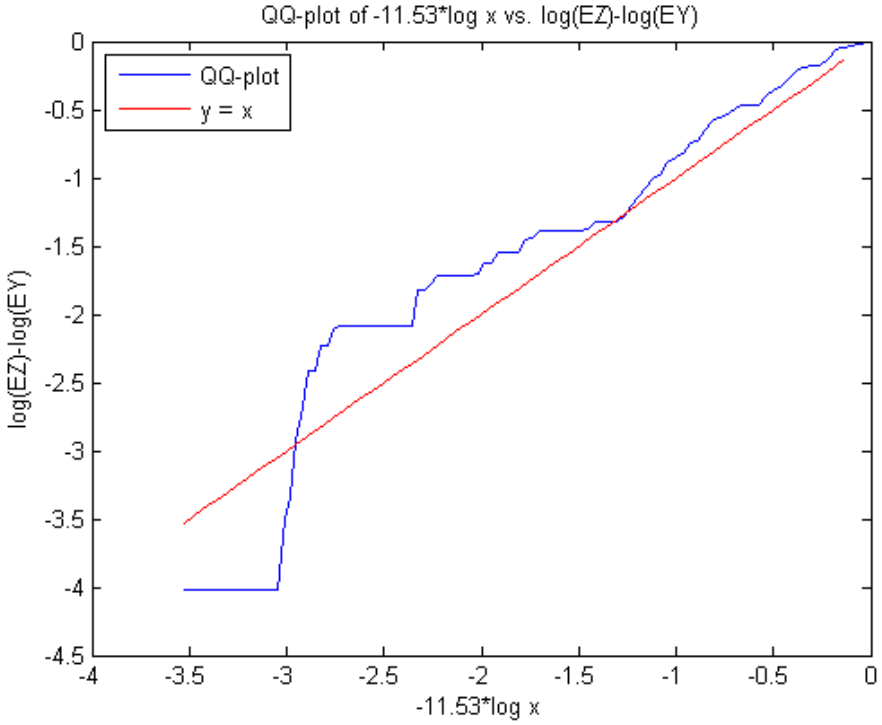
LEG Fit 1:



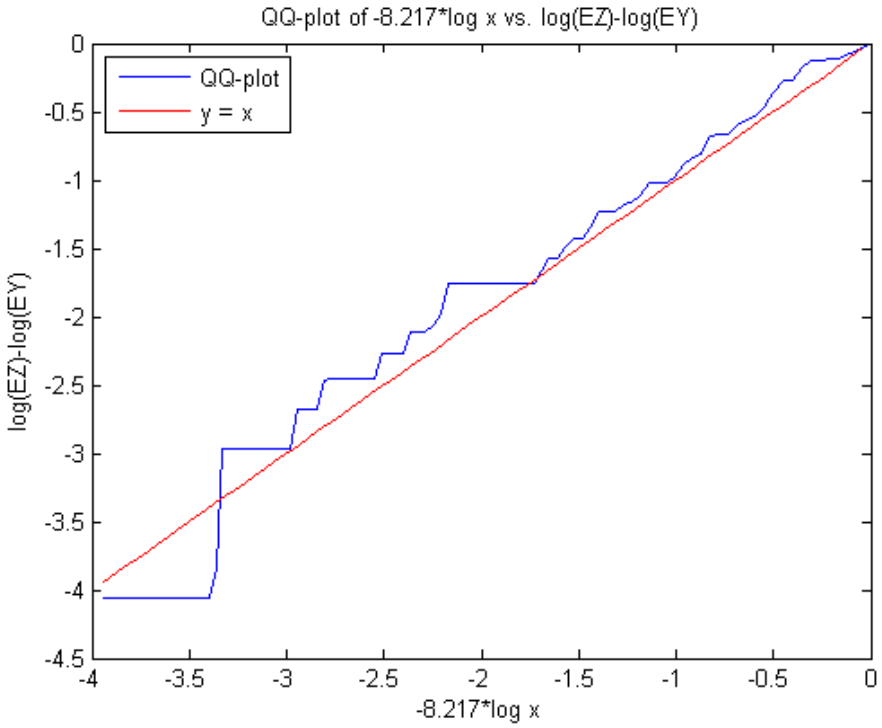
MPN Fit 1:



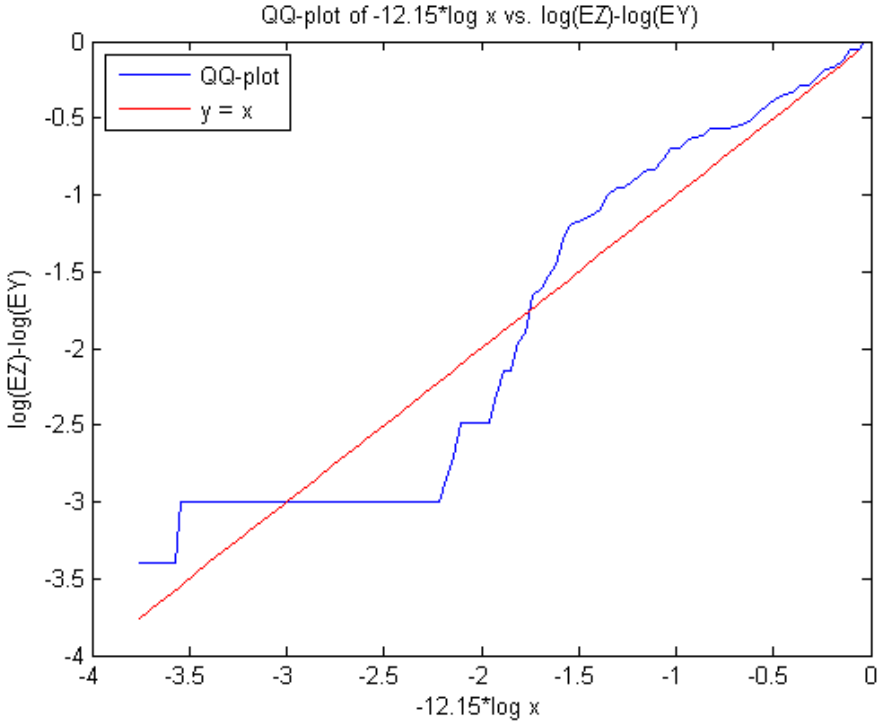
SCW Fit 1:



SON Fit 1:

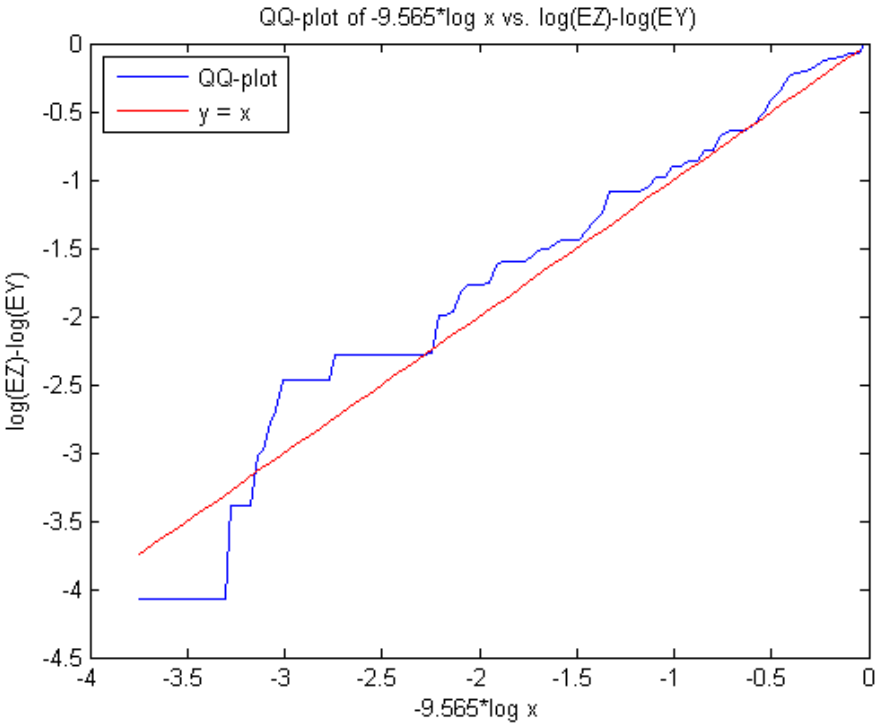

QQ-plot of -8.217*log x vs. log(EZ)-log(EY)
QQ-plot
y = x
log(EZ)-log(EY)
-8.217*log x
0
-0.5
-1
-1.5
-2
-2.5
-3
-3.5
-4
-4.5
-4
-3.5
-3
-2.5
-2
-1.5
-1
-0.5
0

SWB Fit 1:

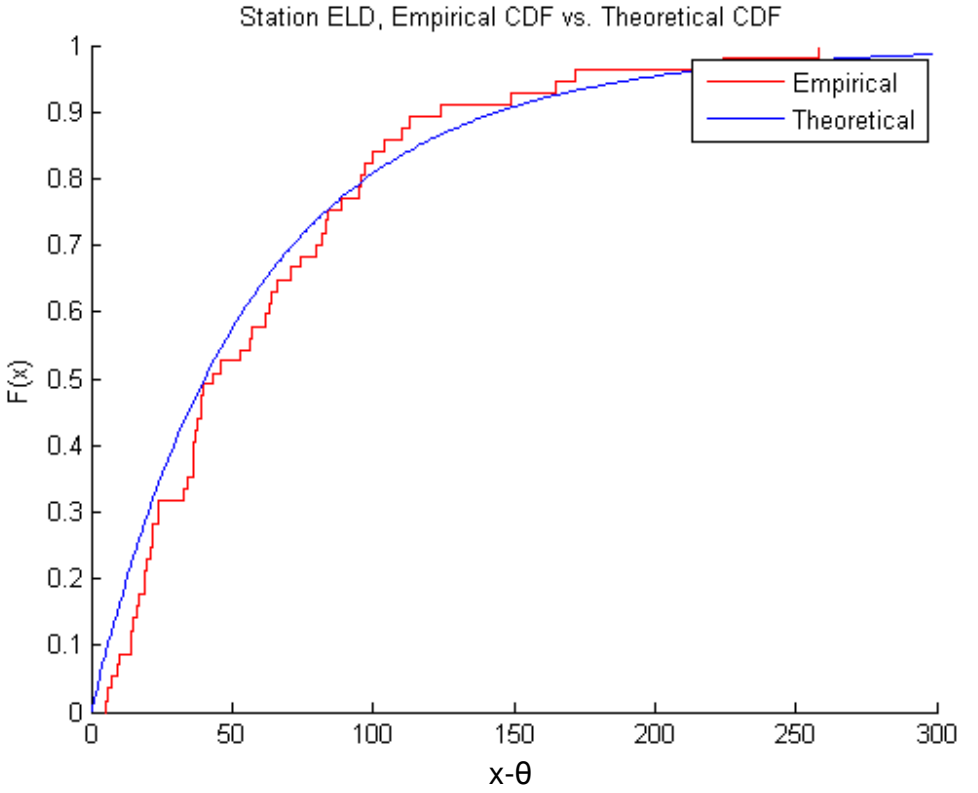

QQ-plot of -12.15*log x vs. log(EZ)-log(EY)
QQ-plot
y = x
log(EZ)-log(EY)
-12.15*log x
0
-0.5
-1
-1.5
-2
-2.5
-3
-3.5
-4
-4
-3.5
-3
-2.5
-2
-1.5
-1
-0.5
0

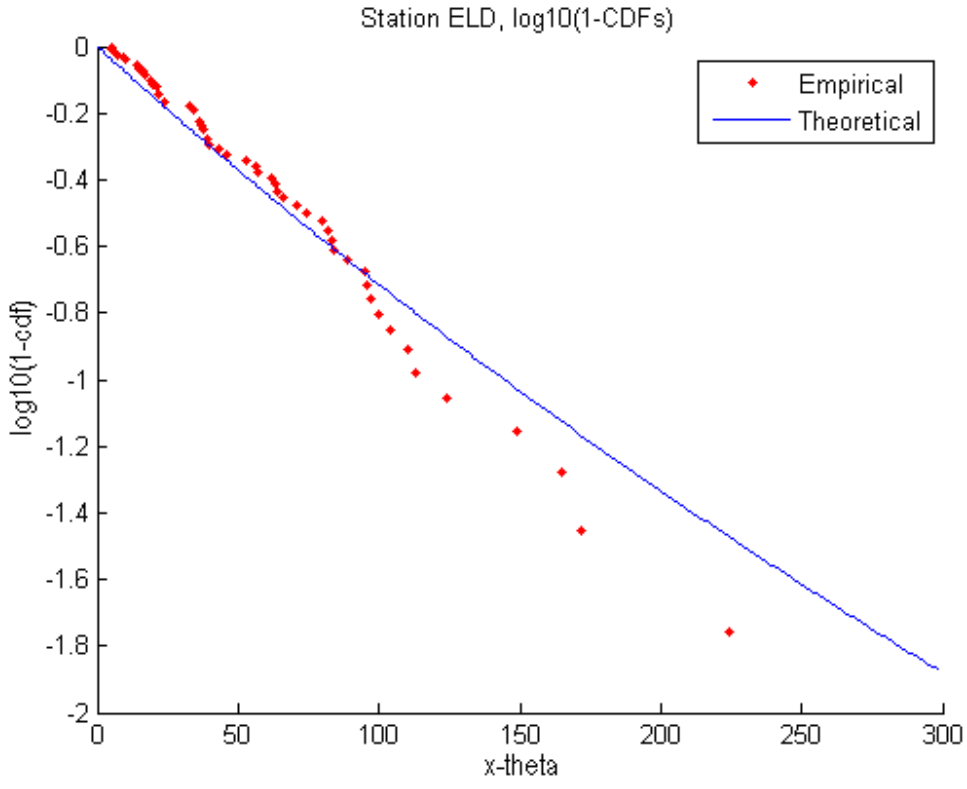
Poorwa Singh
1242326

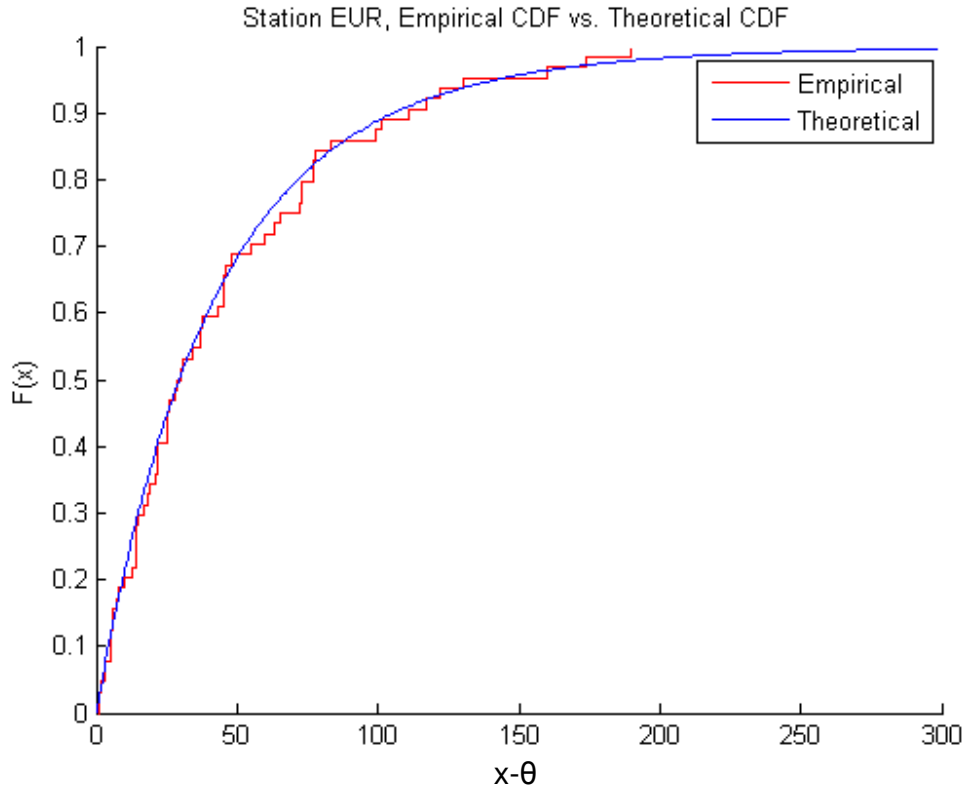
YM6 Fit 1:

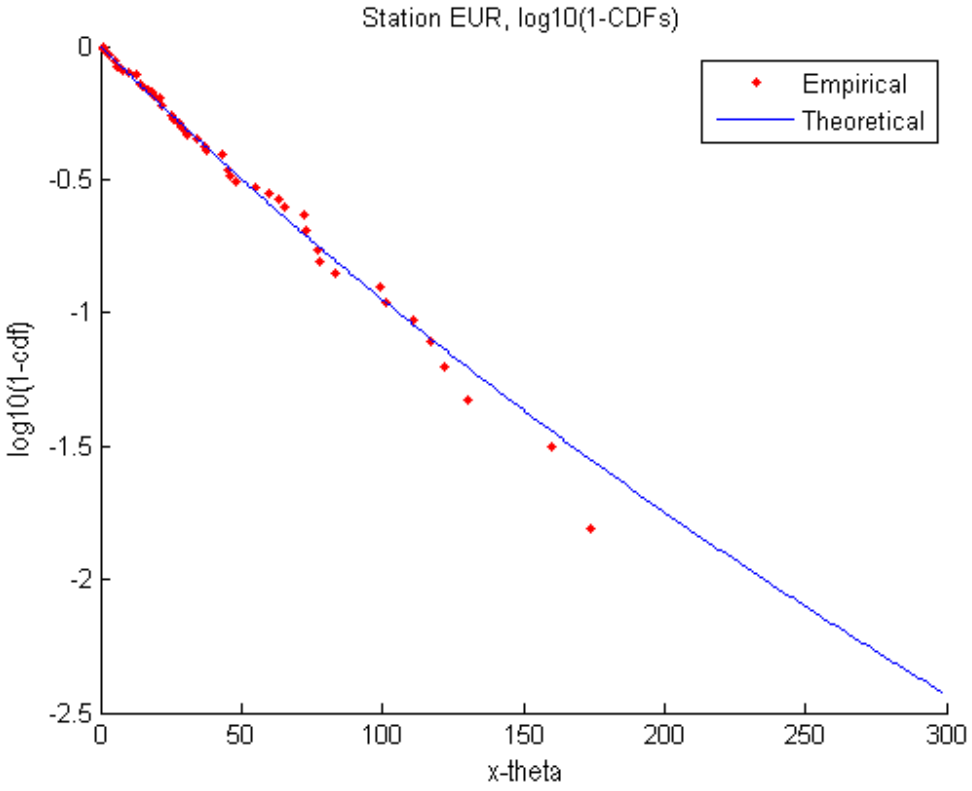


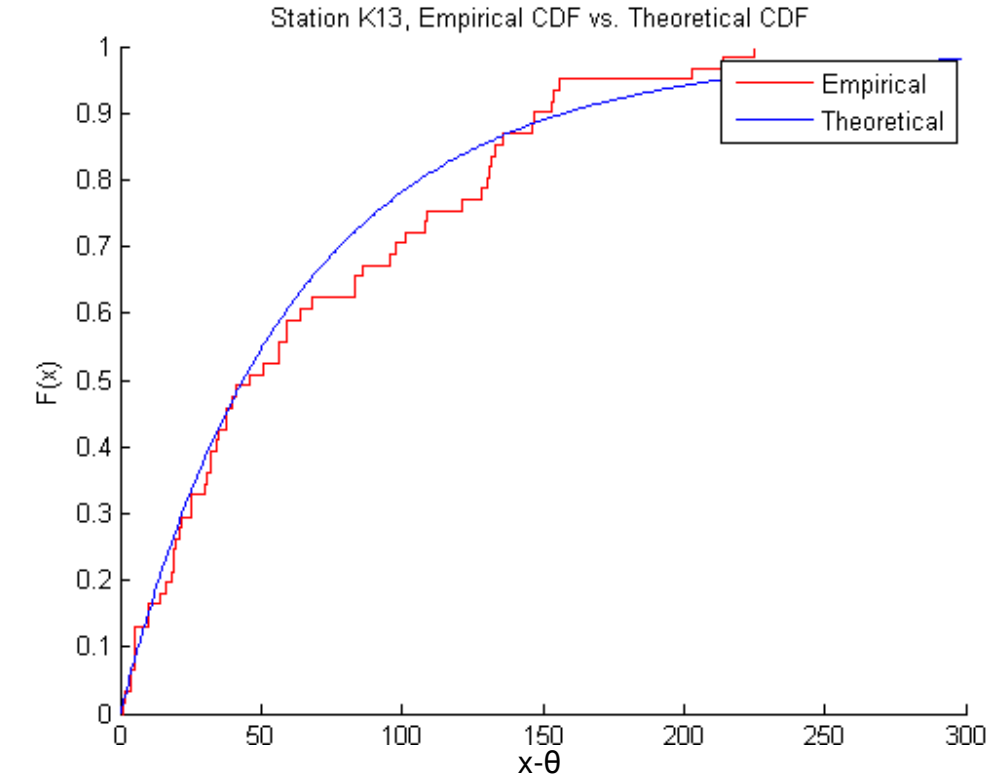
A.8 Comparing the empirical cdf of the peaks over threshold against the theoretical cdf, using parameters, k and σ from method 1

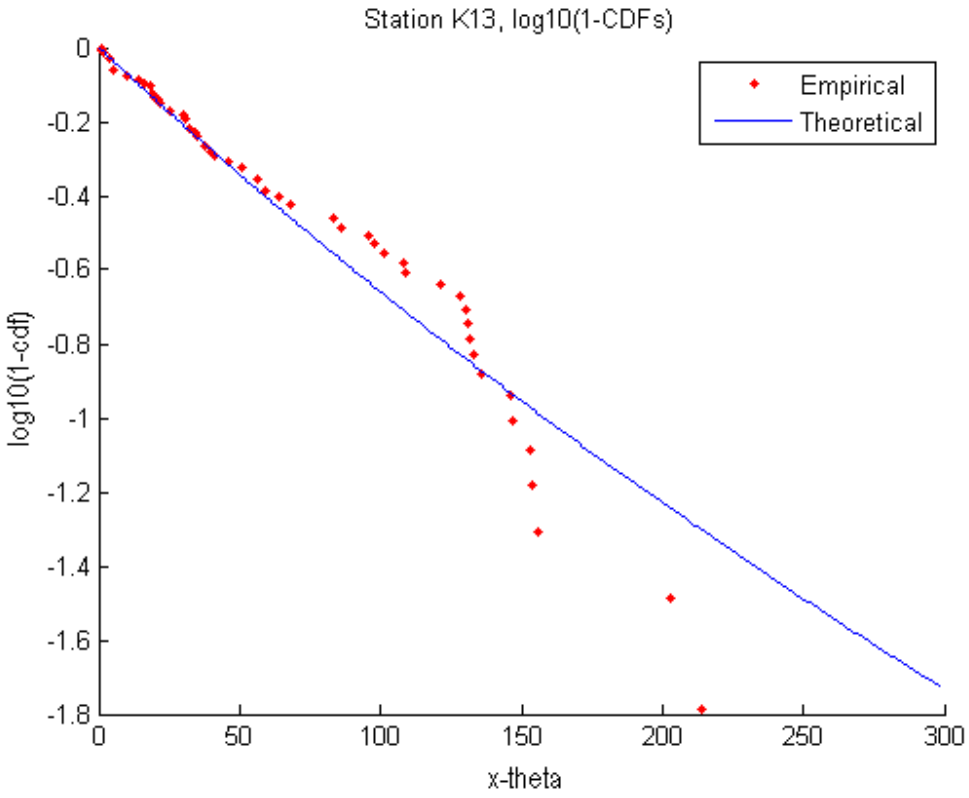


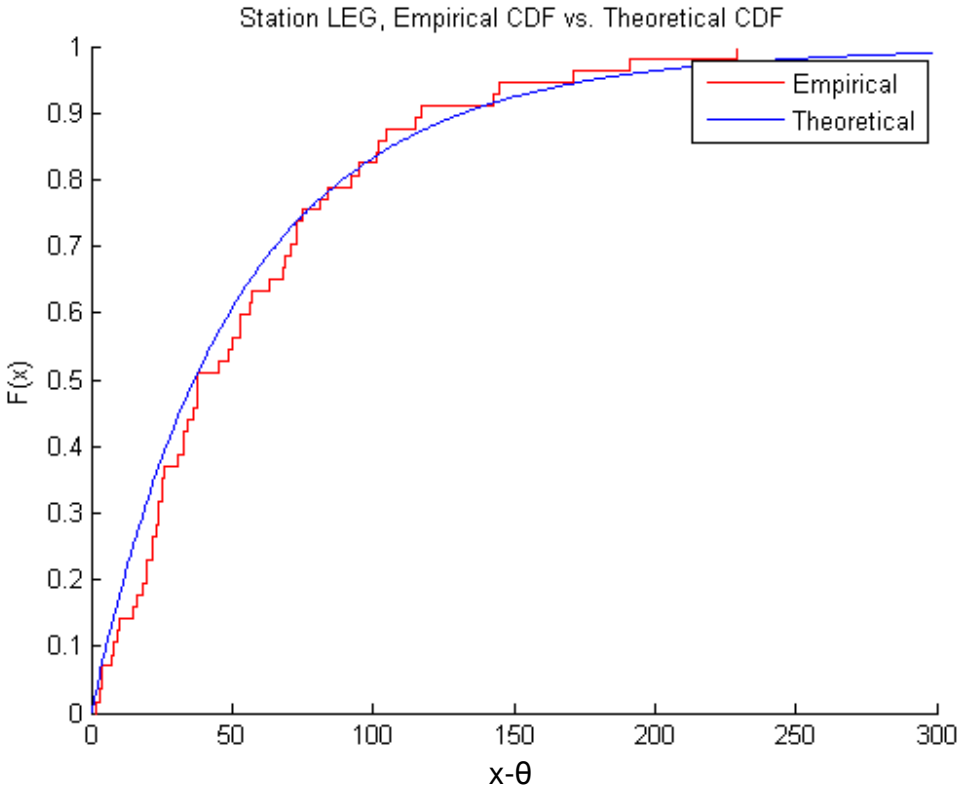


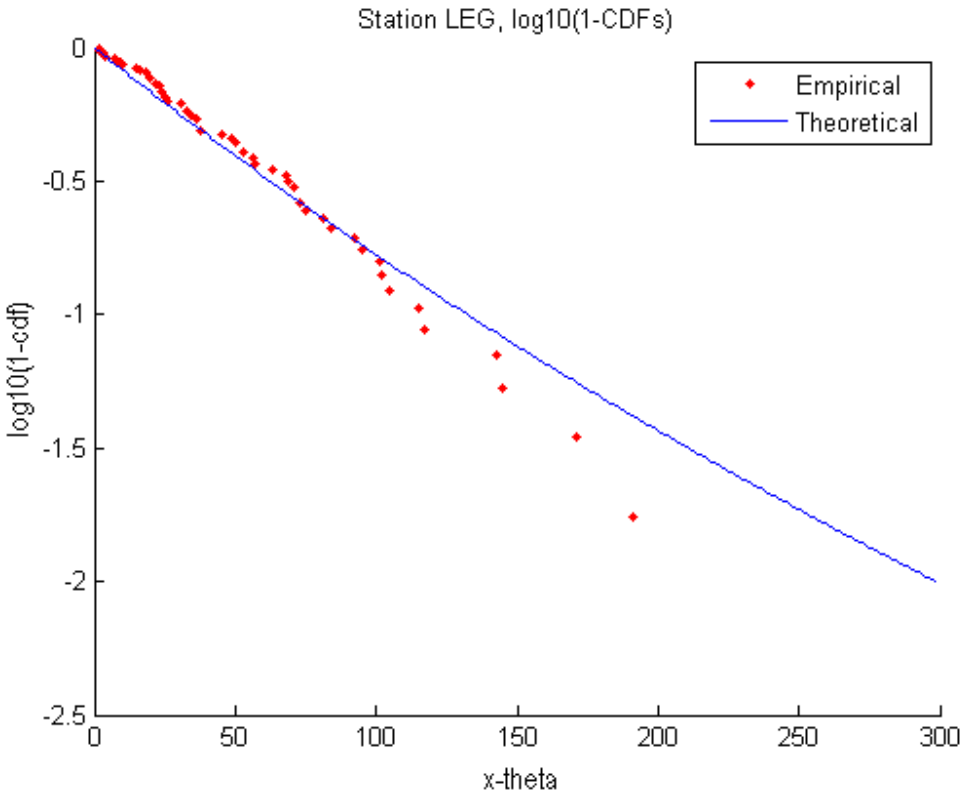

Station EUR, Empirical CDF vs. Theoretical CDF
Empirical
Theoretical
F(x)
x-θ
0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
0.9
1
0
50
100
150
200
250
300


Station EUR, log10(1-CDFs)
Empirical
Theoretical
log10(1-cdf)
x-theta
0
-0.5
-1
-1.5
-2
-2.5
0
50
100
150
200
250
300

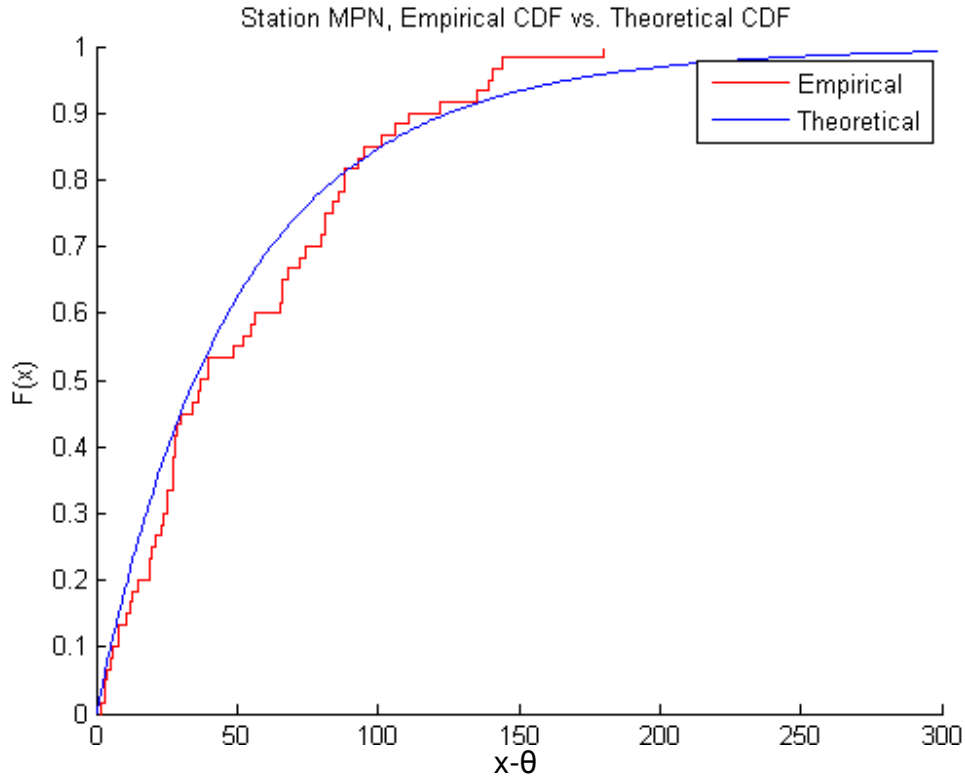

Station K13, Empirical CDF vs. Theoretical CDF
Empirical
Theoretical
F(x)
x-θ
0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
0.9
1
0
50
100
150
200
250
300

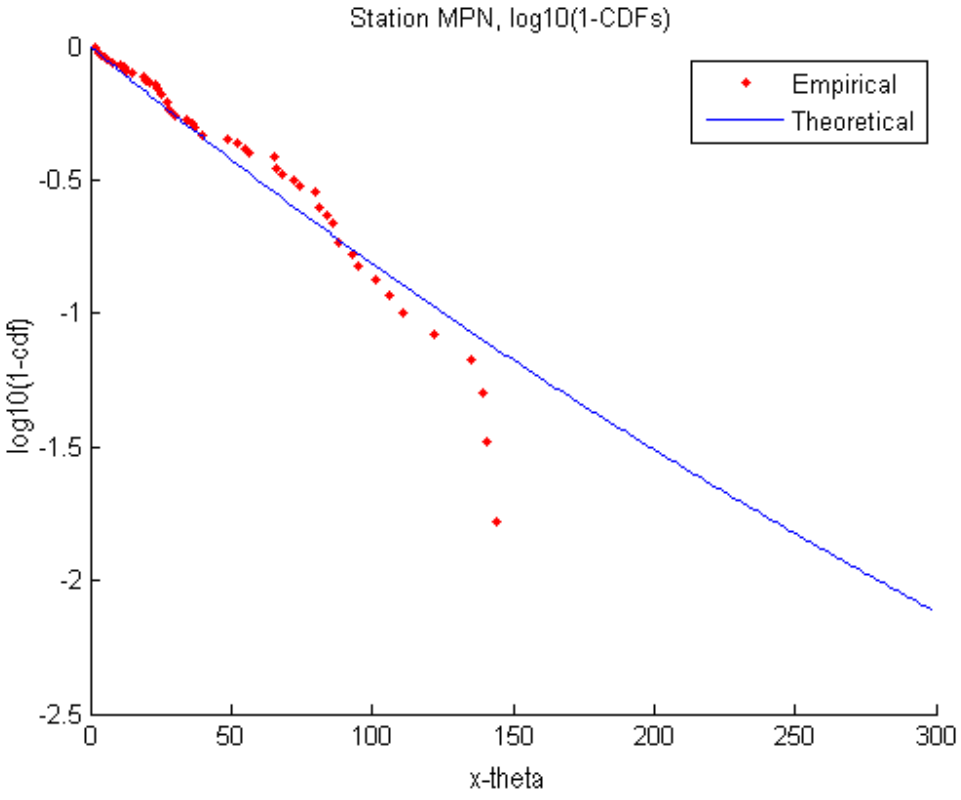

Station K13, log10(1-CDFs)
Empirical
Theoretical
log10(1-cdf)
x-theta
0
-0.2
-0.4
-0.6
-0.8
-1
-1.2
-1.4
-1.6
-1.8
0
50
100
150
200
250
300


Station LEG, Empirical CDF vs. Theoretical CDF
Empirical
Theoretical
F(x)
x-θ
0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
0.9
1
0
50
100
150
200
250
300

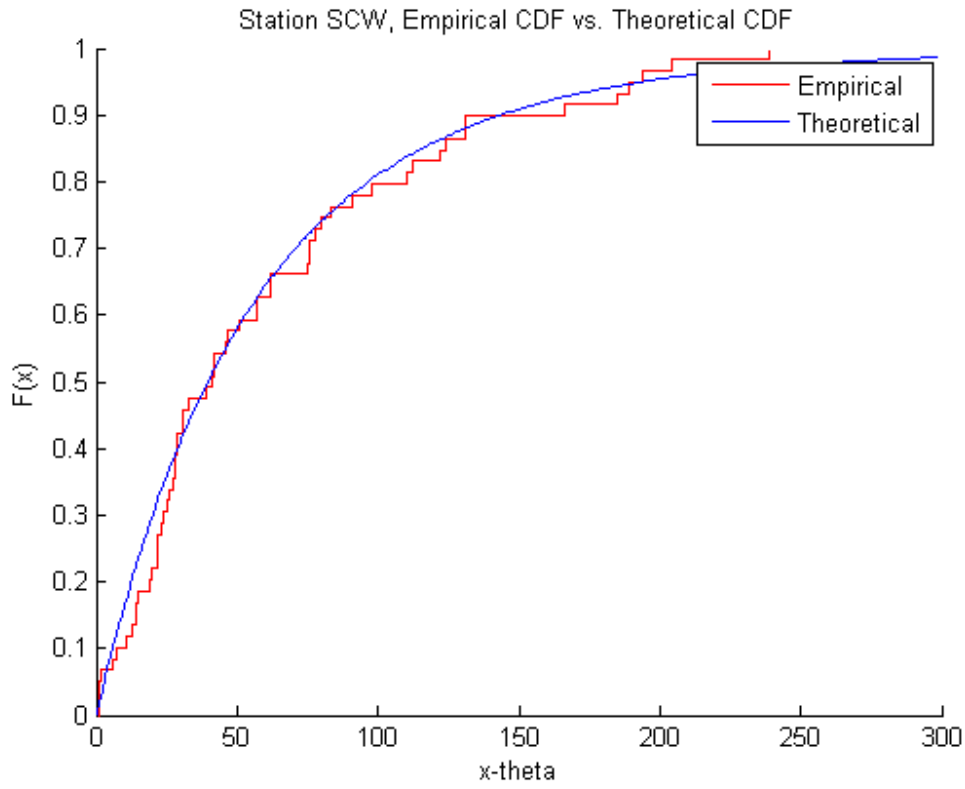

Station LEG, log10(1-CDFs)
Empirical
Theoretical
log10(1-cdf)
x-theta
0
-0.5
-1
-1.5
-2
-2.5
0
50
100
150
200
250
300

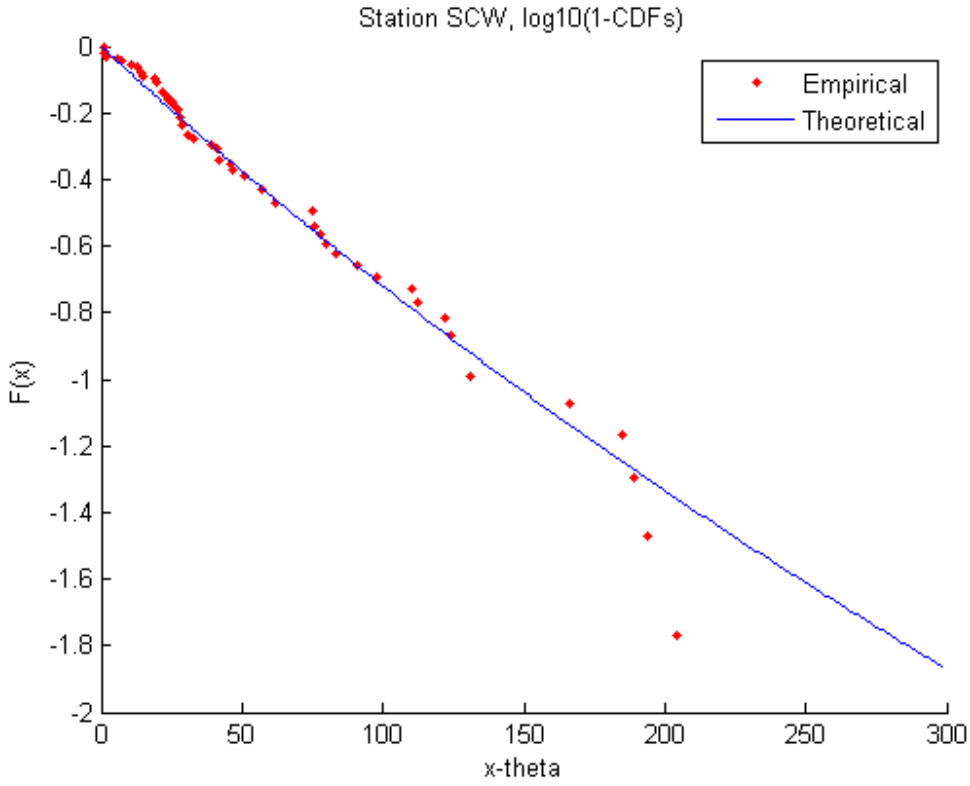
Poorwa Singh
1242326



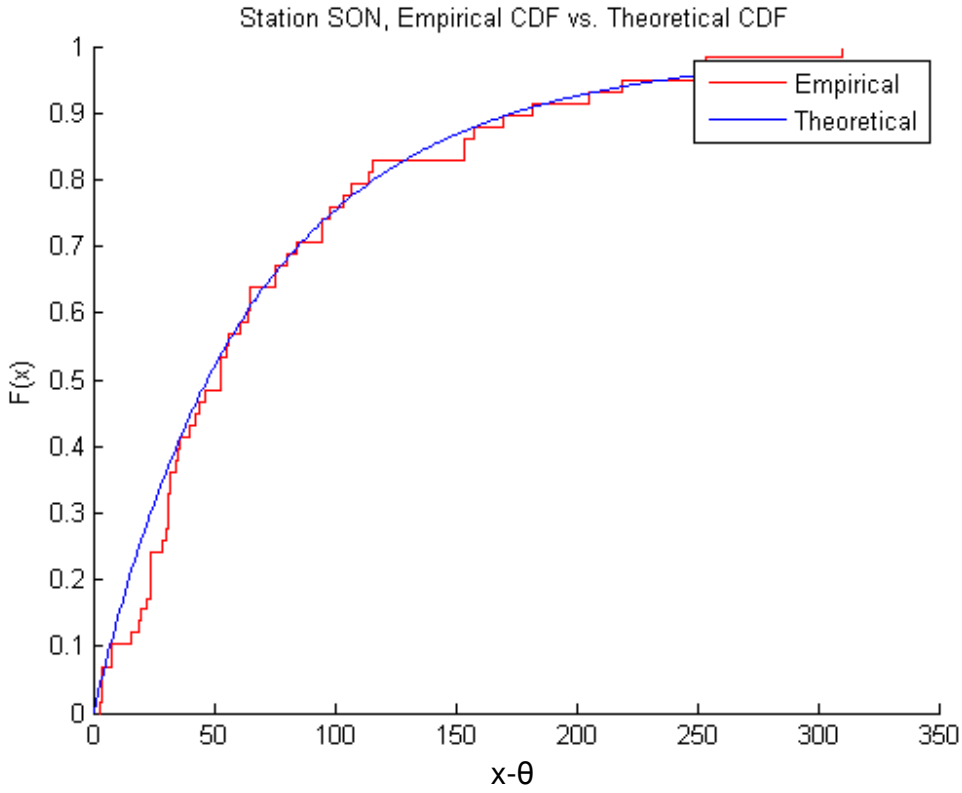


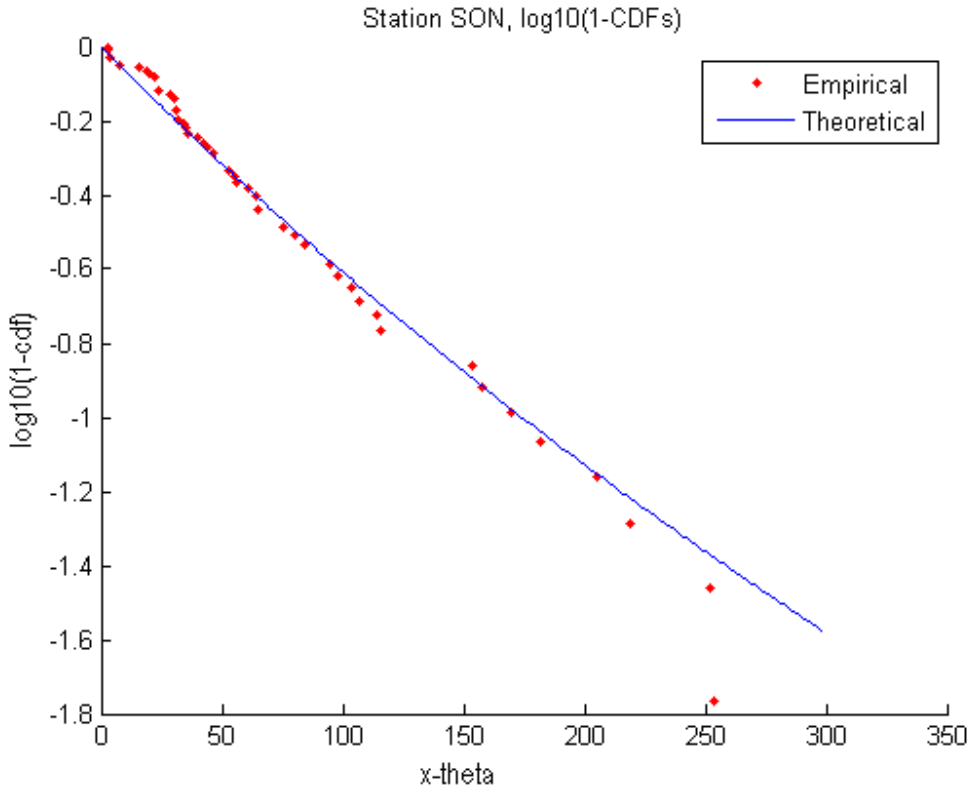
Poorwa Singh
1242326



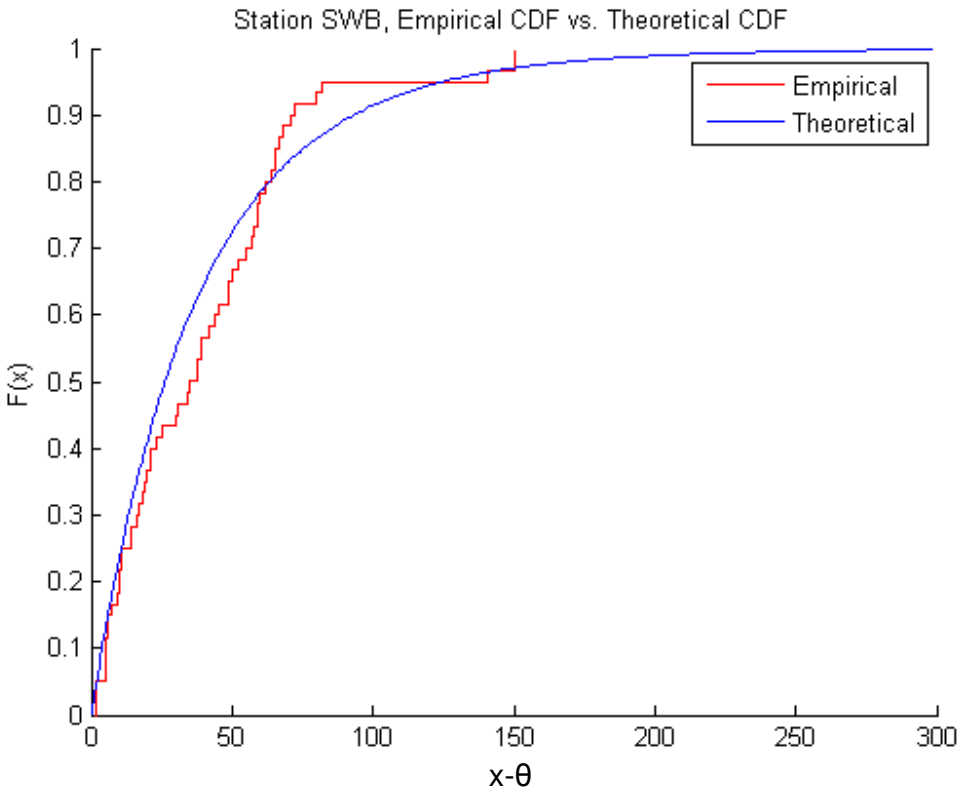


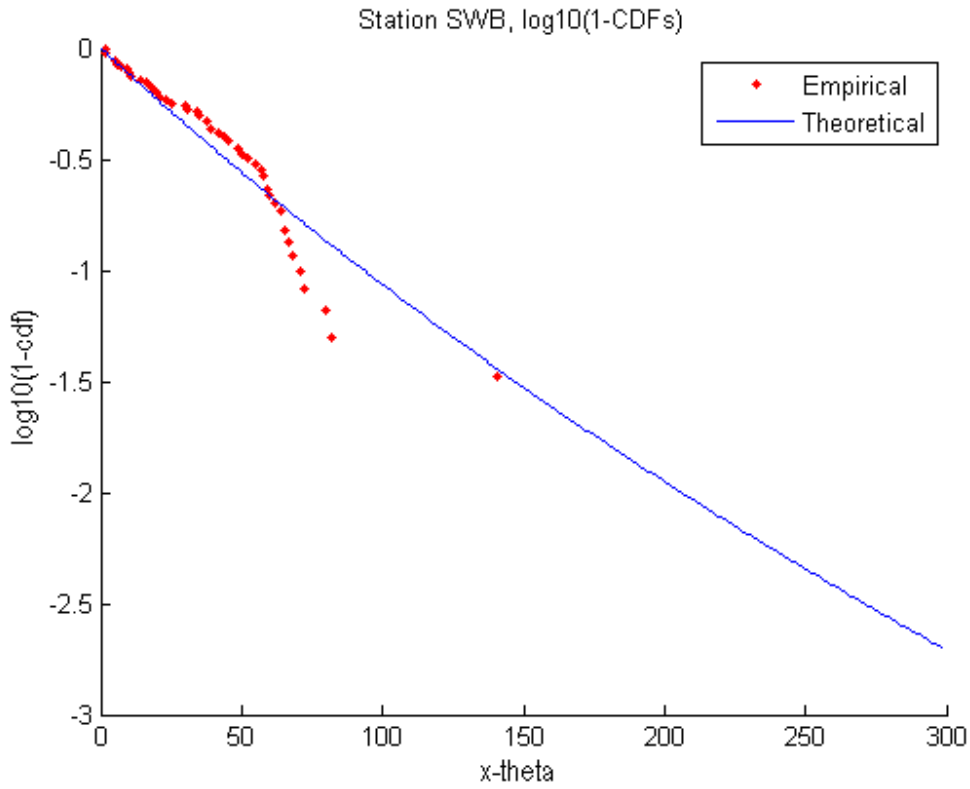
Poorwa Singh
1242326



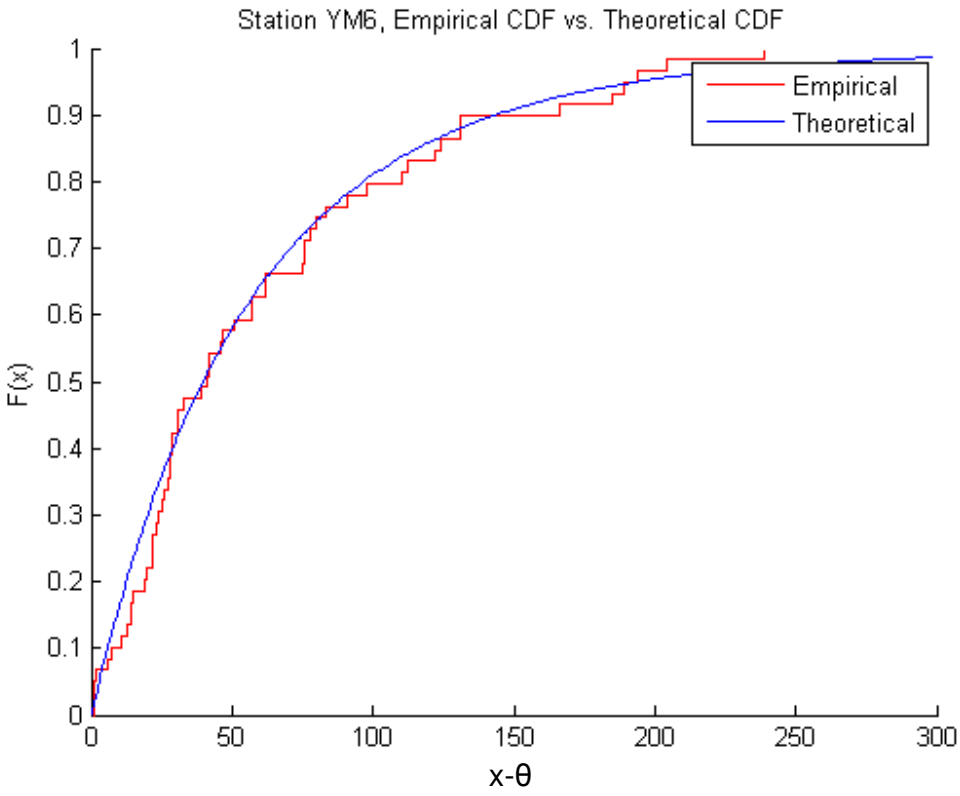


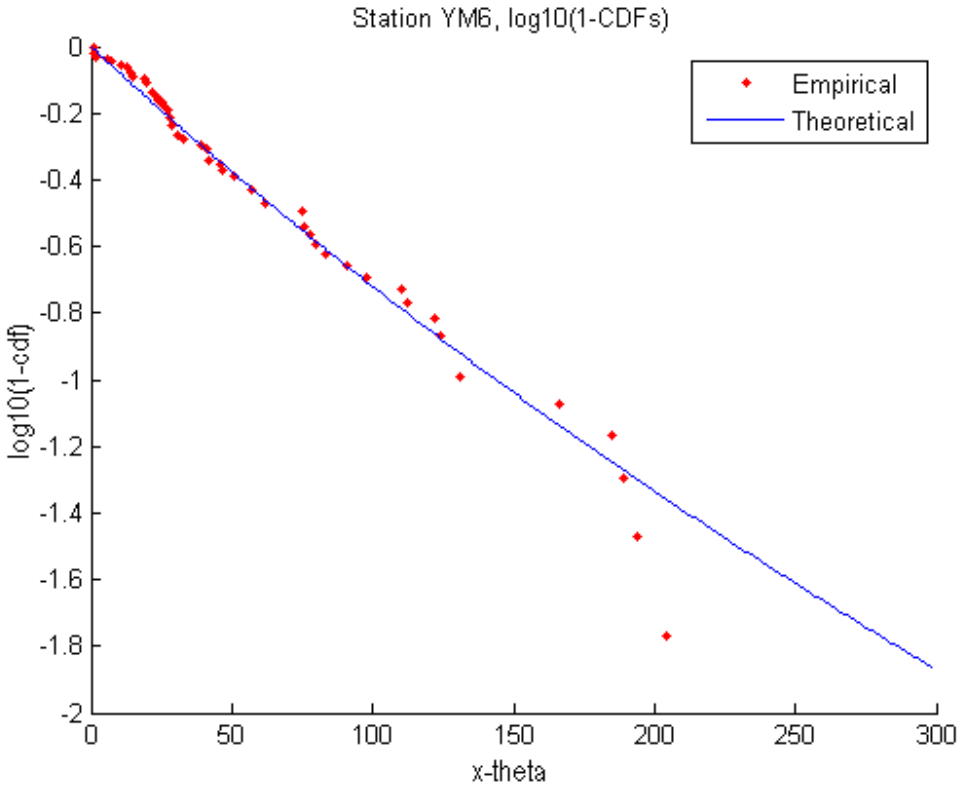
Poorwa Singh
1242326





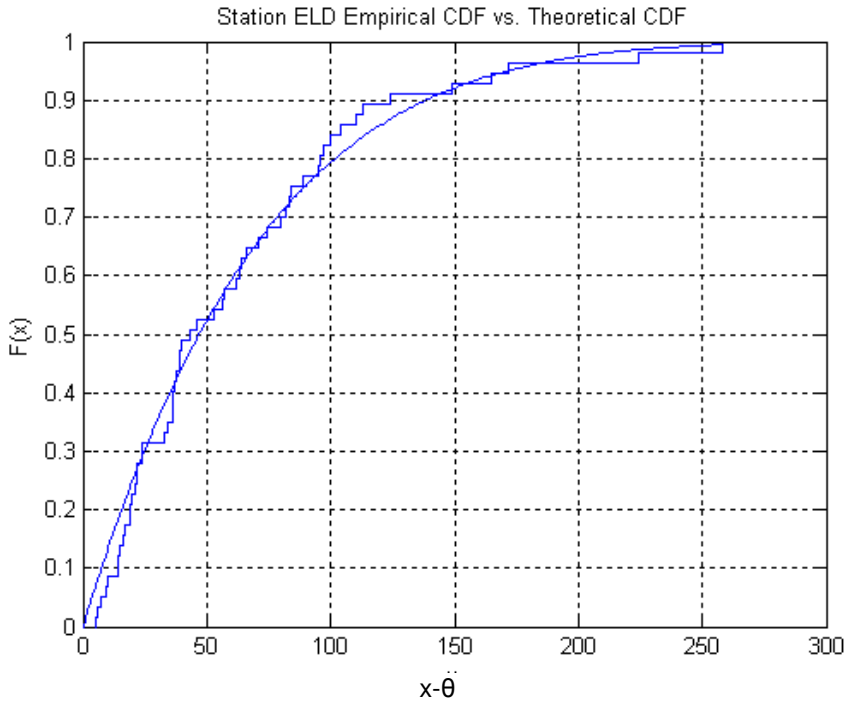
Poorwa Singh
1242326

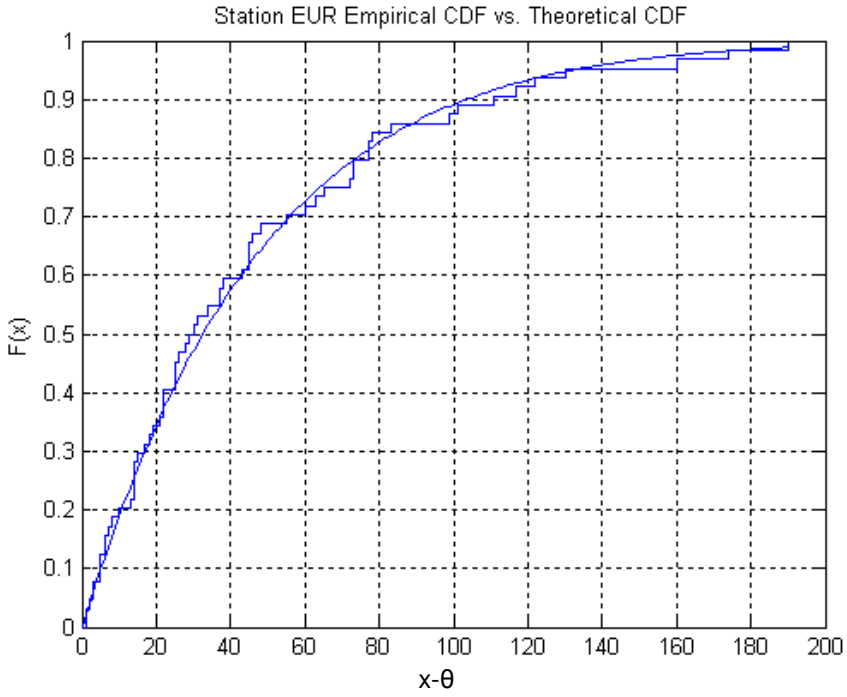


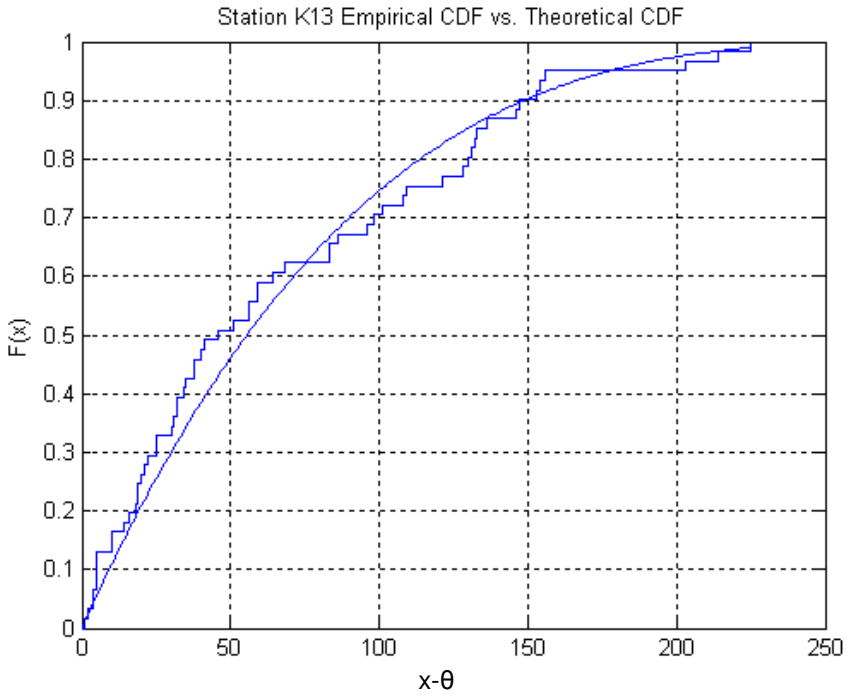


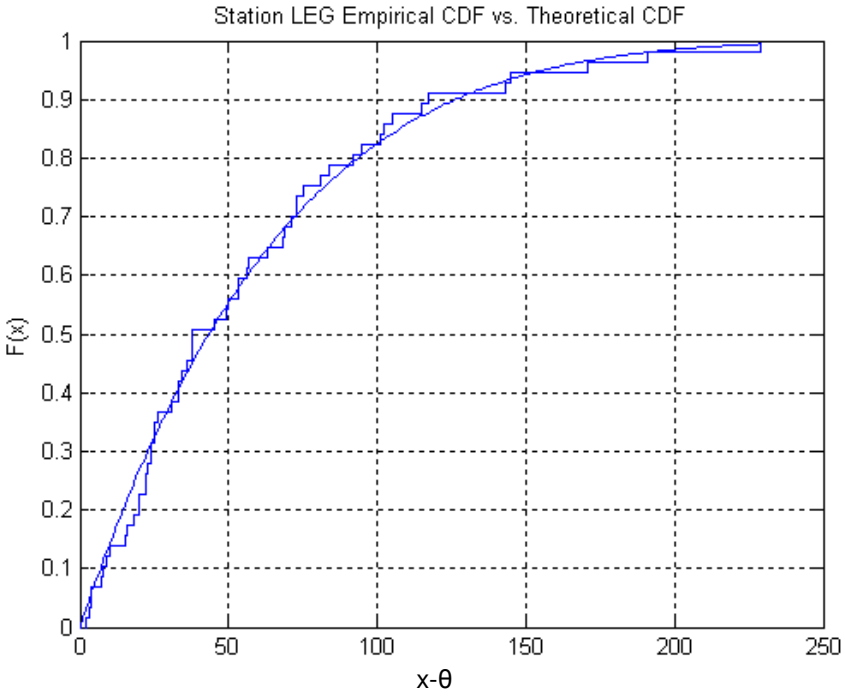
A.9 Comparing the empirical cdf of the peaks over threshold against the theoretical cdf, using parameters, k and σ from method 2—Maximum Likelihood

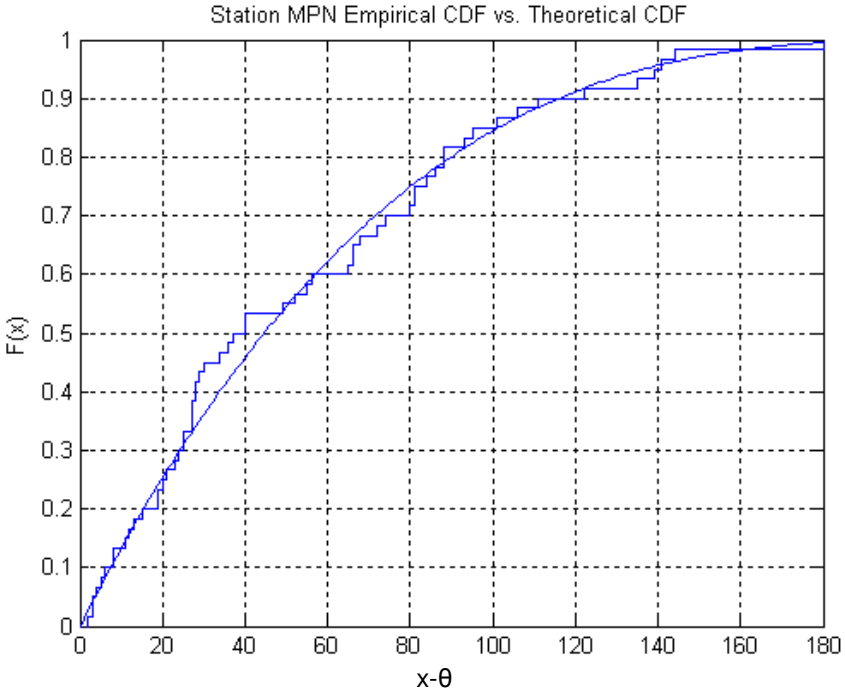
	ELD	EUR	K13	LEG	MPN	SCW	SON	SWB	YM6
p-val	0.6891	0.9973	0.5633	0.9080	0.7078	0.4177	0.7199	0.7614	0.7558
KS-stat	0.0928	0.0490	0.0992	0.0734	0.0891	0.1158	0.0896	0.0849	0.0861

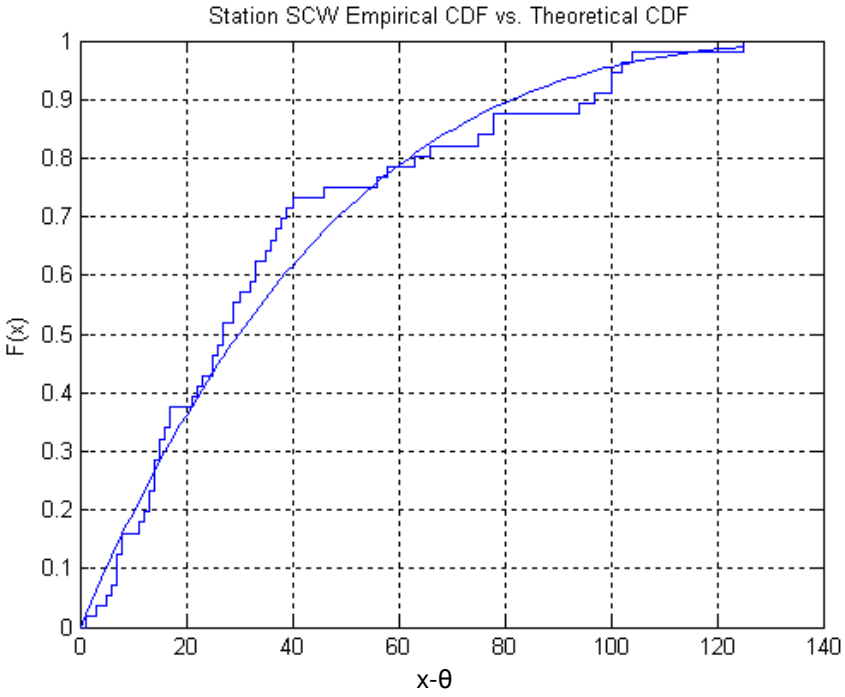


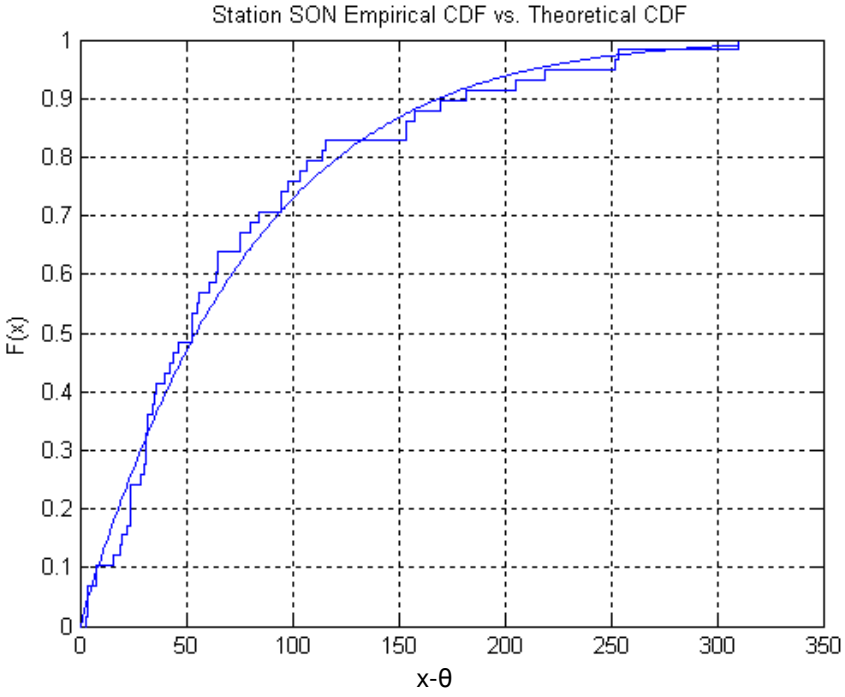

Station EUR Empirical CDF vs. Theoretical CDF
F(x)
x-θ
0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
0.9
1
0
20
40
60
80
100
120
140
160
180
200

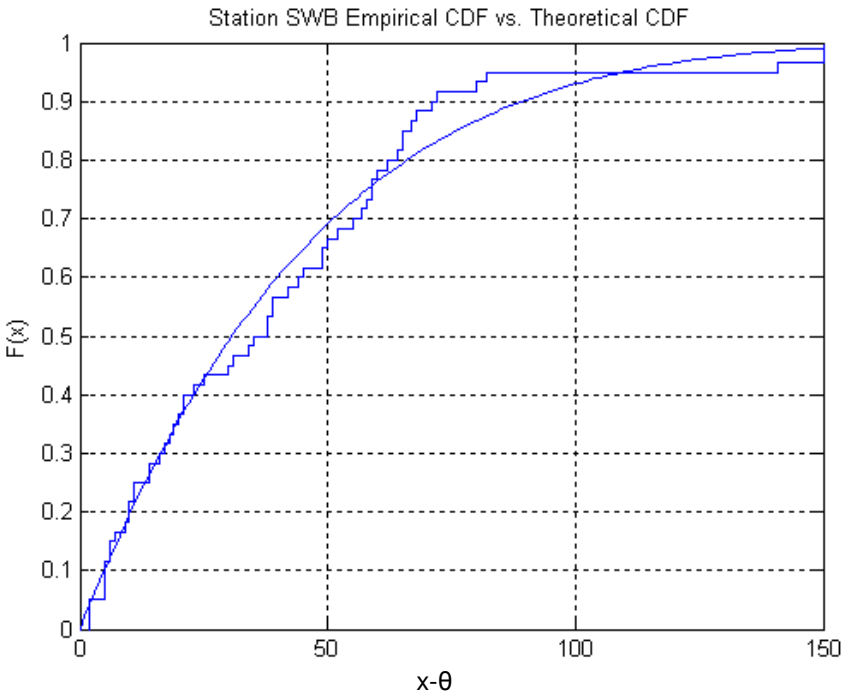

Station K13 Empirical CDF vs. Theoretical CDF
F(x)
x-θ
0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
0.9
1
0
50
100
150
200
250

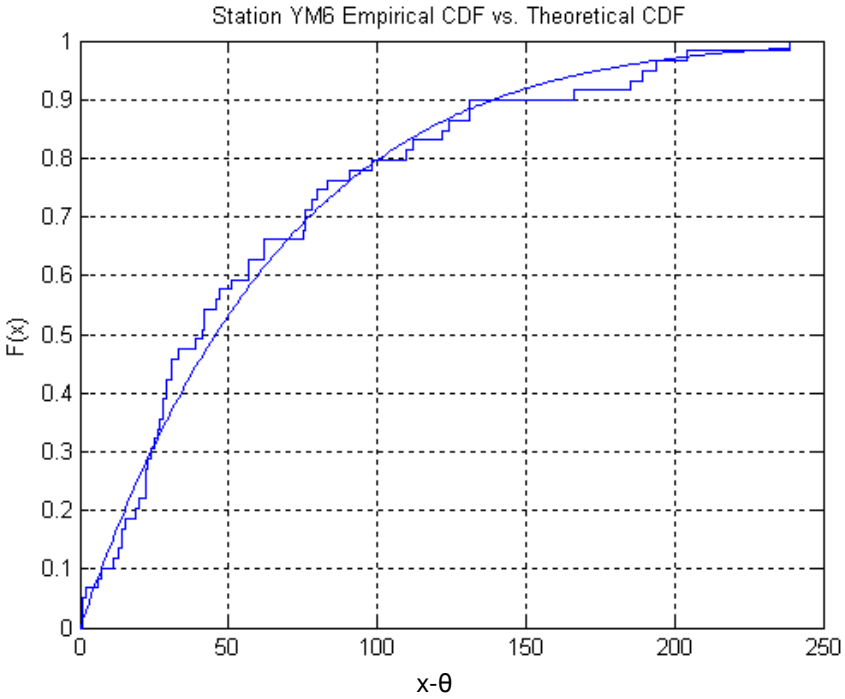

Station LEG Empirical CDF vs. Theoretical CDF
F(x)
x-θ
0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
0.9
1
0
50
100
150
200
250


Station MPN Empirical CDF vs. Theoretical CDF
F(x)
x-θ
0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
0.9
1
0
20
40
60
80
100
120
140
160
180

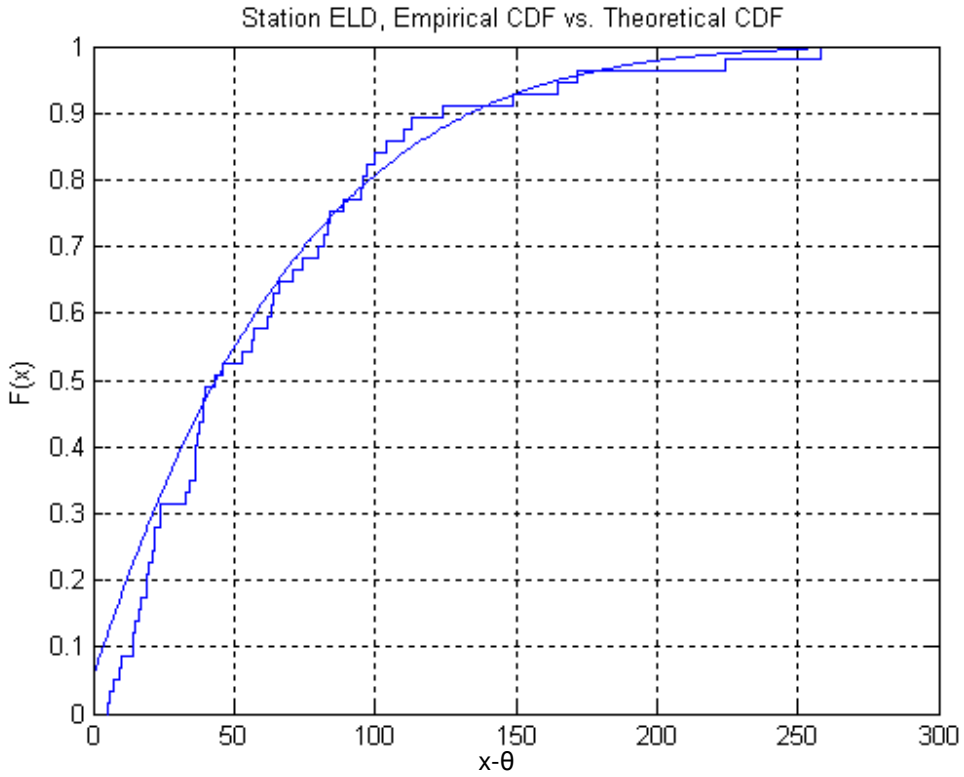

Station SCW Empirical CDF vs. Theoretical CDF
F(x)
x-θ
0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
0.9
1
0
20
40
60
80
100
120
140

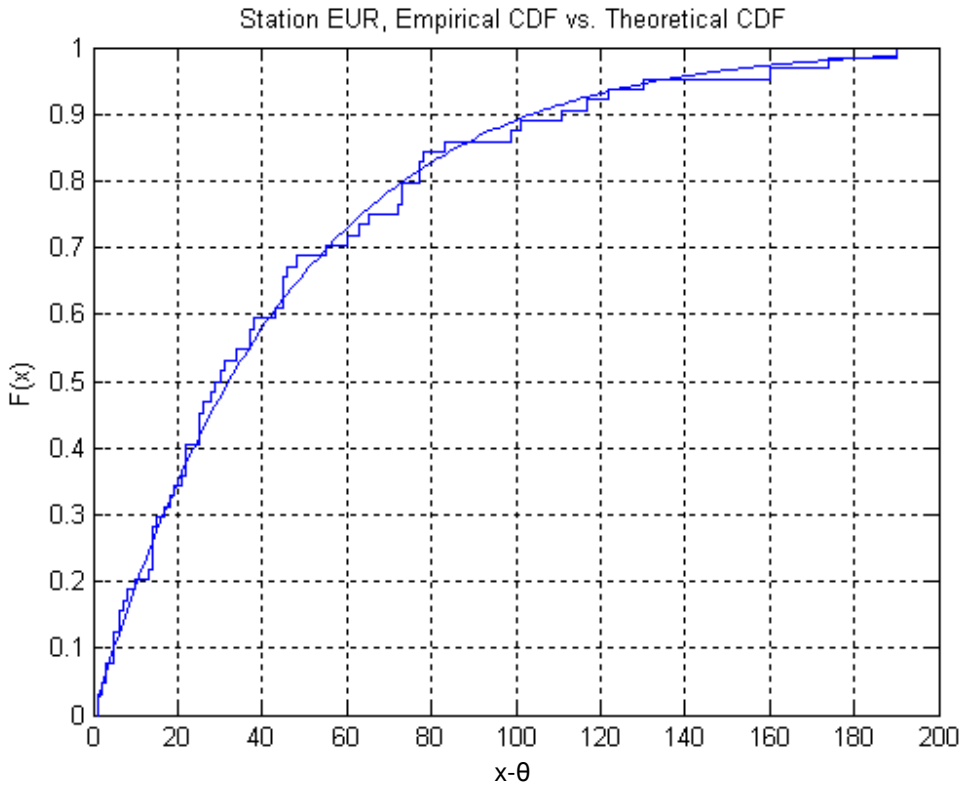

Station SON Empirical CDF vs. Theoretical CDF
F(x)
x-θ
0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
0.9
1
0
50
100
150
200
250
300
350

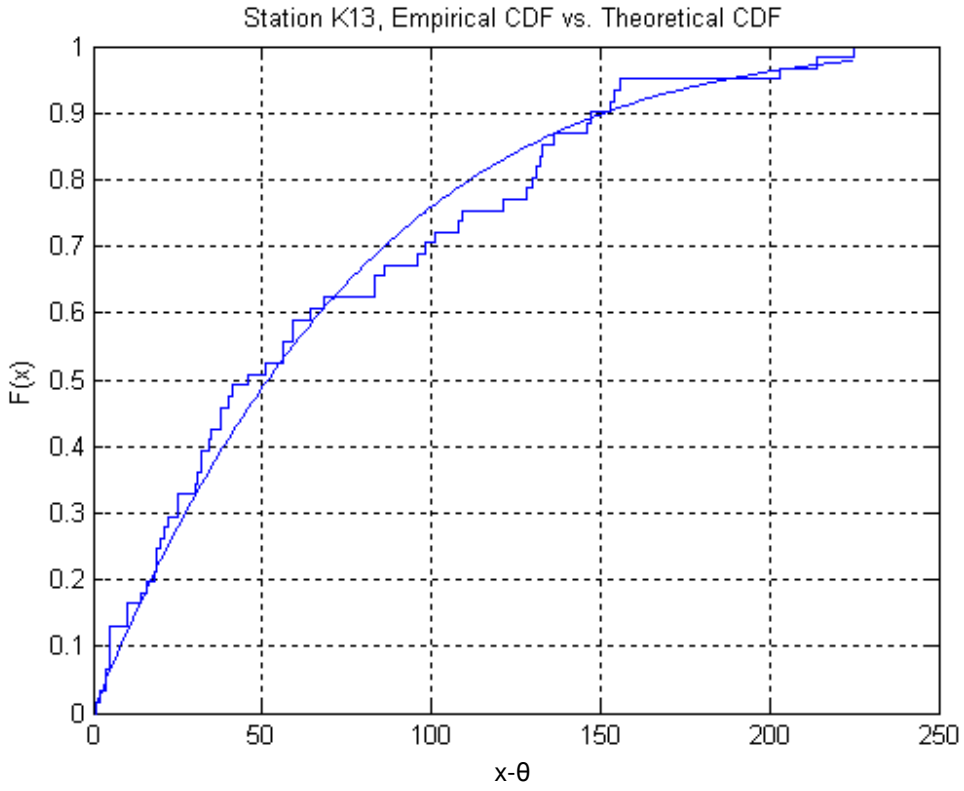

Station SWB Empirical CDF vs. Theoretical CDF
F(x)
x-θ
0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
0.9
1
0
50
100
150

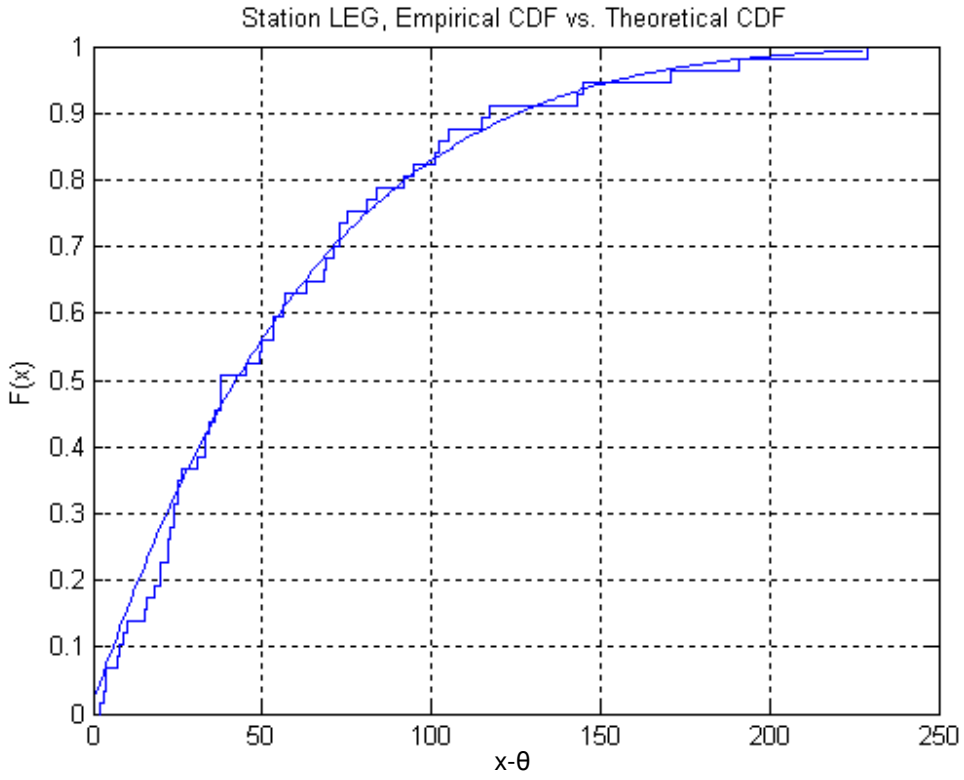

Station YM6 Empirical CDF vs. Theoretical CDF
F(x)
x-θ
0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
0.9
1
0
50
100
150
200
250

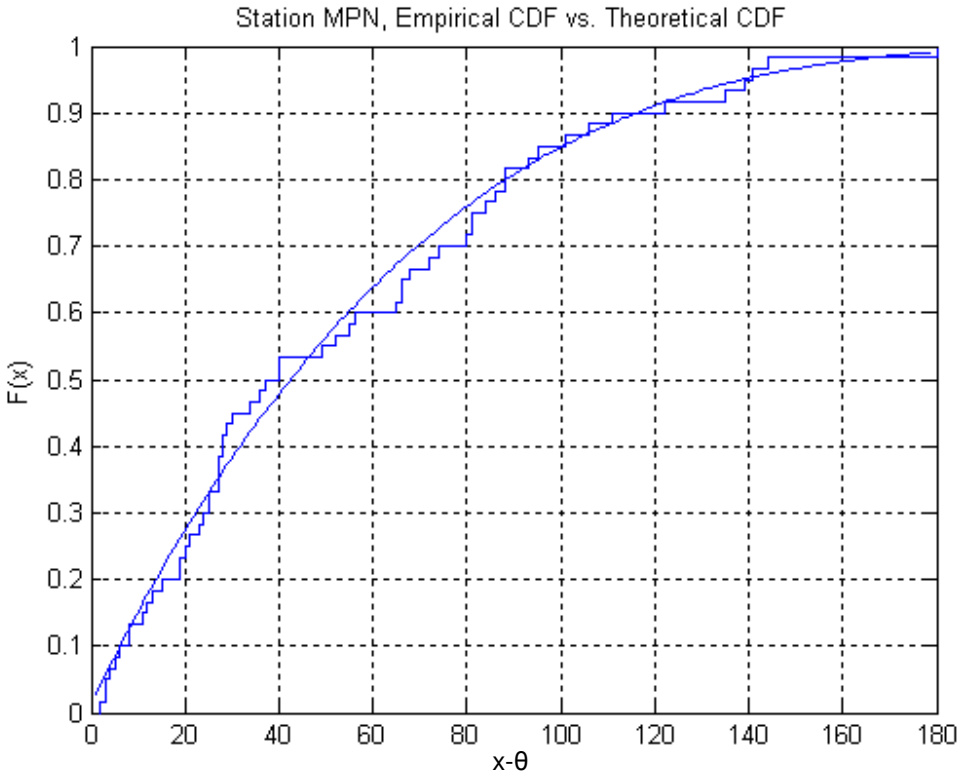
A.10 Comparing the empirical cdf of the peaks over threshold against the theoretical cdf, using parameters, k and σ from method – Method of Moments

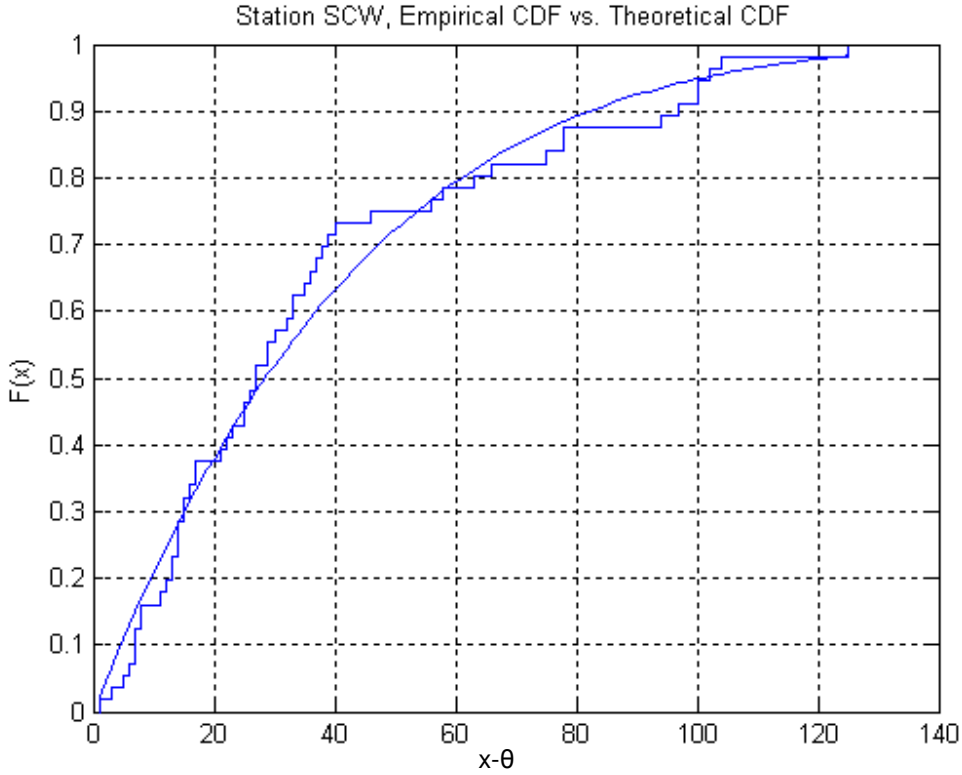


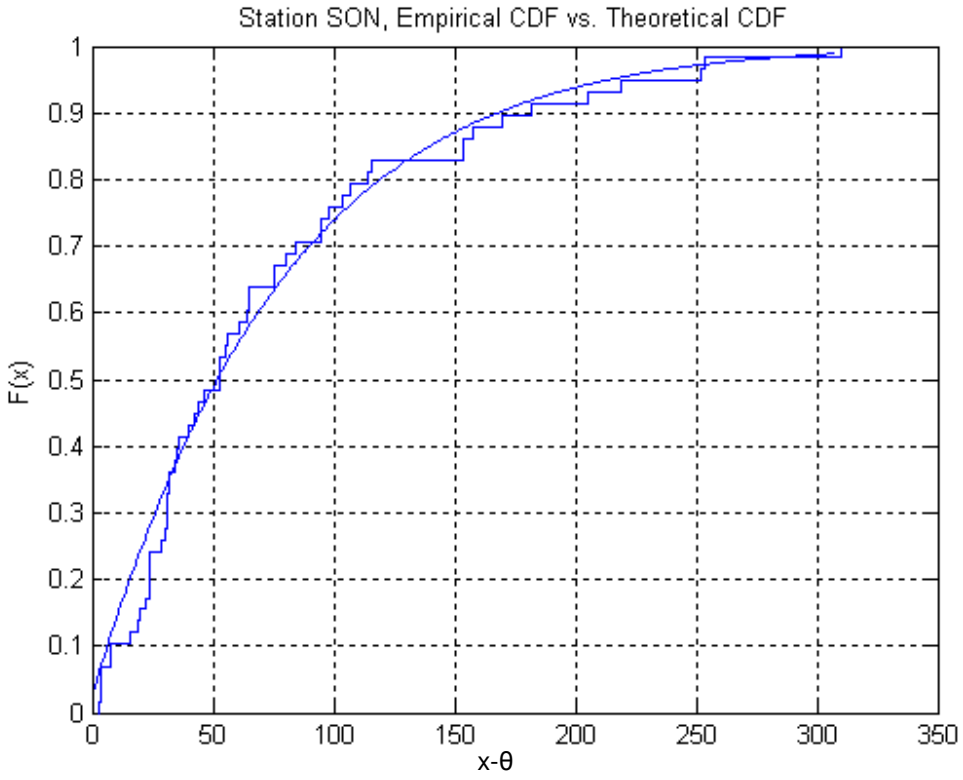


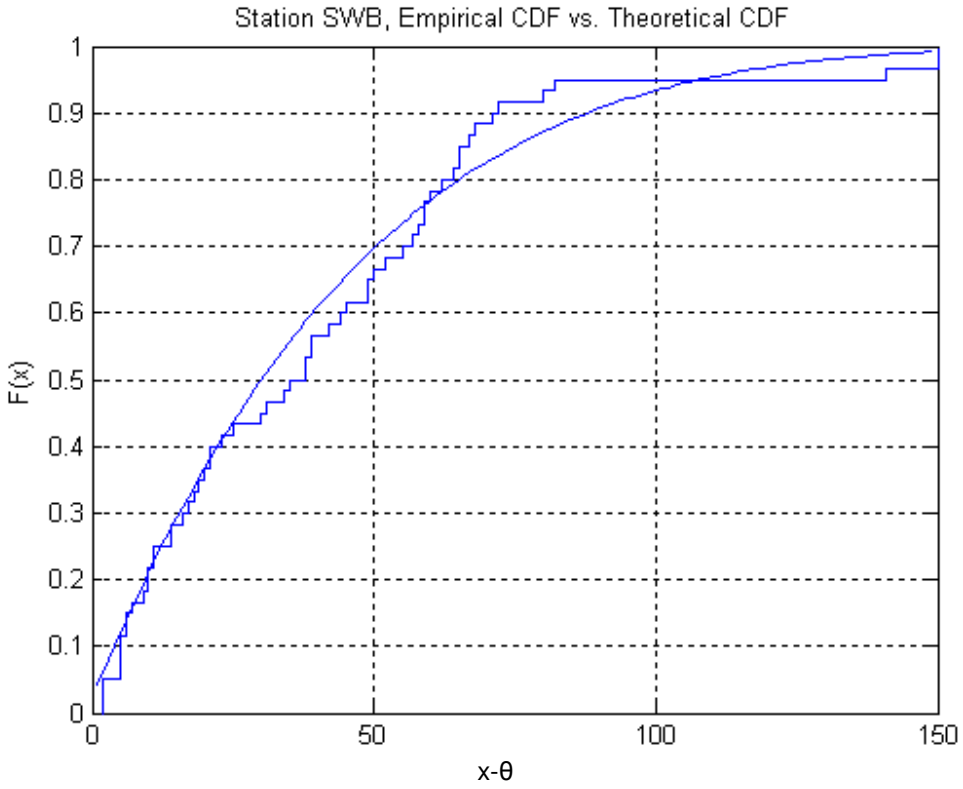

Station K13, Empirical CDF vs. Theoretical CDF
F(x)
x-θ
0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
0.9
1
0
50
100
150
200
250

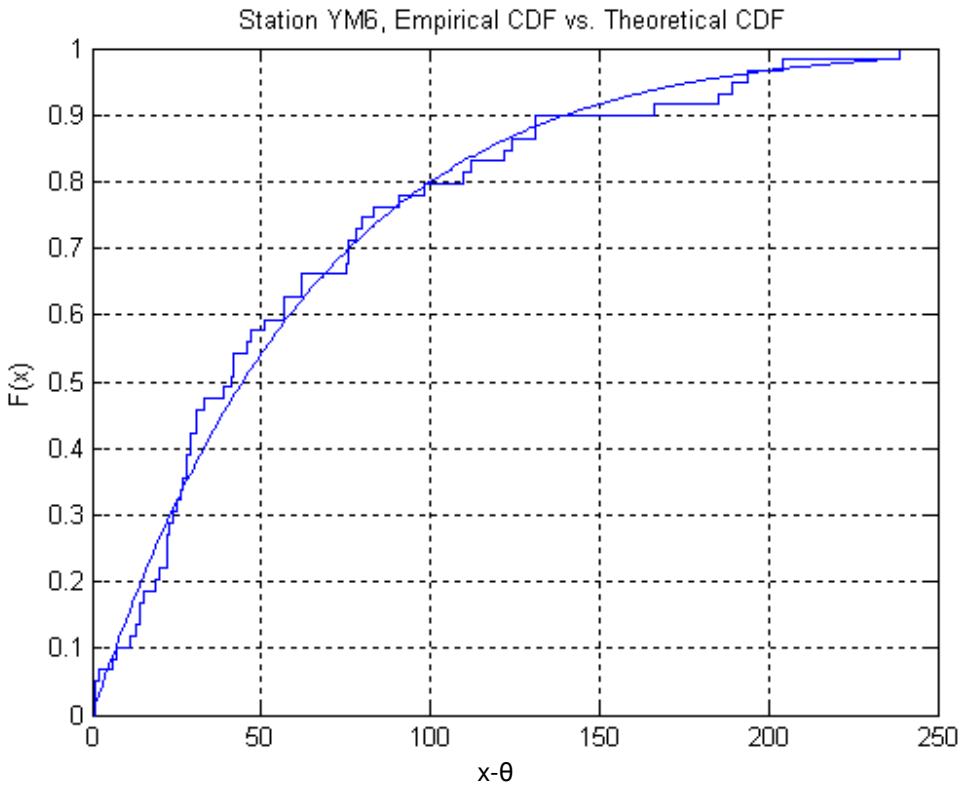

Station LEG, Empirical CDF vs. Theoretical CDF
F(x)
x-θ
0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
0.9
1
0
50
100
150
200
250


Station MPN, Empirical CDF vs. Theoretical CDF
F(x)
x-θ
0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
0.9
1
0
20
40
60
80
100
120
140
160
180


Station SCW, Empirical CDF vs. Theoretical CDF
F(x)
x-θ
0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
0.9
1
0
20
40
60
80
100
120
140


Station SON, Empirical CDF vs. Theoretical CDF
F(x)
x-θ
0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
0.9
1
0
50
100
150
200
250
300
350

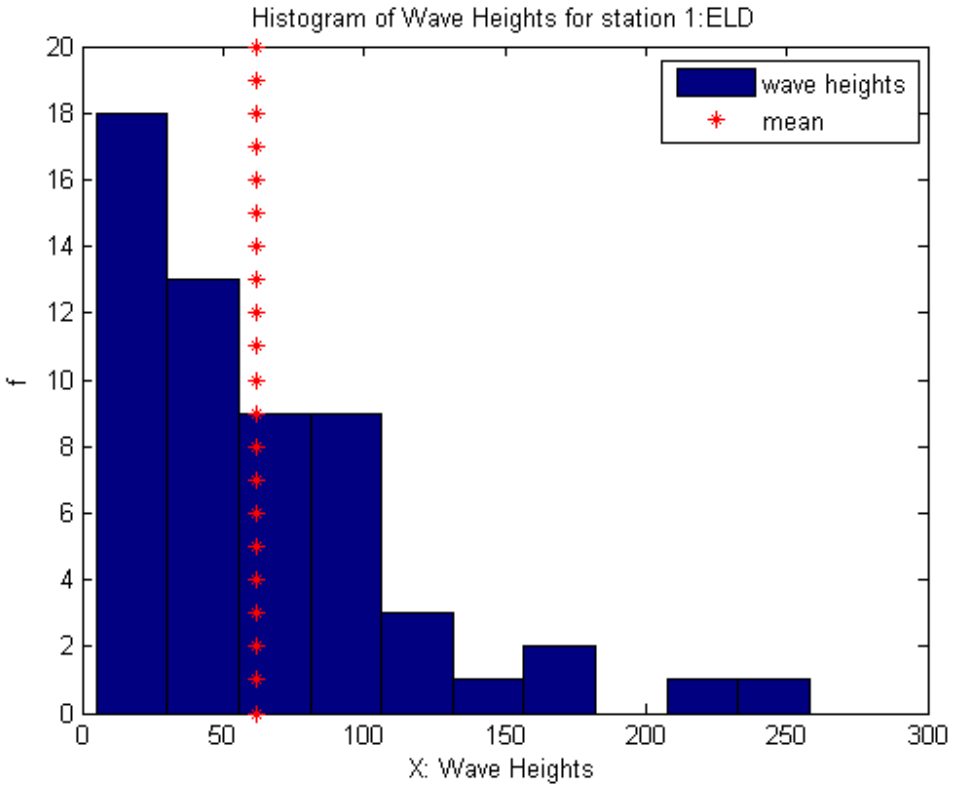

Station SWB, Empirical CDF vs. Theoretical CDF
F(x)
x-θ
0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
0.9
1
0
50
100
150

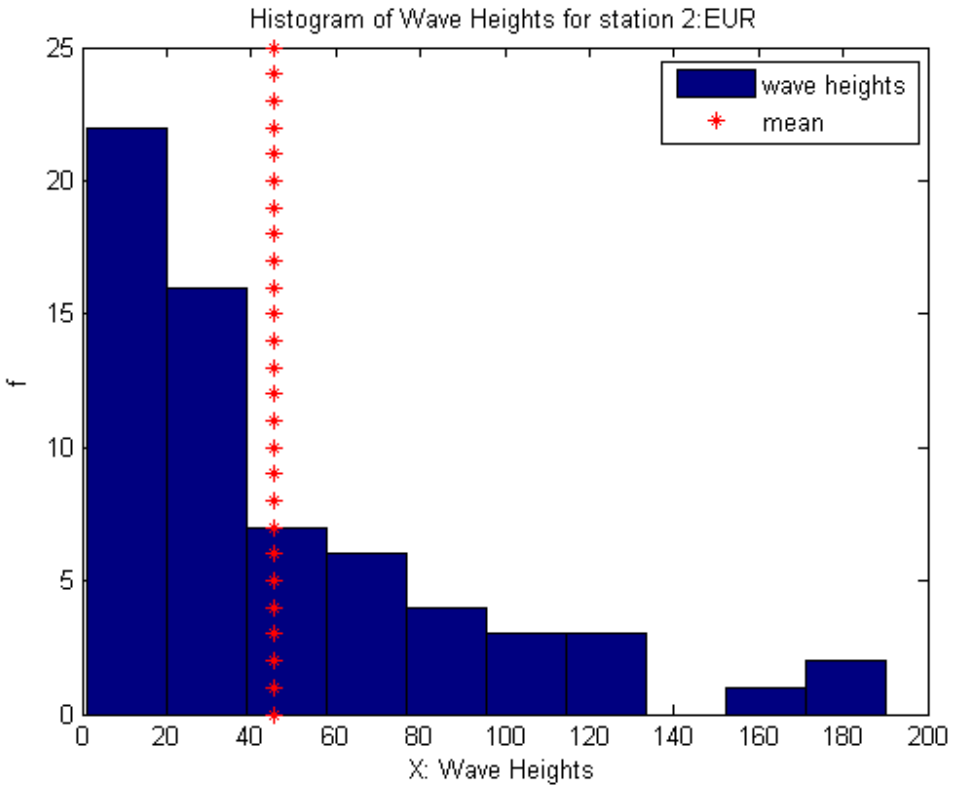


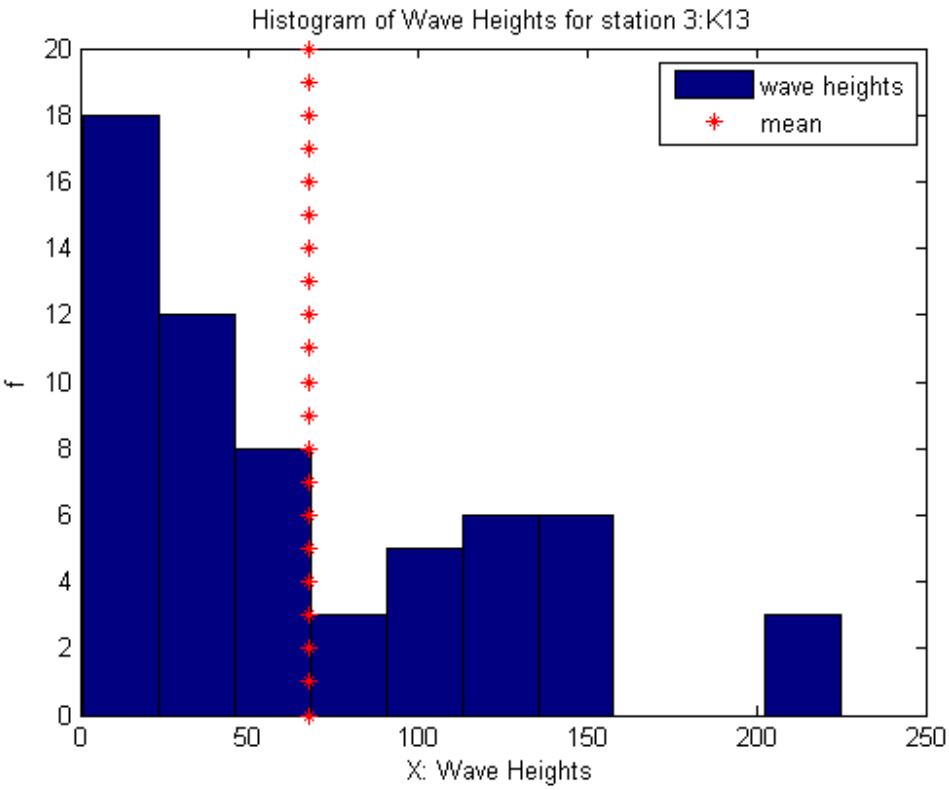
	ELD	EUR	K13	LEG	MPN	SCW	SON	SWB	YM6
h	0	1	0	0	0	1	0	1	1
p-value	0.7261	0.0238	0.7255	0.5155	0.5789	0.0022	0.3701	0.0124	0.0035
ks-stat	0.0899	0.1937	0.0900	0.1064	0.1014	0.2398	0.1193	0.2075	0.2316
crit.val	0.1767	0.1767	0.1767	0.1767	0.1767	0.1767	0.1767	0.1767	0.1767

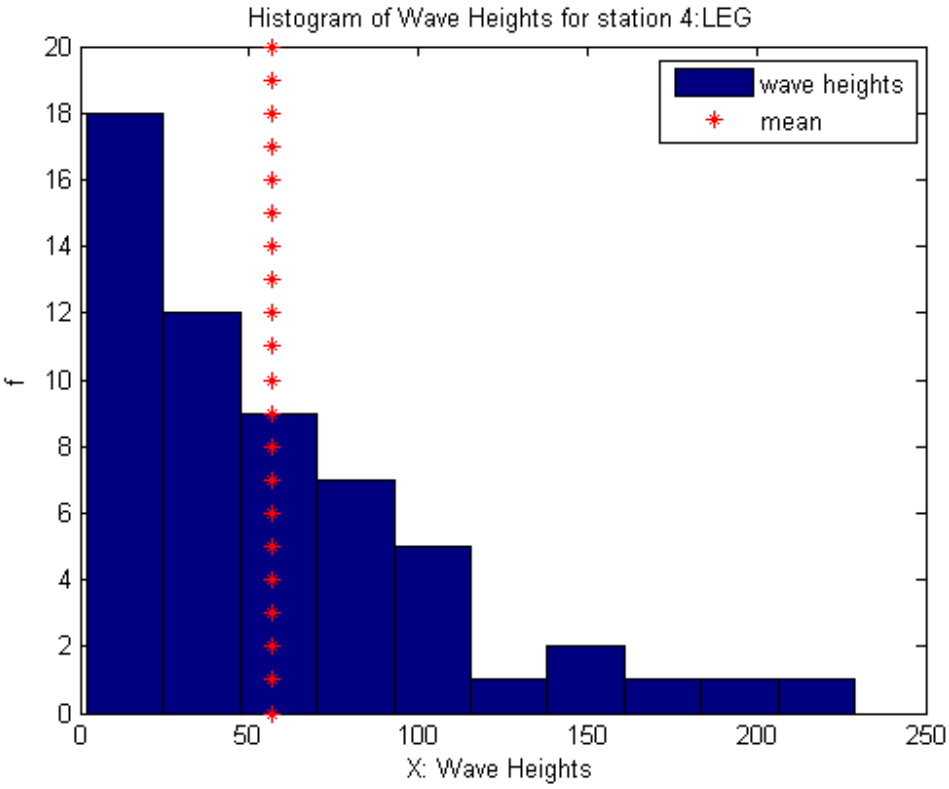
- h returns a 0 or 1, 0 means the populations are equal
- p returns the p-value: the probability that the one of populations is
- drawn from the other, higher means populations are the same
- ksstat is the maximum different at one point between the two cdf's

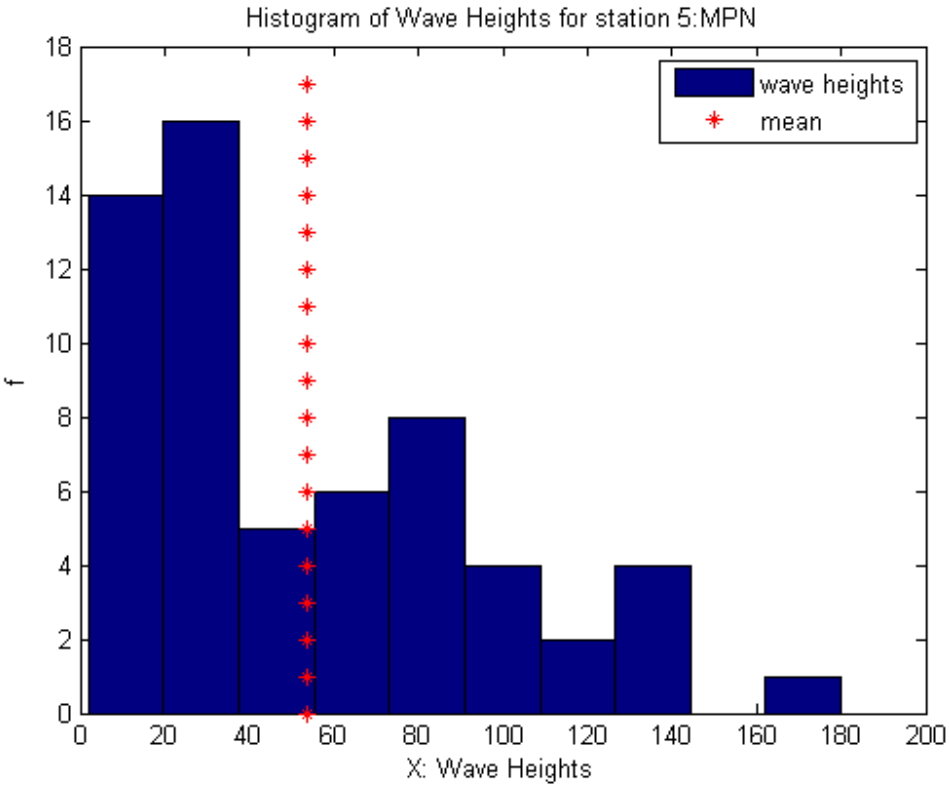
A.11: Histograms of Wave Heights, above θ

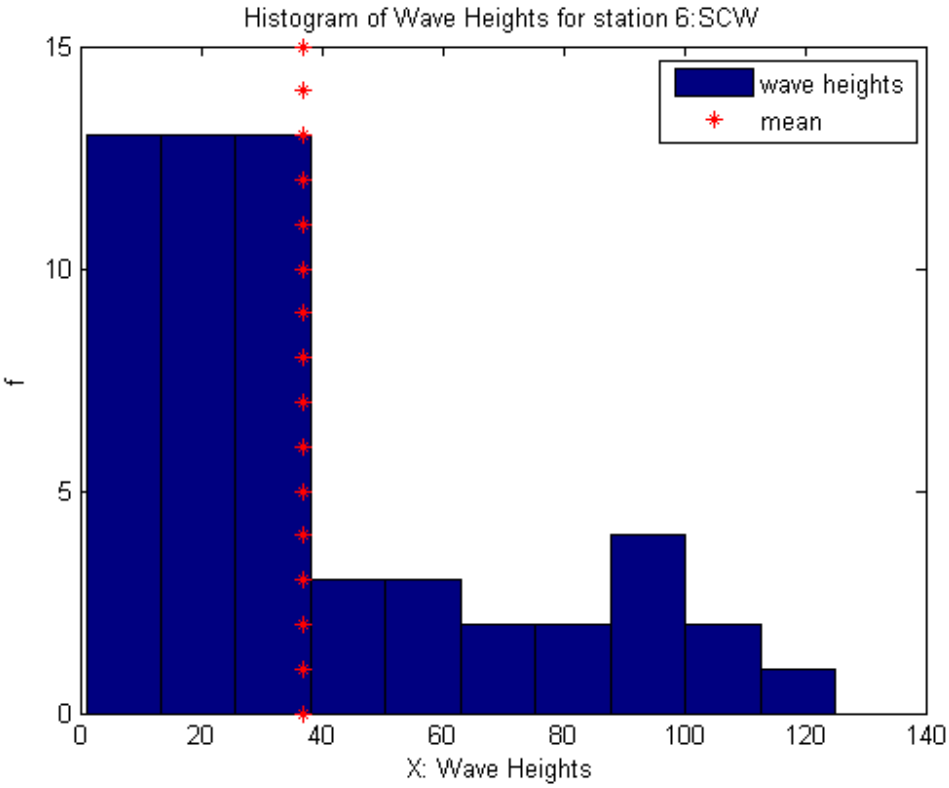


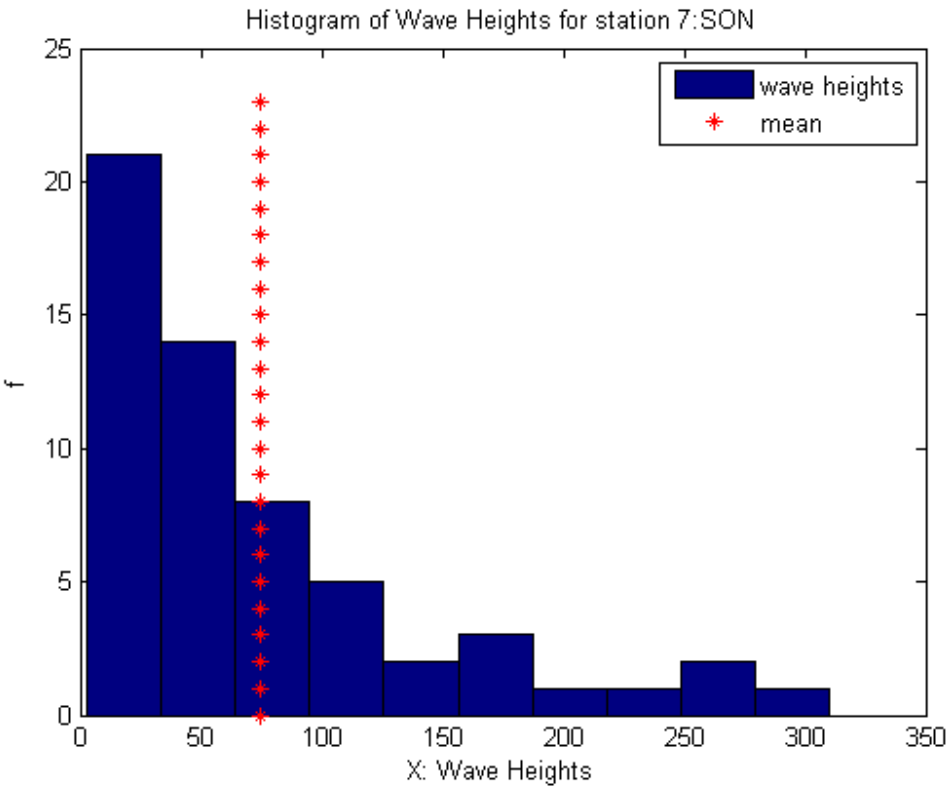


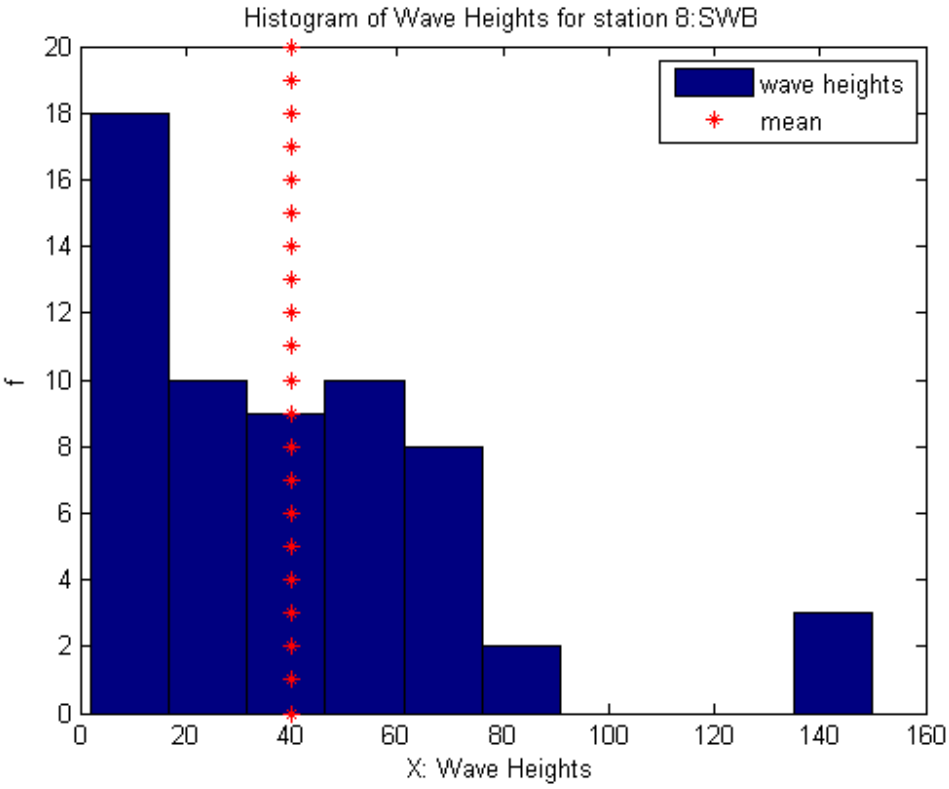

Histogram of Wave Heights for station 3:K13
wave heights
mean
f
X: Wave Heights
0
2
4
6
8
10
12
14
16
18
20
0
50
100
150
200
250


Histogram of Wave Heights for station 4:LEG
wave heights
mean
f
X: Wave Heights
0
2
4
6
8
10
12
14
16
18
20
0
50
100
150
200
250

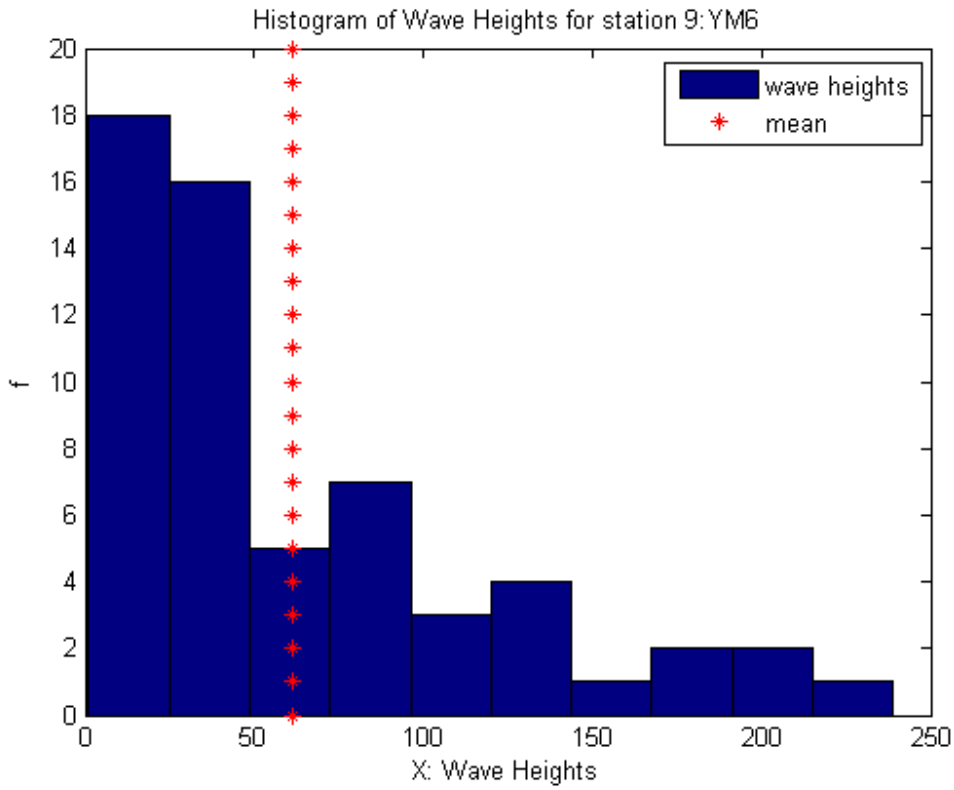

Histogram of Wave Heights for station 5:MPN
wave heights
mean
f
X: Wave Heights
0
2
4
6
8
10
12
14
16
18
0
20
40
60
80
100
120
140
160
180
200

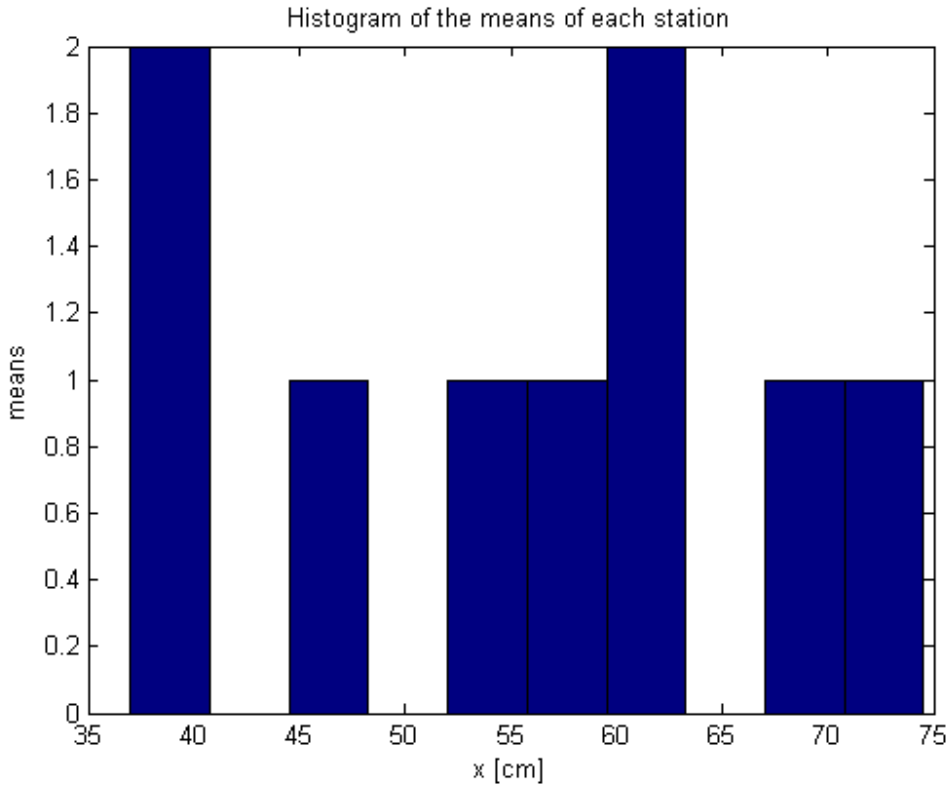

Histogram of Wave Heights for station 6:SCW
wave heights
mean
f
X: Wave Heights
0
5
10
15
0
20
40
60
80
100
120
140


Histogram of Wave Heights for station 7:SON
wave heights
mean
f
X: Wave Heights
0
5
10
15
20
25
0
50
100
150
200
250
300
350


Histogram of Wave Heights for station 8:SWB
wave heights
mean
f
X: Wave Heights
0
2
4
6
8
10
12
14
16
18
20
0
20
40
60
80
100
120
140
160

Poorwa Singh
1242326





A.12 $f(k_1 | X_1, \ldots, X_9)$, wave heights, and log (1-F)

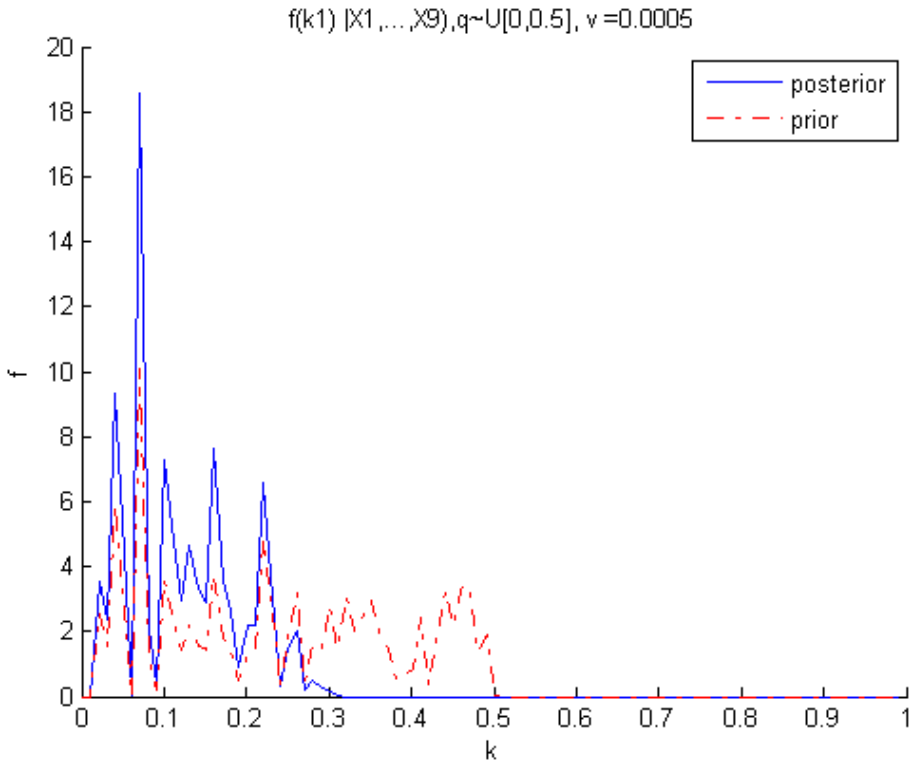
1 in 10,000 Year Wave Heights (above threshold) [cm], using all the Data from each Station

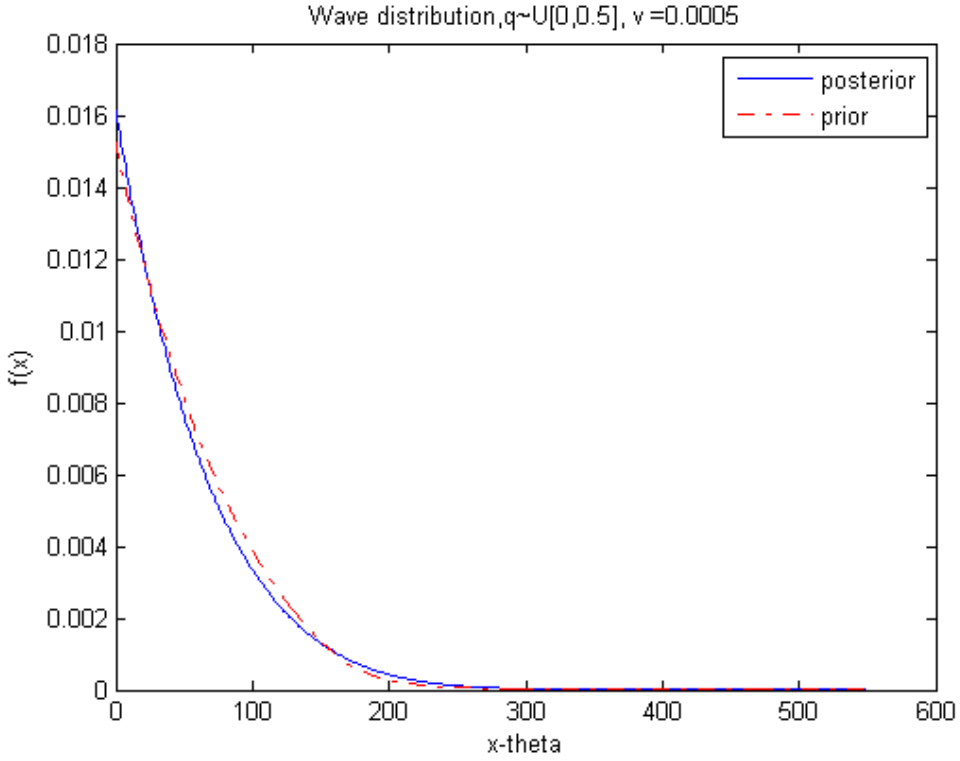
	q~U[0,0.5]		q~U[0,3]	
v =	**prior**	**posterior**	**prior**	**posterior**
0.0005	469.93	610.65	62.29	512.42
0.001	501.49	619.27	79.32	547.03
0.02	421.13	587.55	161.65	727.39
0.04	450.58	602.76	88.71	581.58
0.06	436.42	594.80	96.20	589.68
0.08	449.34	600.05	99.12	600.27
0.1	446.12	595.87	93.61	589.55
0.2	438.58	595.88	98.23	596.74
0.5	489.87	630.93	130.41	620.29
0.8	542.65	674.18	179.81	647.43

1 in 10,000 Year Wave Heights [cm], using all the Data from each Station

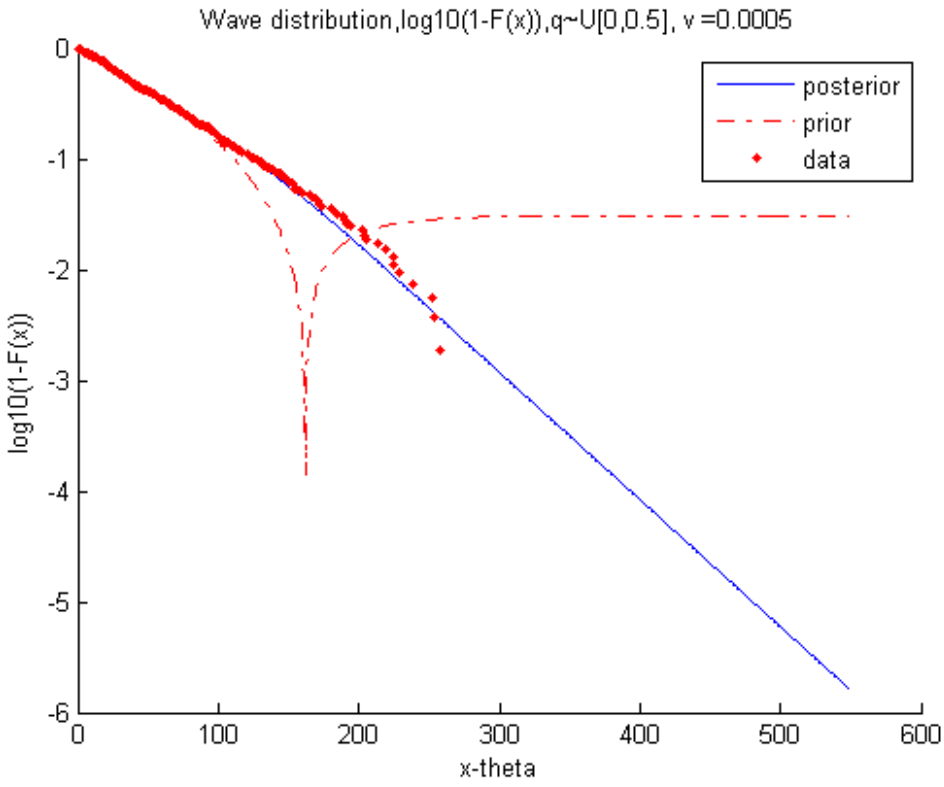
	q~U[0,0.5]		q~U[0,3]	
v =	**prior**	**posterior**	**prior**	**posterior**
0.0005	1008.93	1149.65	601.29	1051.42
0.001	1040.49	1158.27	618.32	1086.03
0.02	960.13	1126.55	700.65	1266.39
0.04	989.58	1141.76	627.71	1120.58
0.06	975.42	1133.8	635.2	1128.68
0.08	988.34	1139.05	638.12	1139.27
0.1	985.12	1134.87	632.61	1128.55
0.2	977.58	1134.88	637.23	1135.74
0.5	1028.87	1169.93	669.41	1159.29
0.8	1081.65	1213.18	718.81	1186.43

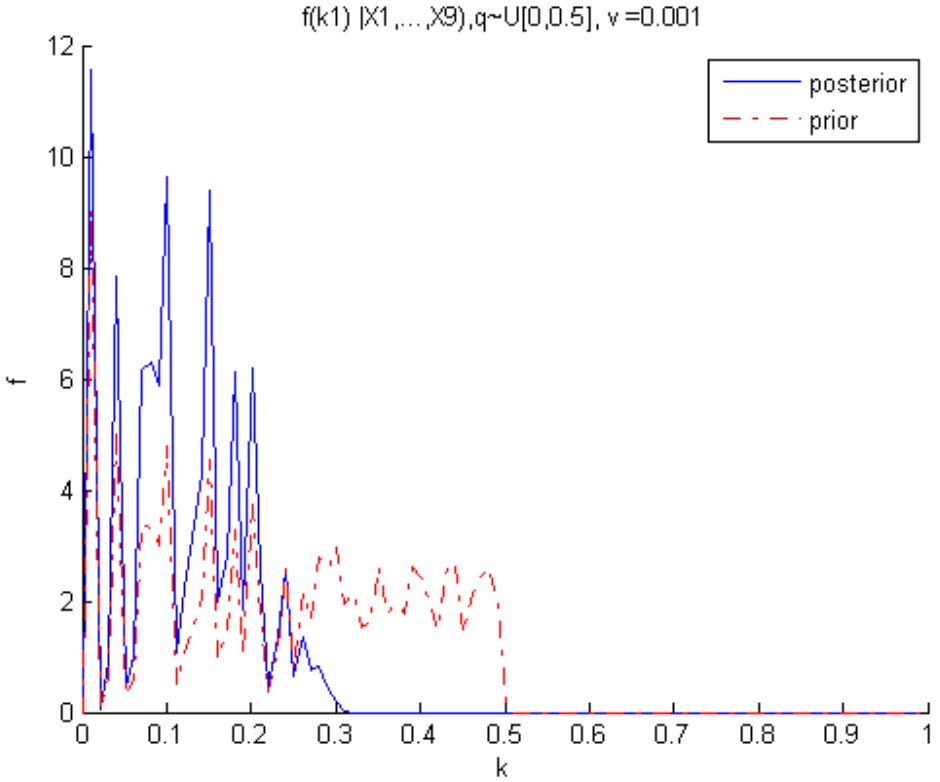
Poorwa Singh
1242326



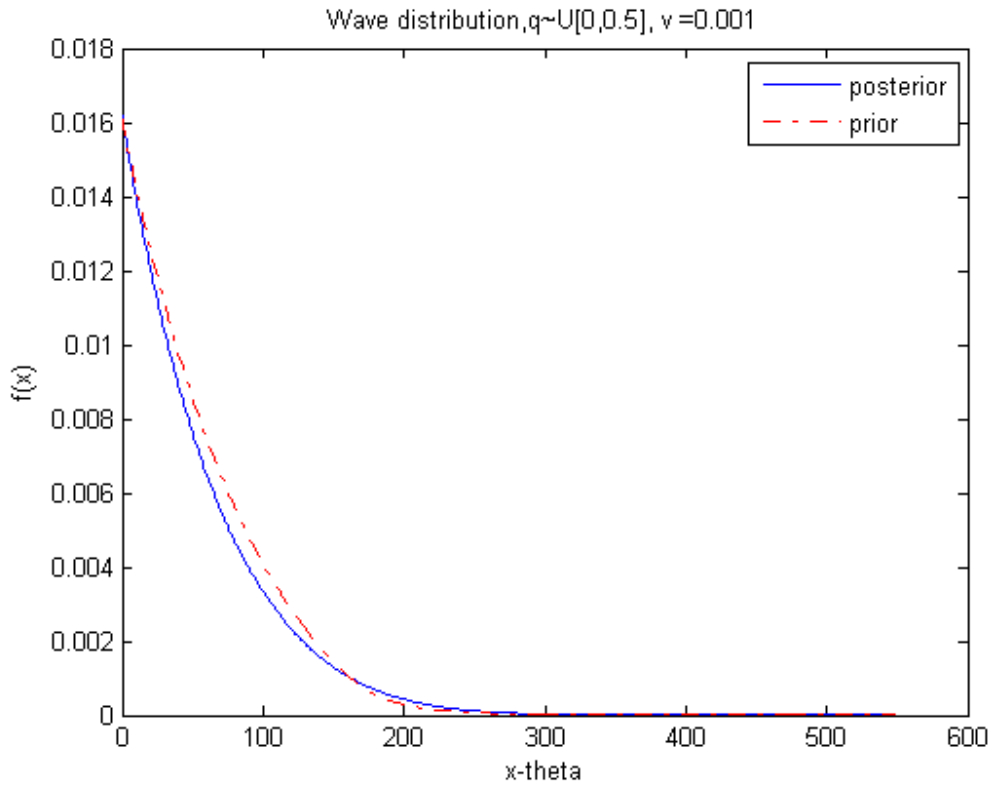


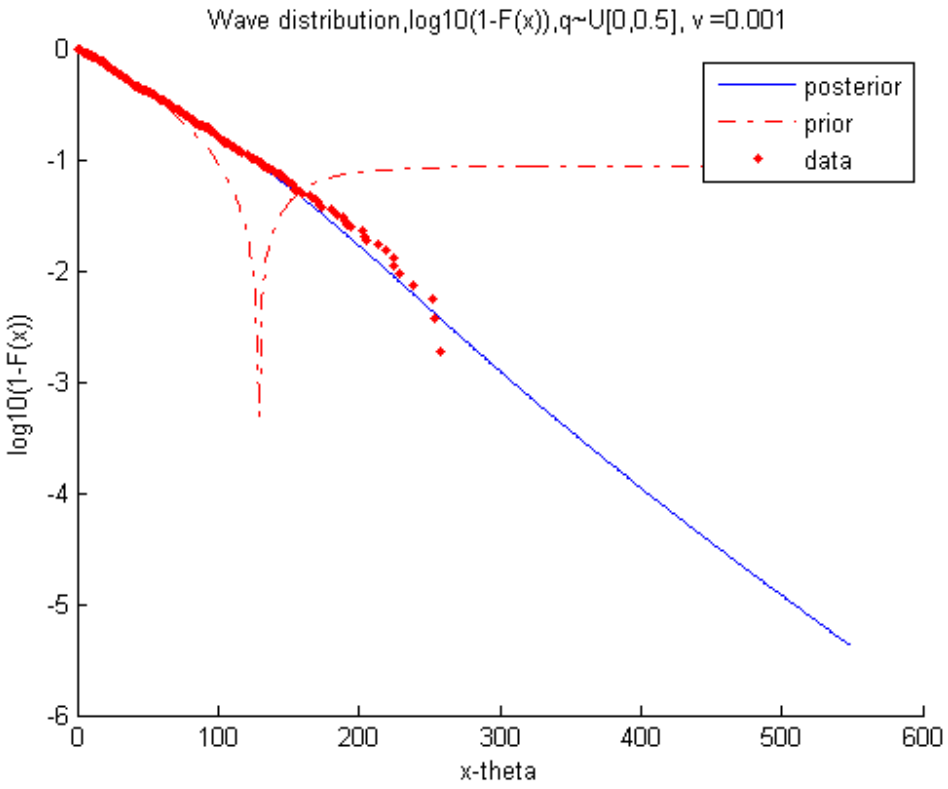
Poorwa Singh
1242326

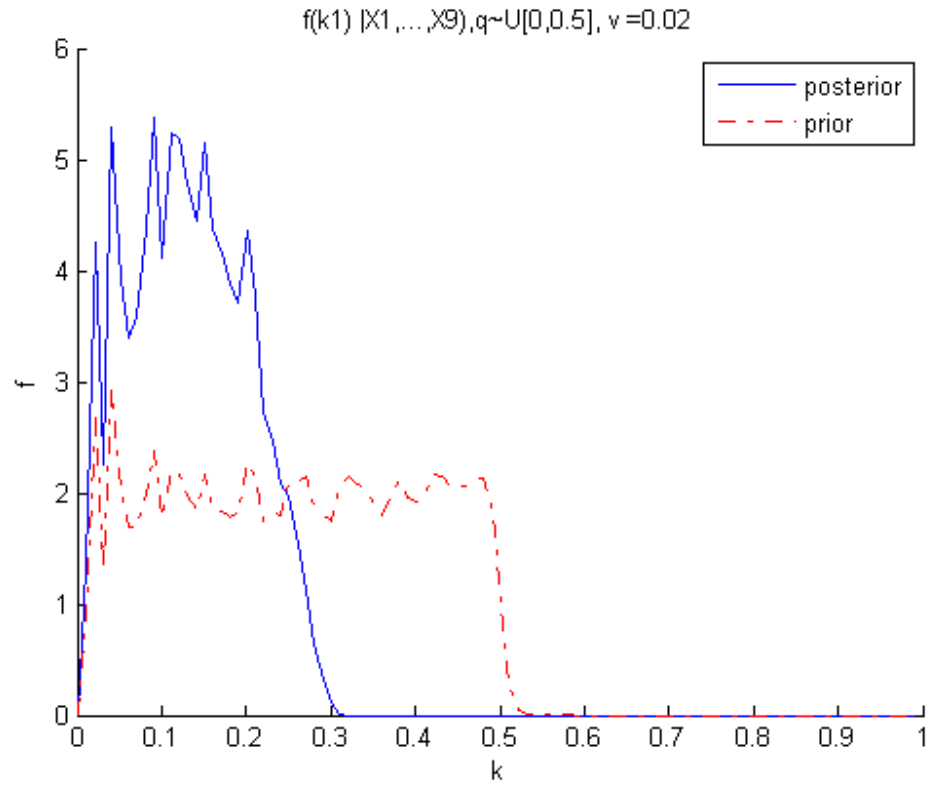




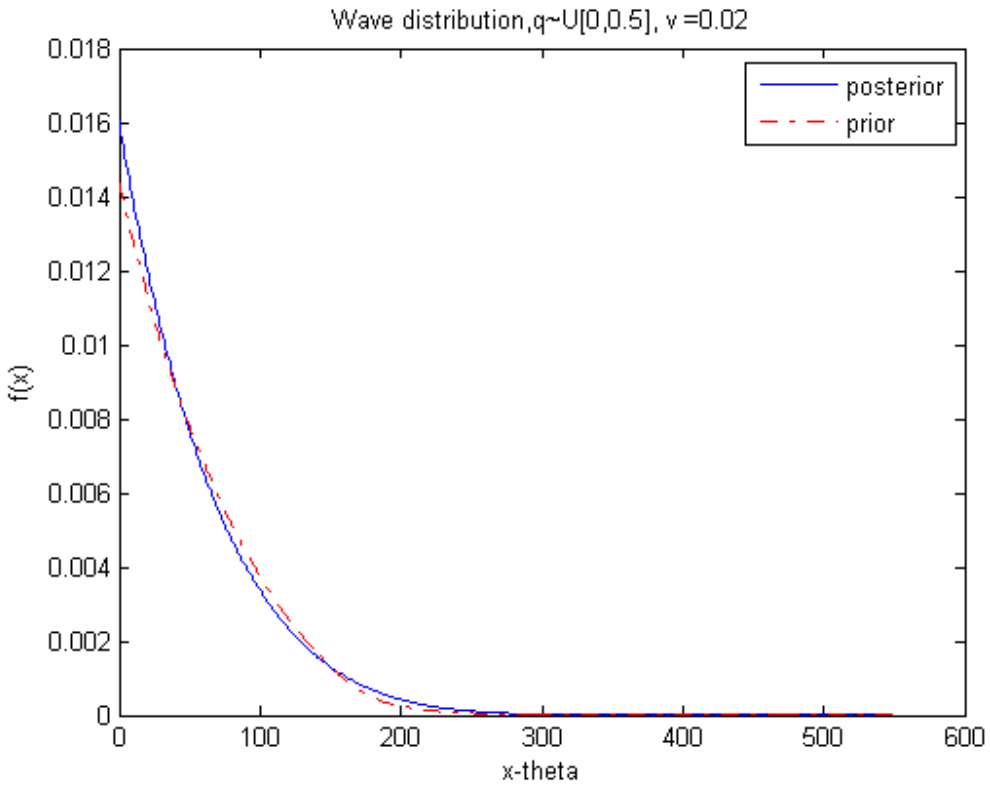
Poorwa Singh
1242326

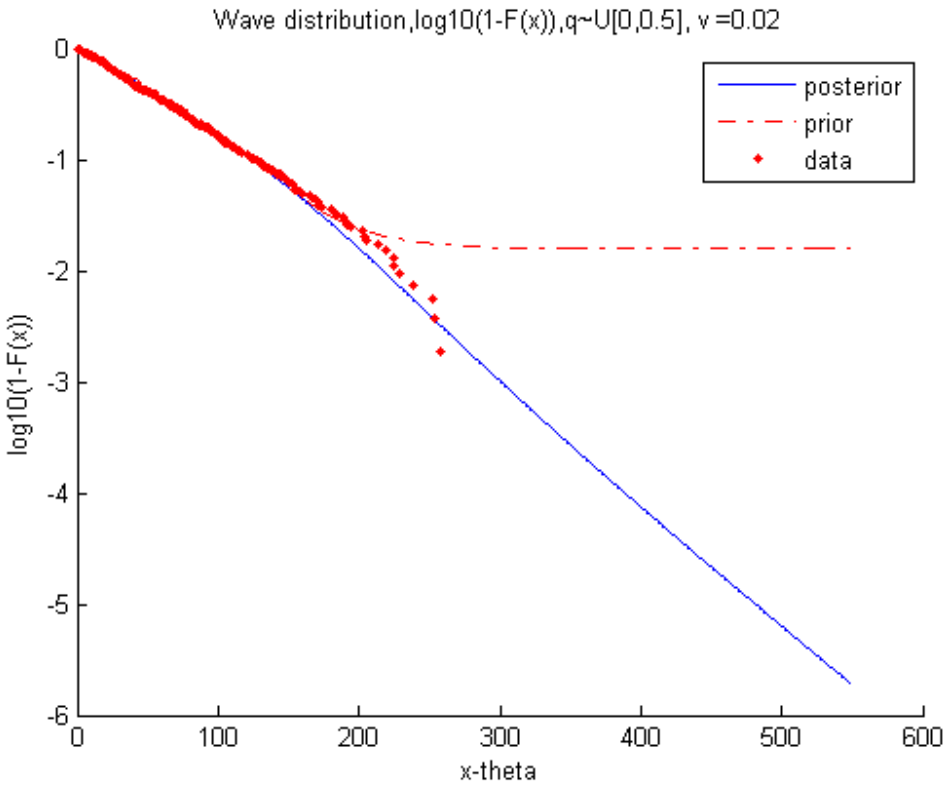



Wave distribution,log10(1-F(x)),q~U[0,0.5], v =0.001
posterior
prior
data
log10(1-F(x))
x-theta
0
-1
-2
-3
-4
-5
-6
0
100
200
300
400
500
600

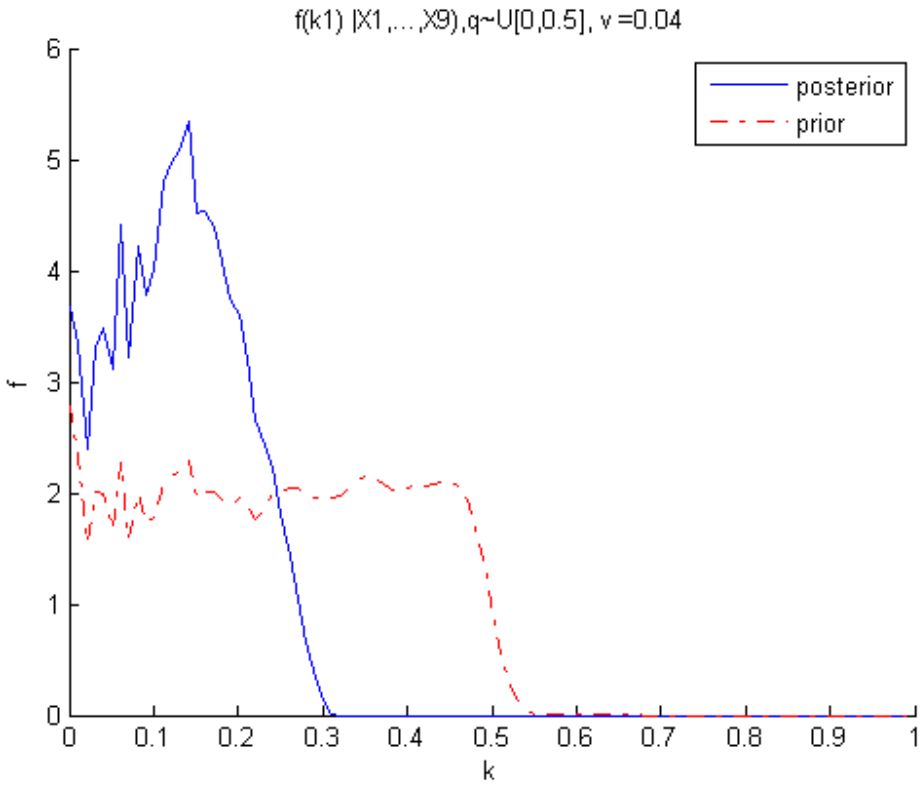

f(k1) |X1,...,X9),q~U[0,0.5], v =0.02
posterior
prior
f
k
6
5
4
3
2
1
0
0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
0.9
1

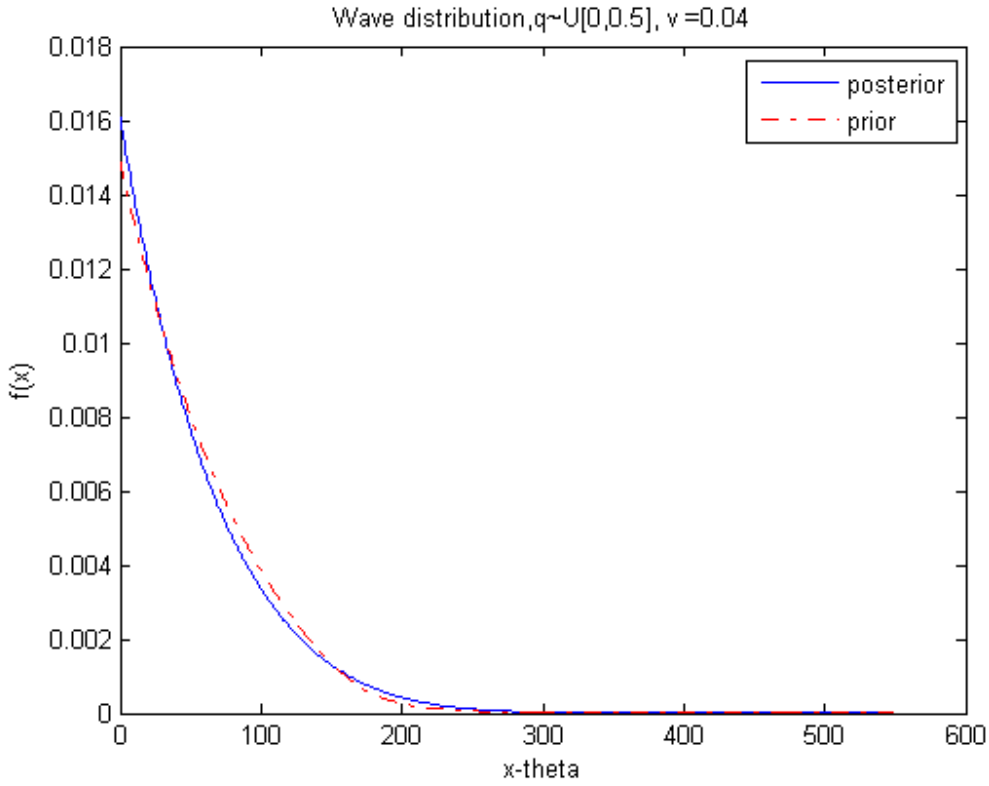
Poorwa Singh
1242326



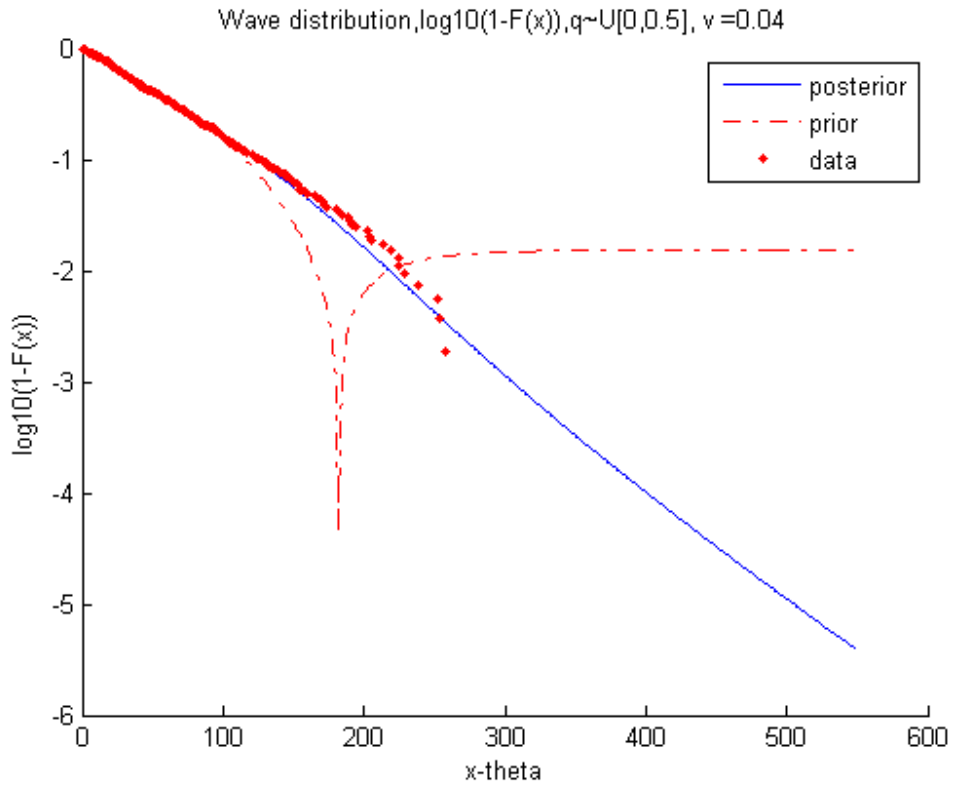


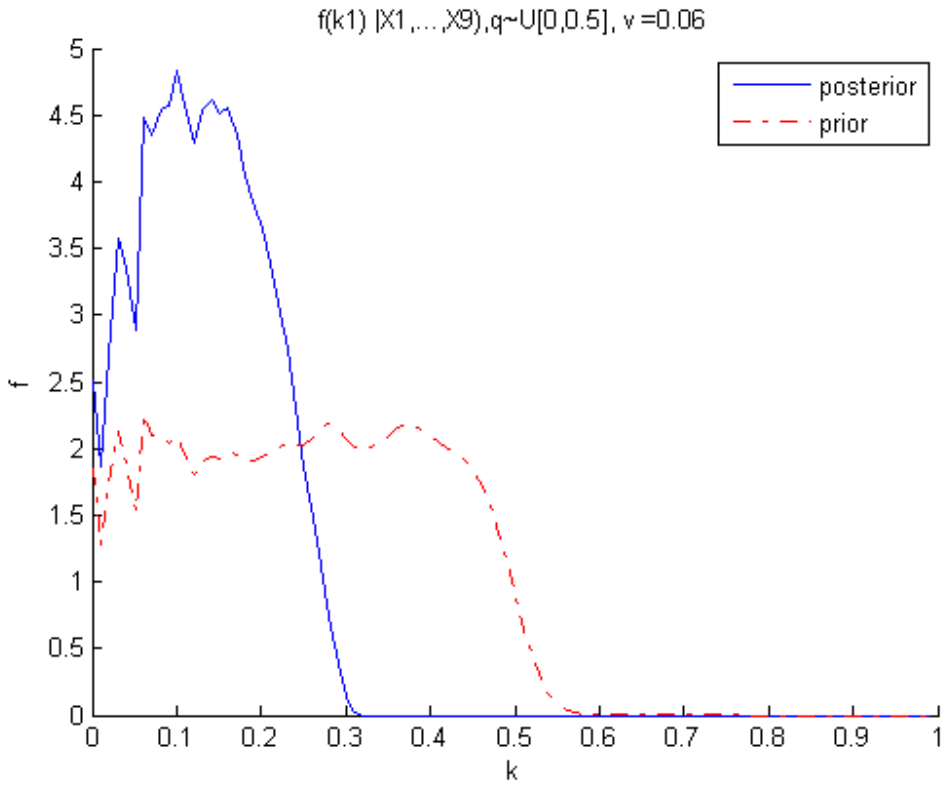
Poorwa Singh
1242326



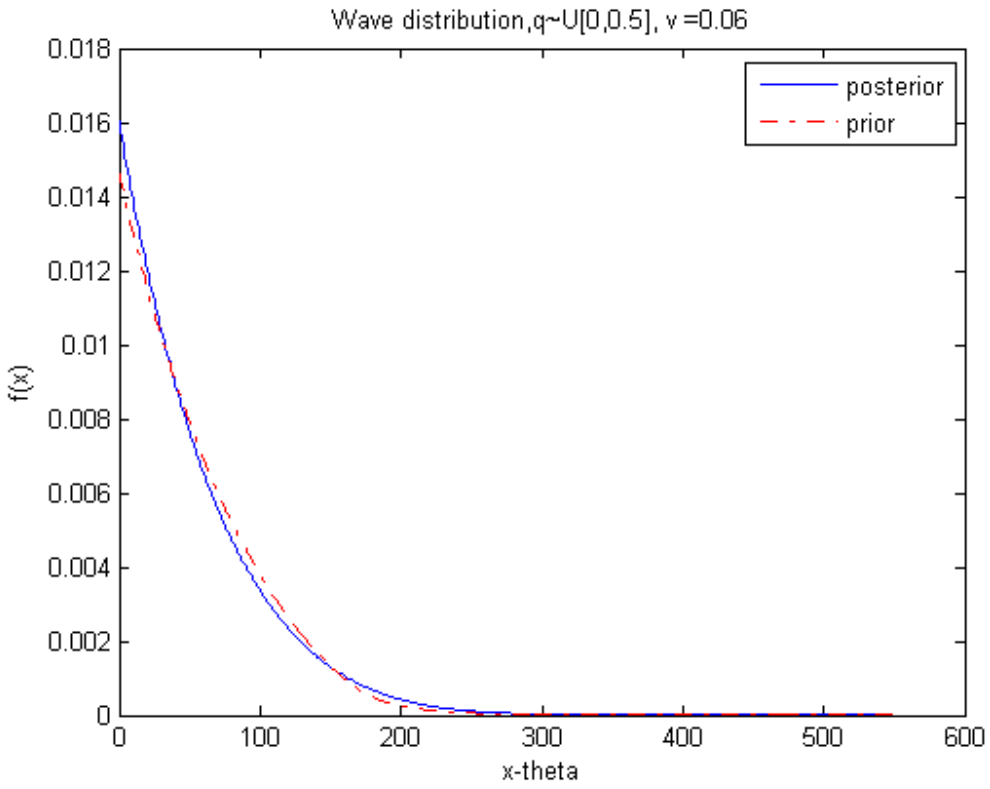


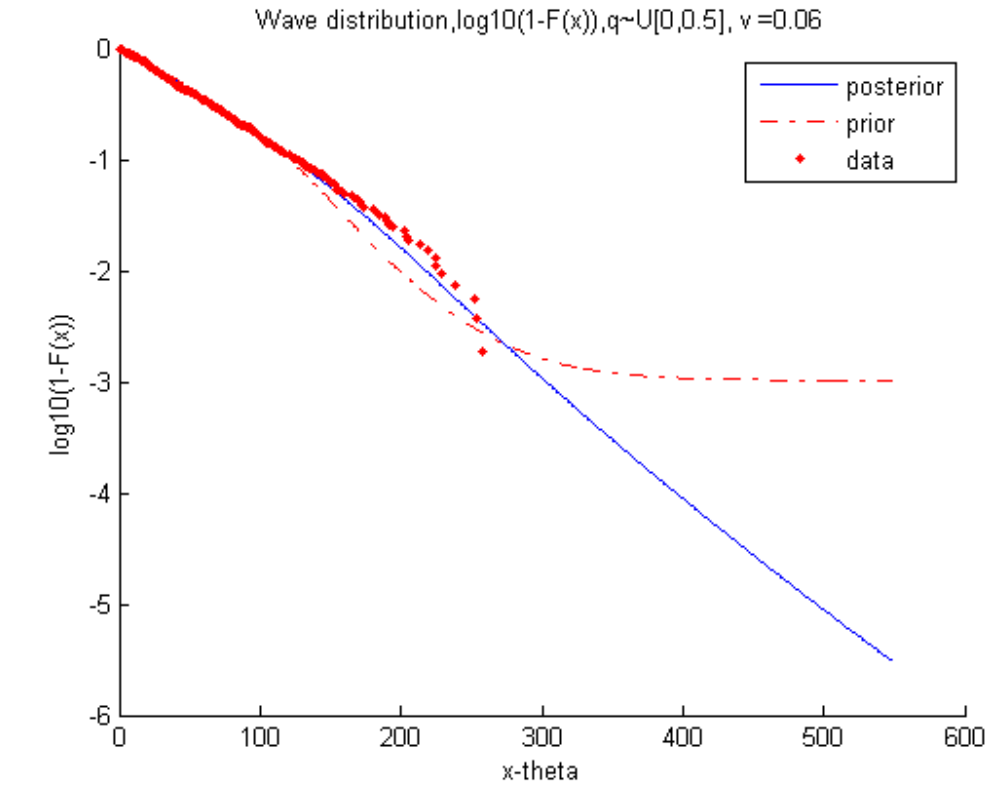
Poorwa Singh
1242326

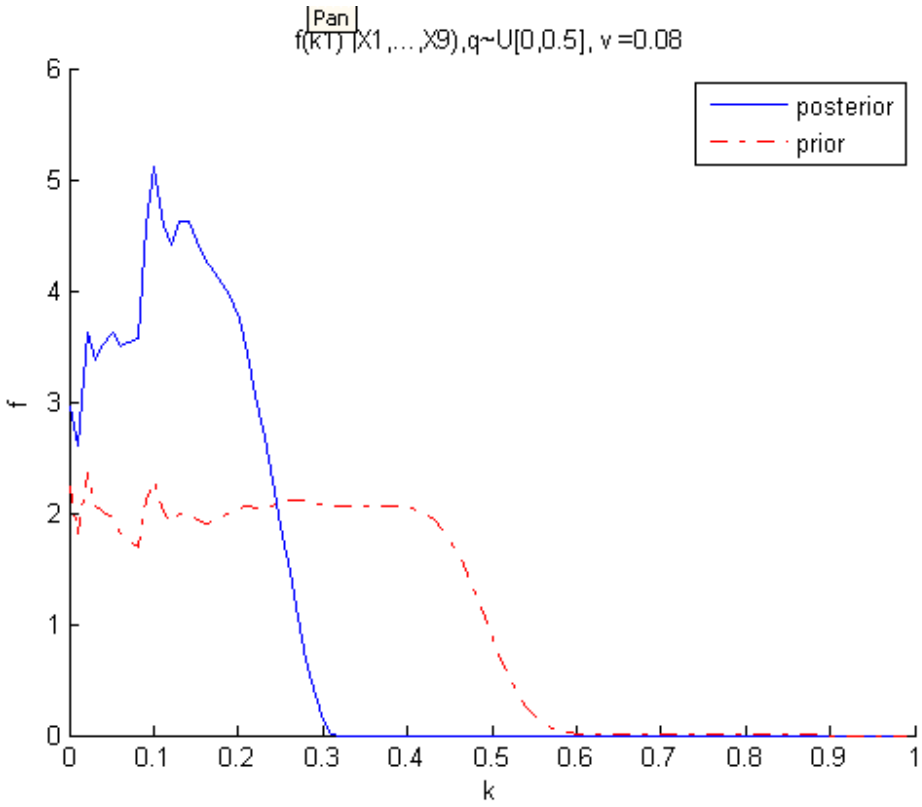


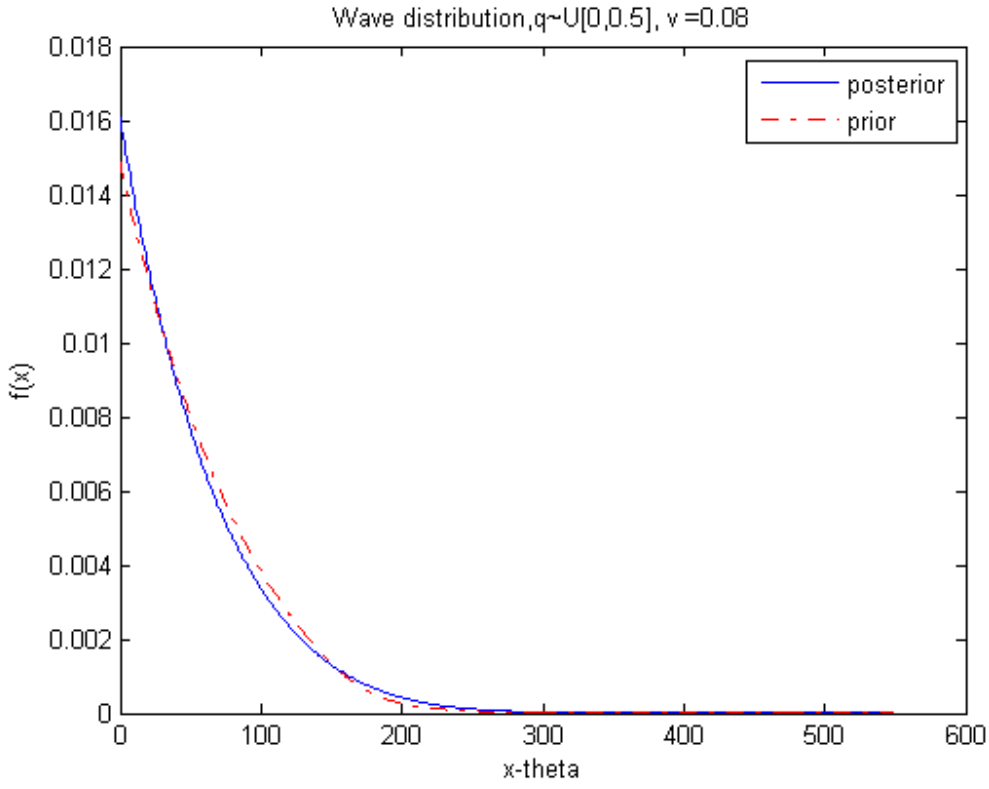


Poorwa Singh
1242326

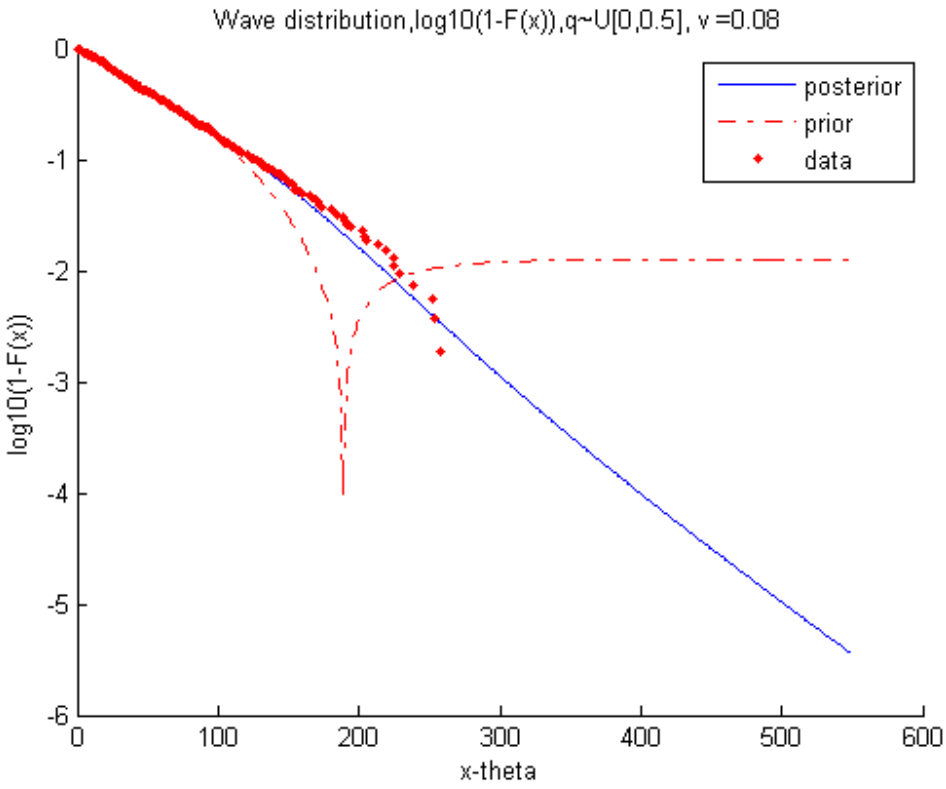




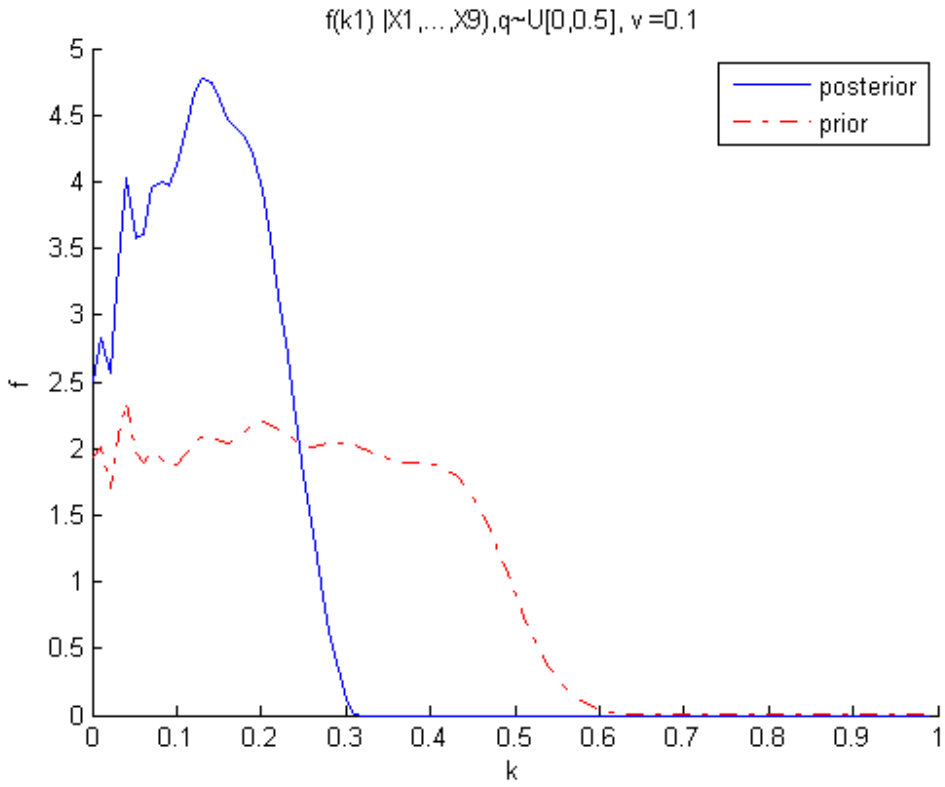

Pan
f(k|T) |X1,...,X9),q~U[0,0.5], v =0.08
posterior
prior
f
k
0
1
2
3
4
5
6
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
0.9

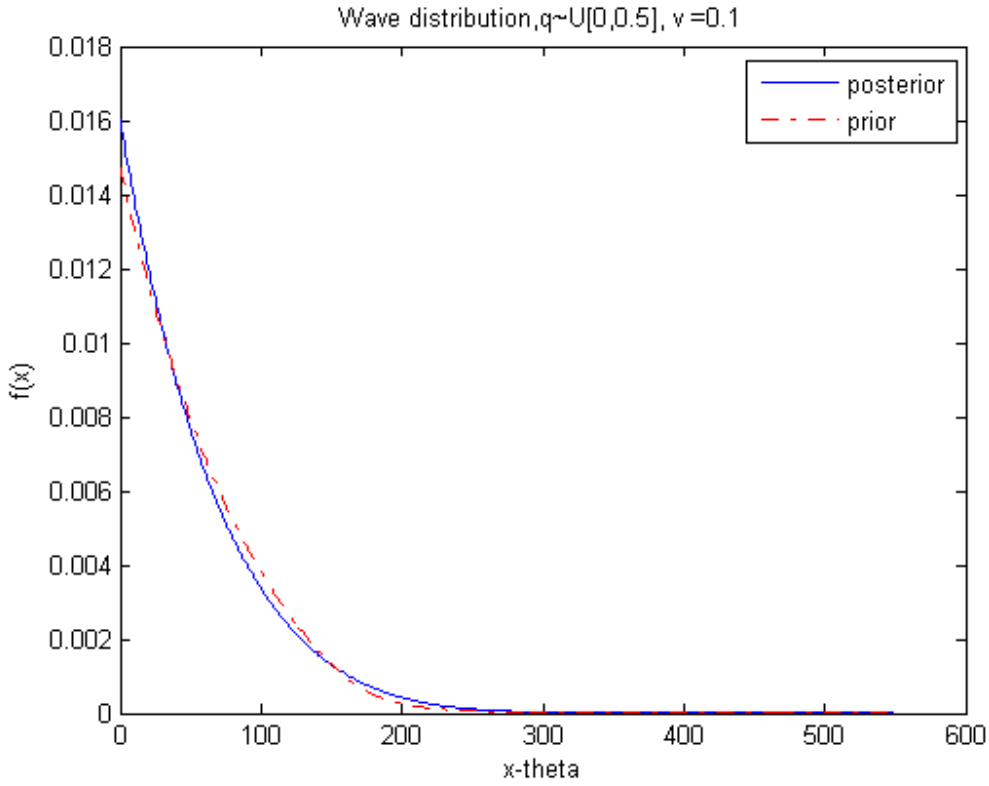

Wave distribution,q~U[0,0.5], v =0.08
posterior
prior
f(x)
x-theta
0.002
0.004
0.006
0.008
0.01
0.012
0.014
0.016
0.018
100
200
300
400
500
600

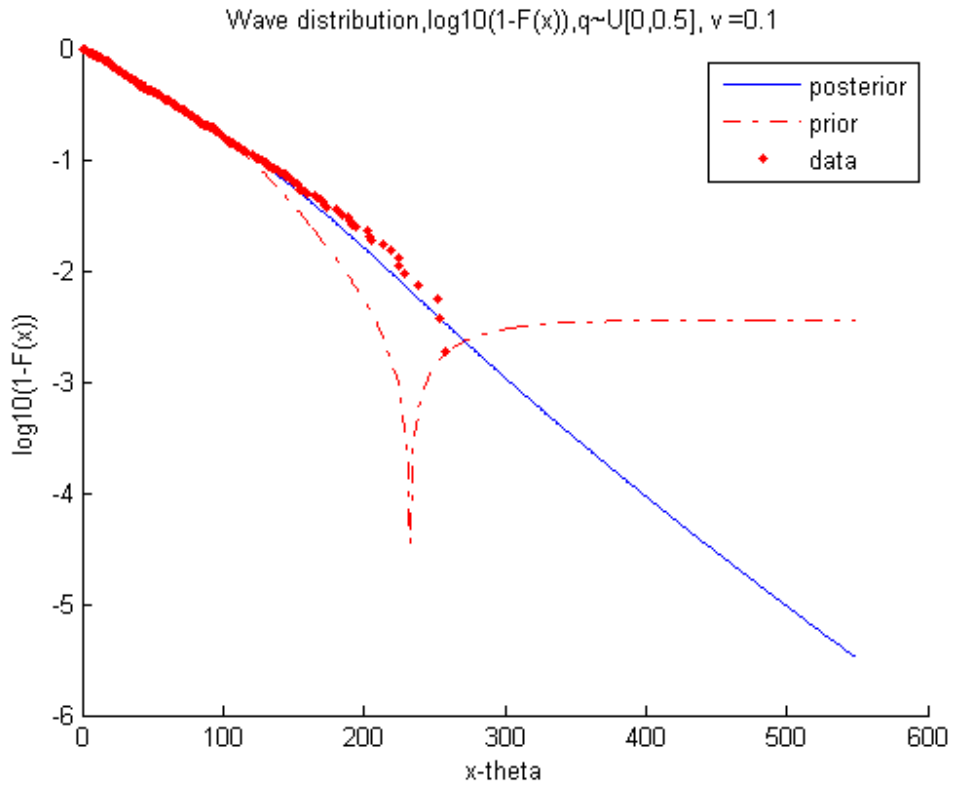
Poorwa Singh
1242326

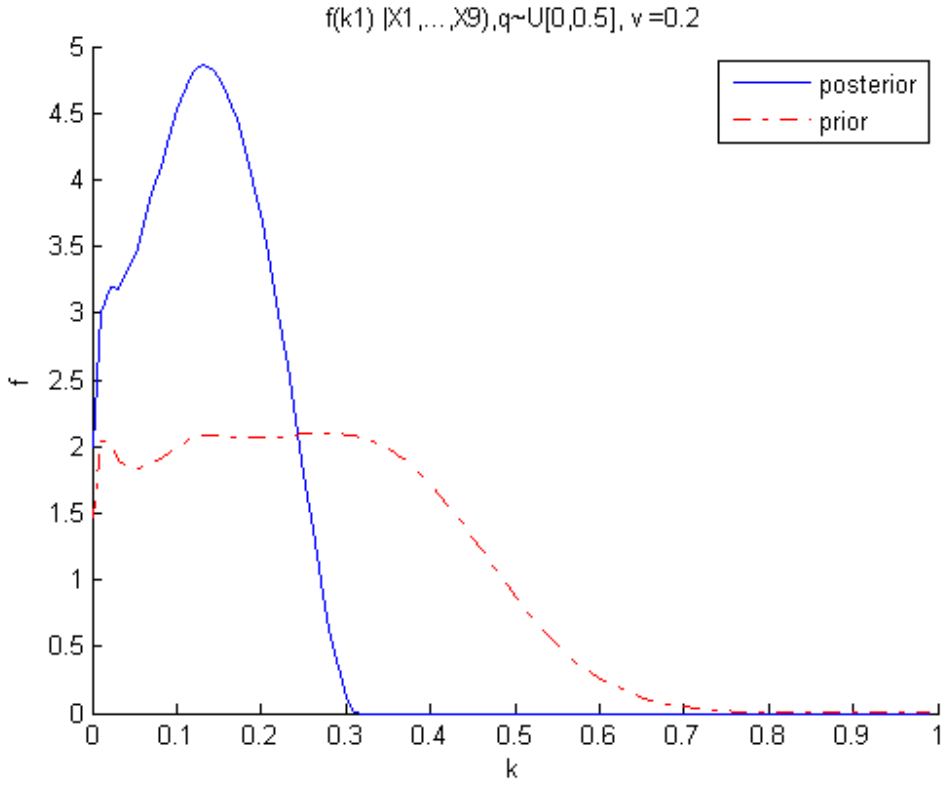


Poorwa Singh
1242326

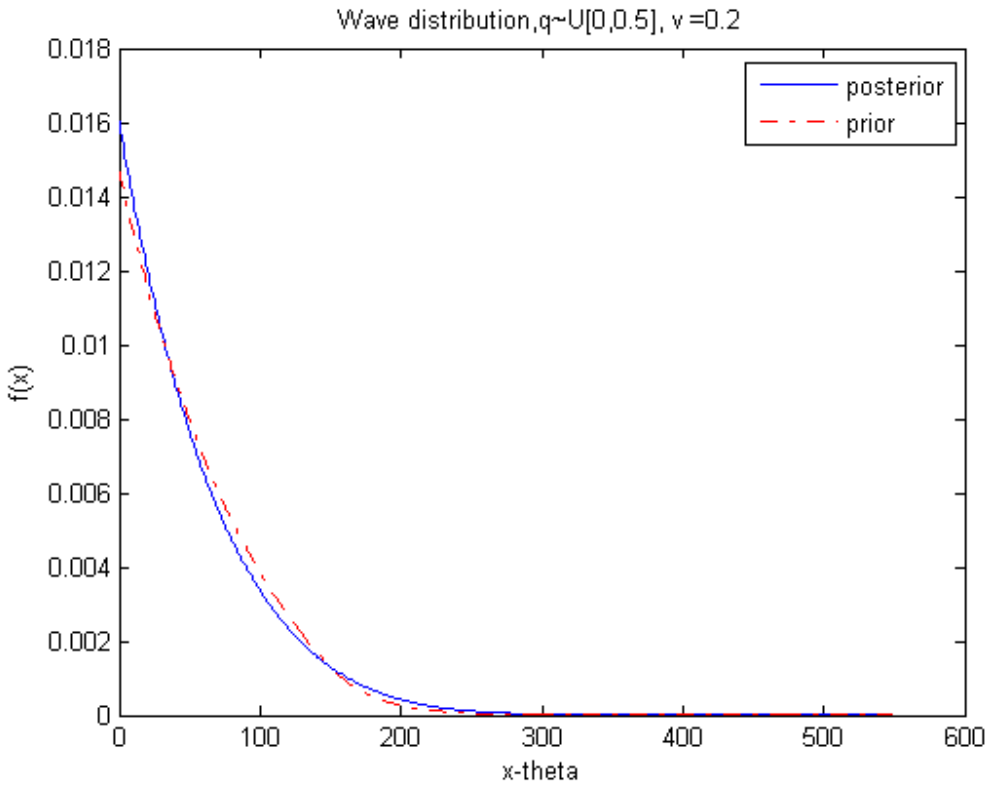


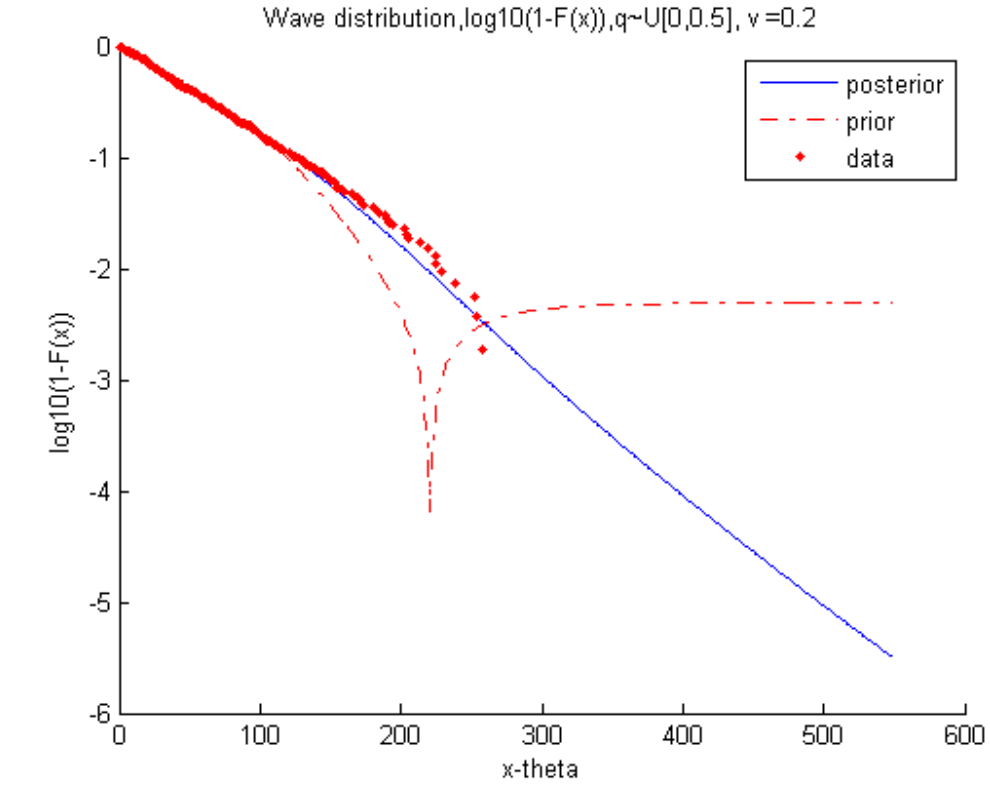



Wave distribution,log10(1-F(x)),q~U[0,0.5], v =0.1
posterior
prior
data
log10(1-F(x))
x-theta
0
-1
-2
-3
-4
-5
-6
0
100
200
300
400
500
600

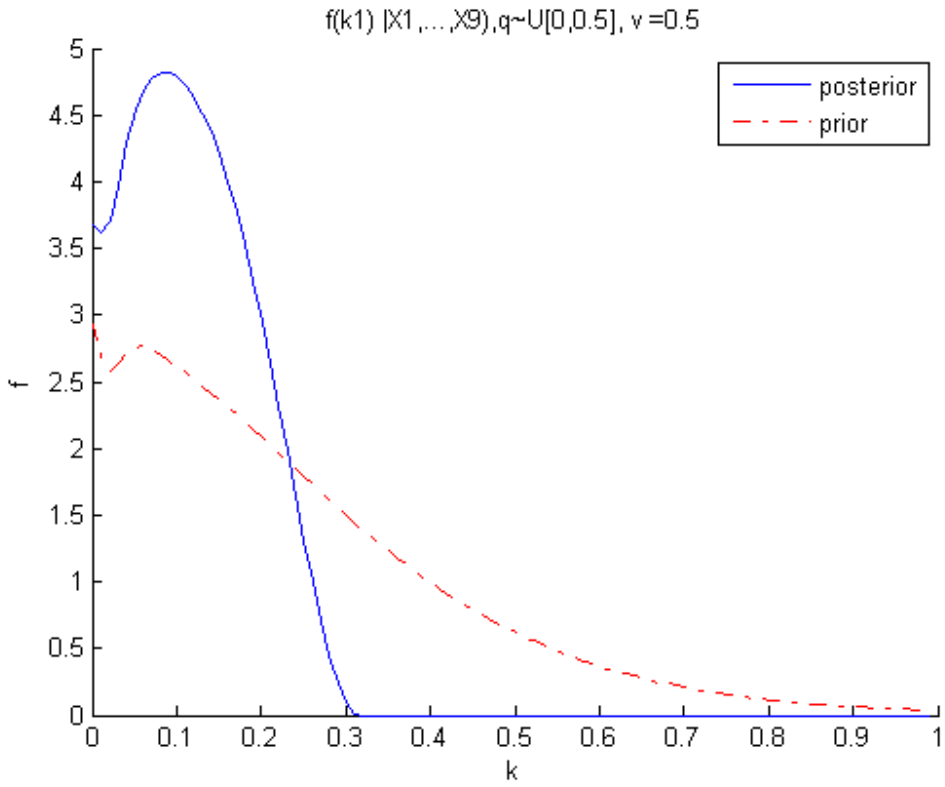

f(k1) |X1,...,X9),q~U[0,0.5], v =0.2
posterior
prior
f
k
5
4.5
4
3.5
3
2.5
2
1.5
1
0.5
0
0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
0.9
1

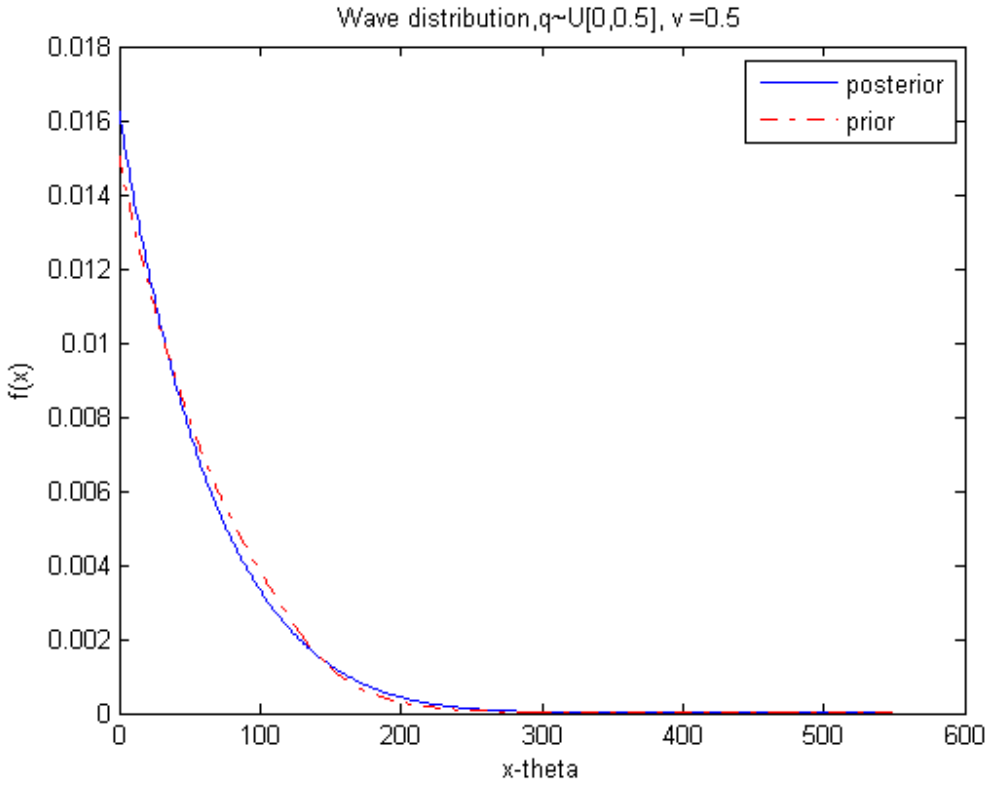
Poorwa Singh
1242326

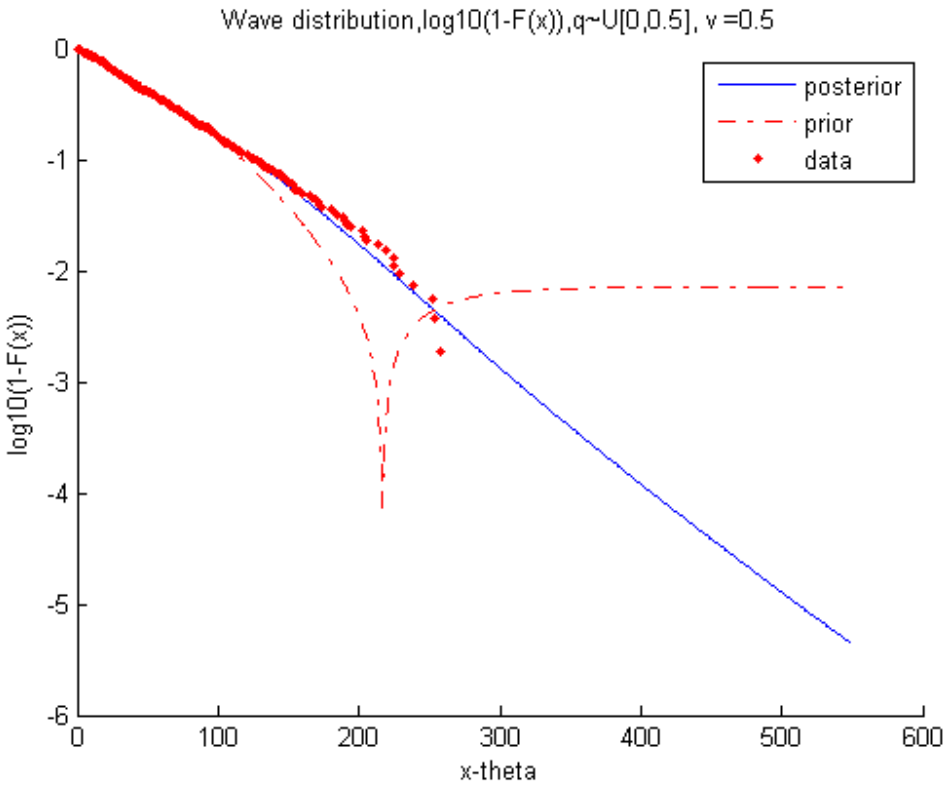


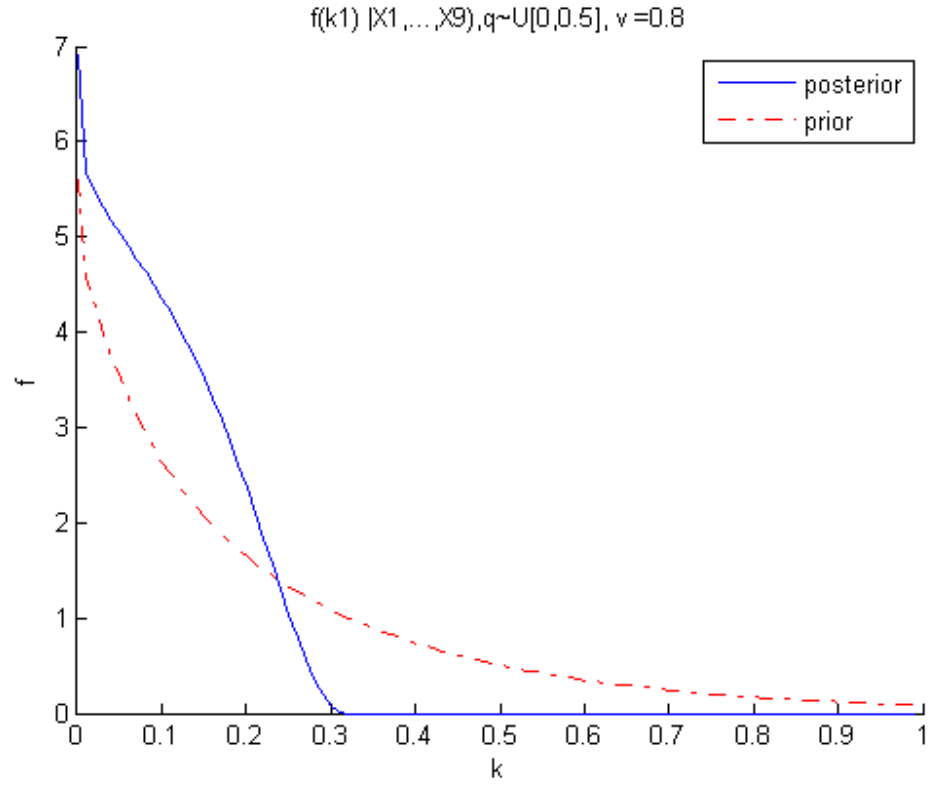


Poorwa Singh
1242326

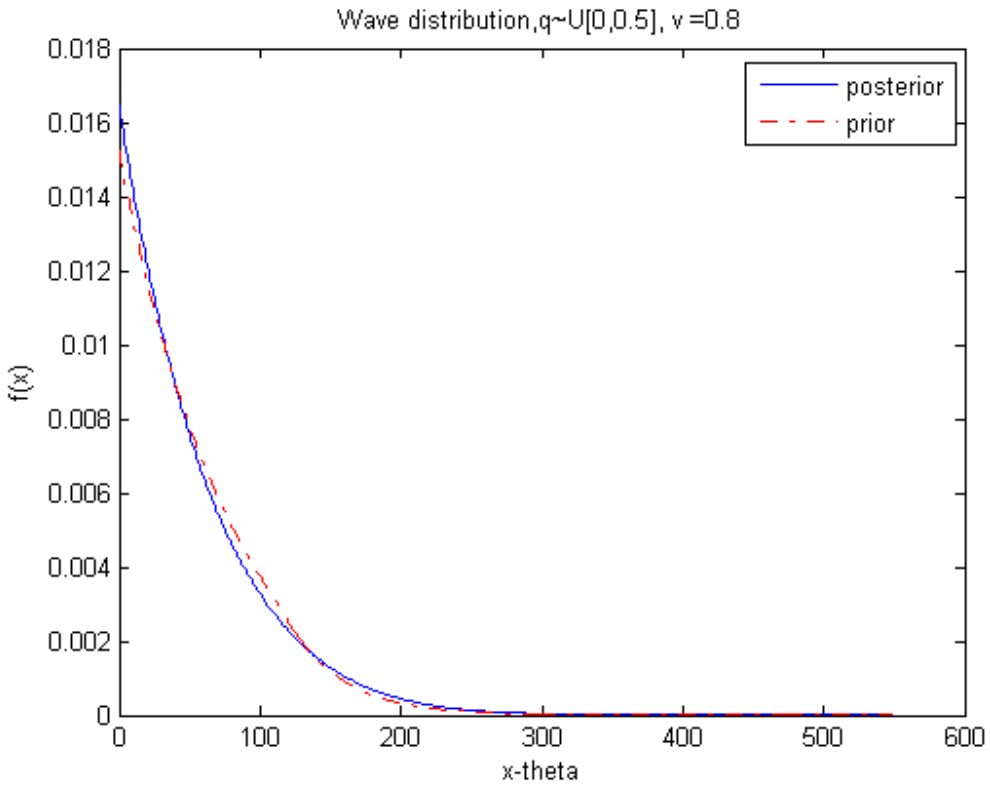


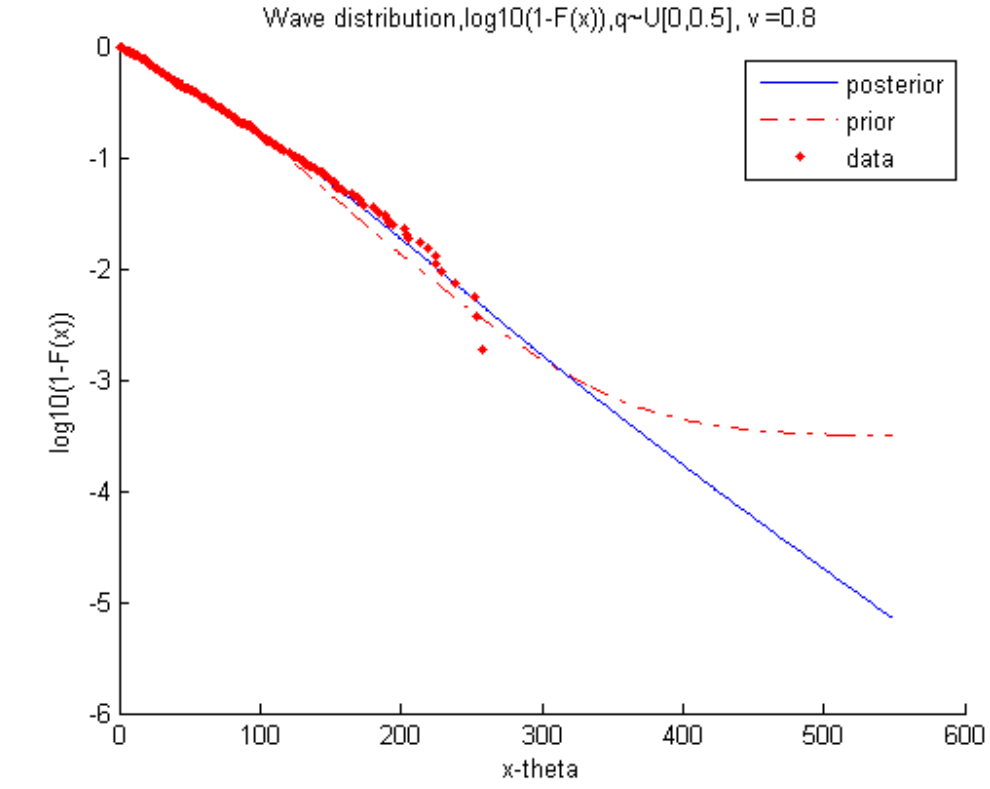



Wave distribution,log10(1-F(x)),q~U[0,0.5], v =0.5
posterior
prior
data
log10(1-F(x))
x-theta
0
-1
-2
-3
-4
-5
-6
0
100
200
300
400
500
600


f(k1) |X1,...,X9),q~U[0,0.5], v =0.8
posterior
prior
f
k
7
6
5
4
3
2
1
0
0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
0.9
1

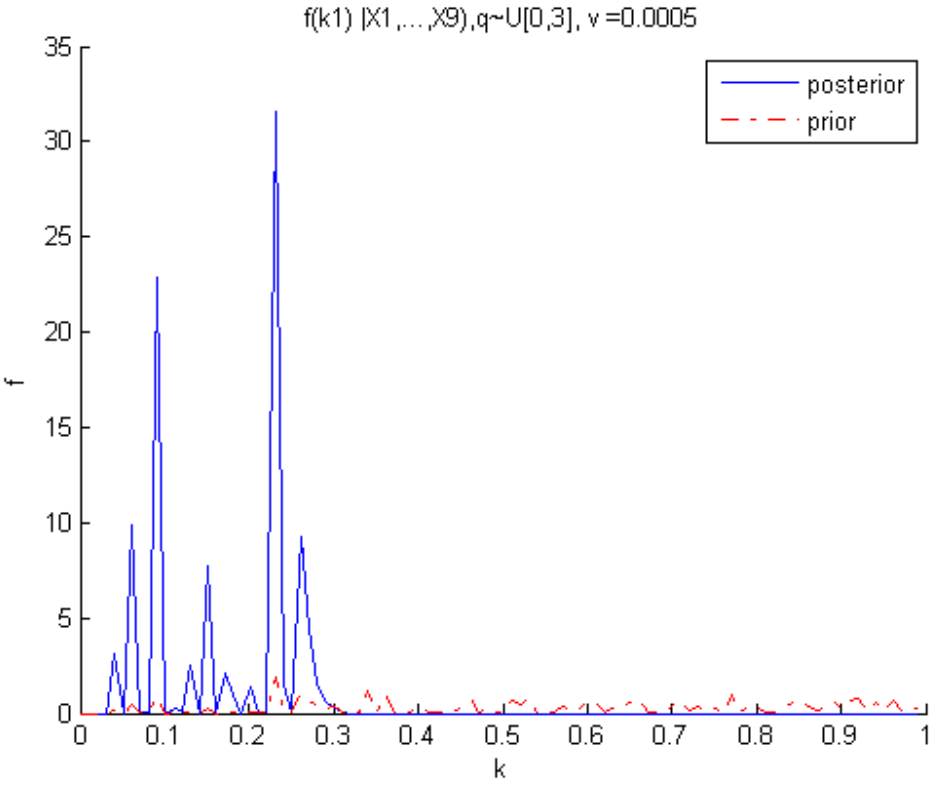
Poorwa Singh
1242326

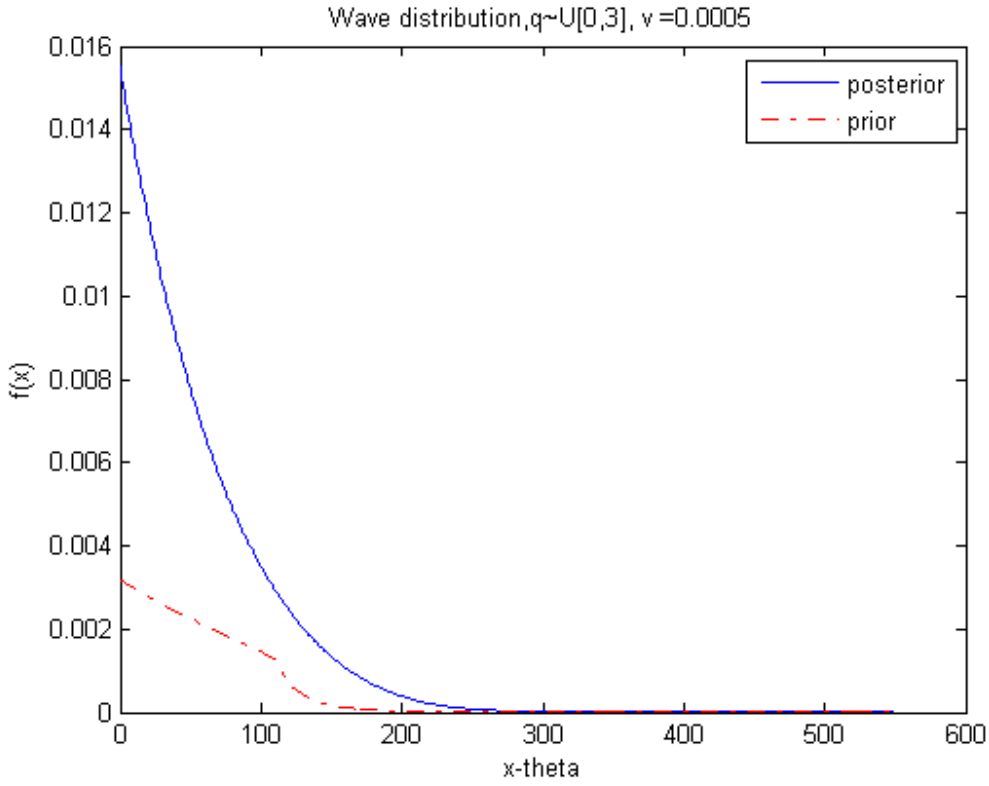




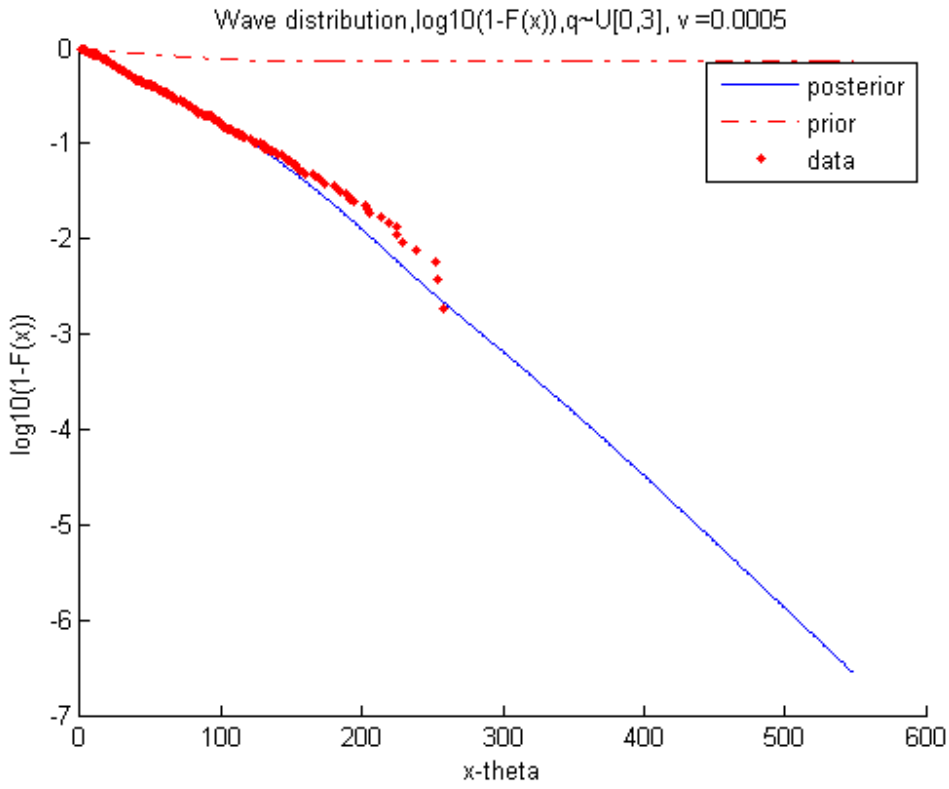
Poorwa Singh
1242326

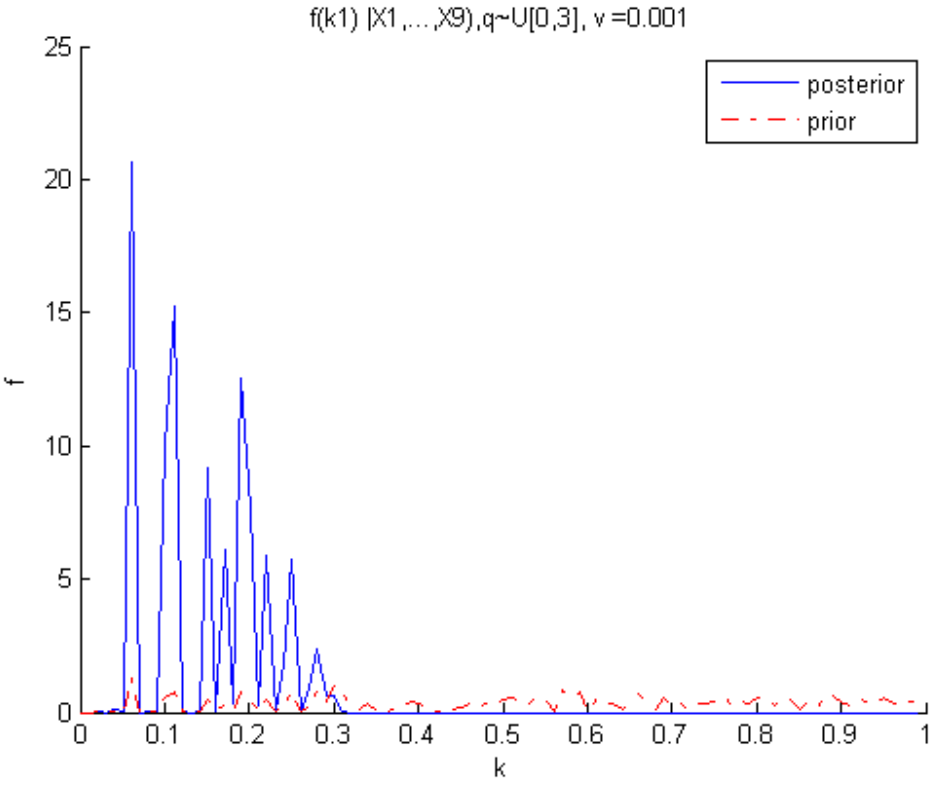
q ~ U[0,3]



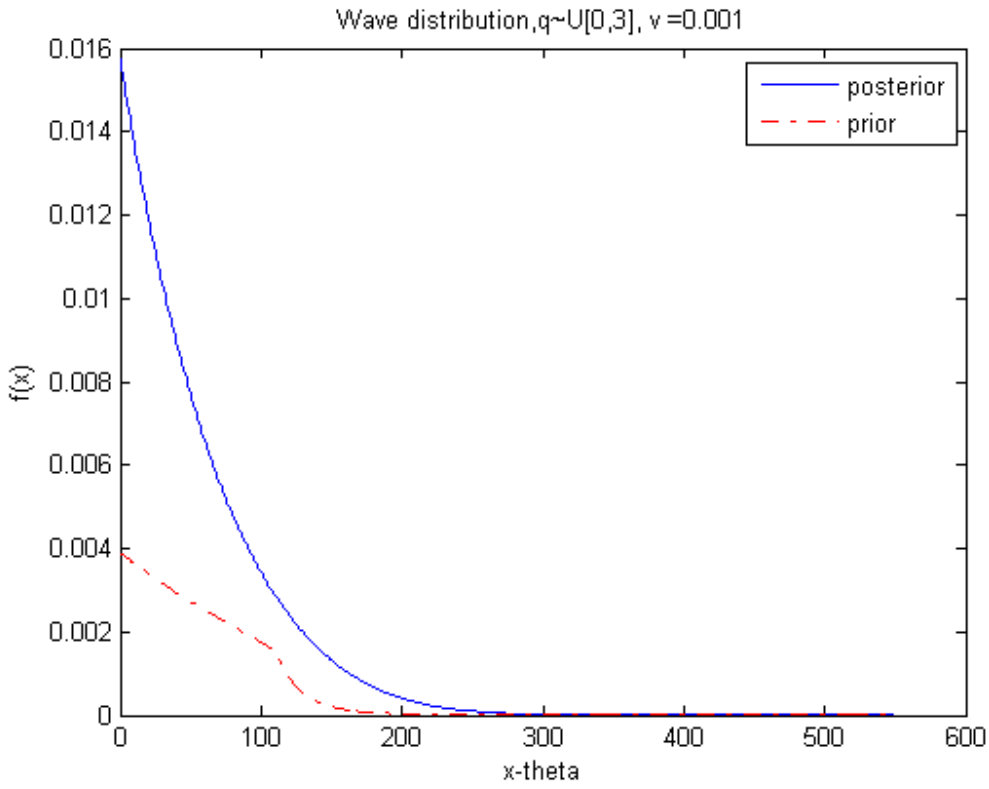


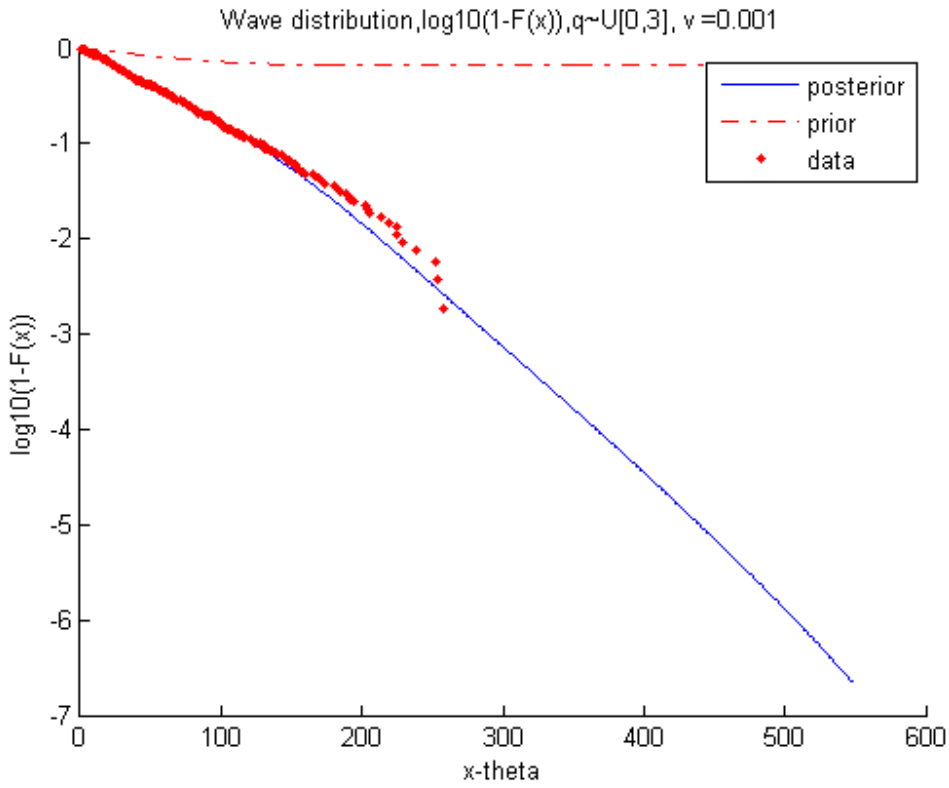
Poorwa Singh
1242326



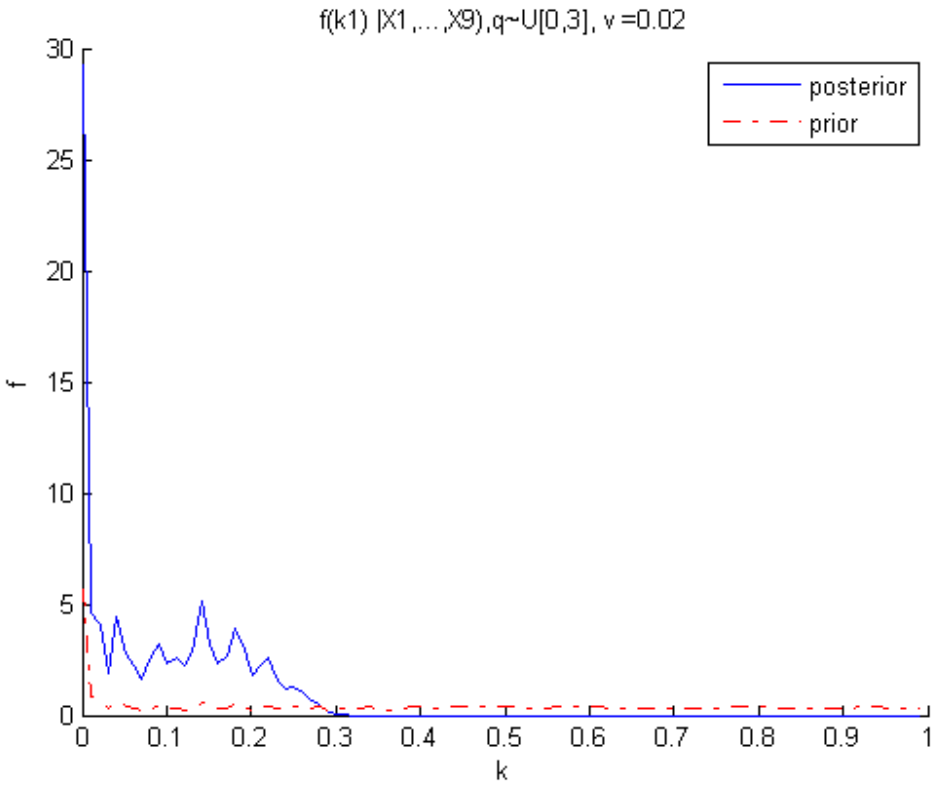


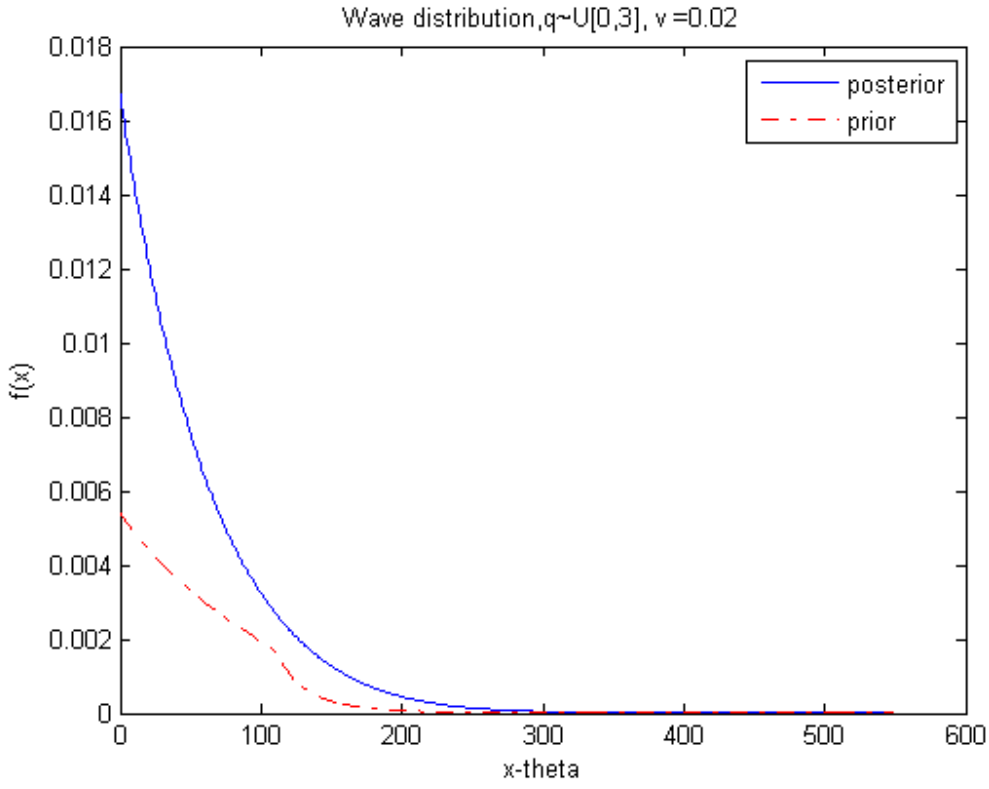
Poorwa Singh
1242326



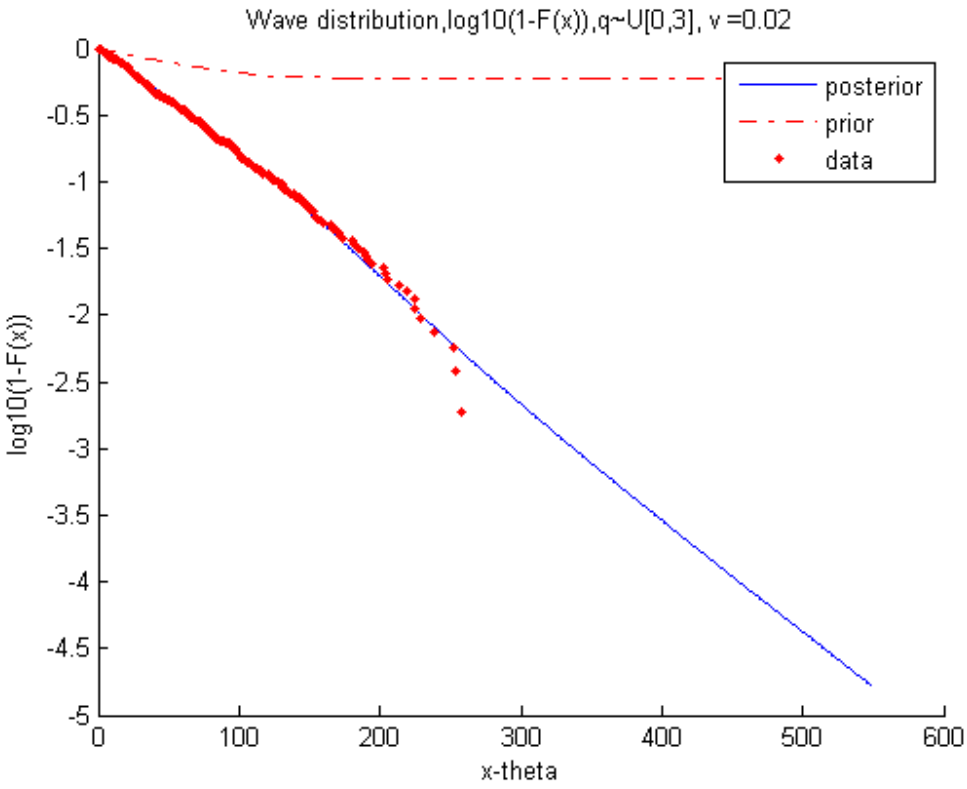


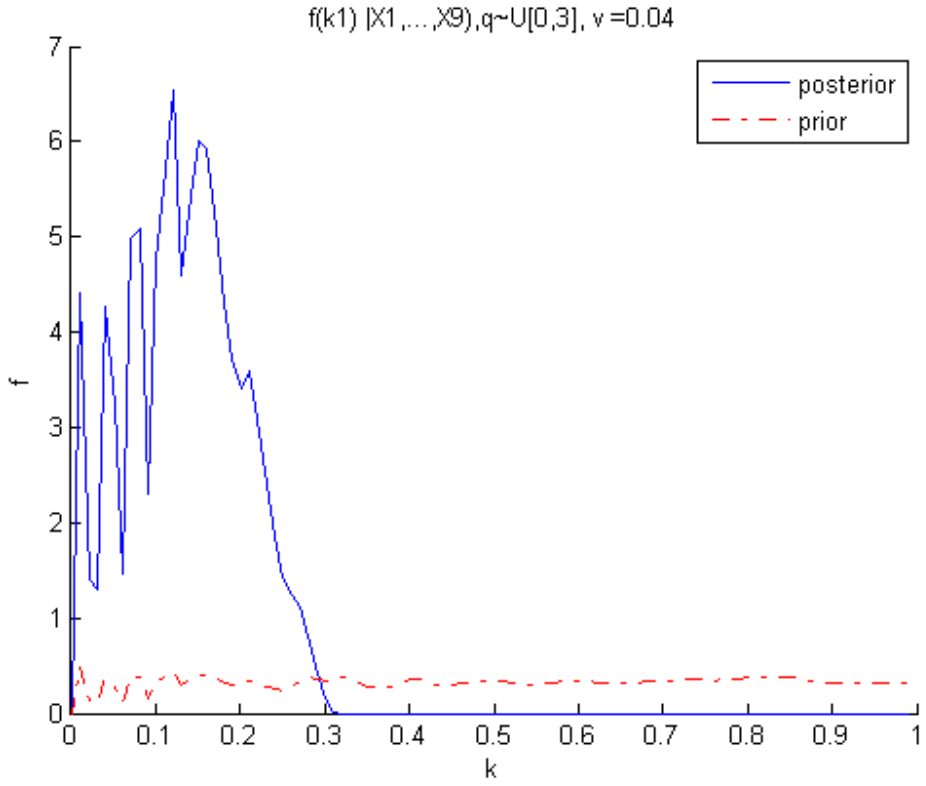
Poorwa Singh
1242326



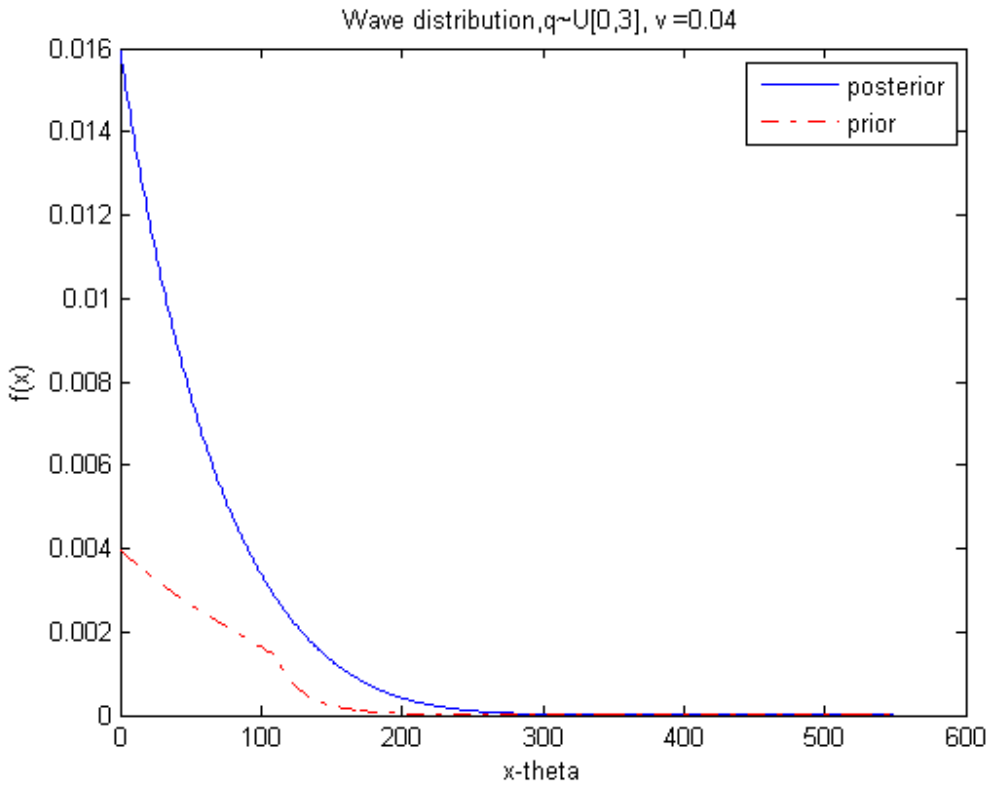


Poorwa Singh
1242326

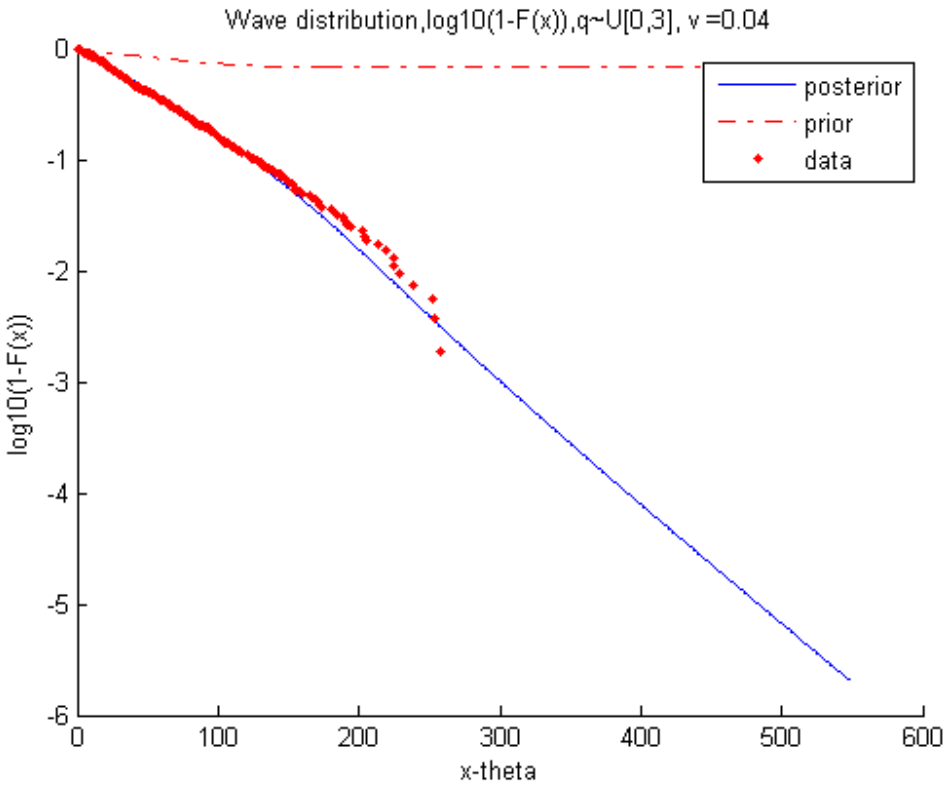


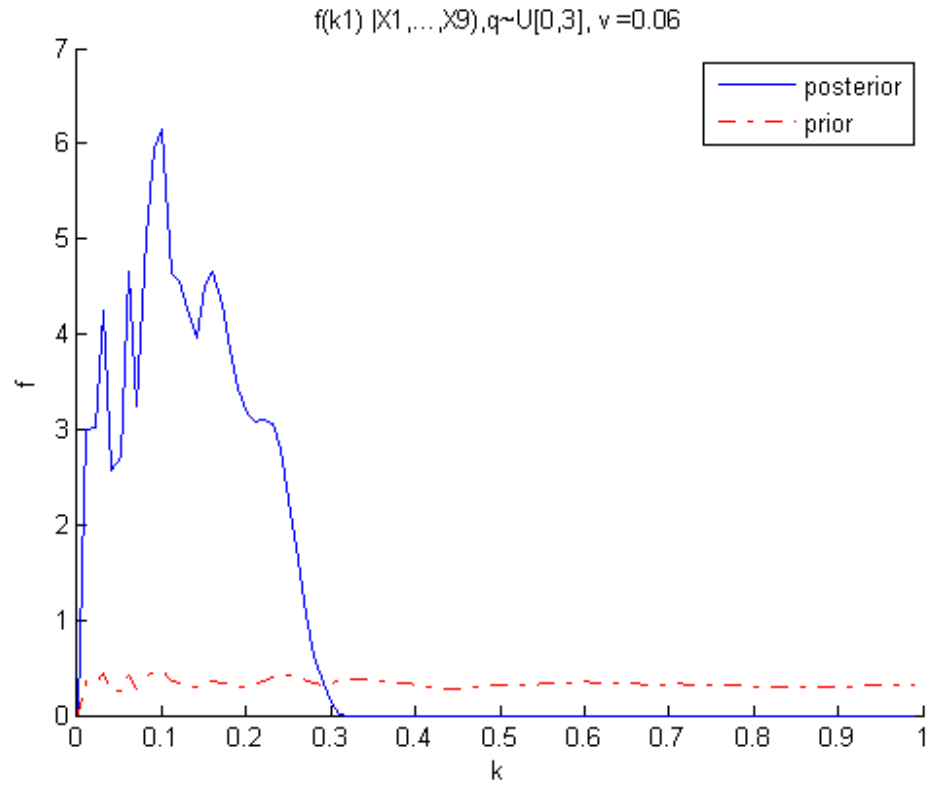


Poorwa Singh
1242326

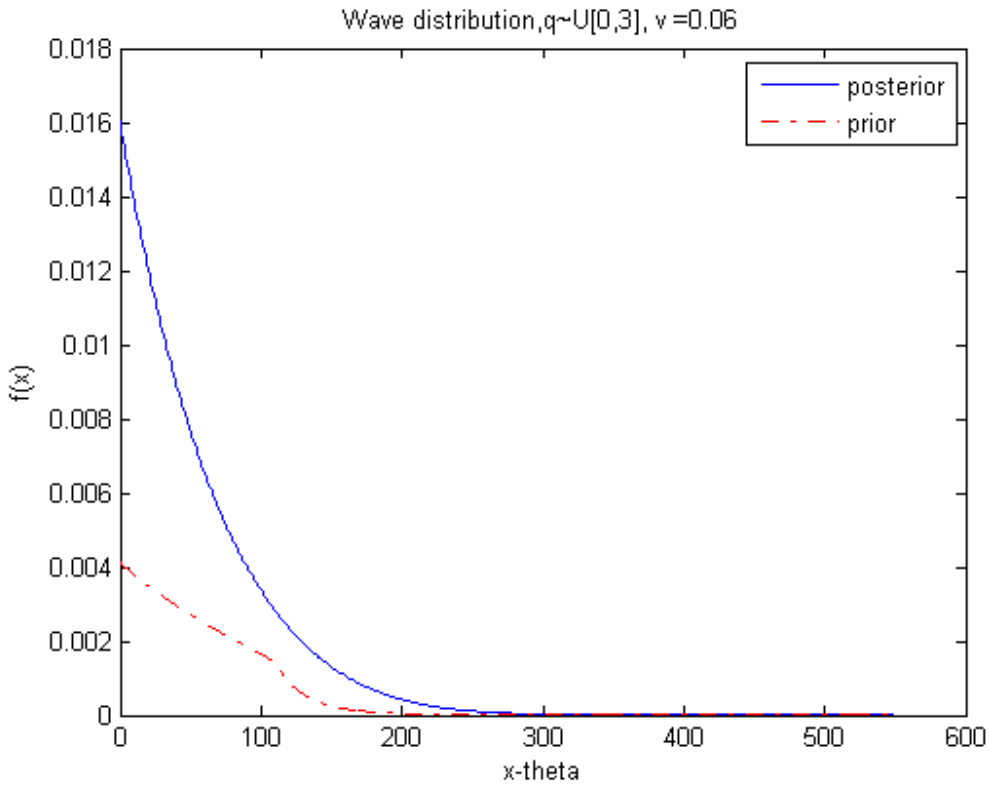


Poorwa Singh
1242326

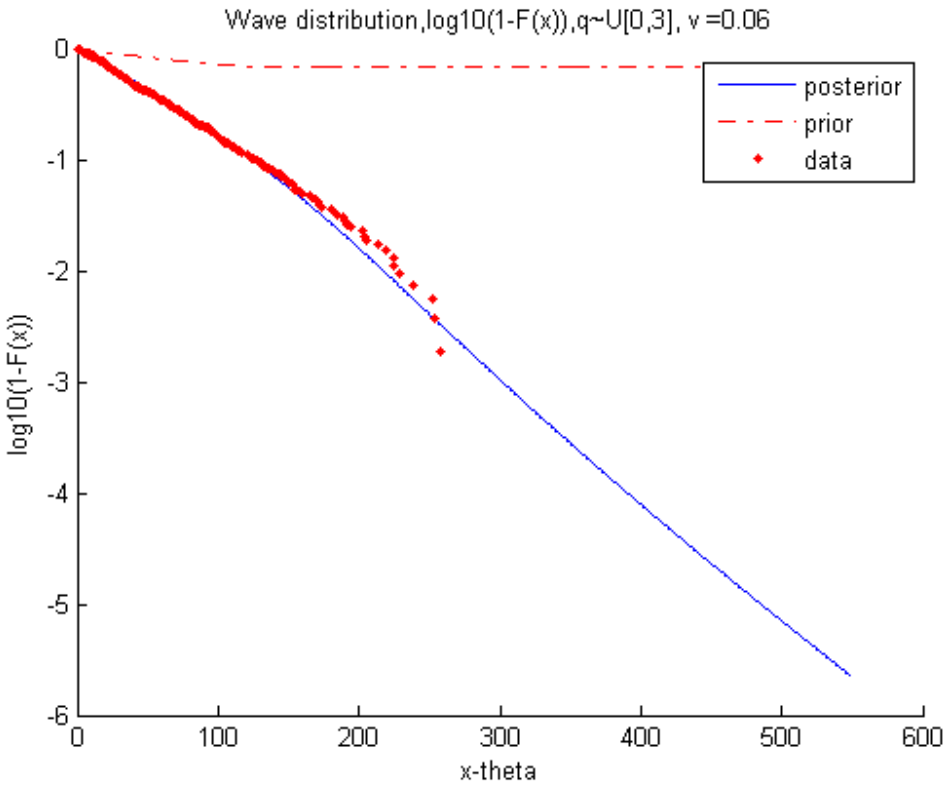


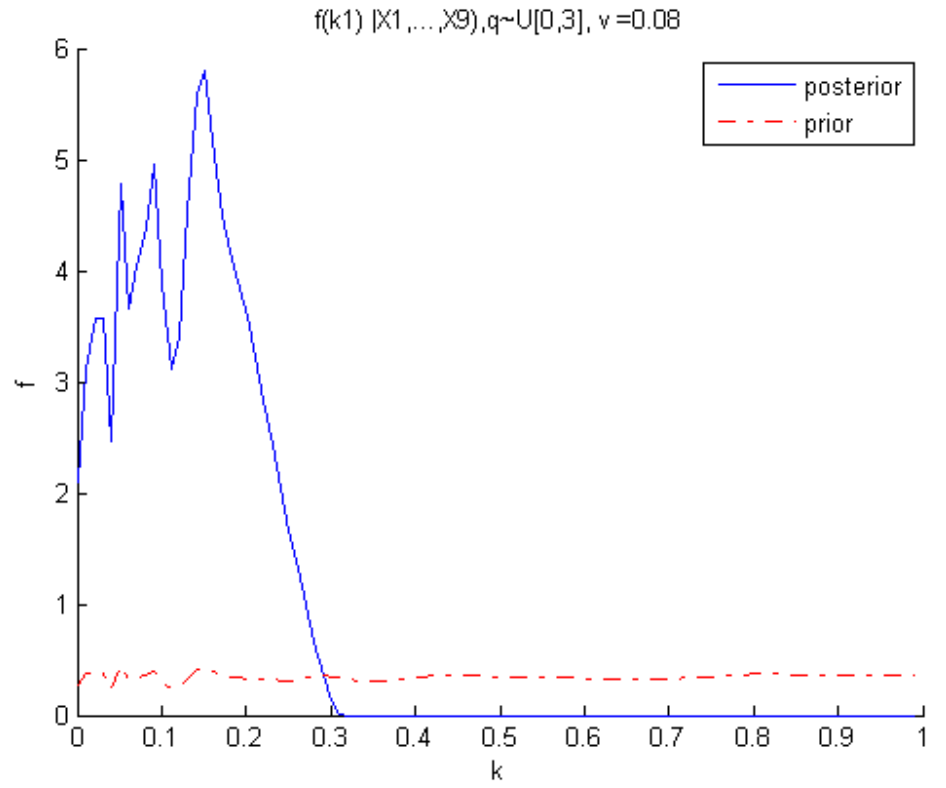


Poorwa Singh
1242326

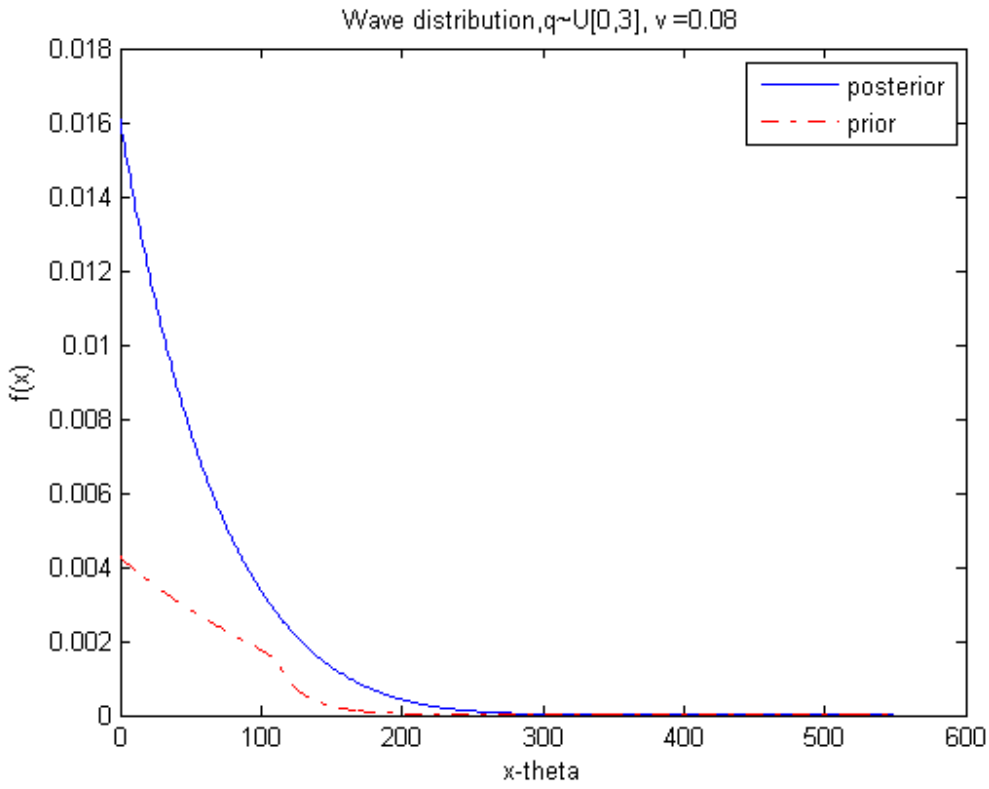


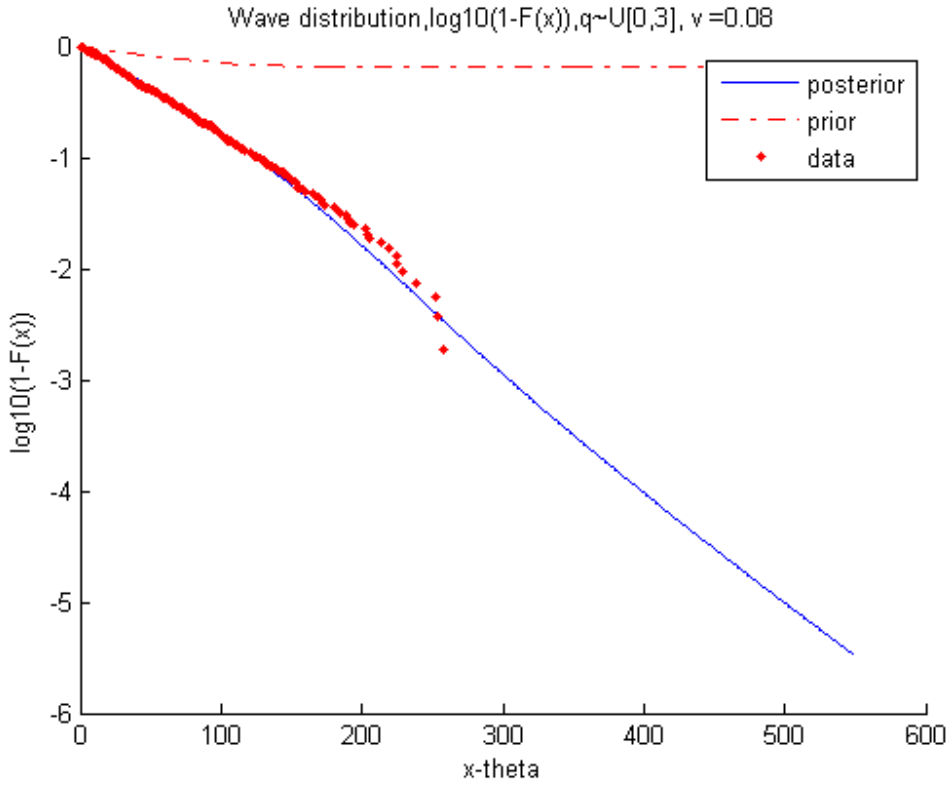
Poorwa Singh
1242326



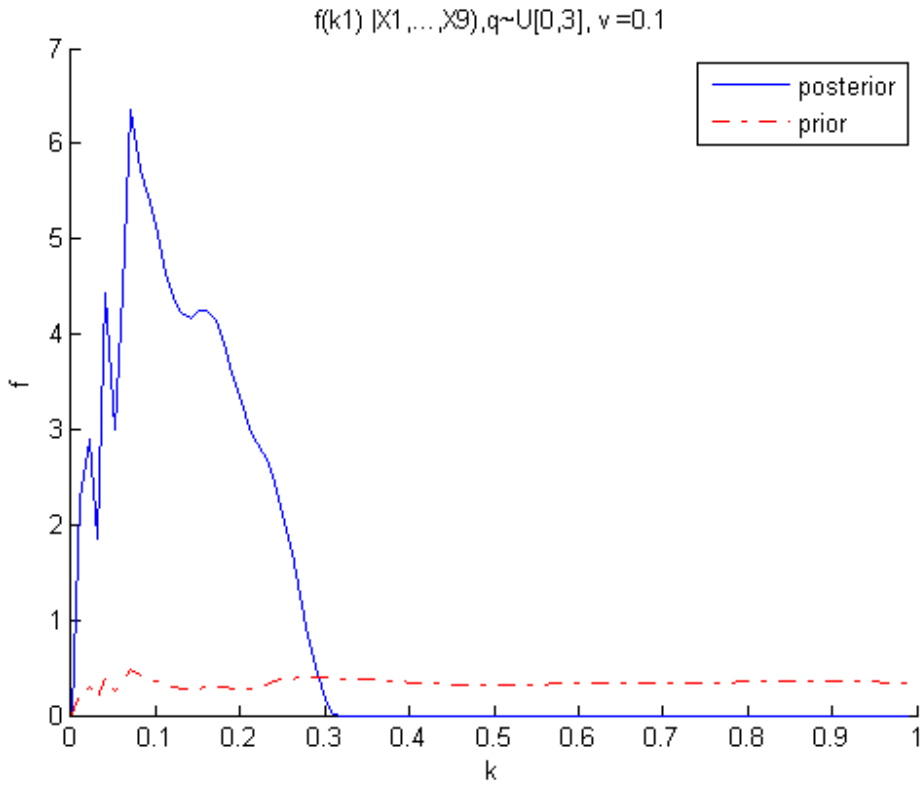


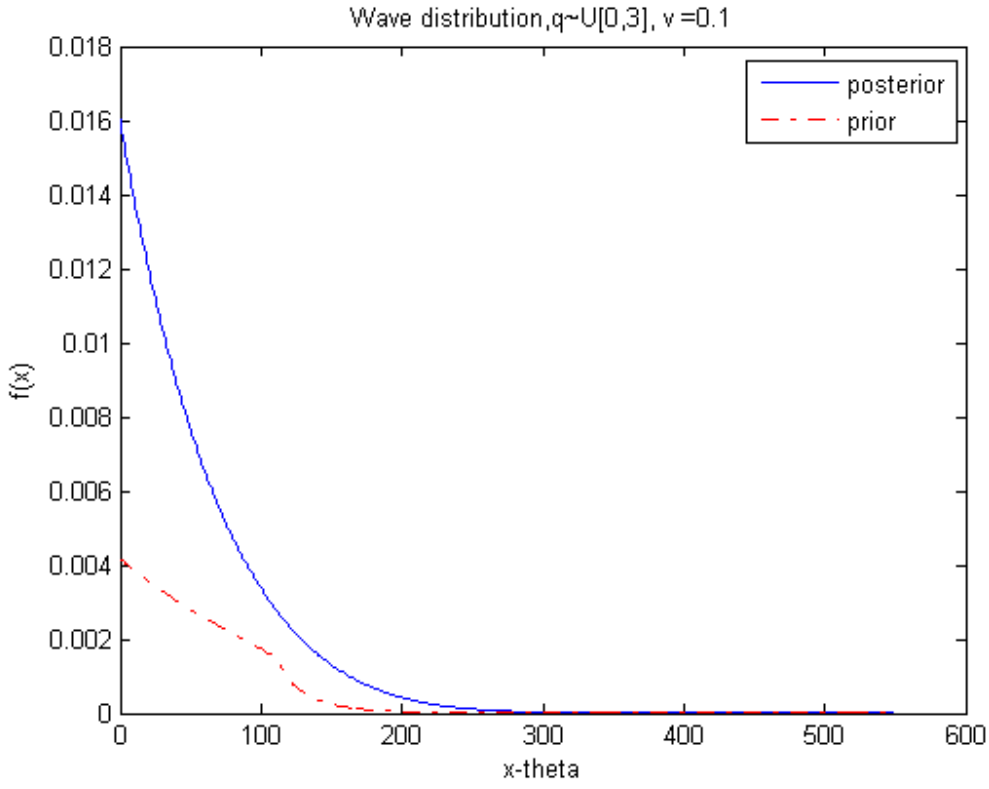
Poorwa Singh
1242326



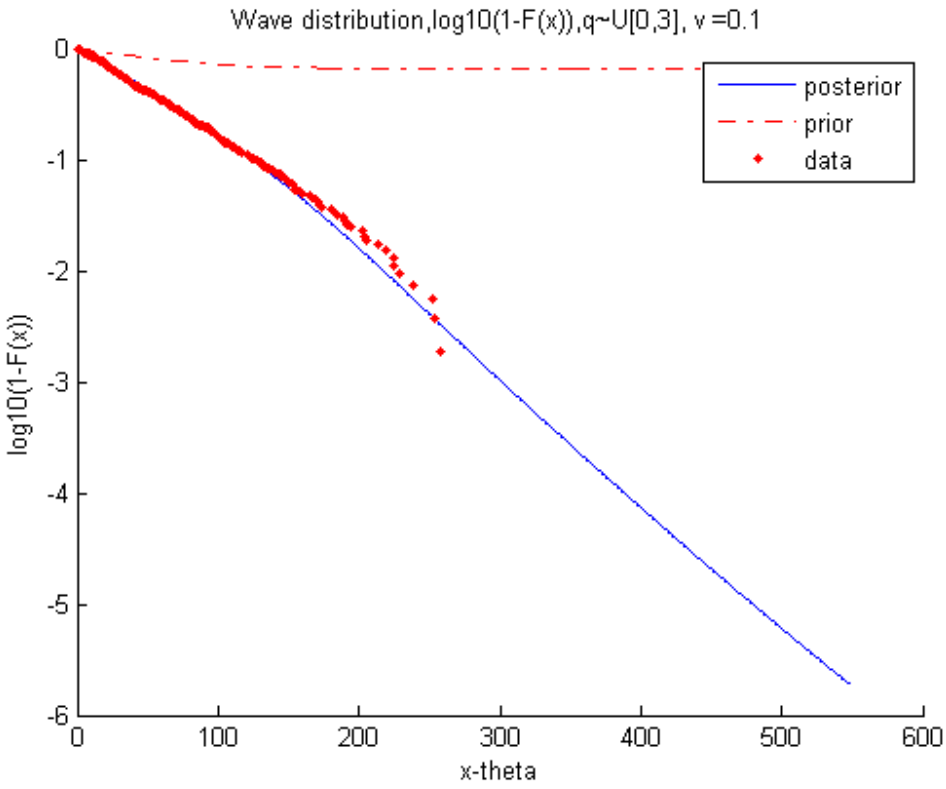


Poorwa Singh
1242326

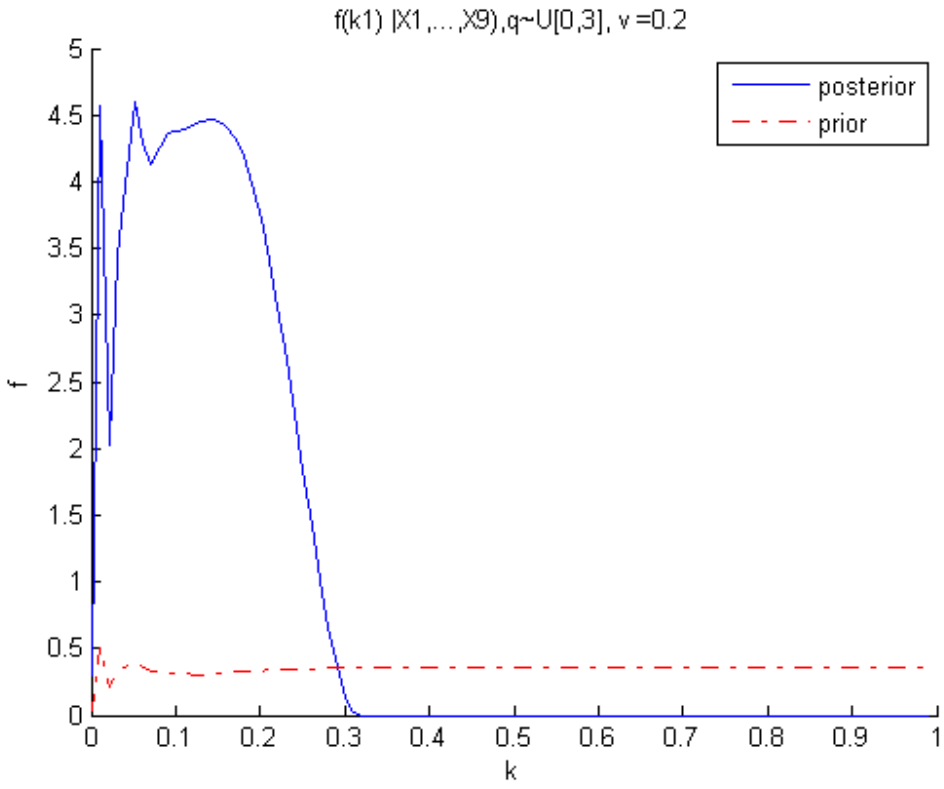


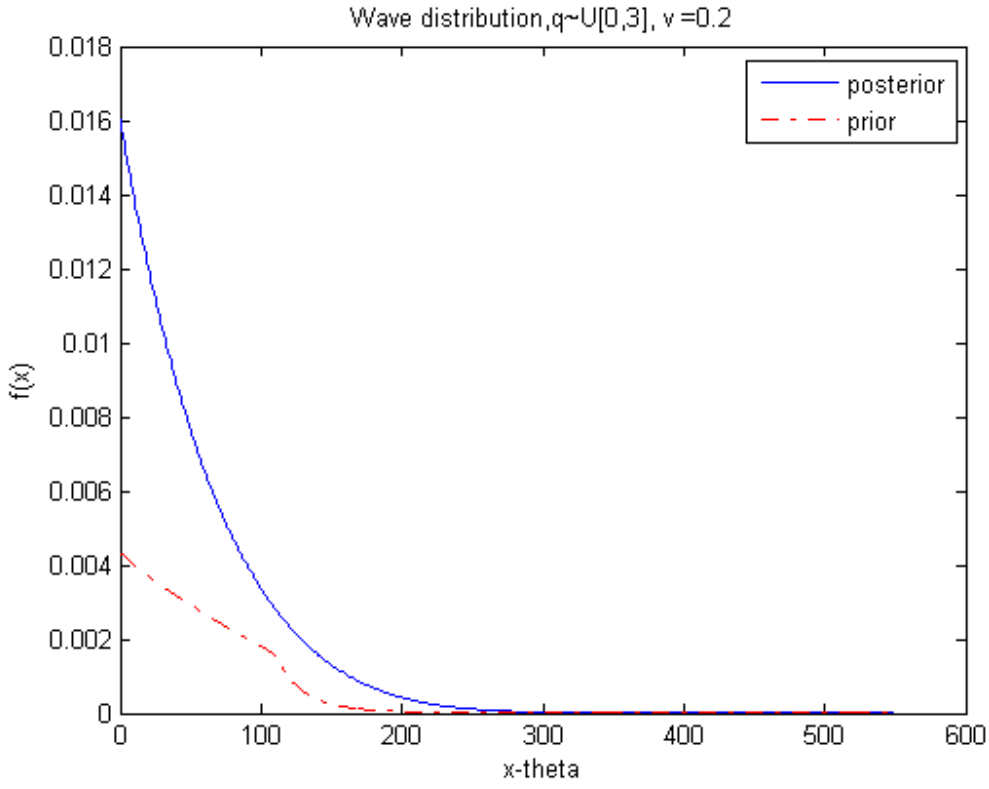


Poorwa Singh
1242326

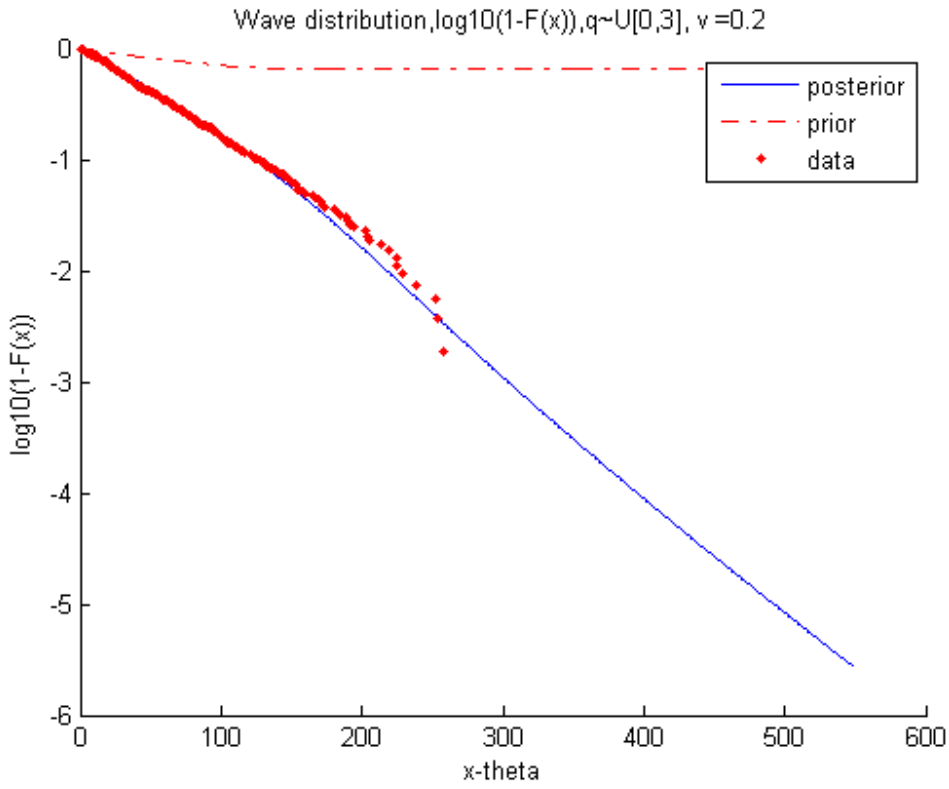


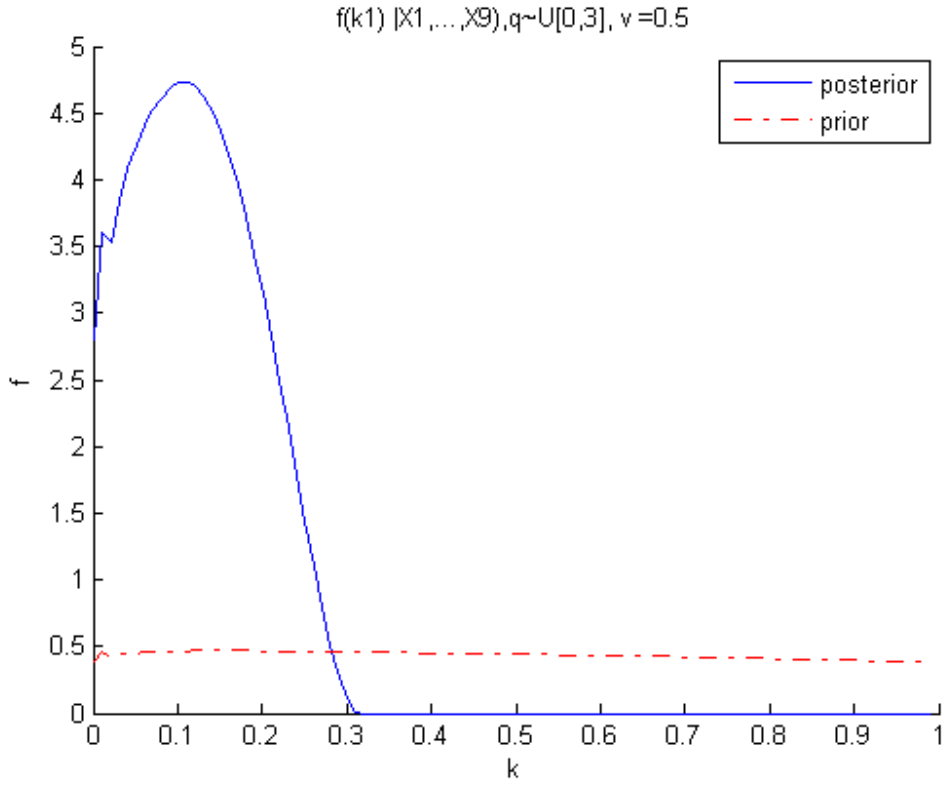
Poorwa Singh
1242326



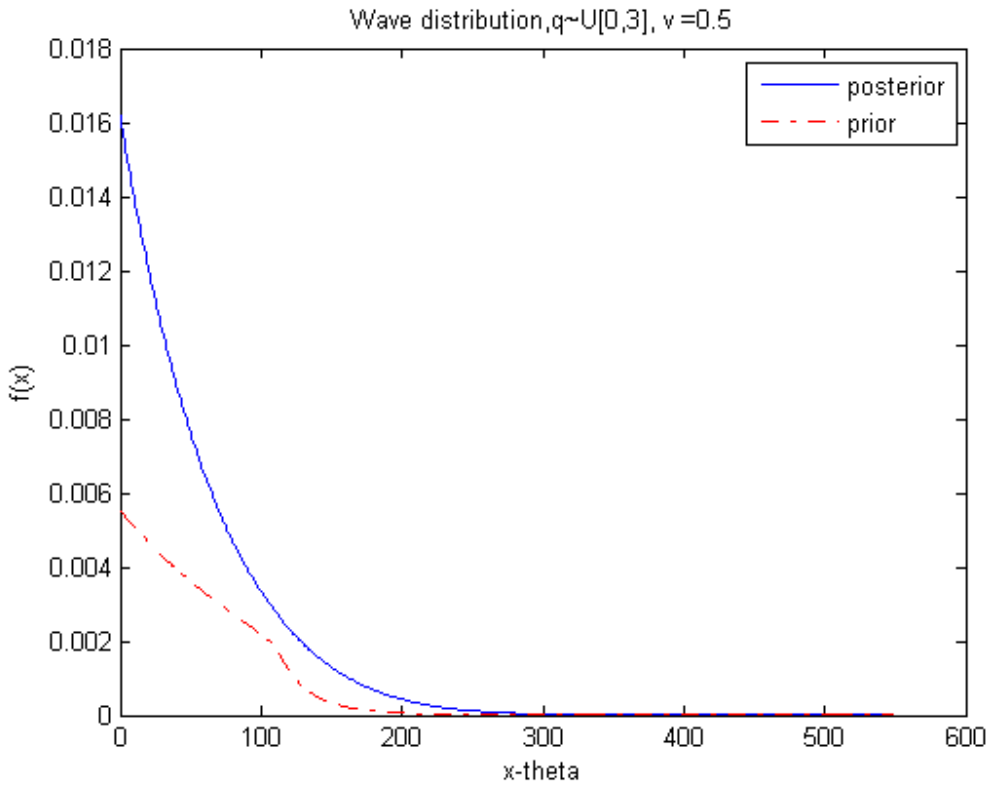


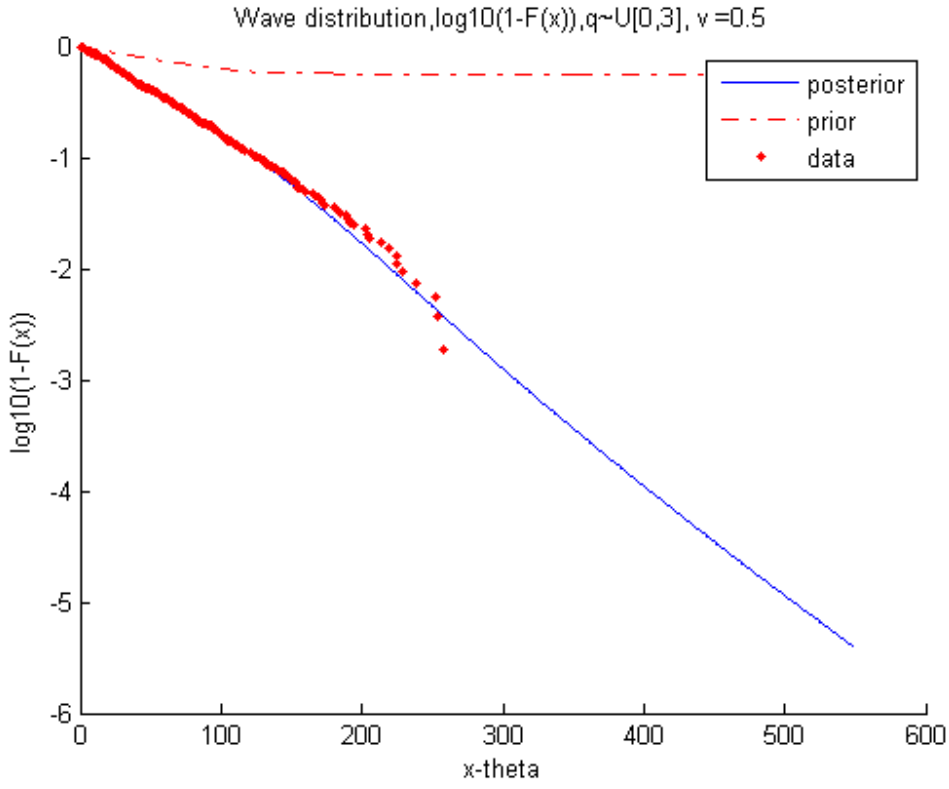
Poorwa Singh
1242326



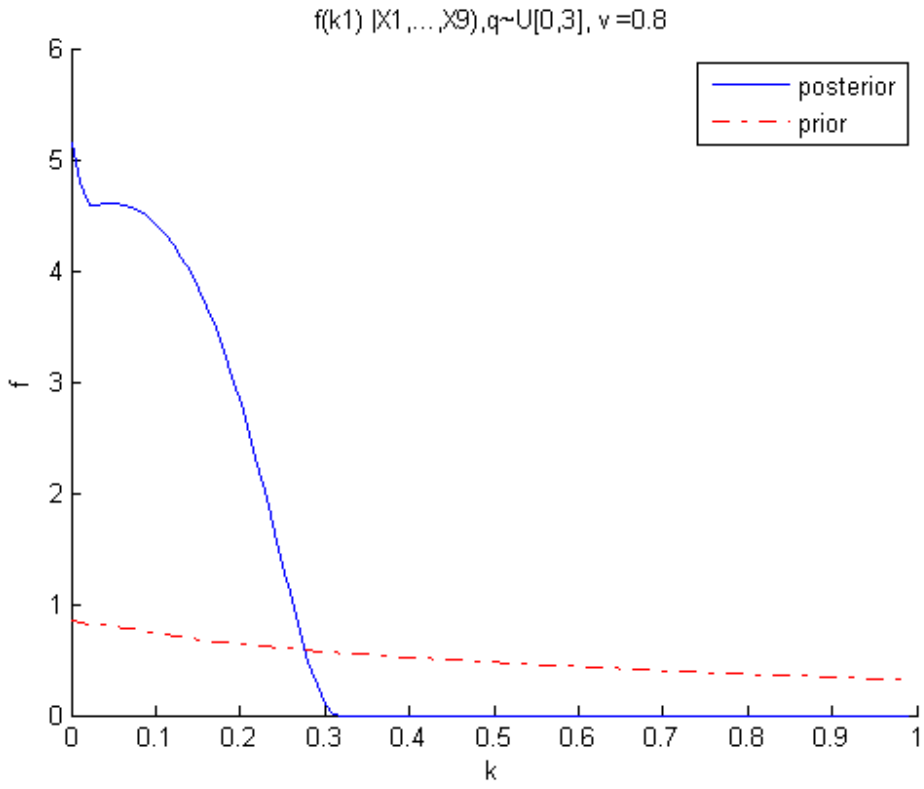


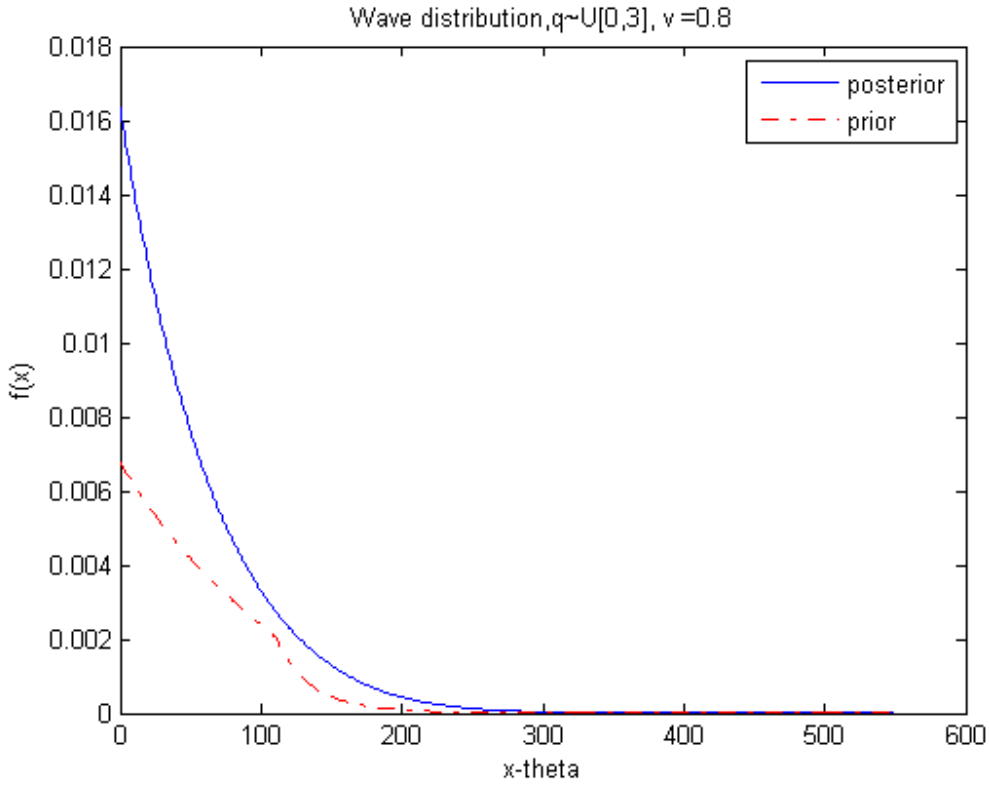
Poorwa Singh
1242326





Poorwa Singh
1242326





Poorwa Singh
1242326