Working report

Jennifer Israelsson

May 22, 2018

1 Preliminaries

1.1 Distributions

Gamma distribution: The gamma distribution is a two-parameter continuous distribution family, characterized by a shape parameter, α, and a scale parameter, θ, both positive. It has got PDF, mean and variance,

$$f(x;\alpha,\theta) = \frac{x^{k-1}e^{-\frac{x}{\theta}}}{\theta^k\Gamma(k)} \tag{1}$$

$$E(X) = \alpha\theta,\ Var[X] = \alpha\theta^2$$

1.2 Statistical test

Wilcoxon signed rank test: Is a non-parametric test that either tests the null hypothesis that a sample has a certain mean or the null hypothesis that two paired samples comes from the same distribution. In the paired case, you take the difference between each pair, rank them from smallest to largest absolut value and then and then add up all the ranks of the positive differences. If this value is greater then the table value for our number of pairs, we reject the null hypothesis and conclude that the data comes from different distributions.[3]
Autocorrelation plot: Autocorrelation describes how correlated the time series data is to itself, i.e how dependent is the data a few days forward on the day today. By plotting correlation as a function on number of lag days, we can see roughly how long a wet spell is in each month. For autocorrelation to work, there can be no missing values, so one needs to remove the months from the data with missing values.

1.3 Extreme value theory

1.3.1 Maximum analysis

In extreme value theory, we only look at the larger values in our data instead of all the data, to get a better understanding of how the tail behaves, and gives us a more accurate way to extrapolate outside our data. Put $M_n = \max(X_1, ..., X_n)$, i.e the maximum value of our n sample points. This is often called a block maxima. Lets also denote $x_F = \sup\{x \in \mathbb{R}; F(x) < 1\}$, the right end point of our distribution F. We are interested in finding the distribution of M_n as n increases. We call a non-degenerated rv X *max-stable* if

$$c_n X + d_n = M_n \tag{2}$$

for all n ≥ 2 and appropriate $c_n > 0$, $d_n \in \mathbb{R}$. This can also be written as

$$\mathbb{P}(M_n \leq x) = F^n(x) \tag{3}$$

with F being the distribution function of the rv X. This leads us to the important *Fisher-Tippet theorem*, which states that; if there exists norming constants c_n and d_n as above and some non-degenerated distribution function H such that

$$c_n^{-1}(M_n - d_n) \xrightarrow{d} H \tag{4}$$

then H belongs to one of the three extreme value distributions; Fréchet, Gumbel and Weibull. These can be summarised in one distribution, called *Generalised extreme value* distribution, GEV.

$$H_\xi(x) = \begin{cases} \exp(-(1+\xi x)^{-1/\xi}), & \text{if } \xi \neq 0 \\ \exp(-(\exp(-x))), & \text{if } \xi = 0 \end{cases} \tag{5}$$

where $1 + \xi x > 0$. With $\xi = 0$, we get Gumbel, $\xi > 0$ Fréchet and $\xi < 0$ Weibull. Closely related to GEV distributions is *maximum domain of attraction*. We say that a rv X belongs to the *maximum domain of attraction* of H, $X \in MDA(H)$, if there exists norming constants $c_n > 0$, $d_n \in \mathbb{R}$ such that

$$c_n^{-1}(M_n - d_n) \xrightarrow{d} H \tag{6}$$

1.3.2 Upper order analysis

Instead of looking at maxima, and thereby throwing away a lot of data, we can look at all values above a threshold, u. We can look at both the distribution and the mean of the data above this threshold

$$F_u(x) = \mathbb{P}(X - u \leq x | X > u) = \frac{F(x+u) - F(u)}{1 - F(u)} \tag{7}$$

$$e(u) = \mathbb{E}(X - u | X > u) \tag{8}$$

$e(u)$ is called *the mean excess function of X*. Just like the block maxima converges to a GEV distribution, the data over a high threshold also converges to a distribution, called the *General pareto distribution*, GPD

$$G_{\xi,\nu,\sigma}(x) = \begin{cases} 1 - (1 + \xi\frac{x-\nu}{\sigma})^{-1/\xi}, & \text{if } 1 + \xi\frac{x-\nu}{\sigma} > 0, \xi \neq 0 \\ 1 - e^{-\frac{x-\nu}{\sigma}}, & \text{if } \xi = 0 \end{cases} \tag{9}$$

where

$$\begin{cases} x \geq 0, & \text{if } \xi \geq 0 \\ 0 \leq x \leq -\sigma/\xi, & \text{if } \xi < 0 \end{cases} \tag{10}$$

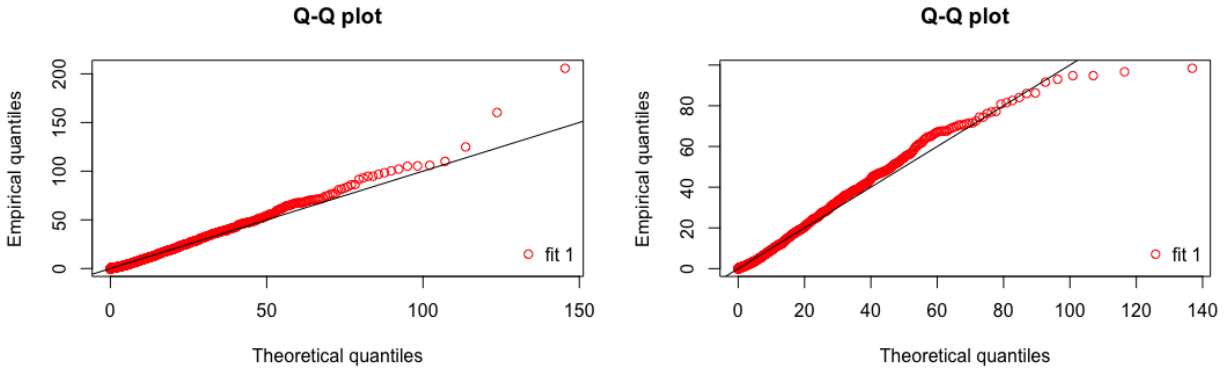
$\nu \in \mathbb{R}$ is called a location parameter and $\sigma > 0$ is a scale parameter. The two distributions are related, such that the distributions for which the block maxima converges to a GEV with parameter ξ, their excess distribution converges to the GPD with same shape parameter ξ.

2 Initial assesment

To get an idea about the general behaviour of the data, and if there are big variations depending on the geography, I slit all the data into the different stations. To get a solid baseline to compare the rest of the data to, I picked out 3 complete data sets; one in the north (ADA), one around the equator (KRA) and one in the south (BOL). As a first attempt to see differences, I picked out only the rainy days from the data and compared the number of days. I then tried to fit this data into a gamma distribution by using Maximum likelihood estimator (MLE). To get a bit deeper understanding of their differences, I split the positiv data into months and looked at the behvaiour in the months; does it have a few extreme values, many or few days of rain, certain years with many days of rain (rainy) or several days with a high amount of rain (wet). As a final thing, I picked out the maximum values in each month and the years for all the extreme values.

2.1 ADA

This station has a lot fewer days of rain compared to the other two station, with only 19% of days rainy, but it is more evenly distributed then for the other stations. The number of days of rain for (Dec, Jan, Feb) is (37, 21, 58), which is higher then for the other. It fits to a **Gamma distriution** with **scale**=21.02 and **shape**=0.553 and has a loglikelihood of -6730. This distribution fits nicely up until 100 quantile (fig 1a), so instead I only fitted the data < 100 and got a **Gamma distribution** with **scale** = 17.590 and **shape** = 0.585 (fig 1b)



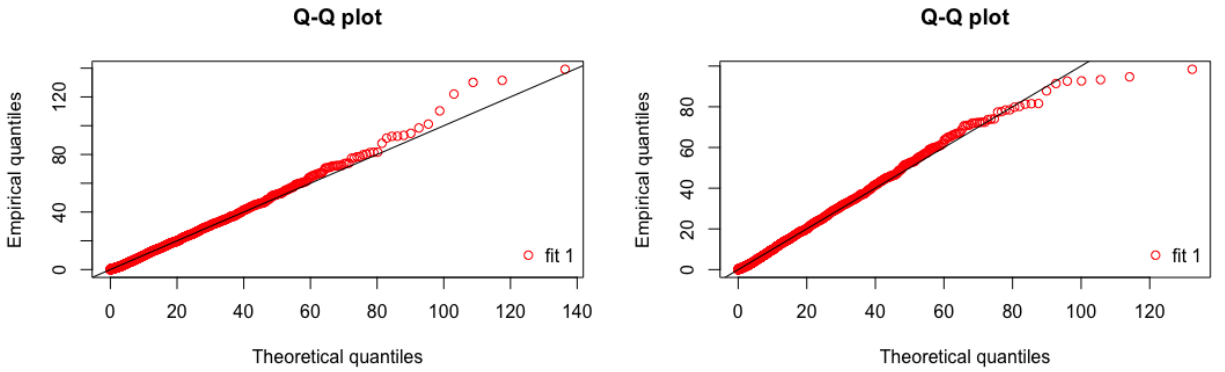
(a) Full data set

(b) Positive values < 100

Figure 1: Q-Q plots for Gamma distribution, ADA

2.2 BOL

This station has approxiamtely 24% of rainy days with very few in (Dec, Jan, Feb), (7,29,21). The **Gamma distribution** with **scale**=17.835, **shape**=0.689 fits well up to 80-100 quantile if fitted with all the data (fig 2a), and if I only fit it with values up to 100, I get a **Gamma distribution** with **scale** = 17.249, **shape** = 0.698 which fits nicely nearly everywhere (fig 2b).



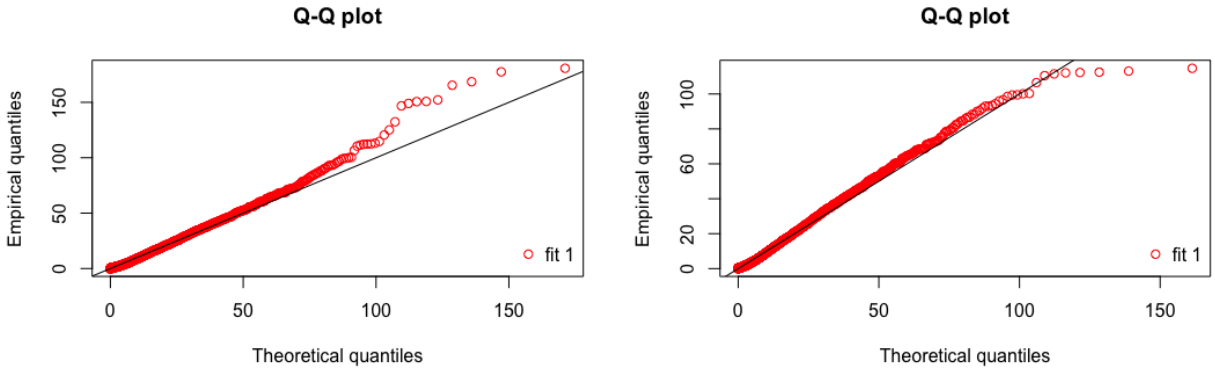
(a) Full data set

(b) Positive values < 100

Figure 2: Q-Q plots for Gamma distribution, BOL

2.3 KRA

This is the rainiest station out of the three, with $\sim$ 27% rainy days, but less well distributed then the ADA. In (Dec, Jan, Feb), it rained (20,35, 30), so values between the other two. The **Gamma distribution** with **scale** = 23.038 and **shape** = 0.595 fits up to around 100 quantile (fig 3a) and if we restrict the data to < 120, it fits nearly perfectly to a **Gamma distribution** with **scale** = 21.774 and **shape** = 0.607 (fig 3b)



(a) Full data set

(b) Positive values < 120

Figure 3: Q-Q plots for Gamma distribution, KRA

2.4 Max values and Extreme years

	Jan	Feb	Mar	Apr	May	Jun	Jul	Aug	Sep	Oct	Nov	Dec
ADA	31.3	74.4	82.6	124.9	80.8	205.7	93	53.6	102.1	100.3	98.4	65
KRA	46.3	52.6	68.4	94.4	165.4	149	150.8	152.2	168.6	180.7	65.8	51.9
BOL	37.1	79.8	110.3	81.3	92.6	101.1	130.1	122	139.2	94.7	74	45.3

Table 1: Table of yearly maximum

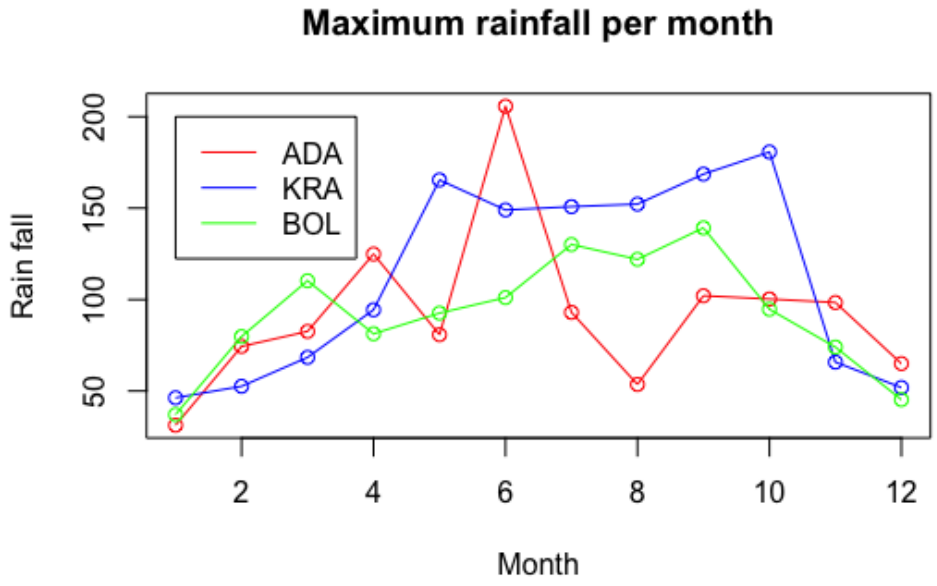
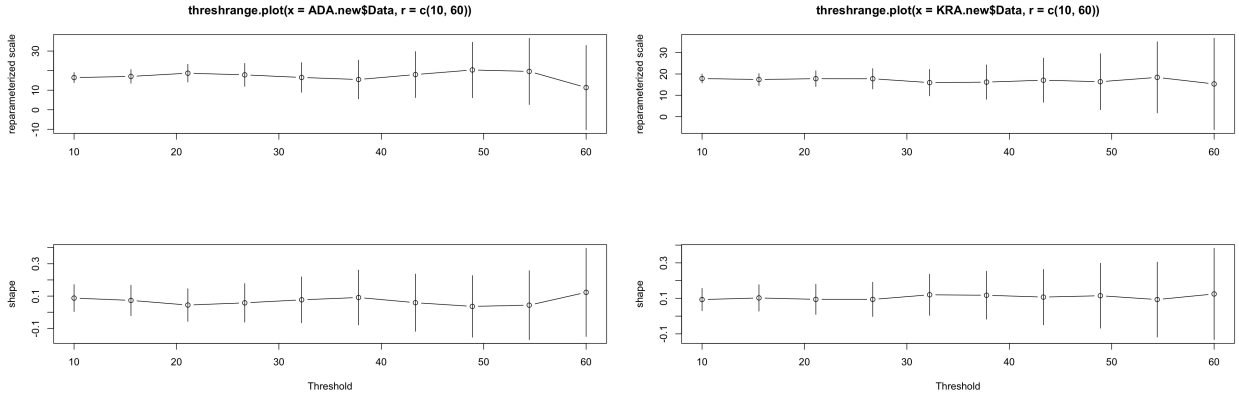


Figure 4: Max values per month over all years

3 Extreme fitting

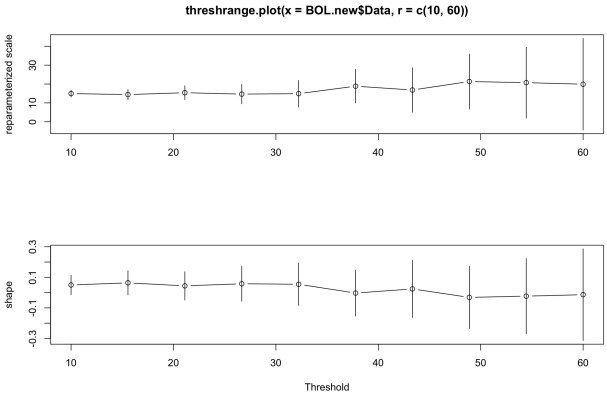
To investigate the extreme bahaviour of the different stations, I started to look at the distribution of the yearly maxima, and ploted it to see if it had an overall behaviour. To determine a good threshold value, I tried to look at **"Hill plots"** and **"Mean excess plots"**, but neither gave a very clear indication for a suitable value. So instead, I looked at threshold parameter plots, starting with a range (10, 80), but moved down to (10, 60) (fig 5 because the varinace of the higher values where too big, and therefore draged out the scale too much.

3.1 Threshrange plots



(a) ADA

(b) KRA



(c) BOL

Figure 5: Threshold range =(10, 60) plots

All of the stations have very similar threshold plots, the variance starts to increase substantially around 30 and is large around 50. For ADA, the variance is already quite large at 40, wheres for the other two it is more around 45 it start to grow rapidly. So suitable thresholds could be 35 for ADA and 40 or 45 for the other 2.

3.2 ADA

The annual maximum for ADA has a rough pattern of getting a new maximum value every 5-7 years until 2005 (fig 6), but it could of course be a new maximum in 2013, which would then continue the pattern. The values fits a **GPD** with parameters (69.649, 23.268, 0.088) relatively well, with a small bump on the density

plot around 150, which is present for all the stations (fig 7). The fitted distribution understimates the rain amount for higher return years.

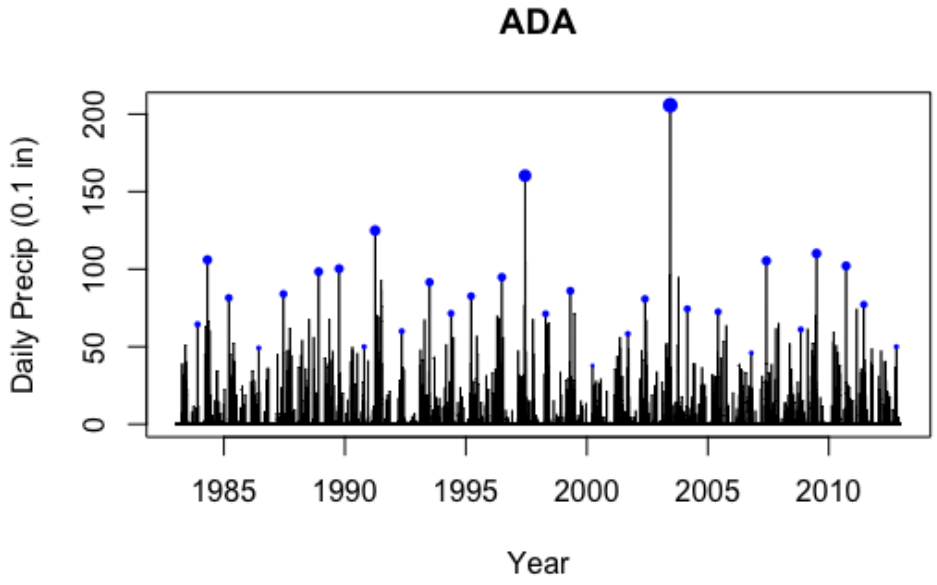


Figure 6: Annual maxima, ADA

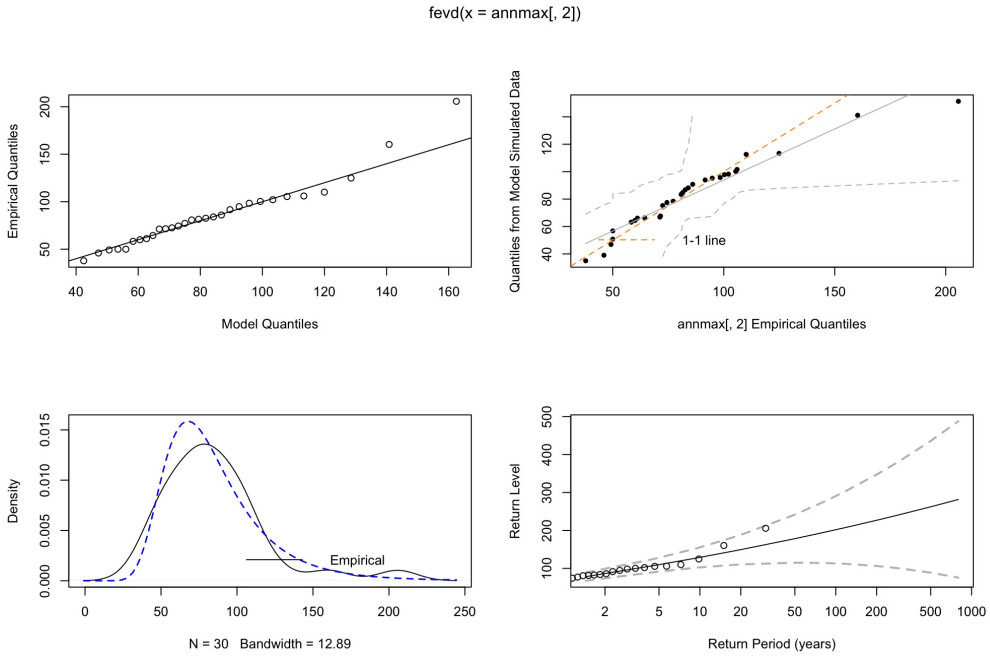


Figure 7: Plots over distribution fit for ADA anuual max

If I tried to fit the data with the threshold set to 40 (140 obs.), only the lower values are well fit, and not the tail which is the one of interest. So instead I picked the threshold to be 50 (84 obs.)to better fit the

tail (fig 8). It is a pretty big bump around 100 in the density plot and a small one around 150. The fitted distribution is with parameters (60.806, 10.714, 0.407).

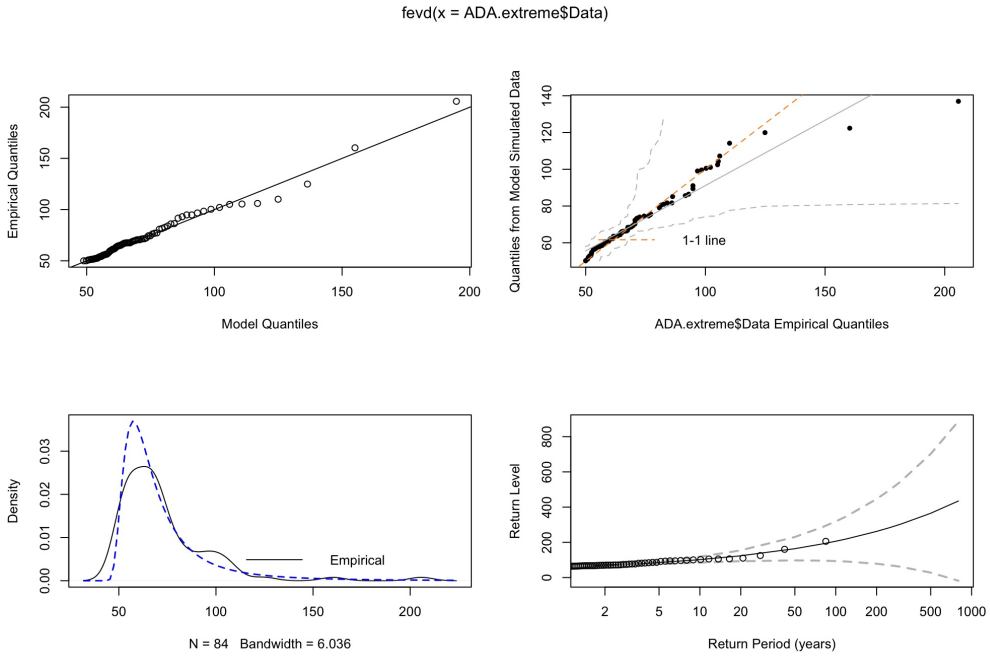


Figure 8: Plots with threshold = 50, ADA

I plot the data, highlighting only the data points that exceeds the threshold, to see if there is a general trend with more extreme values (fig 9). It seems like there where more days with more then 50 mm of rain before 2000 then after.

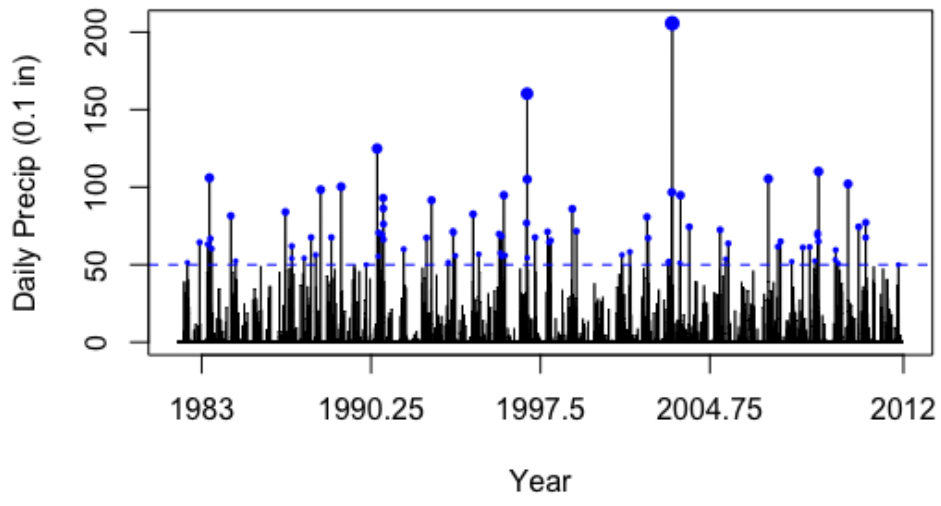


Figure 9: Plots with threshold = 50, ADA

3.3 BOL

The annual maxima for BOL seems to have a two-peak periodic pattern (El nino?) (fig 10), and the maximum values fits fairly good to a **GPD** with parameters (70.312, 16.860, 0.147) (fig 11)., but with a big bump around 130. This distribution matches the return levels much better then for ADA.

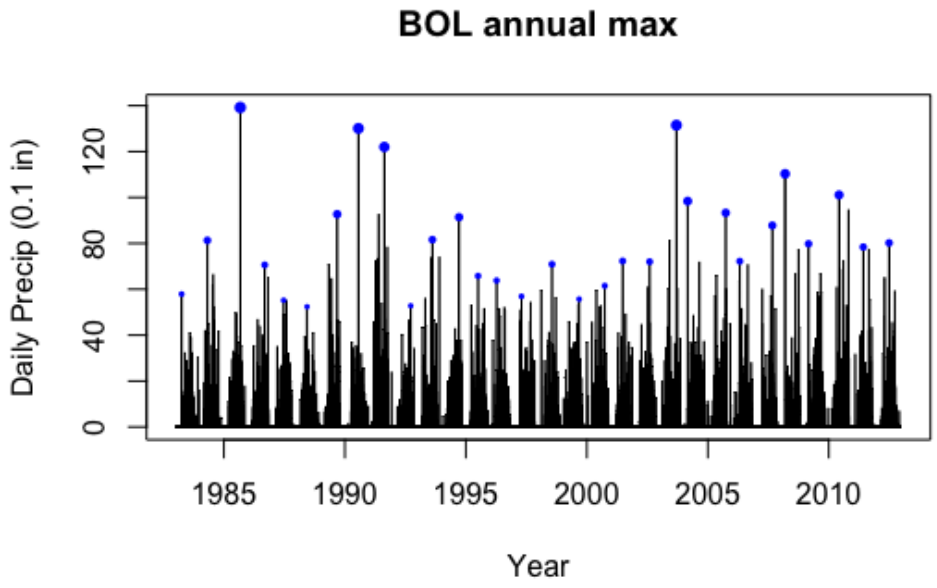


Figure 10: Annual maxima, BOL

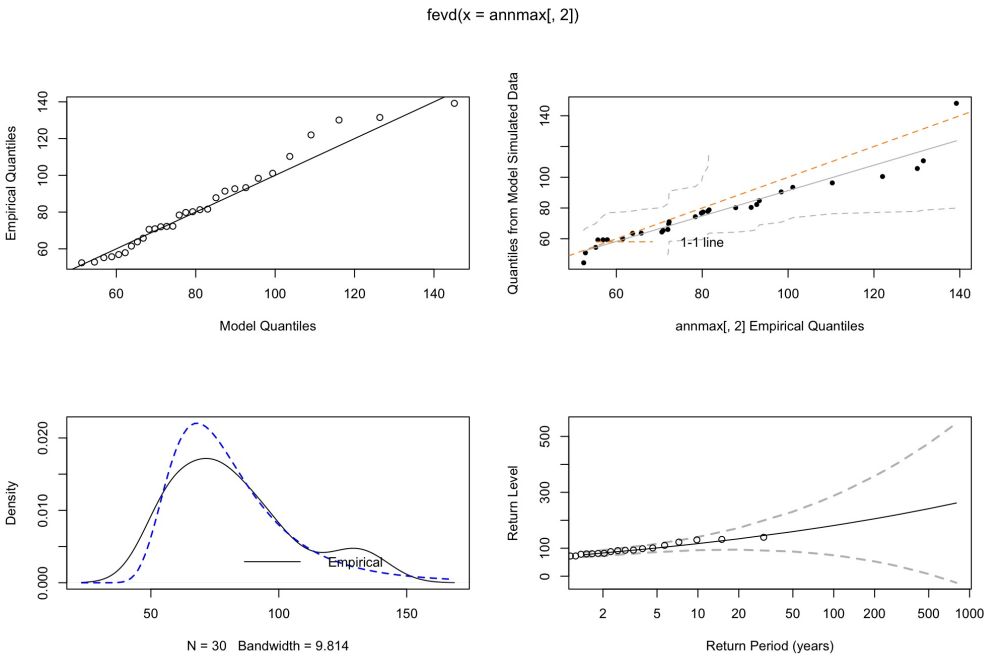


Figure 11: Plots over distribution fit for BOL anuual max

If I use 45 (128 obs.) as threshold, I get a model where the data and the simulated quantiles matches nearly perfect but my data is more heavy tailed and it overestimates the rainfall for higher return levels (fig 12). If i instead pick 60 as my threshold (55 obs.), the tail and the return levels matches much better, but I get a big bump at 100 and 130 (fig 13). The distribution has parameters (68.590, 8.887, 0.423).

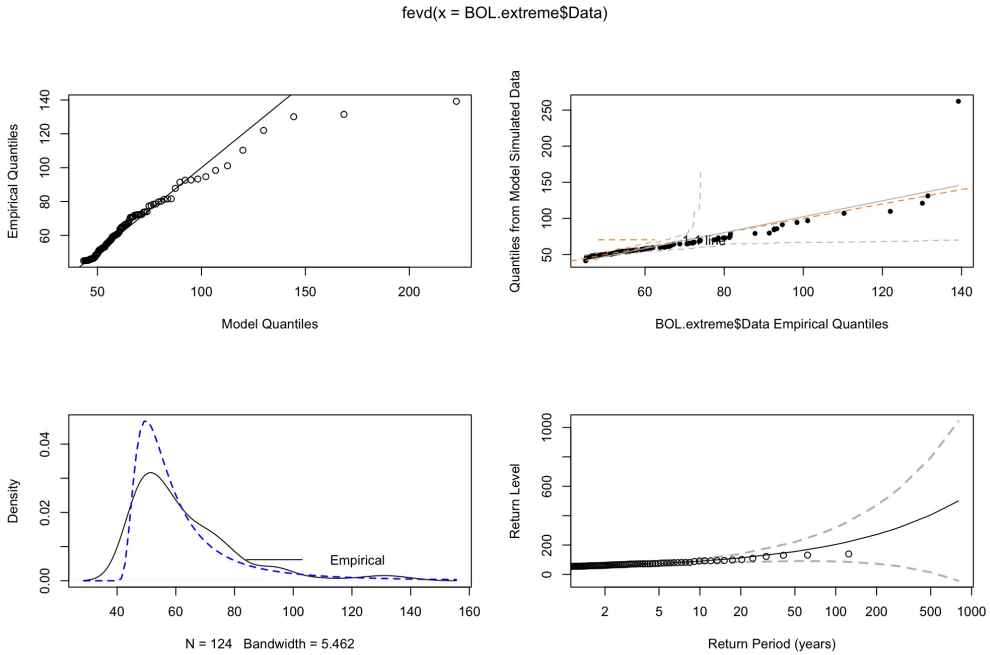


Figure 12: Plots with threshold = 45, BOL

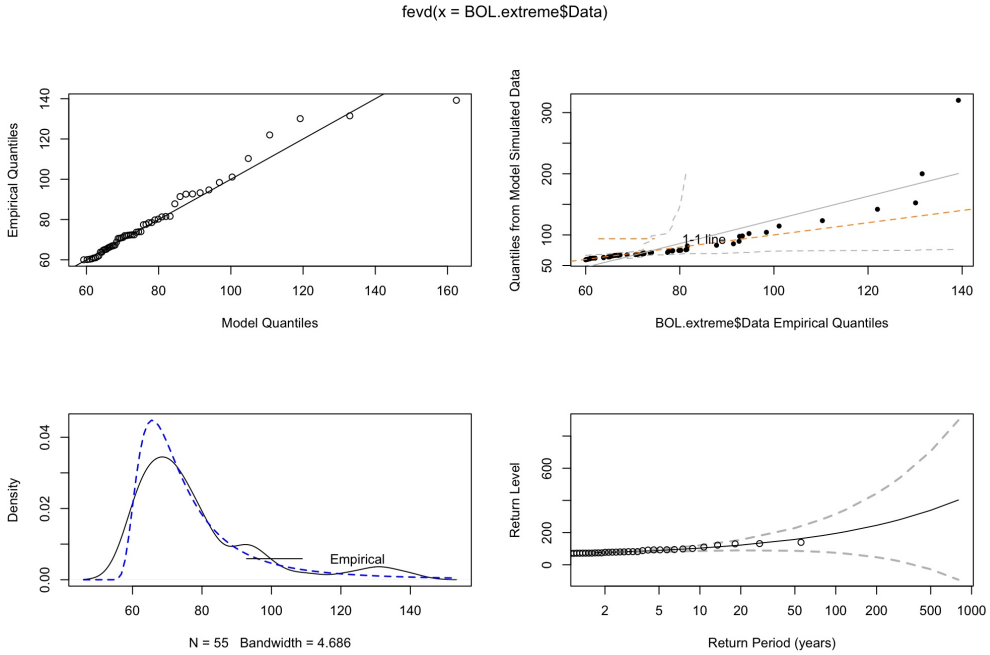


Figure 13: Plots with threshold=60, BOL

If we look at general behaviour of the extreme values, we see opposite bahaviour as for ADA, with much

more large data points after 2000 then before (fig 14).

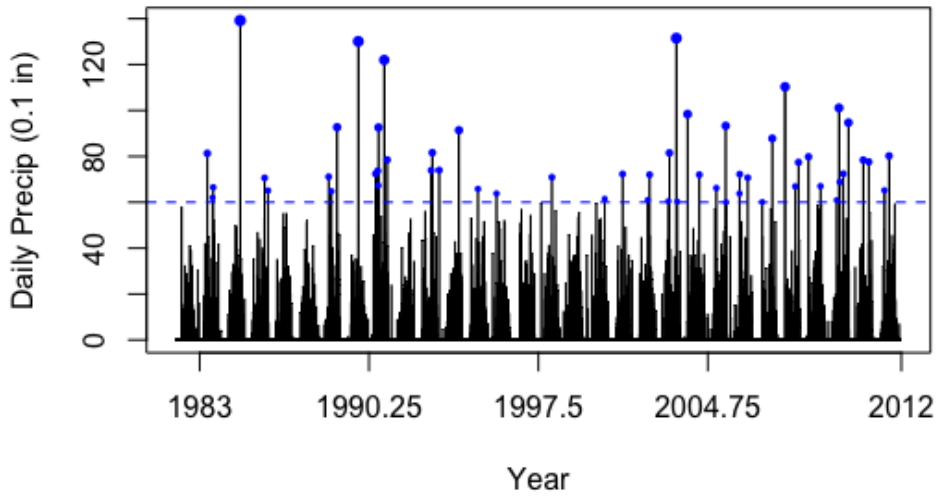


Figure 14: Plots with threshold = 60, BOL

3.4 KRA

The annual maxima of KRA does not really have a shape at all (fig 15) and fits a GPD distribution with parameters (82.566, 4.789, 0.225) (fig 16), but with a large bump around 170.

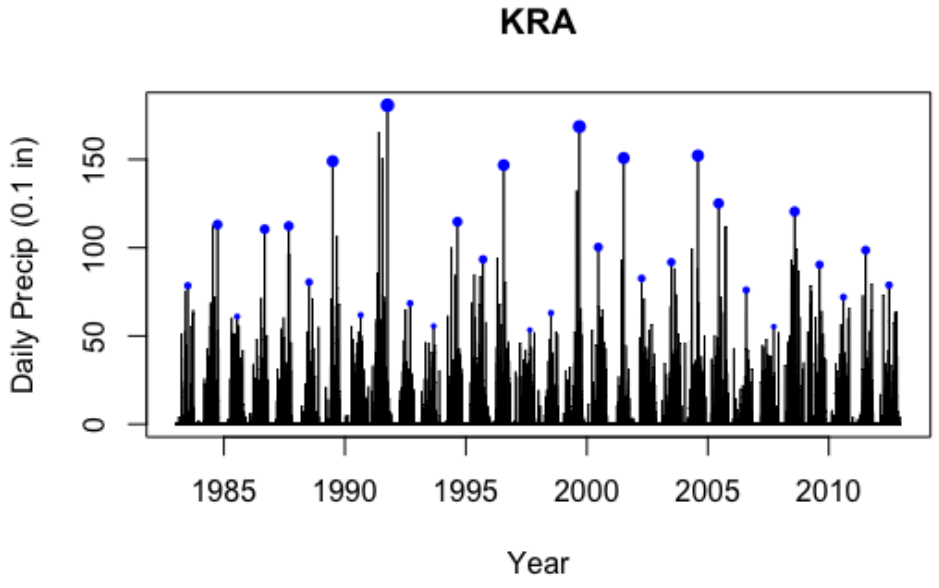


Figure 15: Annual maxima, KRA

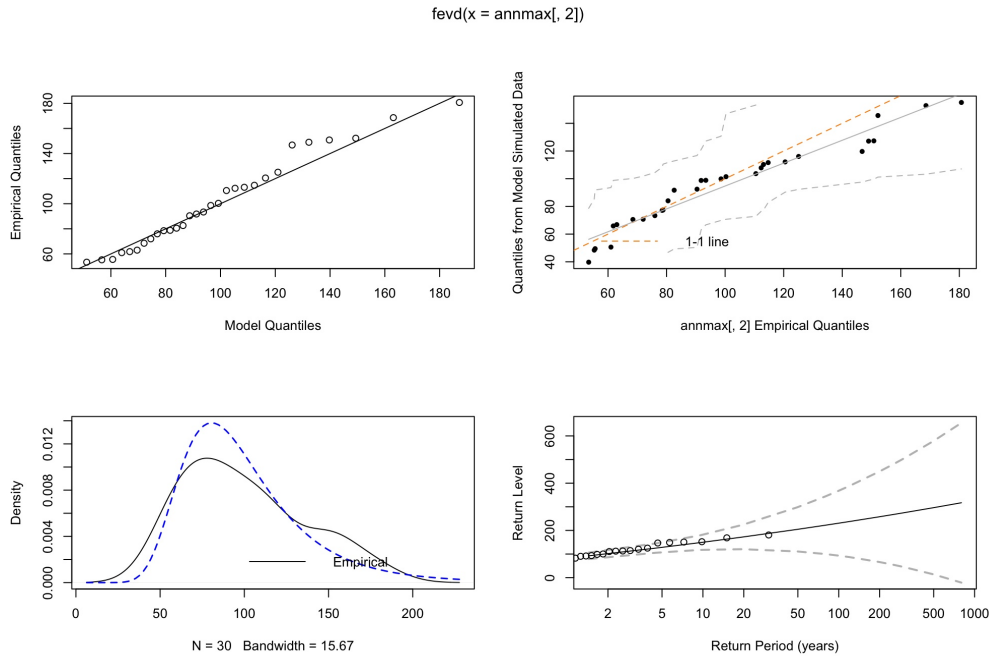


Figure 16: Plots over distribution fit for KRA annual max

KRA behaves very similar to BOL for the threshold 45 (204 obs.), the data has a heavier tail then the fitted distribution, but we get a nearly perfect fit for the quantiles of the simulated data and the density, but overestimates the rainfall for higher return levels (fig 17). But it did not improve very much at all by changing to 60, so instead I used 70 (66 obs.). Now both the tail behvaiour and the return levels matches much better but we see a large bump at 150 (fig 18). The data fits to a model with parameters (82.376, 13.211, 0.564).

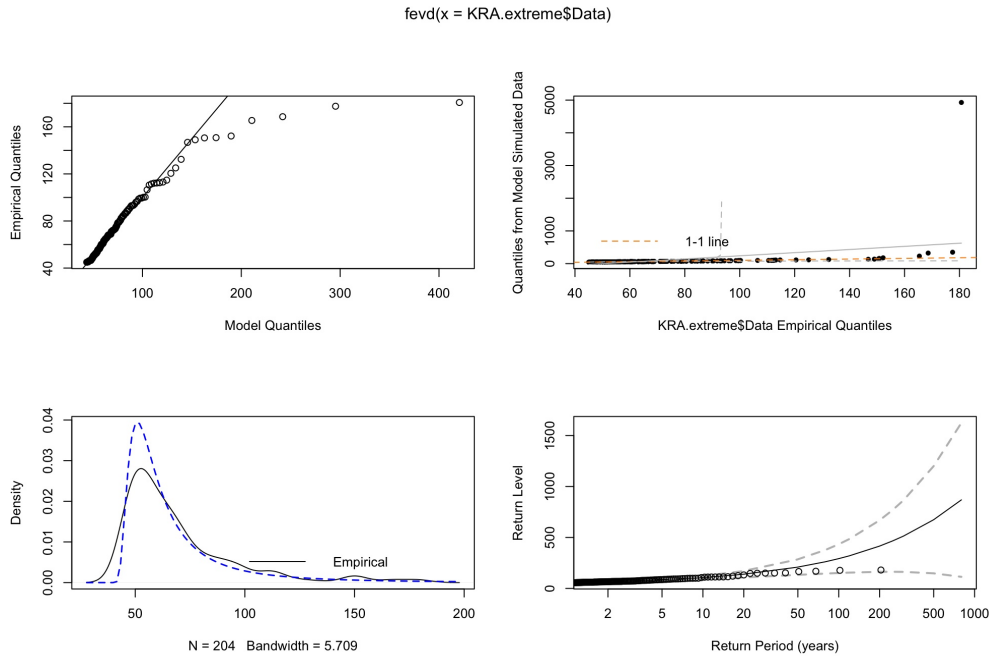


Figure 17: Plots with threshold = 45, KRA

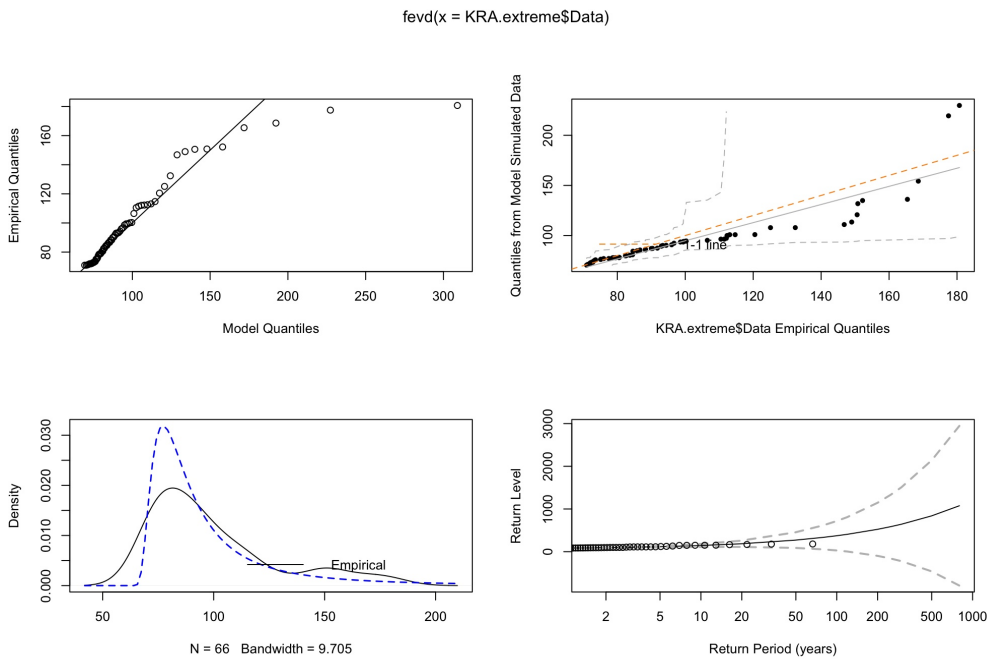


Figure 18: Plots with threshold=70, KRA

KRA has a similar behaviour to BOL, with more extreme days after 2000 hen before and also more spread out instead of in clusters.

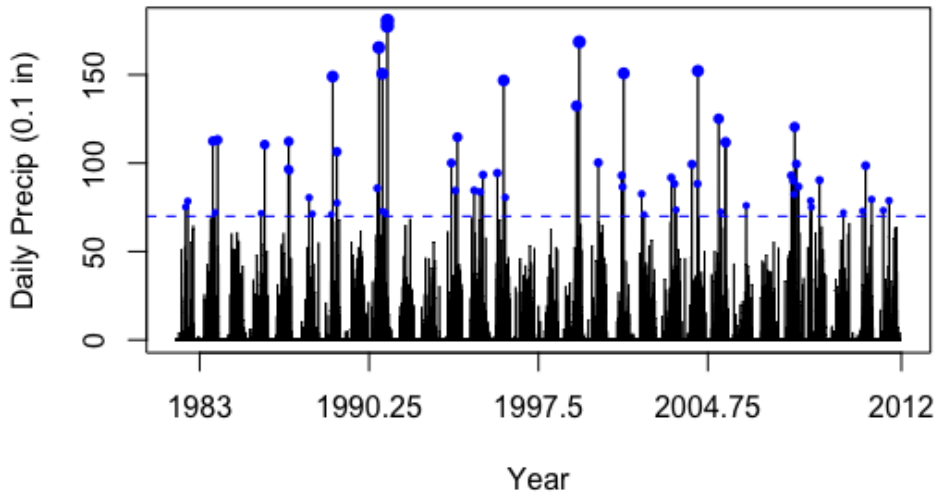


Figure 19: Plots with threshold = 70, KRA

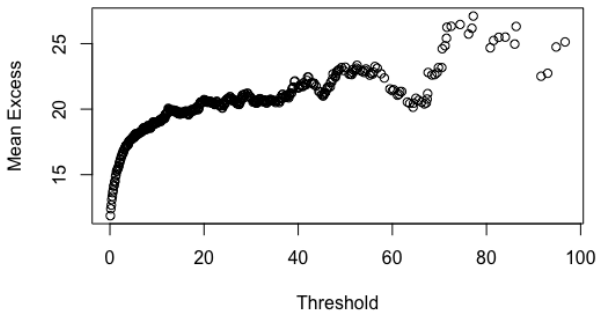
3.5 Summary

	Obs.	Cap	Gamma		Threshold	GPD		
			Scale	Shape		Location	Scale	Shape
ADA	2037	100	17.590	0.585	50	60.806	10.714	0.407
BOL	2688	100	17.249	0.698	60	68.590	8.887	0.423
KRA	2991	120	21.774	0.607	70	82.376	13.211	0.564

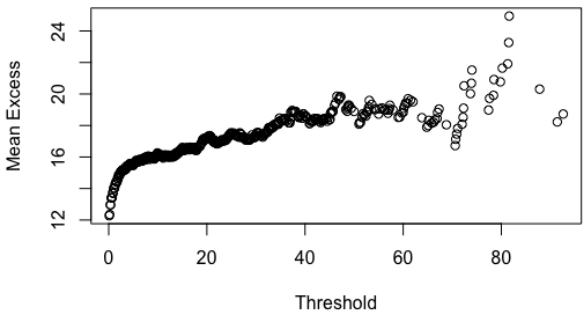
Table 2: Summary of parameters

I can of course not comapare some of the parameters directly since I have used different thresholds and caps, but more the shape of them. ADA and BOL have a similar Gamma distribution, with the difference that BOL has a bigger shape parameter. KRA's shape parameter is between the other two but with a larger scale parameter, possibly since I picked a larger cap. The location parameter in the GPD must be view in relation to the threshold. All of the distributions are shifted veery similarly from the threshold (11, 8, 12) and have similar shape, but differ a bit in their scale. So they do not seem to come from the same distribution.

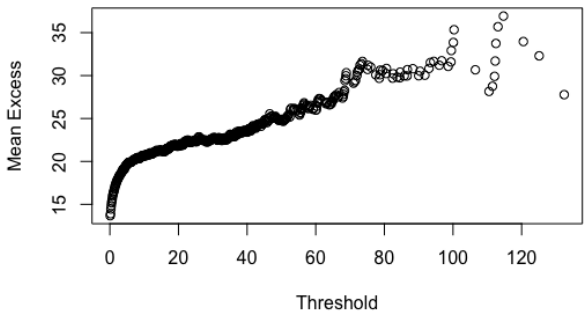
If I look at mean excess plots to confirm my choice of tail distribution, only the plot for KRA shows a clear linear behaviour, the other two behaves as if they should have an exponential tail (fig 20). However, if I try to fit the upper values to an exponential distribution, it is very clear that the data is much more heavy tailed then the exponential, so I choose to stick with the GPD for all of them.



(a) ADA, 10 points omitted



(b) BOL, 10 points omitted



(c) KRA, 9 points omitted

Figure 20: Mean excess plots

4 Comperative assesment

4.1 Data

Station	Long.	Lat.	Pos. Obs.	Missing values	Days (Dec, Jan, Feb)	Annual mean	Rainy day mean	Mode
AXM	-2.23	4.86	4282	58	214, 104, 154	1874	13.13	Semi-bi
ODA	-0.98	5.93	4180	0	140, 60, 151	1407	10.10	Bi
BEK	-2.33	6.2	3851	366	80, 6, 141	1394	10.50	
KDA	-0.25	6.08	3756	30	112 , 81, 167	1293	10.33	Bi
ABE	-0.73	6.65	3415	700	65, 46, 112	1277	10.47	
KSI	-1.6	6.71	3586	0	80 , 51, 136	1347	11.27	Bi
HO	0.46	6.6	3409	0	86, 52, 140	1276	11.23	Bi
TDI	-1.76	4.88	3312	28	113, 37, 92	1079	9.77	Semi-bi
SUN	-2.33	7.33	3187	61	52, 23, 106	1191	11.22	Bi
WEN	-2.1	7.75	3188	2	48, 24, 90	1249	11.75	Bi
KRA	-0.03	7.81	2991	0	30, 20, 35	1366	13.70	Uni
BOL	-2.48	9.03	2688	2	21, 7, 29	1101	12.29	Uni
SAL	-1.06	5.2	2679	123	52, 23, 106	931	10.43	Semi-bi
NAV	-0.01	9.45	2623	62	6, 6, 22	1024	11.71	
WA	-2.5	10.1	2609	62	6, 6, 22	1018	11.70	
TLE	-0.85	8.5	2599	31	7, 6, 23	1017	11.74	Uni
AKA	0.8	6.11	2236	580	74, 23, 56	848	10.99	
NAV1	-1.1	10.9	2184	32	4, 2, 11	988	13.57	Uni
ACC	-0.16	5.6	2130	6	54, 29, 56	747	10.52	Semi-bi
ADA	0.63	5.78	2037	0	37, 21, 58	790	11.63	Semi-bi
TEM	0	5.61	1840	31	42, 23, 49	659	10.75	Semi-bi

Table 3: All stations

All of the missing values comes in clusters corresponding to months, not neccesairly back to back. The three stations with many missing values are all missing a full year. Looking at table 3, I cannot see any direct pattern of the geographically position and the number of rainy days. But what is very evident is that it is a massive spread on the number of rainy days between the stations, and also the spread of number of rainy days in the dry period, even for stations with similar amount of rainy days. One way of splitting up the data could be stations with more than 3000 days of rain and stations with less, since that would split both the interval and the number of stations in half (!) i.e the number of rainy days is uniformly distributed. The annual mean sort of follows the same pattern as the number of positive observations, wheras the rainy day mean seems to follow a different pattern.

There are many ways to split up the data to get the most accurate distribution; number of rainy days, number of rainy days in dry period, annual mean or rainy day mean. Here I have chosen to split it in half on the basis of number of days with rain, which gives the same grouping as if I split it with the annual mean as parameter. A further analysis to do, is to instead use the rainy day mean as parameter and split it. Of course the data will no longer be independet, but it was not truly independent before either.

	Jan	Feb	Mar	Apr	May	Jun	Jul	Aug	Sep	Oct	Nov	Dec
AXM	74.8	71.0	124.2	132.6	182.2	209.5	189	162.8	110.2	180.1	112.9	87.9
ODA	59.1	79.8	127	78.3	166.9	89	101.6	58.8	74.7	99.6	69.4	52.3
BEK	54.3	86.8	93.6	95.7	118.4	104.4	90.2	85	117.8	86	53.9	154
KDA	43	71.2	159.6	105.8	64.7	219.9	101.8	84.8	90.6	65.5	84.7	70.4
ABE	65.2	72.6	91.6	97.8	79.2	145.6	77.5	87.5	121.1	97.5	51.9	67.6
KSI	59.9	60.2	83.3	96.1	125.2	145.8	90	75	111.7	111.5	48.4	84
HO	68.5	61.9	72.4	86.8	128.7	154.2	91.9	92.1	140.3	77.7	68.4	65.1
TDI	34.5	77.8	79.8	90.4	150.5	124.7	158.4	82.5	150.7	152	73.6	82.9
SUN	47.2	70.1	76.1	86	76	121.8	96	71.5	97.8	102.6	45.6	56
WEN	29	55.6	143.3	118	99.4	118.4	102.2	76.9	137.5	94.3	50.2	32.7
KRA	46.3	52.6	68.4	94.4	165.4	149	150.8	152.2	168.6	180.7	65.8	51.9
BOL	37.1	79.8	110.3	81.3	92.6	101.1	130.1	122	139.2	94.7	74	45.3
SAL	47.2	70.1	76.1	86	76	121.8	96	71.5	97.8	102.6	45.6	56
NAV	36.9	67.9	40.6	72.5	126	113.5	87.7	132.9	78.3	68	18.7	37
WA	36.9	67.9	40.6	72.5	126	113.5	86	132.9	101.7	68	18.7	37
TLE	36.9	67.9	77.2	72.5	126	113.5	86	132.9	78.3	68	18.7	37
AKA	50.8	43.8	74.4	116.4	113.2	90.7	77.2	78.9	109.4	77.6	75.6	70.1
NAV1	0.9	22	84.6	70.6	89.6	80.6	116	148.2	133.4	52.2	20.2	31.8
ACC	122.5	71.4	72.3	124.1	157.9	123.3	243.9	57.4	84.6	150.7	66.8	123.7
ADA	31.3	74.4	82.6	124.9	80.8	205.7	93	53.6	102.1	100.3	98.4	65
TEM	39.7	86.7	89.6	116.7	118.2	119.7	129.4	25.9	78.8	71.7	57.9	74.5

Table 4: Table of yearly maximum

WA, NAV and TLE are practically identical with only a few observations differ from each other, which seems to be the dates that they do not have in common due to missing data points. So clearly these three cannot be treated as independently distributed, so I only used TLE in my analysis because it has got the fewest missing data points.

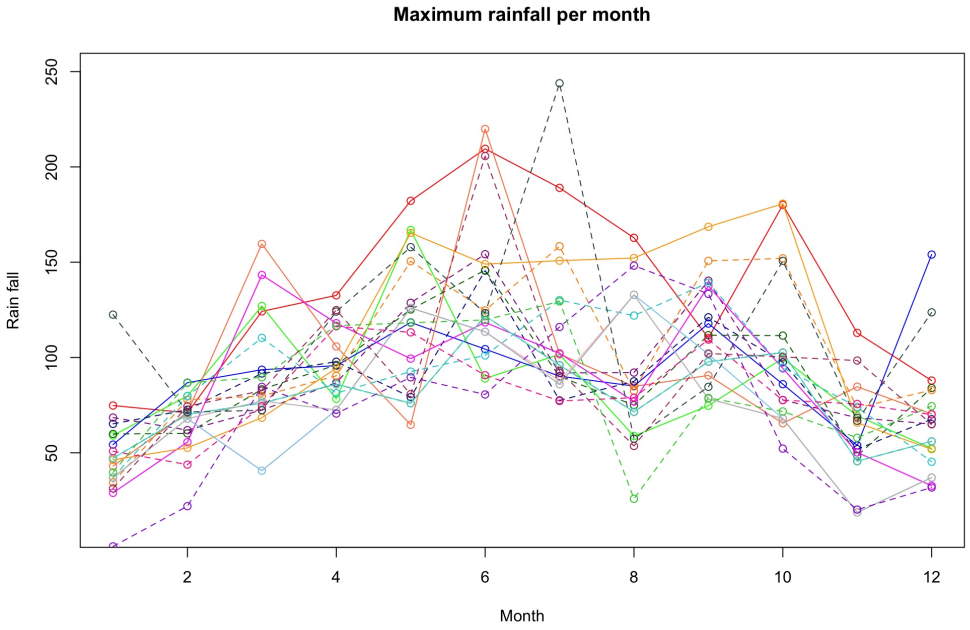
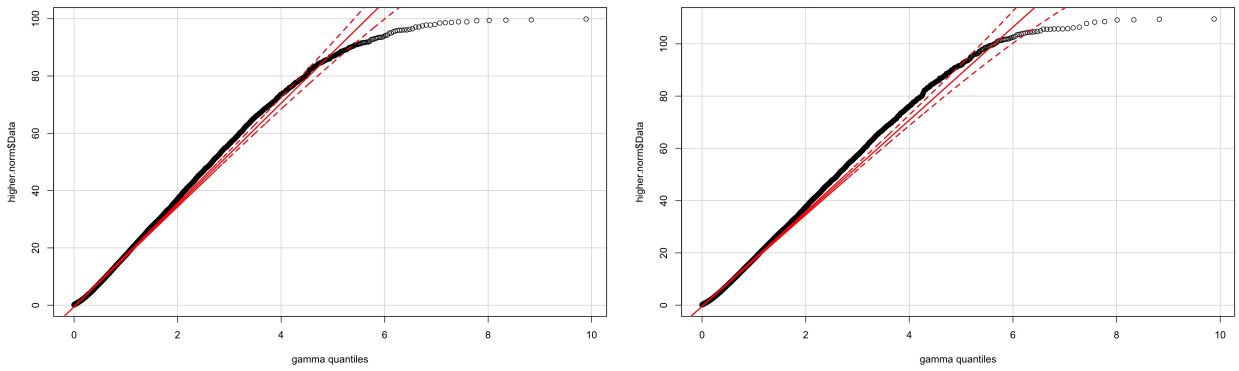


Figure 21: Maximum value in each month, all stations

Even if figure 21 is a bit hard to follow for each station, it is possible to get a view of how the spread and the general behvaiour of the maximum values behaves. We can see that the values in January and December are pretty clusered together, except from the dashed violet which is much lower and dashed gray in January which is much higher and the solid blue and dashed gray in December which both are larger. It is also clear that the spread in April and September is much smaller with no value lying directly outside the others, and February is even closer if we ignore the dashed violet line.

4.2 Distribution fitting

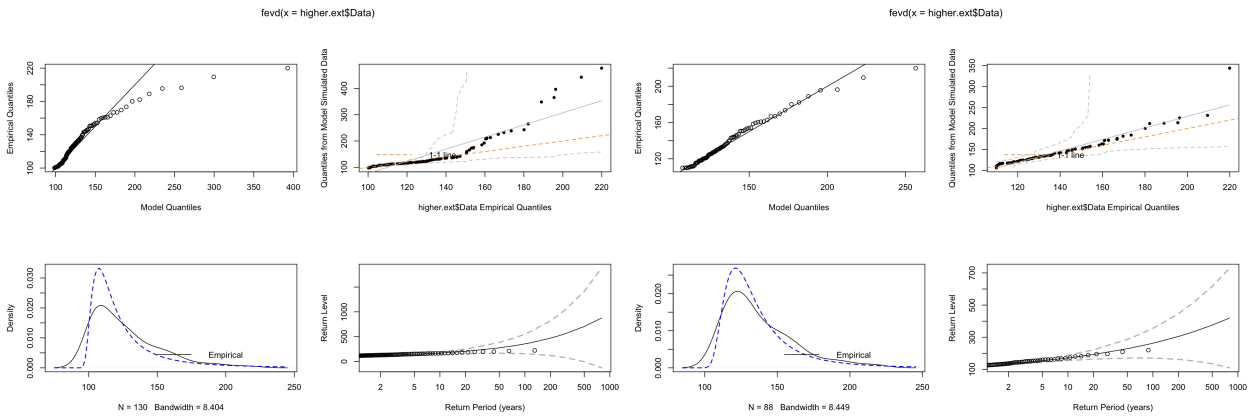
To try and get as accurate distribution fits as possible, I split the data between WEN and KRA and ignored NAV and WA to not have clearly dependent data. For the higher values, I tried both 100 and 110 as splitting value for the gamma and GPD distribution, but I cannot relly see which one is the better fit for the gamma distribution(fig 22)



(a) Data < 100, gamma(17.20, 0.615) (b) Data < 110, gamma(17.47, 0.612)

Figure 22: QQ plot, gamma higher values

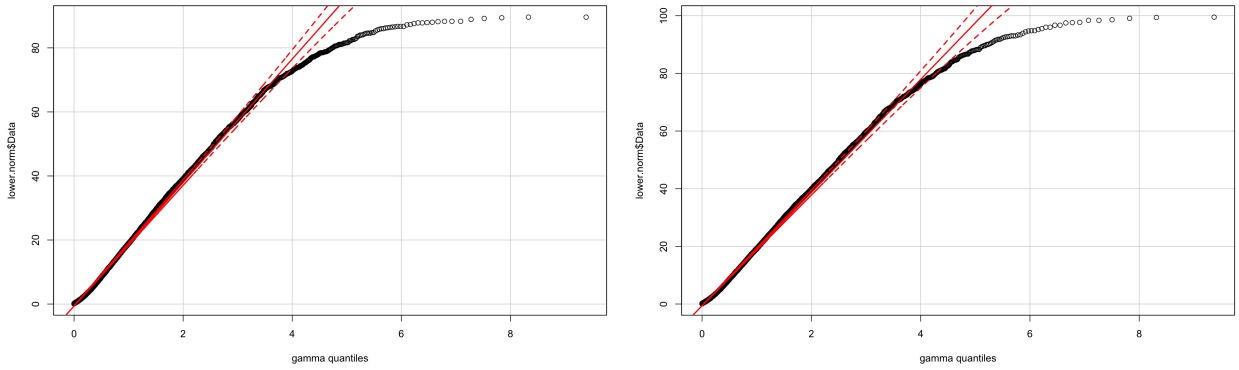
If we insted look at the fit of the GPD for both 100 and 110 as breaking point, it becomes quite clear that 110 is the better choice (fig 23)



(a) Data ≥ 100, GPD(111.92, 12.43, 0.52) (b) Data ≥ 110, GPD(124.60, 14.21, 0.29)

Figure 23: Higher values GPD fit

If I do the same analysis on the lower values, I think the break value should be either 90 or 100 (fig 24)

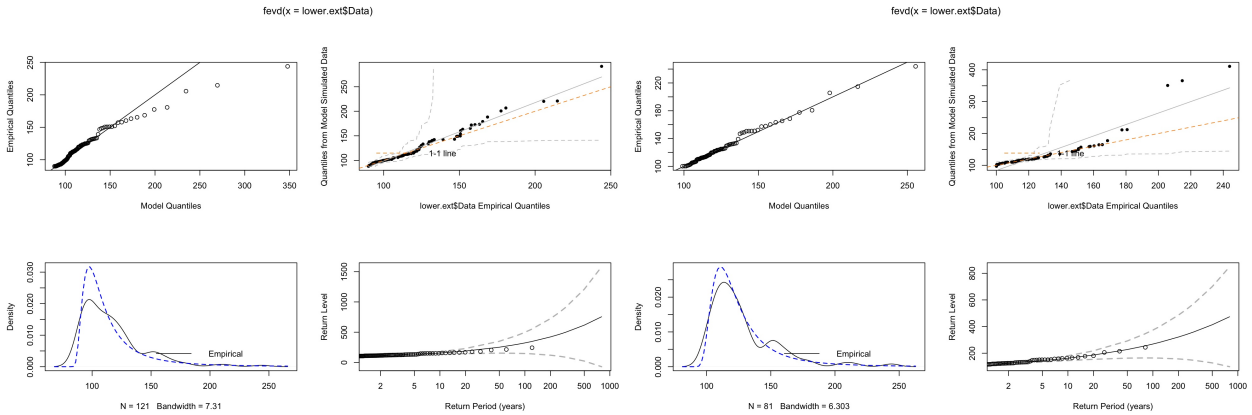


(a) Data < 90, gamma(18.20, 0.616)

(b) Data < 100, gamma(18.59, 0.611)

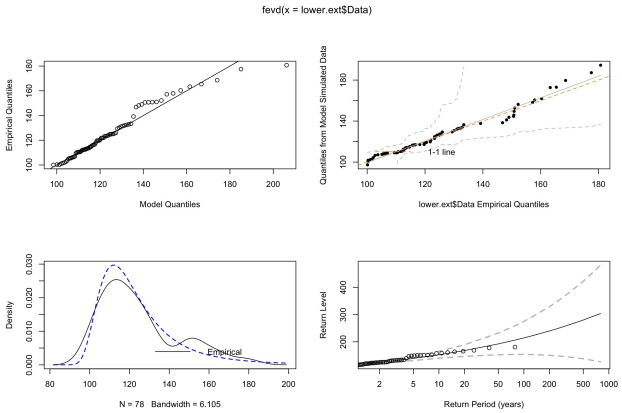
Figure 24: QQ plot, gamma lower values

Now looking at the GPD fit instead, we can once again see that the higher value gives the most accurate extreme fit, and can see that more clearly if we remove the threee largest values to rescale the graphs (fig 25). We can very clearly see the bump in the data around 150, which showed even when I just studied one station at a time. We can see it for the higher value stations as well, but not at all as clear. We can also see that it is worth splitting up the data since we get two different MLE distribution for the two sets if we use the same splitting value, and 100 is a suitible value for the lower data set wheras it is not very good for the higher.



(a) Data ≥ 90, GPD(102.07, 12.83, 0.49)

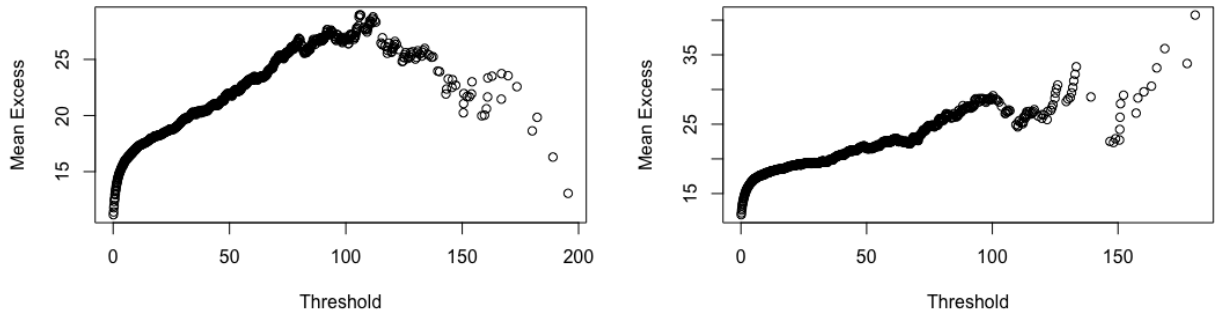
(b) Data ≥ 100, GPD(114.67, 13.52, 0.35)



(c) Data ≥ 100, 3 largest values removed, GPD(114.47, 12.65, 0.22)

Figure 25: Lower values GPD fit

We can also look at the Mean excess plots to confirm that GPD is the correct distribution for the extreme values (fig 26).



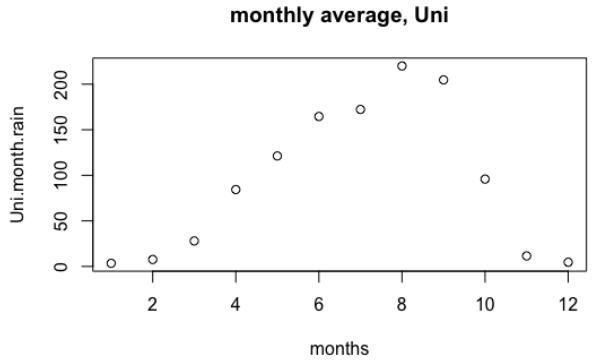
(a) Higher values

(b) Lower values

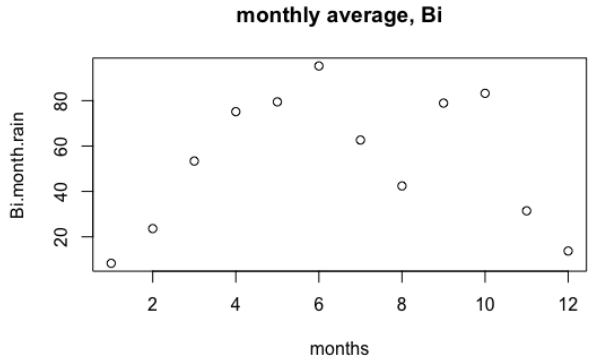
Figure 26: Mean excess plots

5 Rain mode analysis

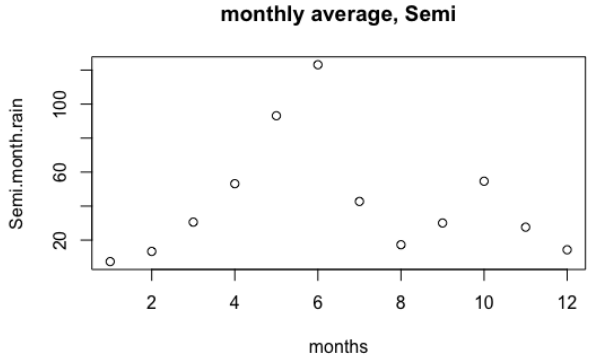
By plotting the monthly average of each station, we can easily see a clear uni modal, bi modal or semi-bi model behaviour for each station. We can therefor split the stations into these three categories and plot these 3 groups monthly averages instead, to get even more data to work with (fig 27)



(a) Uni modal



(b) Bi modal

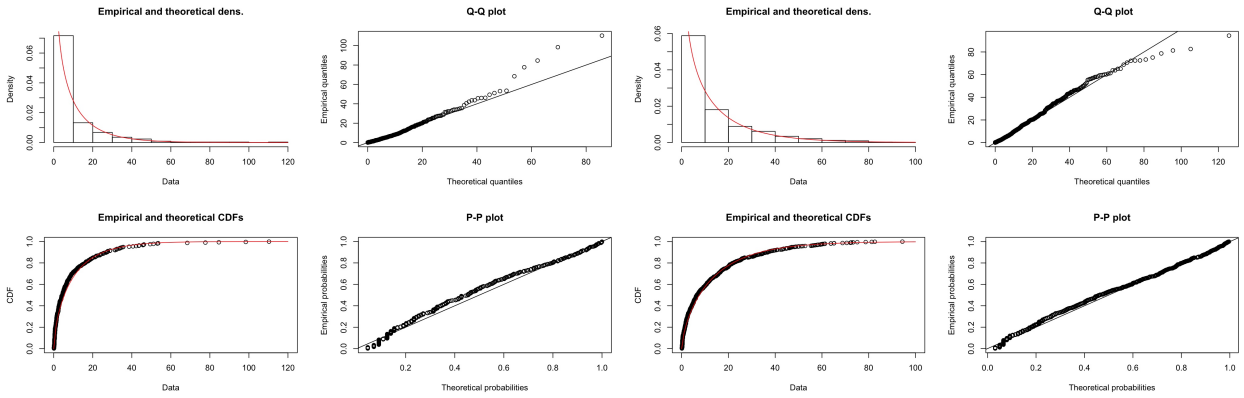


(c) Semi bi modal

Figure 27: Monthly averages for each rain mode group

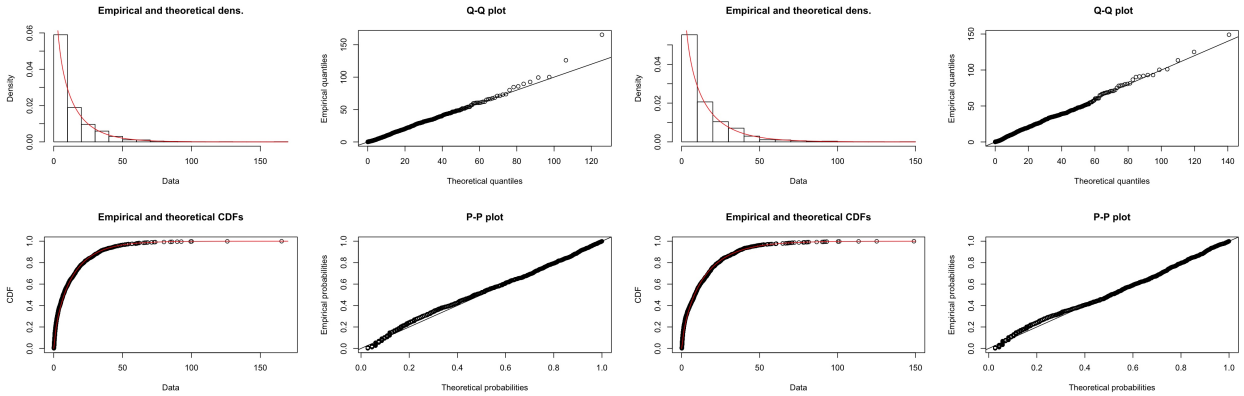
5.1 Uni modal stations (KRA, BOL, TLE, NAV1)

As many other has discovered, the rain distribution in the north only has one rainy season, as can clearly be seen in the histograms (fig 27). It only misses data from one november and one december and it has got proportionaly equally many positiv observations as the semi-bi modal group, but very differently distributed. It has got a much dryer Dec-Feb period then the other two groups, so we do not have enough data in these months to fit a distribution. For the other months, we have enough data to try and fit distributions. By plotting histograms for each month, one can see that all months still seems to follow a gamma distribution, alternatively a lognormal or exponential. By looking at QQ plots of the months, we can see that all months with many data points fits very well up to around 60 or 80 mm, depending on month(fig 28).



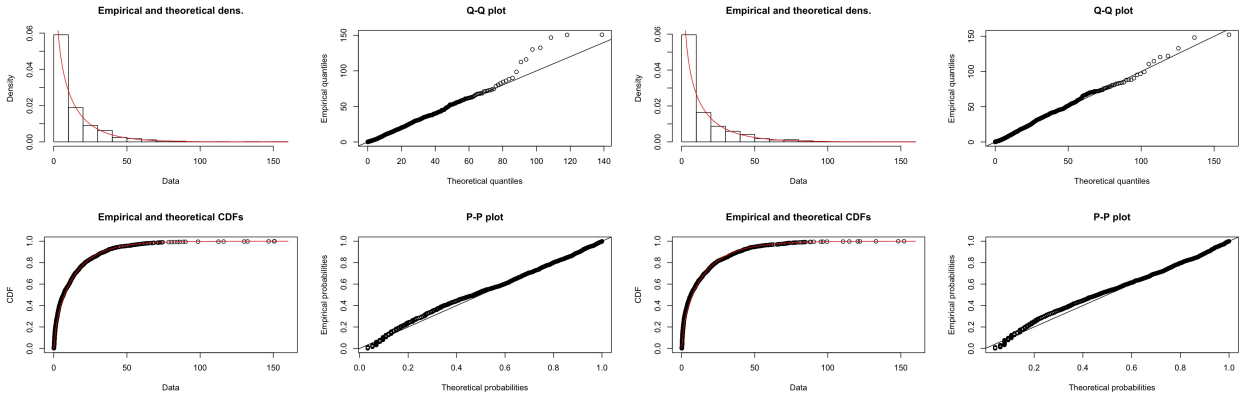
(a) March,(15.633, 0.633)

(b) April, (19.673, 0.674)



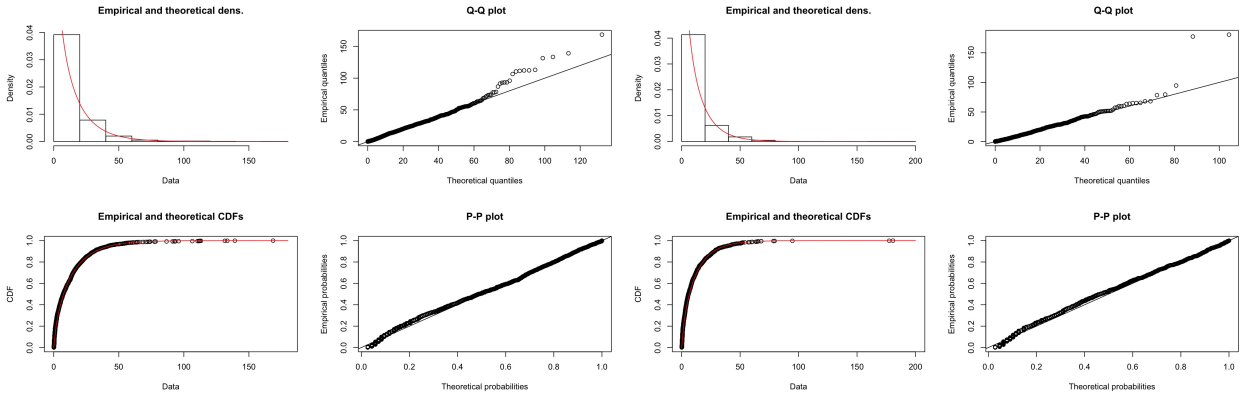
(c) May, (18.326, 0.705)

(d) June, (19.902, 0.703)



(e) July, (19.678, 0.663)

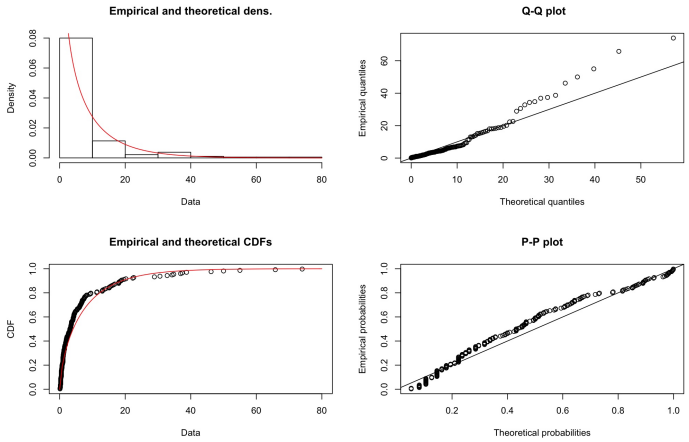
(f) August, (22.687, 0.614)



(g) September, (17.848, 0.716)

(h) October, (15.229, 0.726)

Figure 28a: Gamma distribution fits, Uni modal group



(i) November, (11.537, 0.643)

Figure 28b: Gamma distribution fits, uni modal group

July is one of the clearest months that the fitted model is not suitable for the entire data set. If we instead split the data by 120 mm, the new fit works a lot better. November is a pretty poor fit as well, but we do not have enough data to make it much better. October clearly has a couple of large outliers, but fits nearly perfect else. We can most likely improve the fit of both September and March by splitting the data.

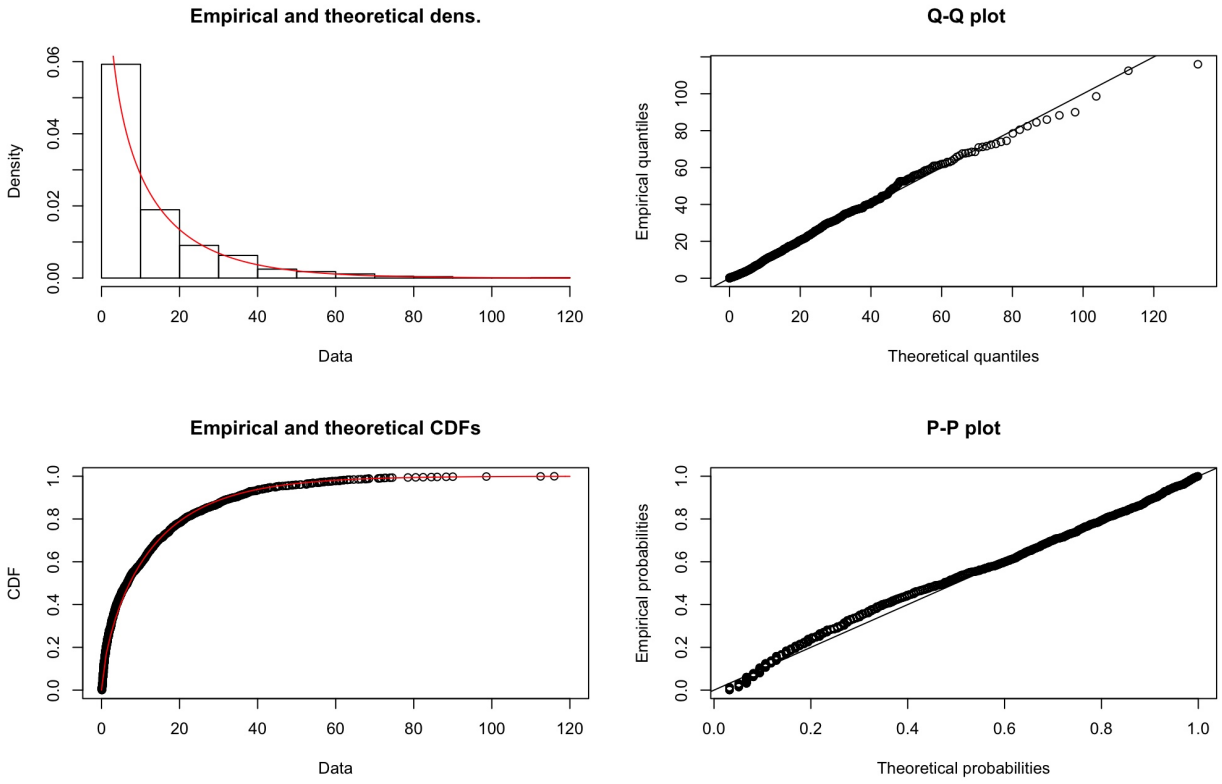
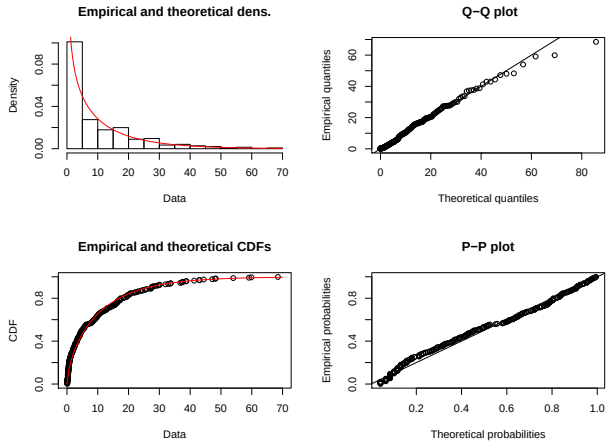
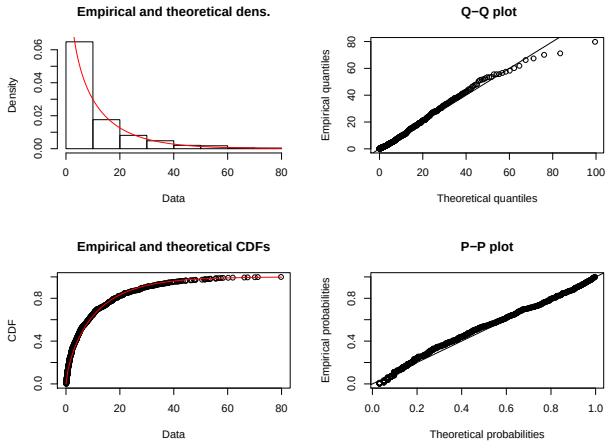


Figure 29: July, Data < 120, (18.668, 0.677)

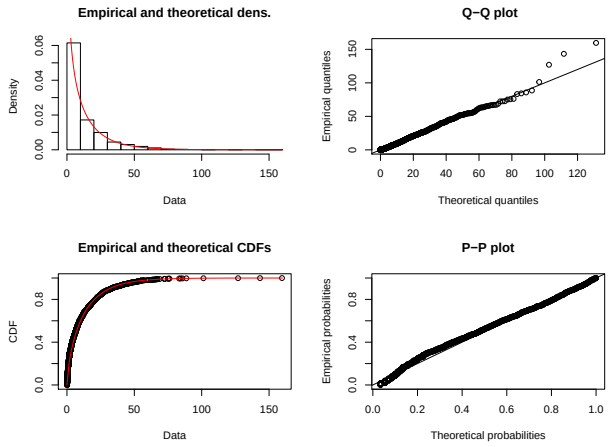
5.2 Bi modal stations (ODA, KDA, KSI, HO, SUN, WEN)



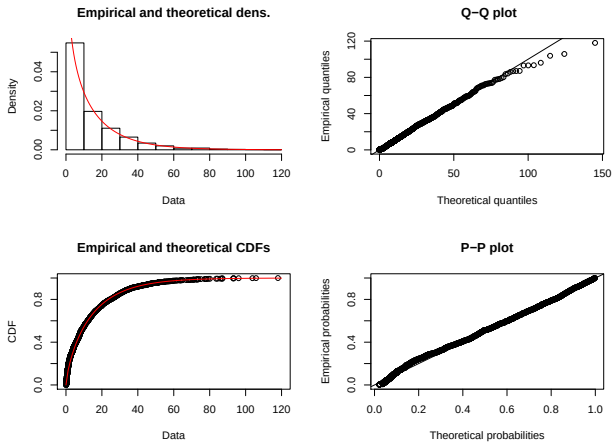
(a) January, (15.983, 0.648)



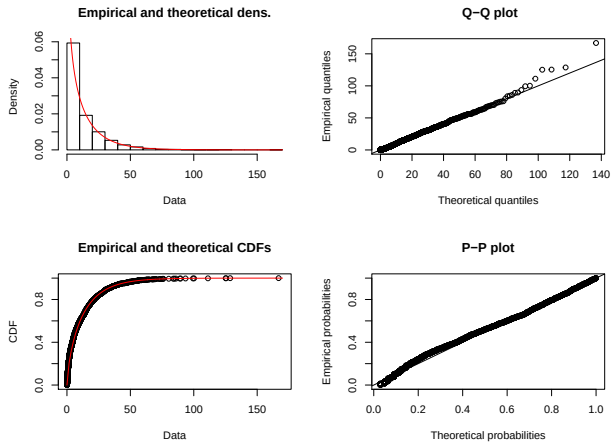
(b) February, (15.275, 0.705)



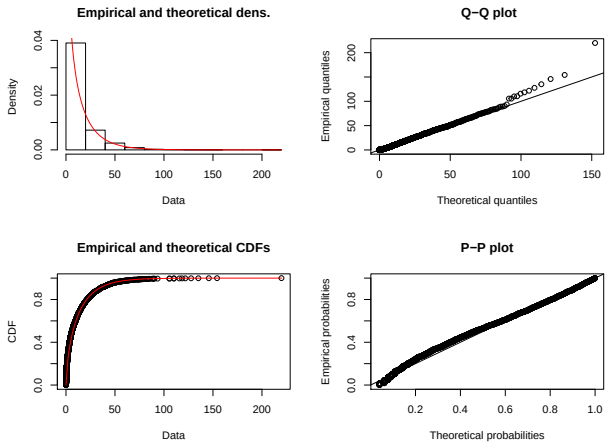
(c) March,(18.644, 0.666)



(d) April, (19.447, 0.737)

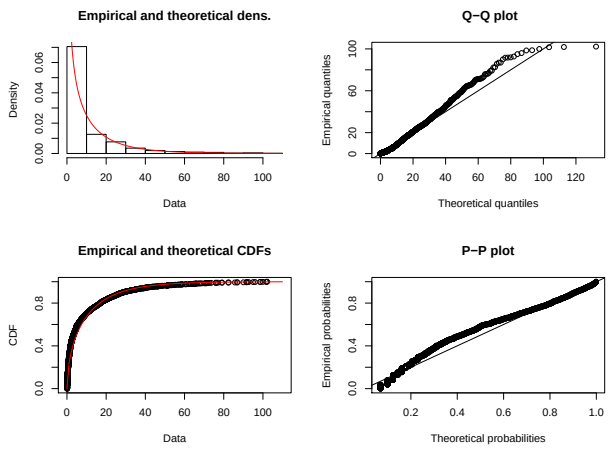


(e) May, (18.246, 0.691)

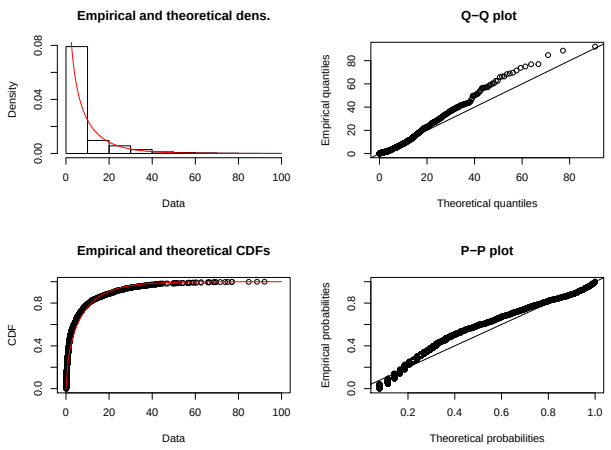


(f) June, (20.498, 0.627)

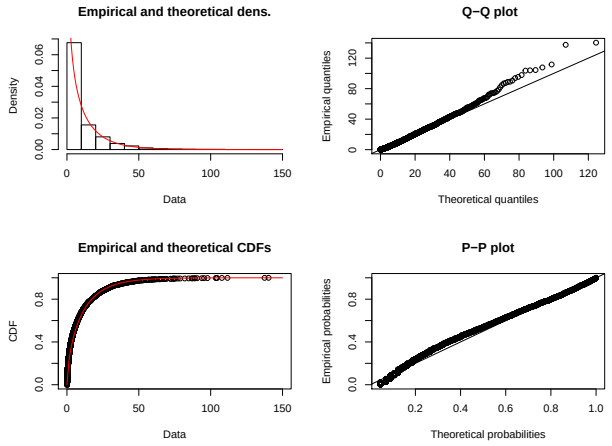
Figure 30a: Gamma distribution fits, Uni modal group



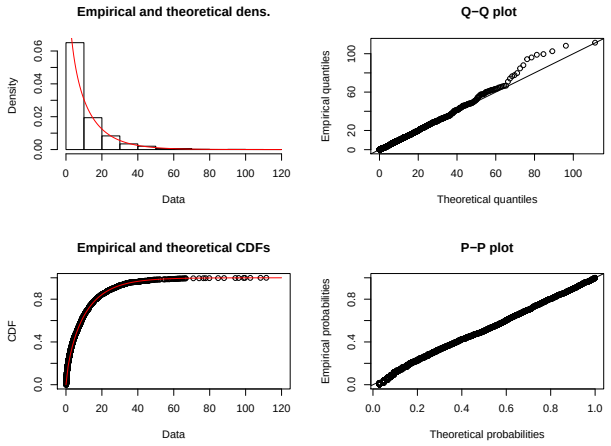
(g) July, (19.122, 0.532)



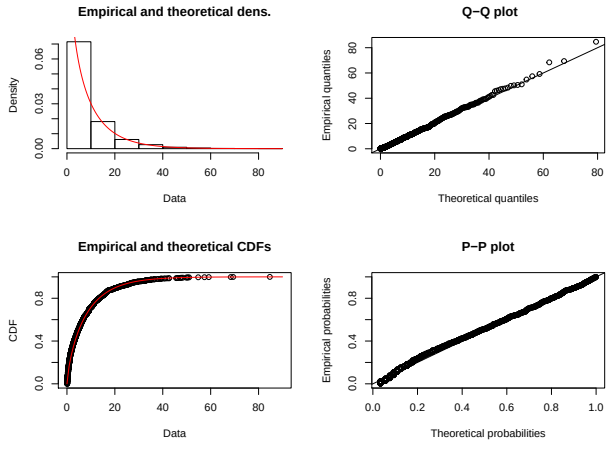
(h) August, (13.068, 0.547)



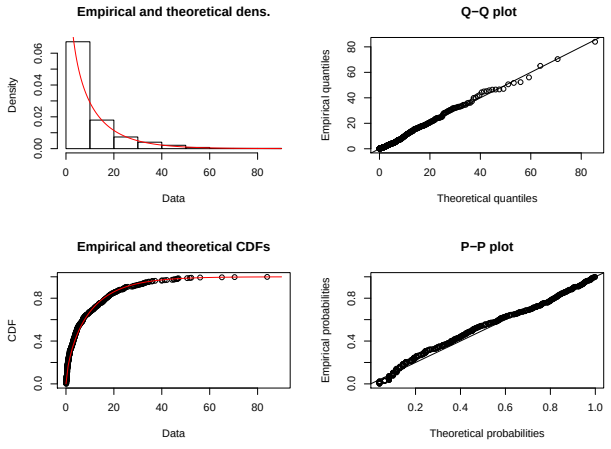
(i) September, (16.732, 0.625)



(j) October, (14.17, 0.73)



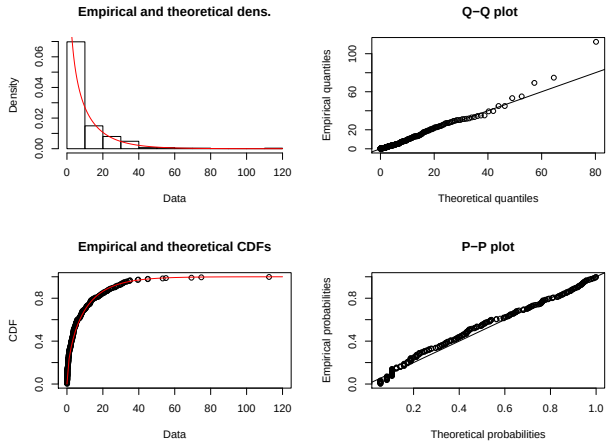
(k) November, (11.099, 0.736)



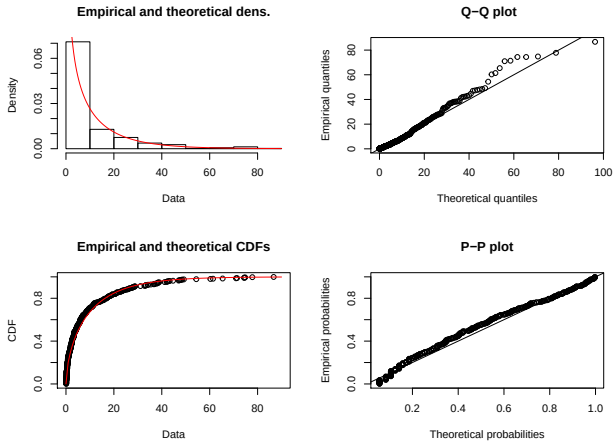
(l) December, (14.223, 0.672)

Figure 30b: Gamma distribution fits, bi modal group

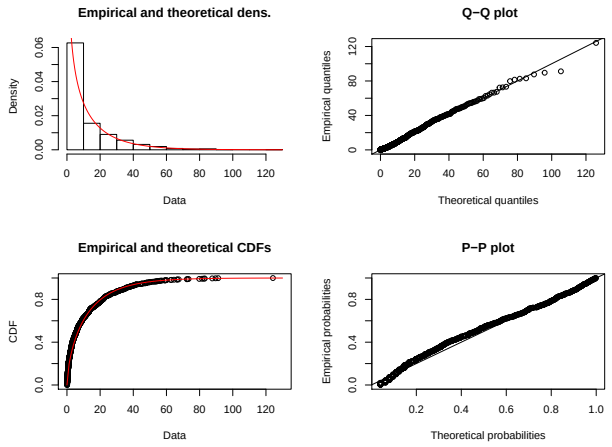
5.3 Semi-bi modal stations (AXM, TDI, SAL, ACC, ADA, TEM)



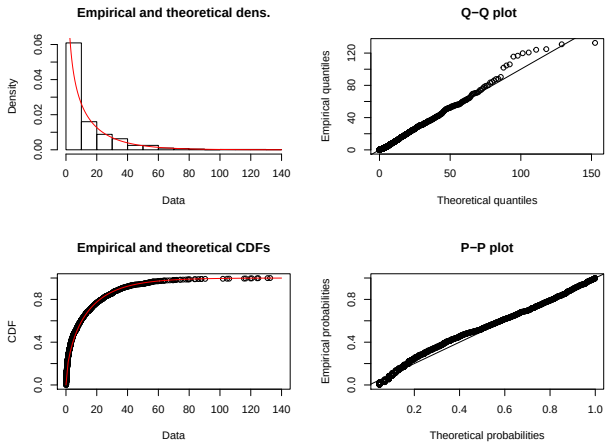
(a) January, (15.346, 0.599)



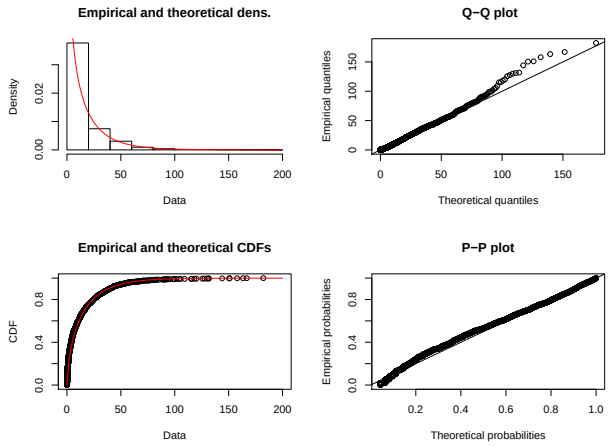
(b) February, (17.008, 0.583)



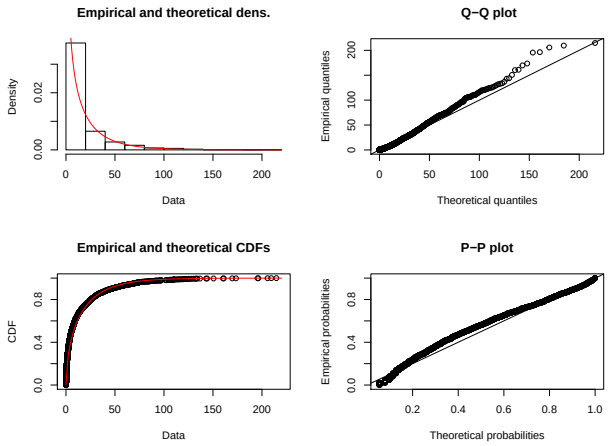
(c) March,(19.727, 0.627)



(d) April, (22.747, 0.588)

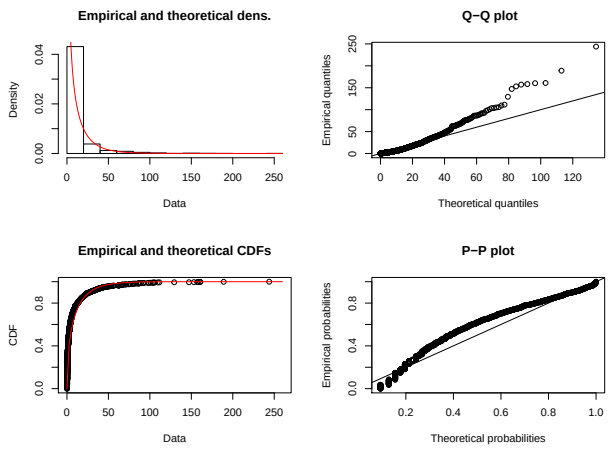


(e) May, (24.726, 0.591)

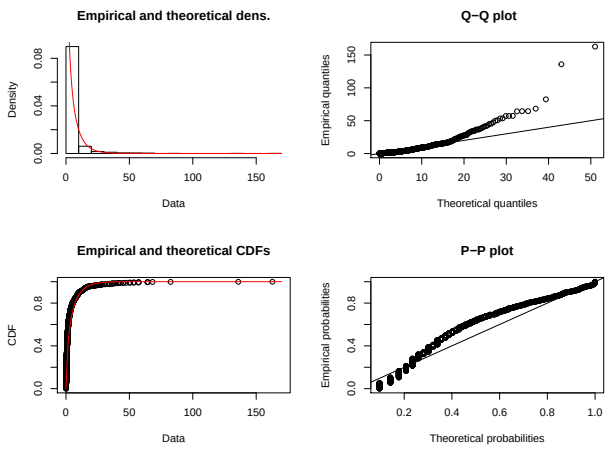


(f) June, (30.324, 0.529)

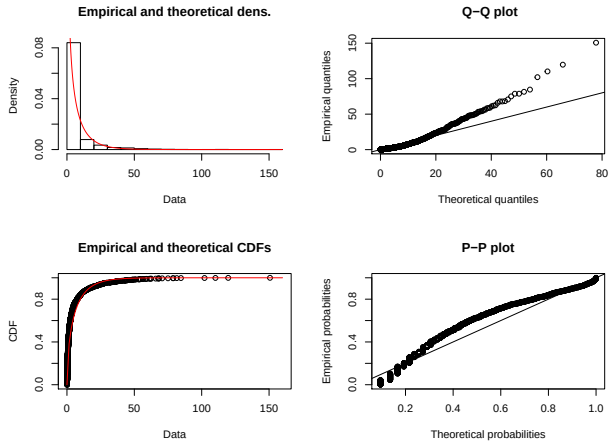
Figure 31a: Gamma distribution fits, semi-bi modal group



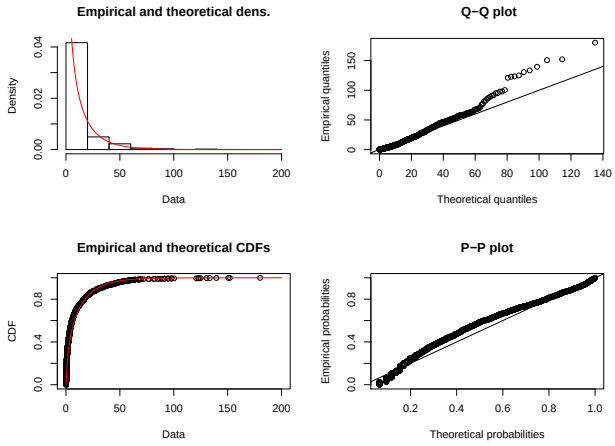
(g) July, (21.233, 0.465)



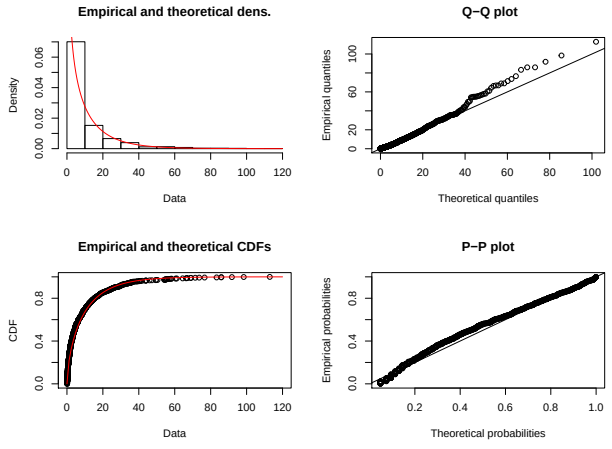
(h) August, (7.705, 0.564)



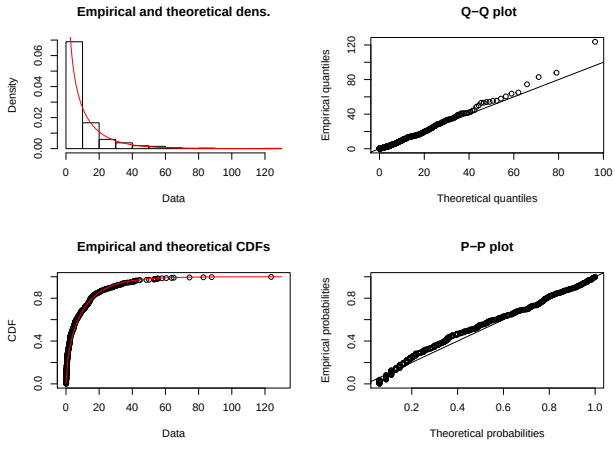
(i) September, (11.671, 0.519)



(j) October, (19.981, 0.538)



(k) November, (15.698, 0.621)



(l) December, (16.805, 0.575)

Figure 31b: Gamma distribution fits, semi-bi modal group

6 Justification of splitting and model selection

To find the most suitable distribution for the data, I started with the threshold 80 to eliminate at least all extreme values, but still keep large values to not change the structure of the data to much. I then fit this smaller data set to; gamma, log-normal and exponential distribution, since all of them where suggested models in other rain prediction papers. I used the AIC, by taking the model with lowest value, to determine the most suitable model, which was the gamma distribution. I then fit the data both using MLE and MGE method and concluded that MLE was the best method, once again with use of AIC.

To determine if the model became better if I only used data up to a certain value, and what value that should be, I instead used Q-Q plots. I fit the gamma distribution to data sets smaller than 40, 50, 60, 70, 80, 90 and 100 mm per day and the full data set. It is very clear that we get a much better model fit if we exclude the most extreme values. The same gamma distribution is fitted if I use 80,90 or 100 as the splitting value, probably since I add very few values in each step.

7 Wilcoxon

By splitting up the data in different ways; in year groups, by annual rainfall or average daliy rainfall, we can use Wilcoxon sign rank test to see if these groups come from the same distribution of if they are different. If we split up the data in decades, and get the average rainfall in each month , we can treat these two data sets as paired. The test is not significant if we run it with consecutive decades (0.30, 0.42) but it is significant on a 10% level between the first decade and the last (0.09). This gives a very small proof that the rain pattern might have changed over the 30 years, but you would need more years to prove such a statement, since you have easily have fluctations in this period.

If we instead calculate the average monthly rainfall between the stations with highest rainfall and lowest(values < 80), we instead get a strongly significant difference with a p-value $= 0.006836$. So clearly there is a stronger variability between the stations then between the decades.

8 Literature review

In general, I have had a hard time finding any work on rainfall distributions in Ghana, but it has been done in other tropical regions and similar work has been done in temperature. Work in Ghana has mainly been focused on the variability in annual or seasonal rainfall.

Ghana is a country in the south of Africas horn, by the coast and shares boarder with Burkina Faso, Côte d'Ivoire and Togo. It has five distinct geographical areas; low plains in the south, the Volta Basin in the centre with the artificial lake 'Lake Volta', the Akwapim-Togo ranges to the east of the Volta Basin with many heights and folded strata, the Ashanti Uplands to the west and high plains in the north[2]. The temperature is peaking around February-March and at its lowest around August. Ghana has three distinct rainfall behvaiours. The northern part experiences an unimodal season with the rainy season between April and September, wheras the rest of the country has a bimodal seanson, first one in April to July and the second September to November. The difference is that some parts of the country has two modes of the same amplitude wheras the other has peaks of different aplitudes. But they all have in common a slowly increasing peak but a rapid decrease in October.[4] The different rain patterns depend on a few wind and preassure phenomenons. It is strongly affected by the position of the **Inter-tropcial convergence zone**, which goes between the nothern and southern tropics every year. The prevailing winds north of the ITCZ is called the Harmmatan and brings hot and dusty air from the Sahara desert between Dacember and March, which gives rise to the very dry season. The prevailing wind south of ITCZ is southwesterly and instead brings humid air from the Atlantic ocean. As ITCZ moves from the nothern position to the souther and back, the opposing prevailing winds gives rise to the West African monsoon which shows as the two rainy seasons. The rainy season in the north corresponds to when the ITCZ are at its most nothern position[4].

To get a better idea of how the proportion of wet and rainy days are distributed several ideas are proposed. In [5] they use the binomial distribution to model dry and wet days and combine it with a continuose distribution for the wet days. The continuouse distributions they look at are; gamma, lognormal, exponential and weibull, which seems to be the standard distributions to test. They determined which was the most

suitable model by looking at the AIC and picked the model with the lowest value. They concluded that the most stations fitted best to a mixed lognormal but some to the gamma, and its was strongly connected with the surrouding topology.

[7] they instead fitted mixted distributions, so two distributions of the same type but with different parameters, and then a weight parameter for the two distributions. This was also done on Malyasian daily rainfall, and they used MLE to fit the distribution. To then identify the most suitable model, they used 7 different goodness-of-fit tests; median absolute difference, Kolmogorov-Smirov, Cramer-von-Mises, Anderson-Darling, New Kolmogorov-smirnov, New Cramer-von-Mises, New Anderson-darling. Since they used so many test, they determined which was the best model by taking the model that performed best in most tests, and did not do any analysis on which tests where the most suitable. Most stations fitted to a mixed weibull and a few to a mixed gamma, once again stronlgy influenced by their geography and topology. Finally, they looked at what model fitted best for each monsoon season and their transitional periods.

In [6] they instead use a Markov chain to look at the probability of rain given a certain number of dry days, and let that probability change with months because of seasonality. This gives a much more descriptive representation of the distribution of rain. They also decided to classify rain less the 2.5 mm as trace instead and made seperate probabilities for that.

A completly different approach is used by [8], where they instead of trying to fit a model to the data points, treat the data points as realisations of a multivariate normal distribution. The data they are working with is measured in mm and does not classify any rain as "traces". This method tries to predict rainfall in both space and time, by letting the observation be a parameter dependent on both. They base their model on a truncated normal model

$$z = \begin{cases} w^{\beta}, & \text{if } w > 0 \\ 0, & \text{if } w \leq 0 \end{cases} \tag{11}$$

where z is the observed rain fall at a specific station and time and w is distributed normally with a known mean and variance. By using a Bayesian method, they incorporate the inceartanty of the parameters into the posterior distribution.

9 West African metorology

The monsoon differs over Africa due to its geographical difference. In west Africa, the northern part consists of land and the southern part of ocean wheras east Africa only consists of land, eventhough the northern part is wider then the southern. In the northern hemisphere summer a thermal low pressure builds up over the continent around 20°N and the ITCZ moves northward to 15°N. At the same time, heat troughs develop over North Atlantic ocean, the anticyclone system **St Helena** builds up ove the South Atlantic ocean and cold water is flowing northward along the south-west African coast. Beacause of all this a pressure gradient is formed between South Atlantic and north Africa which makes the south-easterly flows to recurve and become the south-westerly monsoon flow over west Africa.(McGregor:1998). This flow brings cold and moist oceanic air which is the reason for rain, so the area with maximum rainfall is where we have thickest air masses, i.e closest to the coast. North of the ITCSZ conditions are generally cloudless and dry and south of the zone we find the most cloud cover and maximum rain. Dry years are caused by: retardation of the northward movement of the south-west monsoon, a southward displacement of the intertropical discontinuity,the near-equatorial trough and the zone of maximum surface pressure and anomalously cold water to the nort-west of the line linking SW West Africa and NE Brasil. During the northern hemisphere winter, West Africa is instead under the influence of north-easterly trade winds which bring dry and stable air masses, often containing dust particles. These winds are called **Harmattan**. At this time, the ITCZ is located to the south of the southern west-African coast, hence why there is no rain over west Africa this time of year.

It is not clear wheather ENSO has an impact on the African monsoon or not, but it seems that very intense warm phases of it can reduce the monsoon precipitation, as recorded during the -83 ENSO. The **west African mid-Troposhperic jet** is a thermal wind that occurce due to the temperature gradient between the warm Sahara and the cold Gulf of Guinea. We can find maximum rainfall on the equatorward side of the jet. The **ITCZ** is composed of two zones, one with low pressure, maximum surface temperature and wind confluence and another zone with maximum cloudiness, therefore also maximum rainfall and wind converge. These two zones can be as far aprat as 1000 km. In west Africa, the African easterly jet seems to have a big

influence on the initiation of them. Moisture in the atmosphere is a key thing for creating squall lines, by having moist lower levels and dry mid levels a postive feedback is created over the zone and therfore further developes the zone. There are more squall lines during the beginning and end of the monsoon because the dry mid level is absent during the middle of the monsoon.

Clouds are generally on a higher altitude in the tropics compared to the northern hemisphere. All rainfall is a result of upward movements of moist air. For this to happen, the atmosphere must be in a state of conditional, potential or convective instability. Three types of rainfall can then occure: convectional, cyclonic or orographic. The tropics only experiences convectional rainfall which is characterised by short and intense rainfall. Months with more then 50 mm of rain are sufficiently rainy for crops to grow without irrigation.

10 Dictionary

- **Convergence**: Difference in wind speed forces air to "pile up", which creates a vertical movemnet upwards or downwards depending if the convergence is on the surface or up in the atmosphere.
- **Confluence**: Difference in wind speed gets air together but as wind enters the confluence zone, it speeds up, hence does not converge.
- **Baroclinic**: A atmosphere for which the density depends on both temperature and pressure.
- **Barotropic**: A atmosphere for which the density only depends on pressure.
- **Wind shear**: Change in wind speed and/or direction over a short distance in the atmosphere. It can be both horisontaly or verticaly and usually observed close to weather fronts or thermal winds.
- **Trough (dal)**: Extended region of relatively low atmospheric pressure.
- **Latent heat**: Energy transfer to the atmosphere due to evaporation.
- **Relative humidity**: ratio of amount of water present it the atmosphere realtive to the amount that could be present at that given temperature.
- **Dew formation(dagg)**: Condensation on cool surfaces.
- **Anabatic**: Upward wind motion.
- **Adiabatic**: Any change in internal energy only depends on work such as compression or expansion. Opposite is **diabatic** which is a change in energy between a system and its surrounding due to a temperature gradient.
- **Sensitivity**: An evaluation of how much each input contributes to the model output uncertainty. This can be done in numerous of ways, linear regression being one or running the model and changing one variable at a time.
- **Squall lines** is a linear system of many thunderstorms or clouds that behave as one and therefore can live for 13-15 h instead of just a couple of hourse. They can be hundreds of km long and 30 km wide and are characterised by their explosive growth, rapid propagation and their convex leading ege(?). They peak in the afternoon and usually form west of mountains which implies that surface heating and orographic effects creates them

11 Comparing data to CMIP5

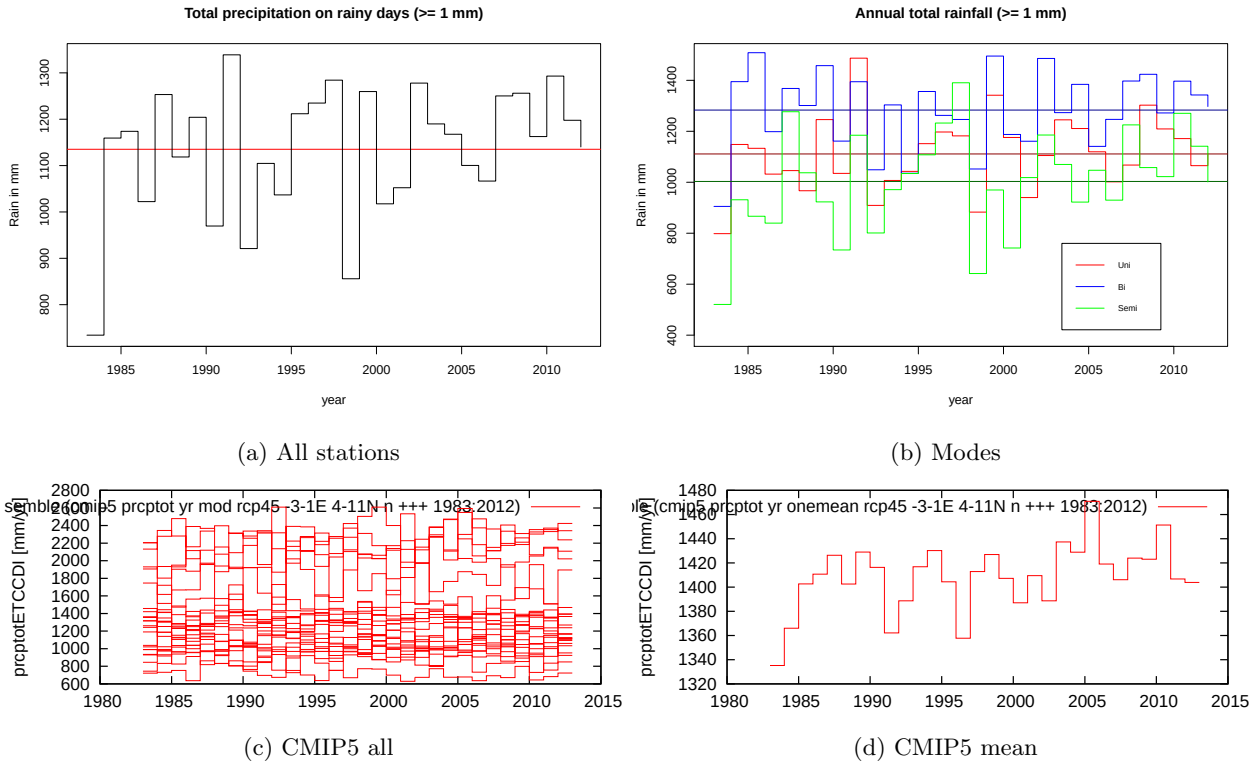
11.1 Rain over a threshold

Studying figure 32, 33, 34 and 35 we can see a very clear pattern in the difference between our data and the GCM (Global Climate Models). The GCMs heavily over estimates the number of rainy days which also leads to a vast over estimation of the annual total rainfall (figure 32, 33). But the span among the GCMs is

massive, ranging from 600 mm up to 2800 mm per year, wheras the range when splitting the data into modes is only between 500 and 1500 mm per year. So it appears that at least a few of the models simulates in the correct range, but looking at the CMIP5 mean it is evient that the majority of the GCMs simulates a much to high annual rainfall. However, the behaviour of the data curve and the CMIP5 mean curve is similar, which could mean that the GCMs can correctly simulate the changes between years even if they cannot correctly simulate number of rainy days.

Looking at figure 33, we can see that all models simulate more rainy days than our data, which leads to the mean curve to be about 100 days per year shifted compare to the data curve. But the range between the different GCMs is once a gain large, ranging from 100-300 days. Looking at the CMIP5 mean curve, we can see that the models are not as good at simulating the variation in rainy days between years as they are at simulating changes in rain amount.

When looking at heavier rainfall, the second known issue becomes clear. The GCMs can simulate the number of days with $\geq$ 10 mm fairly well, the curve is slightly higher then our data but not completly out of range, wheras for $\geq$ 20 mm, they simulate too few days. This is a well know problem, the over simulation of rainy days and the lack of skill to simulate days with very heavy rainfall. Another interesting thing to notice is that CMIP5 mean curve seems to have a uppward pointing trend for both $\geq$ 10mm and $\geq$ 20 mm which is not clear in our data. This is a big issue if we want to use these models to predict future behaviour or precipitation.



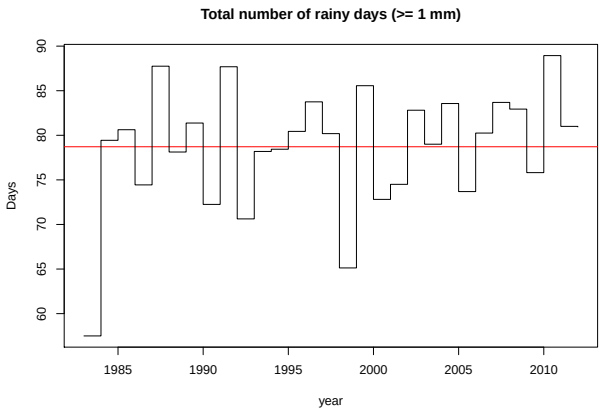
(a) All stations

(b) Modes

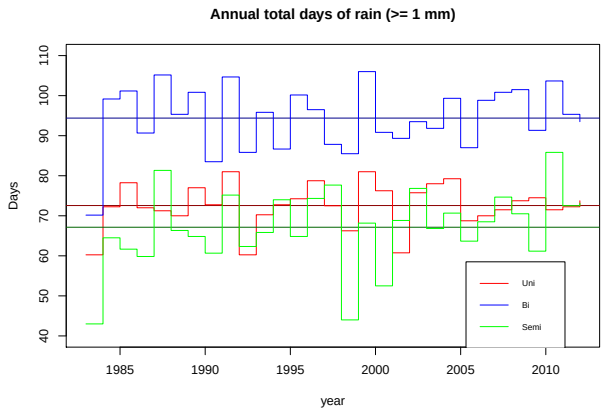
(c) CMIP5 all

(d) CMIP5 mean

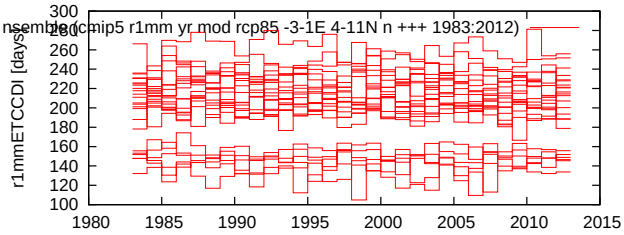
Figure 32: Total annual precipitation ($\geq$1 mm)



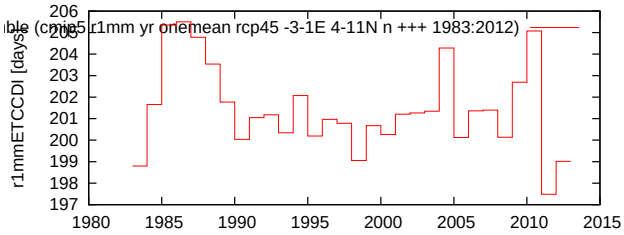
(a) All stations



(b) Modes

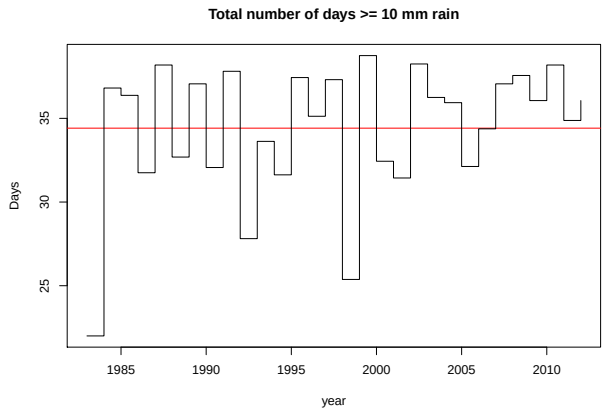


(c) CMIP5 all

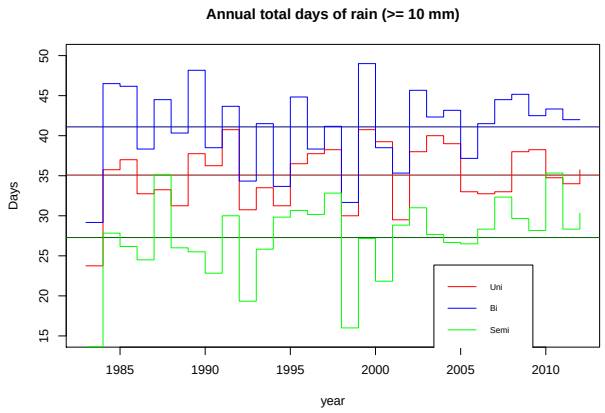


(d) CMIP5 mean

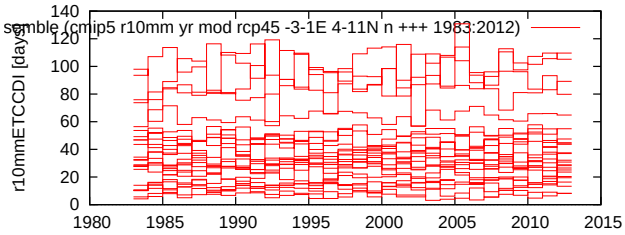
Figure 33: Number of rainy days (≥1 mm)



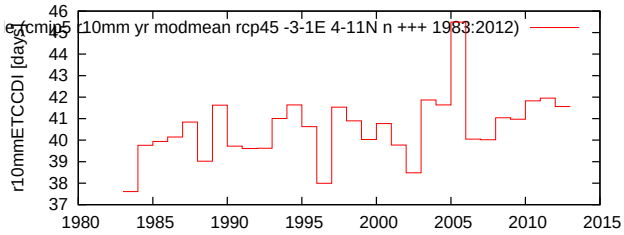
(a) All stations



(b) Modes

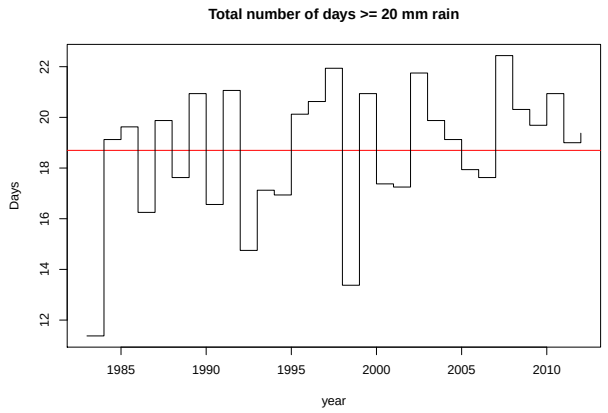


(c) CMIP5 all

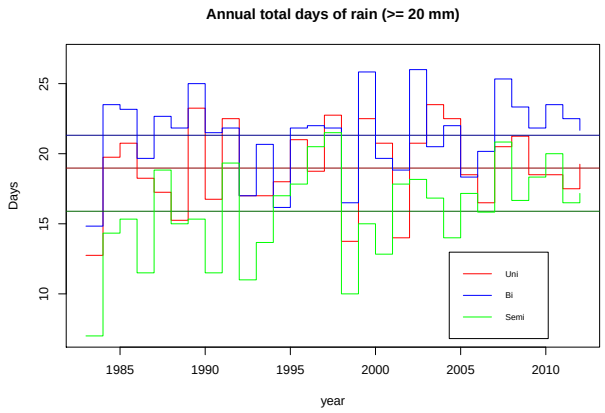


(d) CMIP5 mean

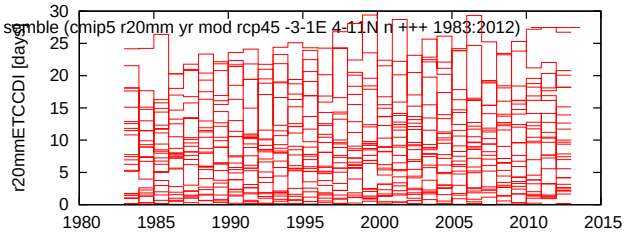
Figure 34: Number of rainy days (≥10 mm)



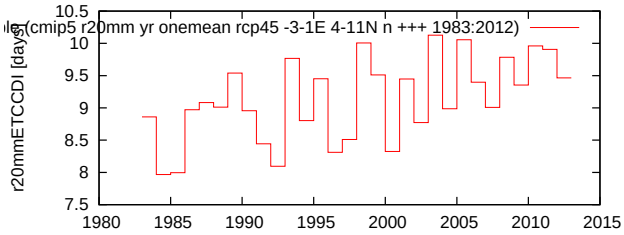
(a) All stations



(b) Modes



(c) CMIP5 all

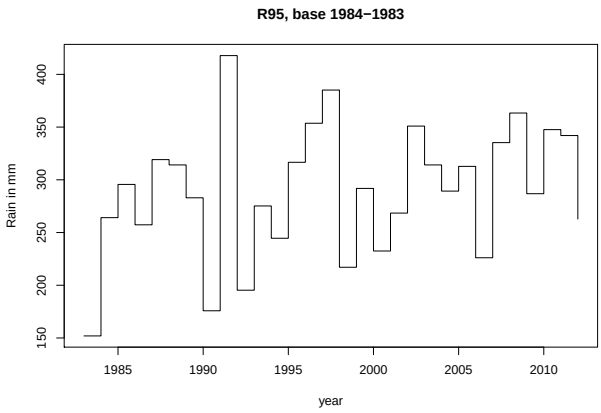


(d) CMIP5 mean

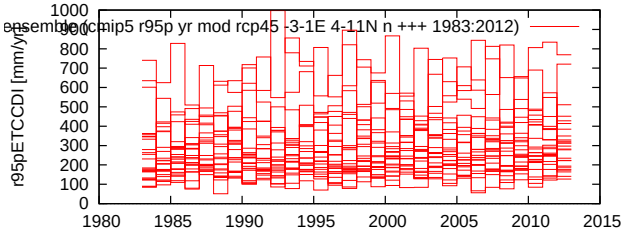
Figure 35: Number of rainy days (≥20 mm)

11.2 Rain quantiles

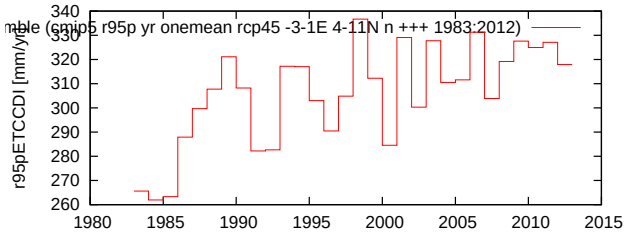
Since CMIP5 uses earlier years then what we have data from as reference period, it is not possible to compare the numbers with each other, but more the behaviour od the curve. 84-93 is picked as a reference period instead of 83-92 to avoid the clearly lower values in -83. In figure 36 both curves seems to exhibit a very similar behaviour, which is a steady increase in the total rainfall on days with heavy rainfall. For days with extreme rainfall (figure 37), the simulated mean is very close to the data mean, but the spread among the models are still very large. The CMIP5 mean is again showing a steady inscrease which is not clearly visible in the data plot. So similar differences can been seen both when looking at the highest percentiles and very heavy rainfall in mm.



(a) All stations

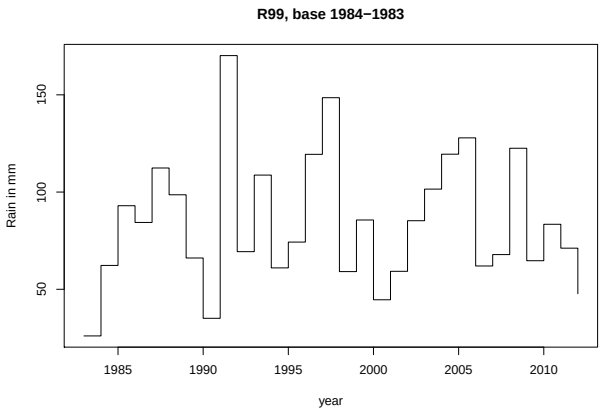


(b) CMIP5 all

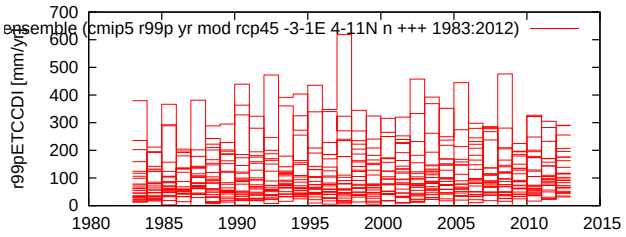


(c) CMIP5 mean

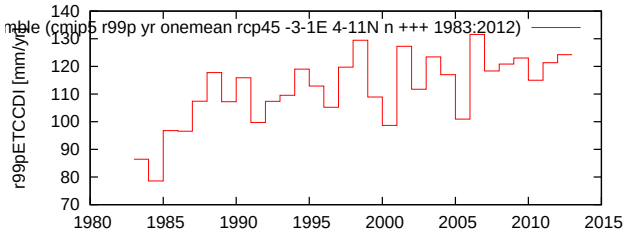
Figure 36: Total rain amount in days above 95% threshold for reference period



(a) All stations



(b) CMIP5 all

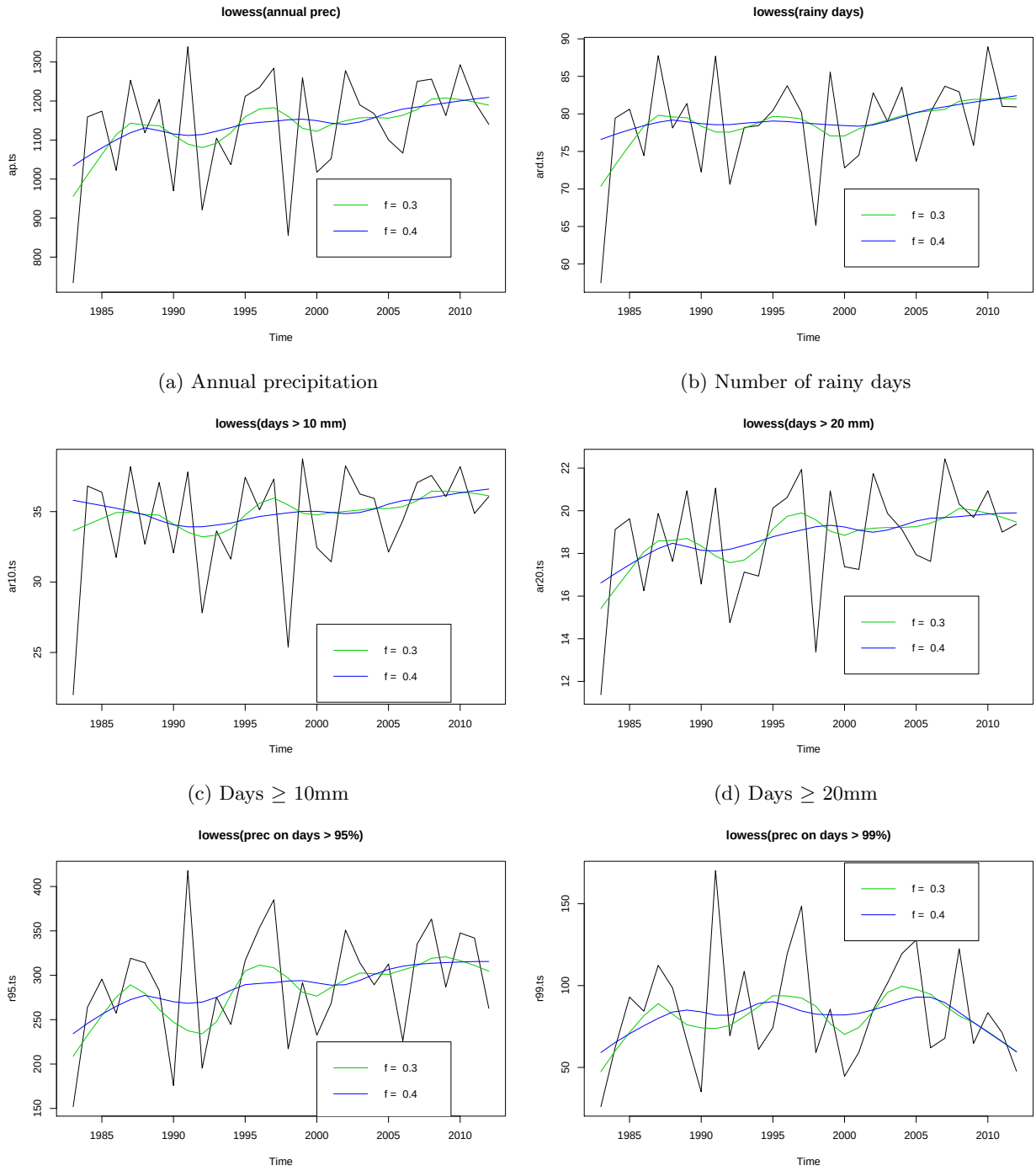


(c) CMIP5 mean

Figure 37: Total rain amount in days above 99% threshold for reference period

12 Trends in timeseries, Lowess

Studying figure 38, it is quite evident that the annual precipitation, very heavy rainfall and the rain amount on R95 (days with more rain then the 95 percentail in the reference period), is increasing over the 30 year period. Number of rainy days is showing a small increase aswell. Number of days with $\geq$ 10 mm is not showing a consistent pattern since it is a small decrease for the first half ot period, to then increase back to the same level as 1983. R99 is showing the opposite pattern, an increase in the amount for the first half of the period to then decrease bac to the 1983 level. This could suggest that there is an increase in the number of rainy days and an increase in days with very hevay rain but a slight decrese in the most extreme rainfalls.



(a) Annual precipitation

(b) Number of rainy days

(c) Days $\geq$ 10mm

(d) Days $\geq$ 20mm

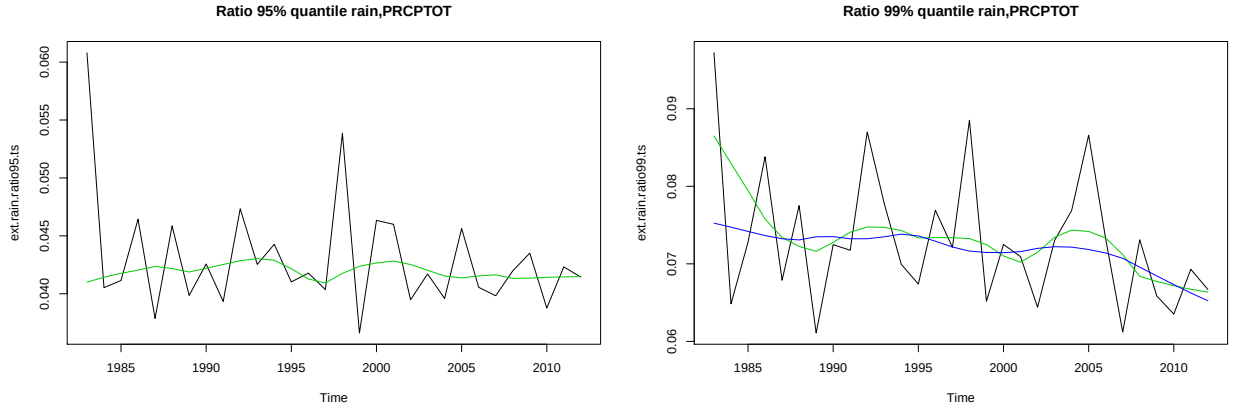
(e) Days $\geq$ 95%

(f) Days $\geq$ 99%

Figure 38: Smooting using lowess

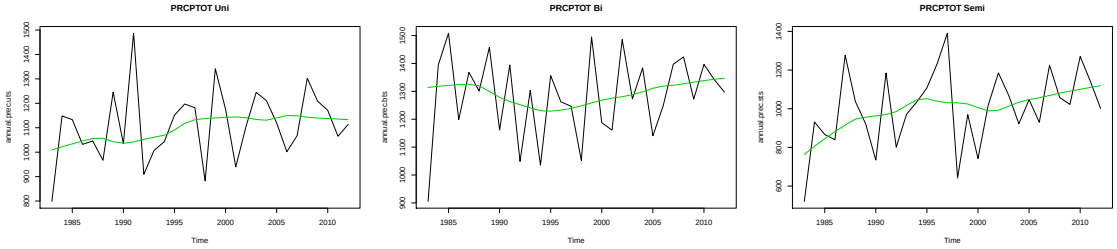
If we leave the changes in extreme indicies and instead look at some extreme values for each year, figure 39 shows that the ratio of very heavy rainfall ($\geq$ 95% quantile) is more or less constant with large flucations around -98. The extreme rainfall ratio is instead showing a downward trend. Since figure 38 (a) show that annual precipitation is increasing over time, this indicated that the extreme rainfall is either constant or not increasing in the same tempo as the total annual rainfall. Autocorrelation plot is insignificant for both time

series.



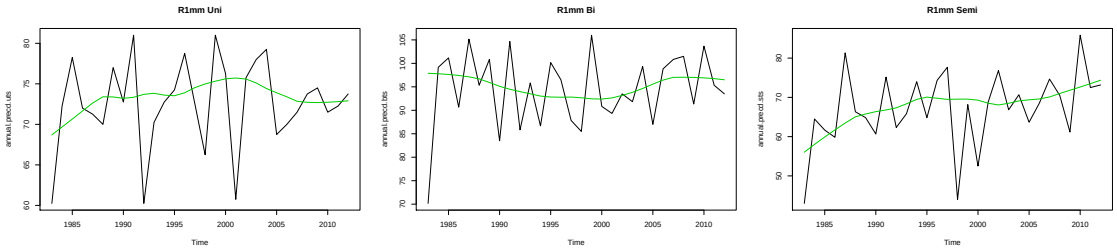
(a) 95% quantile (b) 99% quantile

Figure 39: Ratio extreme rainfall to total annual rainfall



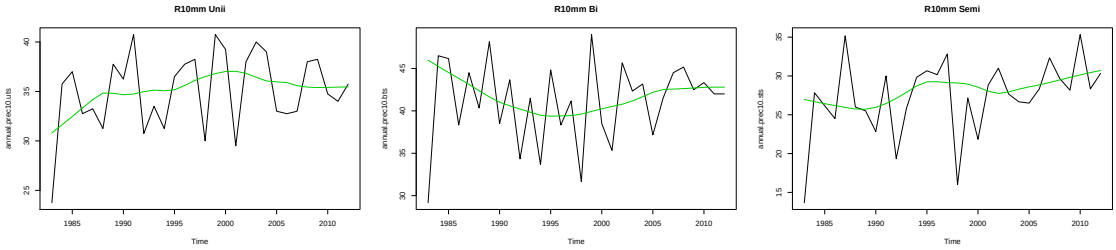
(a) Uni (b) Bi (c) Semi

Figure 40: Total annual precipitation



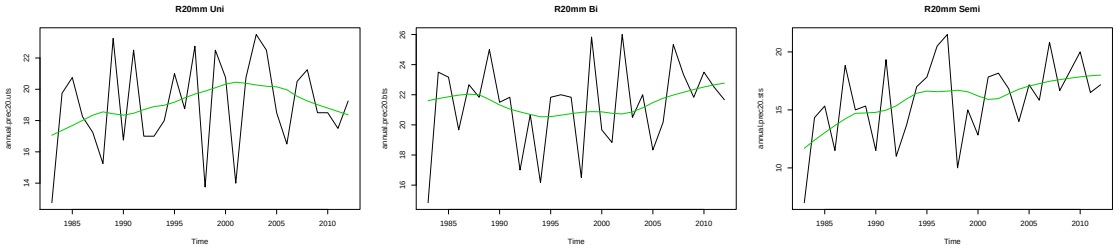
(a) Uni (b) Bi (c) Semi

Figure 41: Total number of rainy days



(a) Uni (b) Bi (c) Semi

Figure 42: Total number of days $\geq$ 10 mm



(a) Uni (b) Bi (c) Semi

Figure 43: Total number of days $\geq$ 20 mm

13 Autocorrelation

Fot the threshold based inidicies there is a weak neagative autocorrelation between consecutive years, with R10 being close to significant. For the quantile based indicies there is instead a positive autocorrelation with a 6 year lag (el niño?).

References

[1] Embrechts, Paul, Klüpperberg, Claudia, Mikosh, Thomas(1997) *Modelling extreme events for Insurance and finance*, New york: Springer-Verlag Berlin Heidelberg

[2] Food and agriculture organisation of the United nations(2005) *Irrigations in Africa in figures - AQUASTAT survey 2005*, Available at: http://www.fao.org/nr/water/aquastat/countries_regions/GHA/GHA-CP_eng.pdf [Accessed 23 Feb. 2018]

[3] Available at: https://onlinecourses.science.psu.edu/stat414/node/319 [Accessed 28 Feb. 2018]

[4] Nkrumah, F., et al. (2014) *Rainfall Variability over Ghana: Model versus Rain Gauge Observation.* International Journal of Geosciences, 5, 673-683. http://dx.doi.org/10.4236/ijg.2014.57060

[5] Open Journal of Modern Hydrology, 2011, 1, 11-22 doi:10.4236/ojmh.2011.12002 Published Online October 2011 (http://www.SciRP.org/journal/ojmh)

[6] Stern, R., & Coe, R. (1984). *A Model Fitting Analysis of Daily Rainfall Data.* Journal of the Royal Statistical Society. Series A (General), 147(1), 1-34. doi:10.2307/2981736

[7] Jamaludin, Shariffah & Aziz Jemain, Abdul. (2007). *Fitting daily rainfall amount in Peninsular Malaysia using several types of Exponential distributions.* Journal of Applied Sciences Research. 3.

[8] Sanso, B., & Guenni, L. (1999). *Venezuelan Rainfall Data Analysed by Using a Bayesian Space-Time Model*. Journal of the Royal Statistical Society. Series C (Applied Statistics), 48(3), 345-362. Retrieved from http://www.jstor.org/stable/2680829

References

[1] Embrechts, Paul, Klüpperberg, Claudia, Mikosh, Thomas(1997) *Modelling extreme events for Insurance and finance*, New york: Springer-Verlag Berlin Heidelberg

[2] Food and agriculture organisation of the United nations(2005) *Irrigations in Africa in figures - AQUASTAT survey 2005*, Available at: http://www.fao.org/nr/water/aquastat/countries_regions/GHA/GHA-CP_eng.pdf [Accessed 23 Feb. 2018]

[3] Available at: https://onlinecourses.science.psu.edu/stat414/node/319 [Accessed 28 Feb. 2018]

[4] Nkrumah, F., et al. (2014) *Rainfall Variability over Ghana: Model versus Rain Gauge Observation.* International Journal of Geosciences, 5, 673-683. http://dx.doi.org/10.4236/ijg.2014.57060

[5] Open Journal of Modern Hydrology, 2011, 1, 11-22 doi:10.4236/ojmh.2011.12002 Published Online October 2011 (http://www.SciRP.org/journal/ojmh)

[6] Stern, R., & Coe, R. (1984). *A Model Fitting Analysis of Daily Rainfall Data.* Journal of the Royal Statistical Society. Series A (General), 147(1), 1-34. doi:10.2307/2981736

[7] Jamaludin, Shariffah & Aziz Jemain, Abdul. (2007). *Fitting daily rainfall amount in Peninsular Malaysia using several types of Exponential distributions.* Journal of Applied Sciences Research. 3.

[8] Sanso, B., & Guenni, L. (1999). *Venezuelan Rainfall Data Analysed by Using a Bayesian Space-Time Model*. Journal of the Royal Statistical Society. Series C (Applied Statistics), 48(3), 345-362. Retrieved from http://www.jstor.org/stable/2680829

[9] Deidda, R.: *A multiple threshold method for fitting the generalized Pareto distribution to rainfall time series*, Hydrol. Earth Syst. Sci., 14, 2559-2575, https://doi.org/10.5194/hess-14-2559-2010, 2010.

[10] Li, C., V. P. Singh, and A. K. Mishra (2012), *Simulation of the entire range of daily precipitation using a hybrid probability distribution*, Water Resour. Res., 48, W03521, doi:10.1029/2011WR011446

[11] http://blog.minitab.com/blog/statistics-and-quality-data-analysis/large-samples-too-much-of-a-good-thing [Accessed 14 March 2018]

[12] https://onlinecourses.science.psu.edu/stat414/node/322