Advanced SWAT Course Notes: Basic Rainfall Analysis of INUMET data

Willem Vervoort & Flora Mer

21-11-2017

Introduction

This is part of a series of teaching documents related to the "How do I use satellite and global reanalysis data for hydrological simulations in SWAT?" jointly organised by the University of Sydney, IRI (the University of Columbia) and INIA, Uruguay.

This specific part explains a basic rainfall analysis of INUMET data. It is a basic comparison of INUMET Stations in the Santa Lucia subcatchment.

Load INUMET Weather Station data

This part will show how we load the INUMET Weather Station data and how we select the stations with enough data to realize the comparison analysis.

Packages for the script

The following packages are needed for the all script.

```
# Some useful packages to install
library(tidyverse)
library(zoo)
library(rgeos)
library(sp)
library(rgdal)
library(lubridate)
```

Download INUMET data and selection

The following script will show :

- how to load the INUMET rainfall data
- how to select the stations with good amount of data
- how to plot the localization of the INUMET rainfall stations in the Santa Lucia subcatchment

```
# Read in the INUMET station locations
Stations <- read.csv("rainfall stations-INUMET-subcuencasantalucia.csv")

# Read in the precipitation data for each INUMET stations
Pdata <- read.csv("Precipitacion SantaLucia_inumet_stations_prcp.csv",
    na.strings = "NaN")


Dates2 <- as.Date(paste(Pdata[, 1], Pdata[, 2], Pdata[, 3], sep = "-"))

# Find the INUMET stations that have 90% of data after 2001
Pdata_2000 <- Pdata[Dates2 >= "2000-01-01", ]
Pdata_2000 <- Pdata_2000[-(1:5), ]
result <- rep(0, (ncol(Pdata_2000) - 3))
for (i in 4:ncol(Pdata_2000)) {
    result[i - 3] <- sum(ifelse(is.na(Pdata_2000[, i]), 1, 0))/nrow(Pdata_2000)
}

# result indicates the fraction of NA data for the stations
# Throw out all the columns and rows where result >0.1
Pdata_new <- Pdata_2000[, -(which(result > 0.1) + 3)]
Dates2000 <- Dates2[Dates2 >= as.Date("2000-01-01")]


Pdata_z <- zoo(Pdata_new[, 4:ncol(Pdata_new)], order.by = Dates2000,
    frequency = 1)
head(Pdata_z)
```

```
##            X2673 X2826 X2819 X2748 X2725 X2715 X2683 X2680 X2670 X2498
## 2000-01-01     0     0     0     0     0     0     0     0     0     0
## 2000-01-02     0     0     0     0     0     0     0     0     0     0
## 2000-01-03     0     0     0     0     0     0     0     0     0     0
## 2000-01-04     0     0     0     0     0     0     0     0     0     0
## 2000-01-05     0     0     0     0     0     0     0     0     0     0
## 2000-01-06     0     0     0     0     0     0     0     0     0     0
##            X2549 X86545 X2588 X2632
## 2000-01-01     0      0     0     0
## 2000-01-02     0      0     0     0
## 2000-01-03     0      0     0     0
## 2000-01-04     0      0     0     0
## 2000-01-05     0      0     0     0
## 2000-01-06     0      0     0     0
```

```
# read in the shapefile
SL <- readOGR("sl_shape/subcuencaSantaLuciahastariostaluciachico.shp")
```

```
## OGR data source with driver: ESRI Shapefile
## Source: "sl_shape/subcuencaSantaLuciahastariostaluciachico.shp", layer: "subcuencaSantaLuciahastarios
## with 1 features
## It has 1 fields
```

```
map <- ggplot() + geom_polygon(data = SL, aes(x = long, y = lat,
    group = group), colour = "black", fill = NA)
```

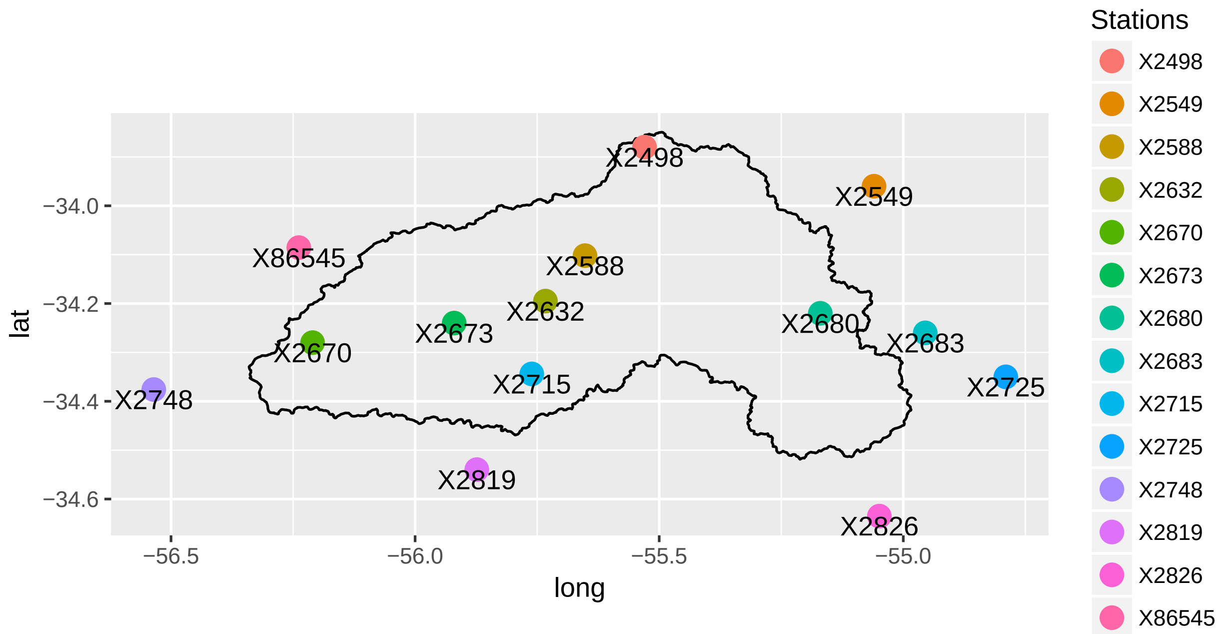
```
## Regions defined for each Polygons
```

```
# station Df
Stat_df <- data.frame(Stations = colnames(Pdata_z), Long = Stations$POINT_X,
    Lat = Stations$POINT_Y)

pl <- ggplot() + geom_polygon(data = SL, aes(x = long, y = lat,
    group = group), colour = "black", fill = NA) + coord_equal() +
    geom_point(data = Stat_df, aes(Long, Lat, colour = Stations),
        size = 4) + geom_text(data = Stat_df, aes(Long, Lat,
    label = Stations, vjust = 1))
```

```
## Regions defined for each Polygons
```

```
pl
```



After selecting the INUMET Stations with good amount of datat (>90% of data available), we plotted the spatial map with the location of the INUMET Stations within and close to the Santa Lucia subcatchment.

Calculate basic statistics based on daily and annual step

We will do some basic statistics to understand better the rainfall pattern accross all INUMET statioons.

Log mean daily rainfall

First, we will compare the mean daily rainfall of the INUMET stations. More specifically, we will calculate the logarithm of the daily mean for each INUMET station. The R script is the following:

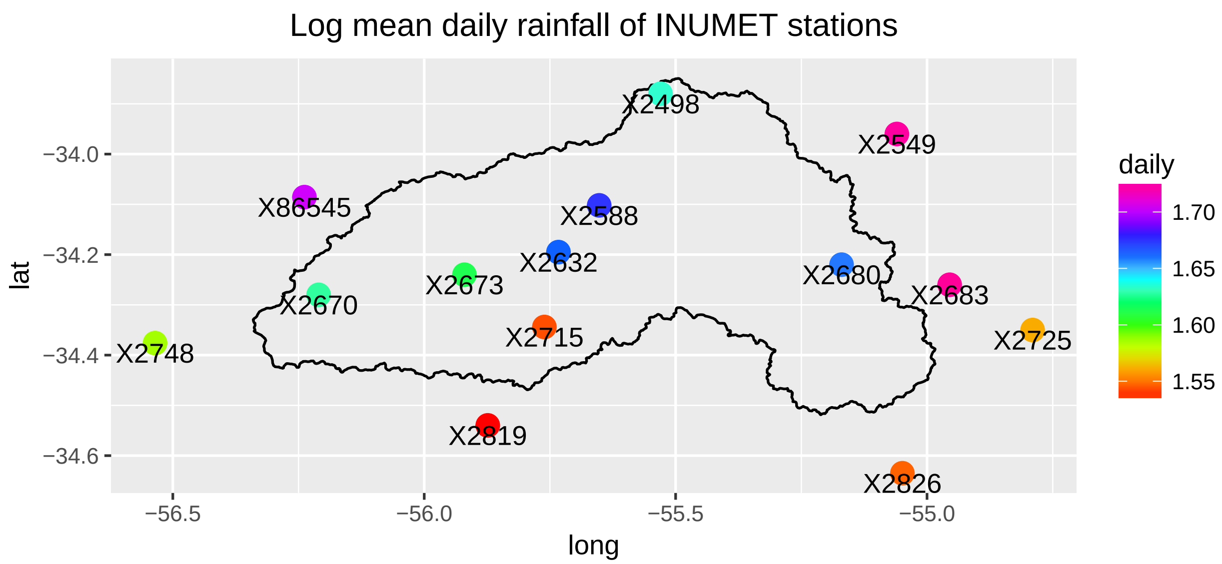
```
# log mean daily rainfall
mean_d_P <- apply(Pdata_z, 2, function(x) exp(mean(log(x + 1),
    na.rm = T)))

mean_df <- Stat_df
# adding a column to the dataframe mean_df with the mean
# values
```

```
mean_df$daily <- mean_d_P
# spatial plot of the mean values of INUMET stations
pl <- ggplot() + geom_polygon(data = SL, aes(x = long, y = lat,
    group = group), colour = "black", fill = NA) + coord_equal() +
    geom_point(data = mean_df, aes(Long, Lat, colour = daily),
        size = 4) + geom_text(data = mean_df, aes(Long, Lat,
    label = Stations, vjust = 1)) + scale_colour_gradientn(colors = rainbow(10)) +
    ggtitle("Log mean daily rainfall of INUMET stations") + theme(plot.title = element_text(hjust = 0.5
```

```
## Regions defined for each Polygons
```

```
pl
```



From the map "Log mean daily rainfall of INUMET stations", we can see there is a small increasing gradient from south to north regarding the mean daily rainfall. Indeed, INUMET stations located in the north of the subcatchment have higher mean daily values of rainfall (x2549, x86545) compare to INUMET stations located in the south of the subcatchment (X2819, X2826, X2715). However, the mean daily rainfall variation does not seem very high, it varies from a bit less than 1,55 to a bit more than 1,70.

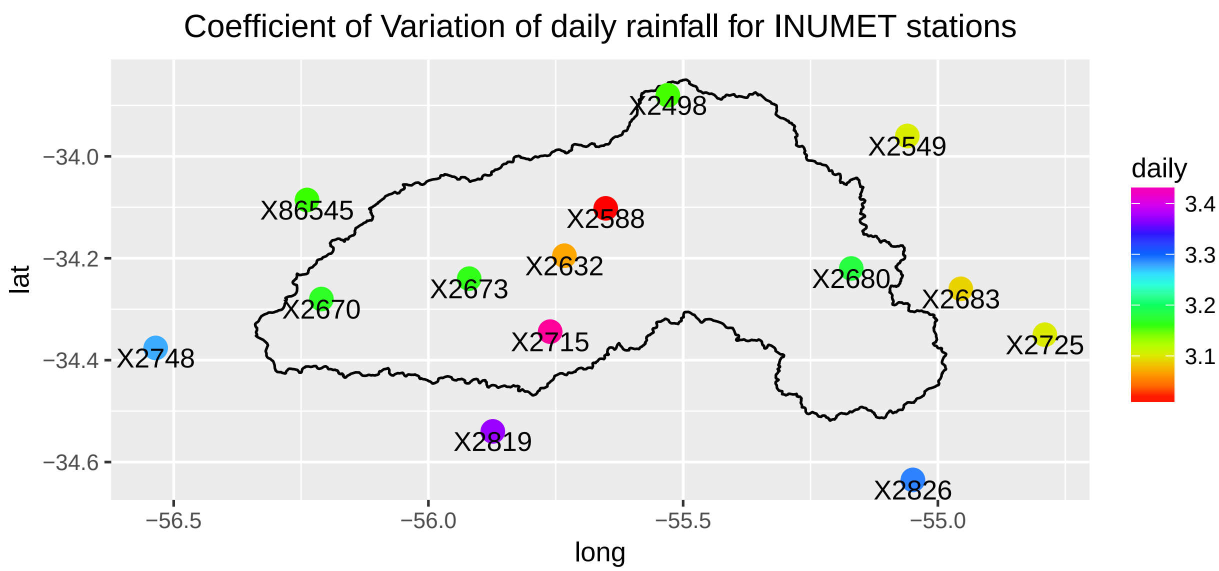
Coefficient of variation of daily rainfall

Now we will have a look at the coefficient of variation (CV) of daily rainfall of each INUMET stations. The CV is defined as the ratio of the standard deviation (sd) to the mean. In this case, it shows the extent of variability in relation to the mean of the daily rainfall.

```r
# CV daily rainfall
CV_d_P <- apply(Pdata_z, 2, function(x) sd(x, na.rm = T)/mean(x,
    na.rm = T))
CV_df <- Stat_df
# adding a column to the dataframe CV_df with the CV values
CV_df$daily <- CV_d_P
# spatial plot of the CV of daily rainfall of INUMET stations
pl <- ggplot() + geom_polygon(data = SL, aes(x = long, y = lat,
    group = group), colour = "black", fill = NA) + coord_equal() +
    geom_point(data = CV_df, aes(Long, Lat, colour = daily),
        size = 4) + geom_text(data = CV_df, aes(Long, Lat, label = Stations,
    vjust = 1)) + scale_colour_gradientn(colors = rainbow(10)) +
    ggtitle("Coefficient of Variation of daily rainfall for INUMET stations") +
    theme(plot.title = element_text(hjust = 0.5))
```

```
## Regions defined for each Polygons
```

```r
pl
```



Contrary to the mean daily rainfall pattern, south INUMET Stations (ex: X2819, X2826, x2748) show a higher variability pattern of daily rainfall than north INUMET Stations (x2588, x2498). We can see the variability slightly increase from north to south in the subcatchment. However, the variability interval for all INUMET stations is not very different.

Mean annual rainfall

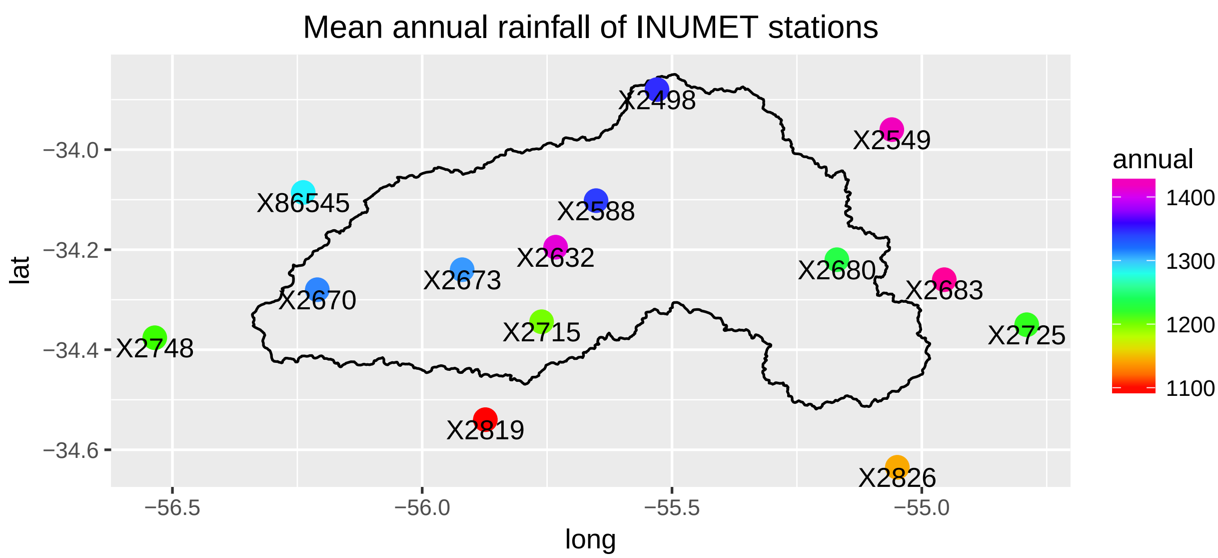
After analyzing daily rainfall, we will compare the mean annual rainfall data of INUMET stations. The script R is the following:

```
# mean annual rainfall
annual <- aggregate(Pdata_z, list(year = format(time(Pdata_z),
    "%Y")), sum, na.rm = T)
mean_a_P <- apply(annual, 2, mean, na.rm = T)
# adding a column to the dataframe mean_df with the mean
# annual values
mean_df$annual <- mean_a_P
# spatial plot of the mean annual rainfall of INUMET stations
pl <- ggplot() + geom_polygon(data = SL, aes(x = long, y = lat,
    group = group), colour = "black", fill = NA) + coord_equal() +
```

```r
    geom_point(data = mean_df, aes(Long, Lat, colour = annual),
        size = 4) + geom_text(data = mean_df, aes(Long, Lat,
    label = Stations, vjust = 1)) + scale_colour_gradientn(colors = rainbow(10)) +
    ggtitle("Mean annual rainfall of INUMET stations") + theme(plot.title = element_text(hjust = 0.5))
```

```
## Regions defined for each Polygons
```

```r
pl
```



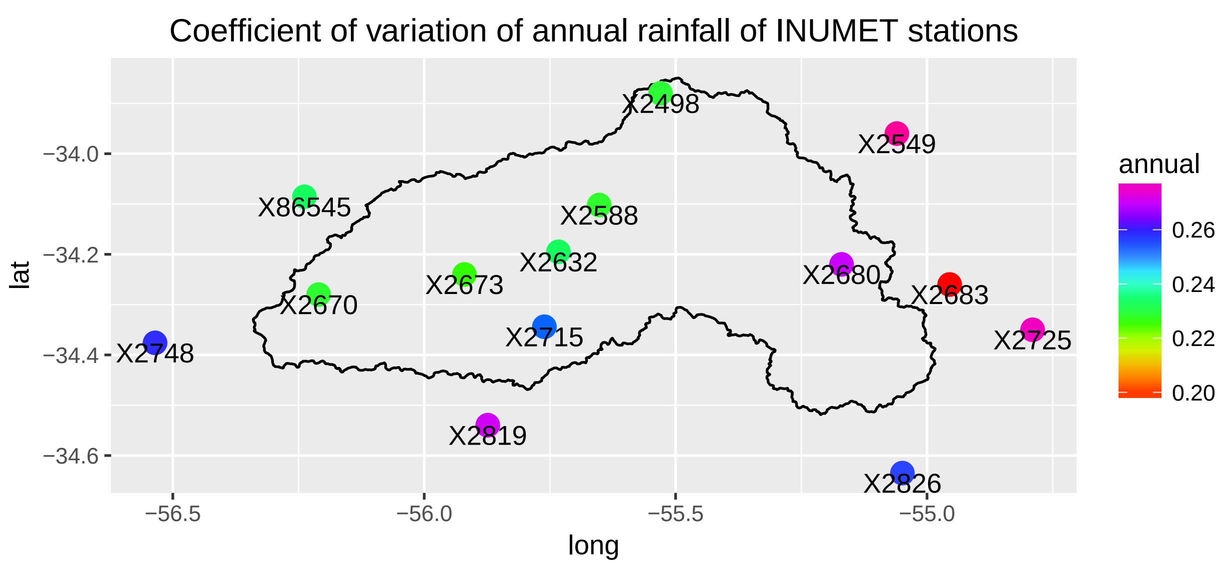
The mean annual rainfall map also show a higher value of annual rainfall in the north than in the south of the subcatchment, such as the log mean daily rainfall map. It confirms our previous results.
Coefficient of variation of annual rainfall We will have a look at the coefficient of variation (CV) of annual rainfall. The definition of the coefficient of variation was described in the section "Coefficient of variation of daily rainfall". The script R is the following:

```r
# CV annual rainfall
CV_a_P <- apply(annual, 2, function(x) sd(x)/mean(x))
# adding a column to the dataframe CV_df with the CV annual
# rainfall values
CV_df$annual <- CV_a_P
# spatial plot of the CV of annual rainfall of INUMET
```

```r
# stations
pl <- ggplot() + geom_polygon(data = SL, aes(x = long, y = lat,
    group = group), colour = "black", fill = NA) + coord_equal() +
    geom_point(data = CV_df, aes(Long, Lat, colour = annual),
        size = 4) + geom_text(data = CV_df, aes(Long, Lat, label = Stations,
    vjust = 1)) + scale_colour_gradientn(colors = rainbow(10)) +
    ggtitle("Coefficient of variation of annual rainfall of INUMET stations") +
    theme(plot.title = element_text(hjust = 0.5))
```

```
## Regions defined for each Polygons
```

```r
pl
```



The CV annual rainfall map also show a higher variability of annual rainfall in the south than in the north of the subcatchment, such as the CV daily rainfall map. It confirms our previous results.

Calculate basic statistics based on seasonality

After analyzing the rainfall data at daily and annual step, we will do some statistic calculation at seasonal step.

Function for seasonal analysis

The following R script define the function with the division of the 4 seasons.

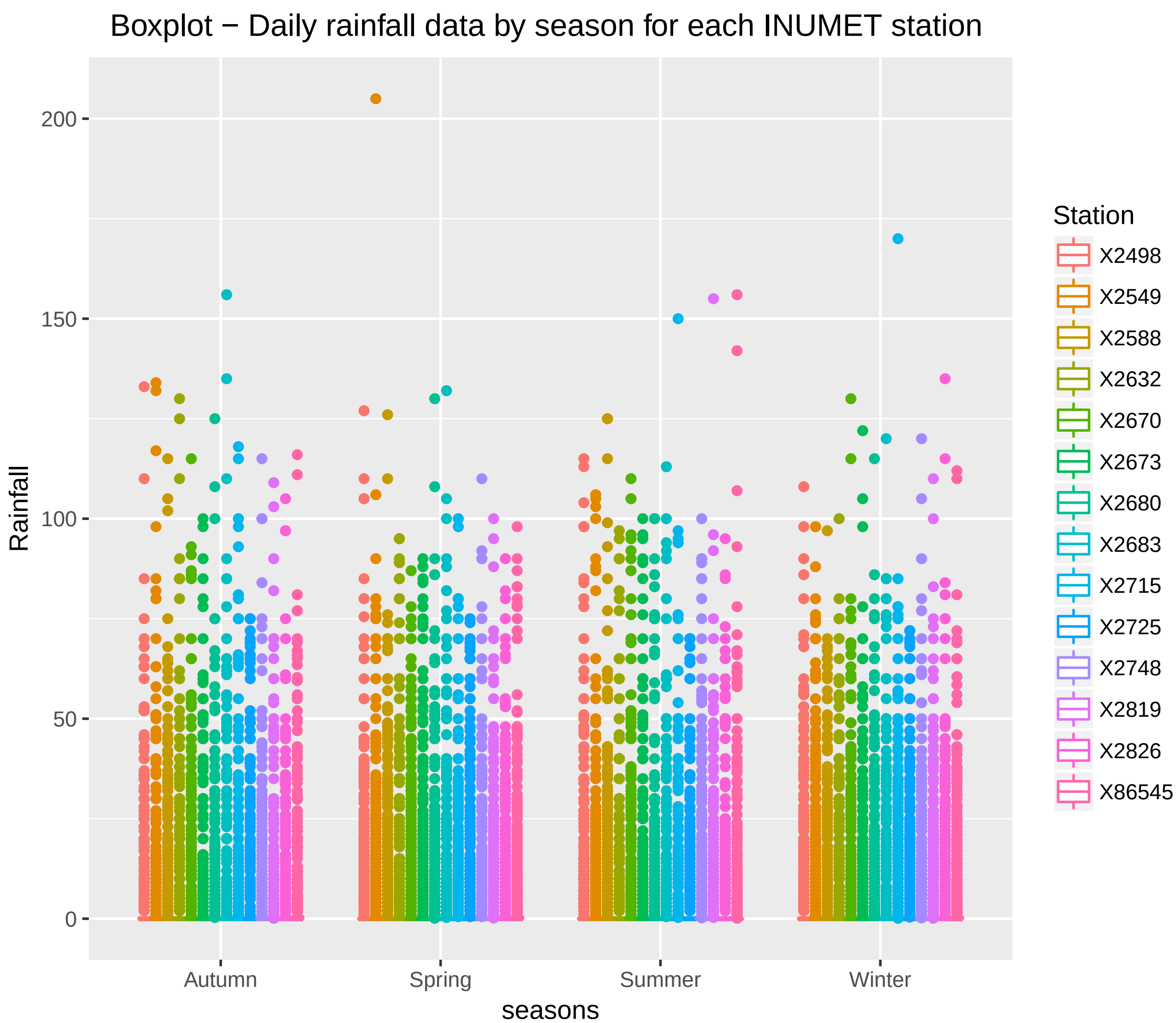
```
# Seasonal analysis
seasons <- sapply(time(Pdata_z), function(x) ifelse(month(x) >
    11 || month(x) < 3, "Summer", ifelse(month(x) < 5, "Autumn",
    ifelse(month(x) < 8, "Winter", "Spring"))))
```

Seasonal analysis based on daily data

We will first have a look at seasonal analysis based on daily data. The R script is the following:

```
# based on daily data
Pseason <- data.frame(coredata(Pdata_z), seasons)
Pseason_pl <- gather(Pseason, key = Station, value = Rainfall,
    X2673:X2632)
# boxplot of daily rainfall data for each season and each
# INUMET station
Sp <- ggplot(Pseason_pl, aes(seasons, Rainfall)) + geom_boxplot(aes(colour = Station)) +
    ggtitle("Boxplot - Daily rainfall data by season for each INUMET station ") +
    theme(plot.title = element_text(hjust = 0.5))
Sp
```

```
## Warning: Removed 390 rows containing non-finite values (stat_boxplot).
```

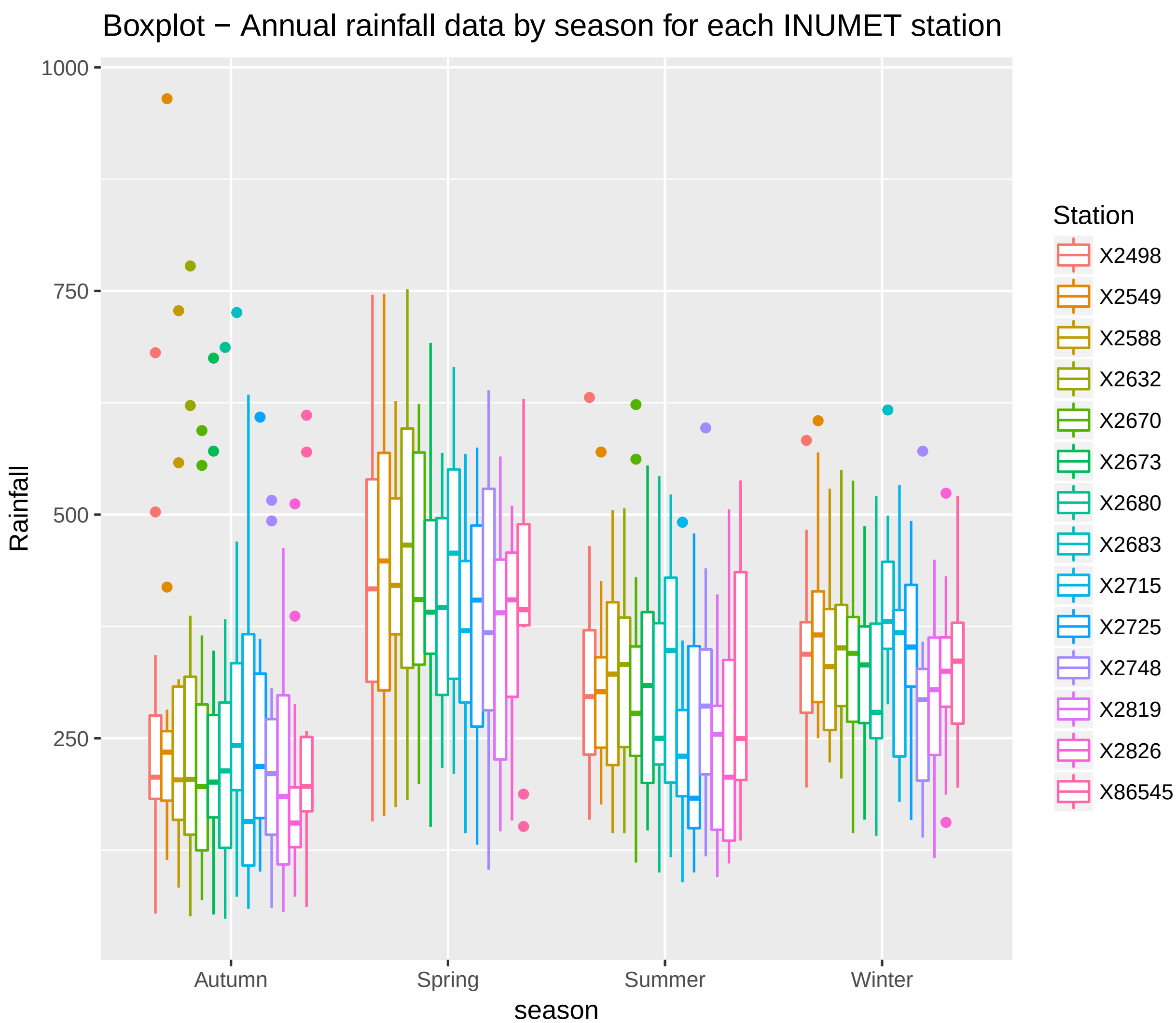


We can't see well the boxplot as there is a lot of variability within the daily rainfall data.

Summarise annual rainfall by season

To improve the result of the above analysis, we will plot the annual rainfall for each INUMET stations.

```
# summarise annual rainfall by season
season_P <- aggregate(coredata(Pdata_z), list(year = format(time(Pdata_z),
    "%Y"), season = seasons), sum, na.rm = T)
Pseason_pl <- gather(season_P, key = Station, value = Rainfall,
    X2673:X2632)
# boxplot
Sp <- ggplot(Pseason_pl, aes(season, Rainfall)) + geom_boxplot(aes(colour = Station)) +
    ggtitle("Boxplot - Annual rainfall data by season for each INUMET station ") +
    theme(plot.title = element_text(hjust = 0.5))
Sp
```



The boxplot shows we have more rainfall during spring but also more variability regarding annual rainfall data. Autumn is the season with less rainfall where the variability is lower.

Calculate mean by season

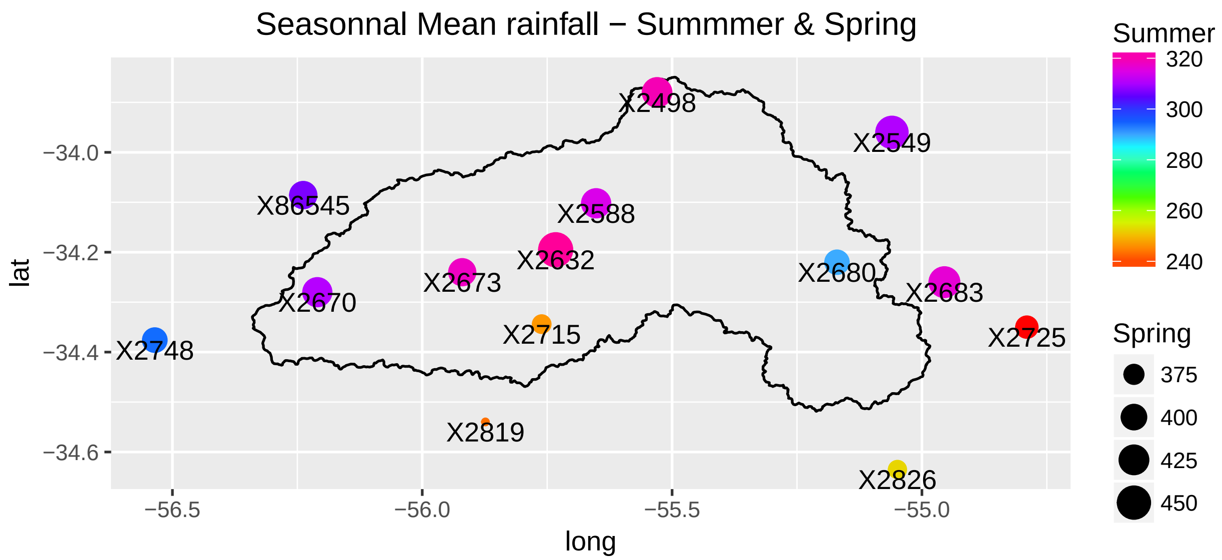
Another seasonal analysis can be done by calculating the mean rainfall by season. The R script is the following:

```
# calculate mean by season
mean_s_P <- aggregate(season_P[, 3:16], list(season = season_P$season),
    mean)
for (i in 1:nrow(mean_s_P)) {
    mean_df[, i + 5] <- t(mean_s_P[i, 2:15])
    colnames(mean_df)[i + 5] <- mean_s_P$season[i]
}
# Spatial plot of the mean rainfall by season Summer and
# Spring
pl <- ggplot() + geom_polygon(data = SL, aes(x = long, y = lat,
    group = group), colour = "black", fill = NA) + coord_equal() +
    geom_point(data = mean_df, aes(Long, Lat, colour = Summer,
```

```
    size = Spring)) + geom_text(data = mean_df, aes(Long,
Lat, label = Stations, vjust = 1)) + scale_colour_gradientn(colors = rainbow(10)) +
ggtitle("Seasonnal Mean rainfall - Summmer & Spring ") +
theme(plot.title = element_text(hjust = 0.5))
```

```
## Regions defined for each Polygons
```

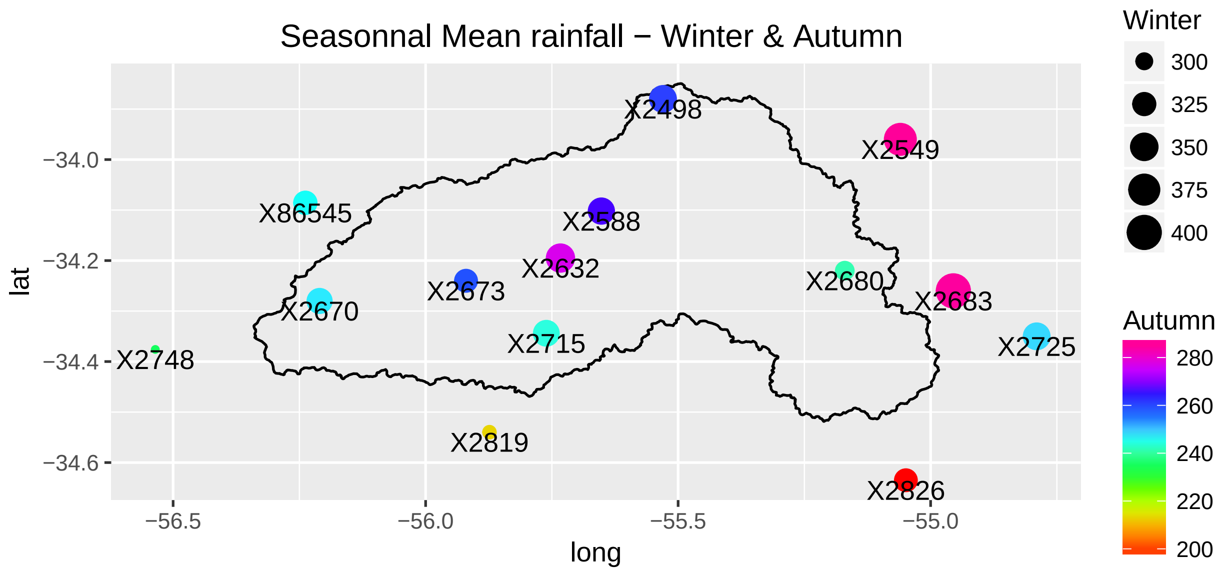
```
pl
```




```
## Autumn & Winter
pl <- ggplot() + geom_polygon(data = SL, aes(x = long, y = lat,
    group = group), colour = "black", fill = NA) + coord_equal() +
    geom_point(data = mean_df, aes(Long, Lat, colour = Autumn,
        size = Winter)) + geom_text(data = mean_df, aes(Long,
    Lat, label = Stations, vjust = 1)) + scale_colour_gradientn(colors = rainbow(10)) +
    ggtitle("Seasonnal Mean rainfall - Winter & Autumn ") + theme(plot.title = element_text(hjust = 0.5
```

```
## Regions defined for each Polygons
```

```
pl
```



During summer and spring, mean rainfall are higher for INUMET stations located in the north of the subcatchment than in the south. As showed previously, the mean rainfall is higher during spring than during summer.
The same pattern is showed during winter and autumn. Rainfall are higher during winter than during autumn.

Calculate double mass curves

The double mass analysis is a common data analysis approach for investigating the behaviour of records made of meteoroligical data at a number of different locations. The R script is the following:

```
# double mass curves set all missing data to 0
Pdata_z_cor <- Pdata_z
Pdata_NA <- Pdata_z
Pdata_NA[] <- 0
for (i in 1:14) {
    Pdata_NA[is.na(Pdata_z_cor[, i]), i] <- 1
    Pdata_z_cor[is.na(Pdata_z_cor[, i]), i] <- 0
}
```

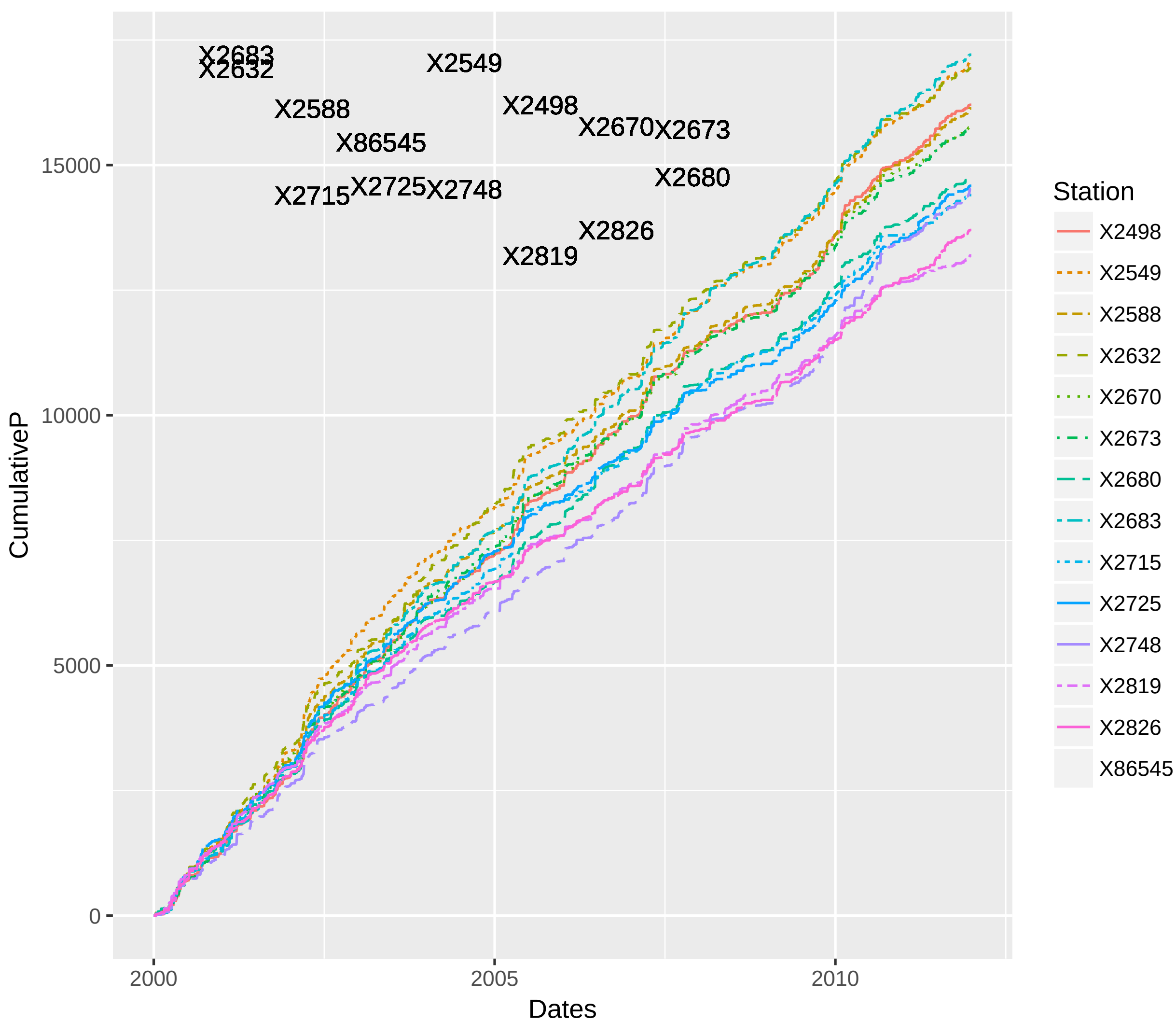
```r
NAcount <- apply(Pdata_NA, 2, sum)

# step 1 calculate cumulative mass curves
Cum_mass <- apply(Pdata_z_cor, 2, cumsum)

# plot the cumulative mass curves
plot_df <- gather(as.data.frame(Cum_mass), key = Station, value = CumulativeP,
    X2673:X2632)
plot_df$Dates <- rep(time(Pdata_z), 14)
maxP <- Cum_mass[nrow(Cum_mass), 1:14]
plot_df$maxP <- rep(maxP, each = nrow(Pdata_z))

pl <- ggplot(plot_df, aes(Dates, CumulativeP)) + geom_line(aes(colour = Station,
    linetype = Station)) + geom_text(aes(as.Date("2004-01-01"),
    maxP, label = Station, hjust = rep(rep(c(-3, -2, -1, 0, 1,
        2, 3), 2), each = nrow(Pdata_z))))

pl
```



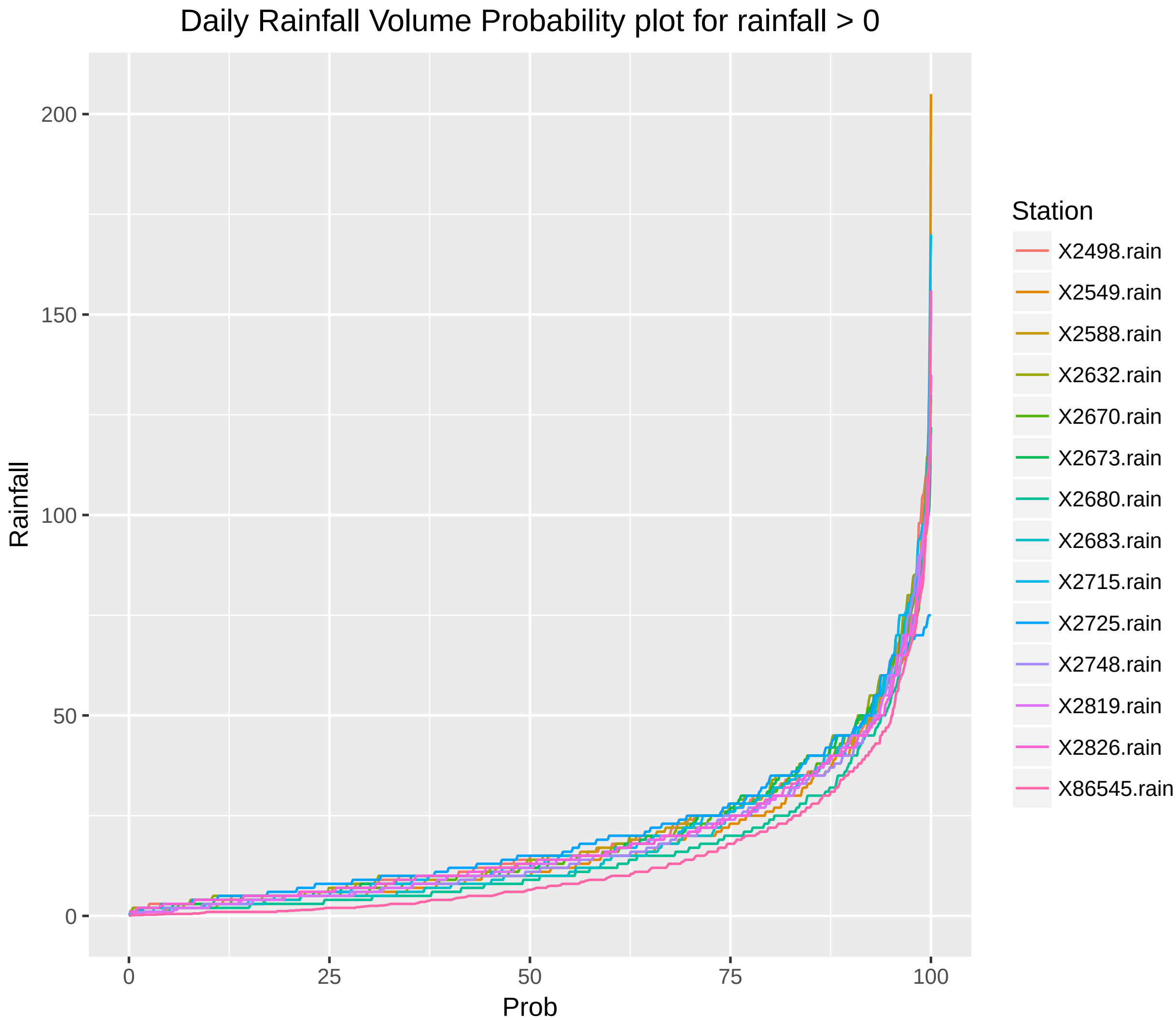
The cumulative mass curves plot shows for each INUMET stations the cumulative of rainfall accross year. It also shows the light increase of rainfall from south to north accross the subcatchment. We can also notice that some mass curves are sticked together which means some stations are very similar.

The following R script is for plotting the frequency curves for rainfall greater than 0.The frequency curves for rainfall give us the probability of daily rain volume when rainfall is greater than 0.

```
# frequency curves for rainfall > 0
FDC_gen <- function(DATA) {
    FDC <- data.frame(probs = seq(0, 1, length = 1000) * 100,
        rain = quantile(DATA[DATA > 0], probs = seq(0, 1, length = 1000),
            na.rm = T))
    return(FDC)
}

Fcurves <- apply(Pdata_z, 2, FDC_gen)
# this is a list with a data.frame for each station
# str(Fcurves)
F_df <- do.call(cbind, Fcurves)
F_df <- F_df[, -seq(3, 27, by = 2)]

F_df_pl <- gather(F_df, value = Rainfall, key = Station, X2673.rain:X2632.rain)
colnames(F_df_pl)[1] <- "Prob"
F_df_pl$Station <- sapply(F_df_pl$Station, function(x) gsub("s.rain",
    "", x))
# plot of frequency curves for rainfall > 0
f_plot <- ggplot(F_df_pl, aes(Prob, Rainfall)) + geom_line(aes(colour = Station)) +
    ggtitle("Daily Rainfall Volume Probability plot for rainfall > 0") +
    theme(plot.title = element_text(hjust = 0.5))
f_plot
```



From the daily rainfall volume probability plot we can observe all INUMET stations are following the same pattern which can indicate the consistency of the rainfall data between stations.