Consulting Report

Te Jung Chen

March 03, 2020

Abstract

Climate change has been a major issue in the world. The streamflow of a watershed at a particular station changes in climatic attributes such as temperatrue, rainfall and snowfall. This study explores the impact of the climatic change by analyzing which variables a watershed's streamflow at a particular location is most sensitive to, quantify each variable's contribution to streamflow change and whether a temporal change in temperature have a greater impact on streamflow than changes in rainfall.

Introduction

The streamflow of a watershed at a particular station changes due to climate change. There are many factors that can impact the streamflow of a watershed such as temperature, rainfall and snowfall. Hence, the goal of this analysis is to determine which factor creates more impact on a watershed's streamflow and quantify each factor's contribution to streamflow change. Moverover, the analysis will be divided into three categories: data cleaning and processing, EDA (exploratory data analysis) and modeling.

For data cleaning and processing, the current year is adjusted to match the client's specification of a year to begin on October 1st and end on September 30th. The years that are missing more than 10 years of data is removed as specified by the client (Table 2). Since the data have a lot of missing values, the data is aggregated (average) by year such that one year corresponds to a single observation. Hence, it can remove the seasonality effect and fix the problem of division by 0 (Snowfall/Total Precipitation).

Data Description

The data consists of 17,990 real daily observations of daily temperature, daily rainfall, daily snowfall, daily total precipitation (rain + snow), and daily streamflow (height) for a single station between the year of 1963 and 2012 at a natural watershed in Canada. The term "daily" implies that the values provided for each observation have been either averaged (temperature) or accumulated (rainful/snowfall amount) across the day. Moreover, the stations are all within Prairie Pothole region.

Table 1: First 6 rows of a the streaflow dataset

Year	Streamflow (mm/day)	T_min (C)	T_max (C)	Snowmelt (mm/day)	Precipitation (mm/day)	Station
1969	NA	NA	NA	NA	NA	station_100
1970	NA	NA	NA	NA	NA	station_100
1971	0.63962	NA	NA	NA	NA	station_100
1972	0.50637	NA	NA	NA	NA	station_100
1973	0.40065	NA	NA	NA	NA	station_100
1974	0.49415	NA	NA	NA	NA	station_100

Table 2: First 6 rows of a the streaflow dataset without missing values

	Year	Streamflow (mm/day)	T_min (C)	T_max (C)	Snowmelt (mm/day)	Precipitation (mm/day)	Station
151	1946	1.5335	-0.72534	10.8600	1.22280	3.2192	station_1003
152	1947	1.8432	-2.20660	10.2330	0.90575	2.8340	station_1003
153	1948	2.7499	-1.50390	9.7727	1.06890	3.4645	station_1003
154	1949	1.0952	-3.90360	9.3862	0.99945	2.6156	station_1003
156	1951	2.7059	-2.48700	9.6769	1.25210	3.9589	station_1003
157	1952	1.5702	-2.41870	9.0940	1.28000	2.8008	station_1003

Methods

After aggregating the data points by year, the observations are still not independent. For instance, next year's weather may depend on this year's weather because of climate change. The effects of independent variables would also vary per station. Thus, we will be using a mixed effects (random slope) model with auto correlated error terms (AR1), where the random effect is the station and accounts for location variability and the AR1 error terms account for dependency between the years.

In order to determine which factors are most important, the metric used is the standardized coefficients of the linear model above that models streamflow as a function of temperature, snow-to-total-precipitation, rainfall, and the interaction between temperature and each of, snow-to-total-precipitation ratio, and rainfall.

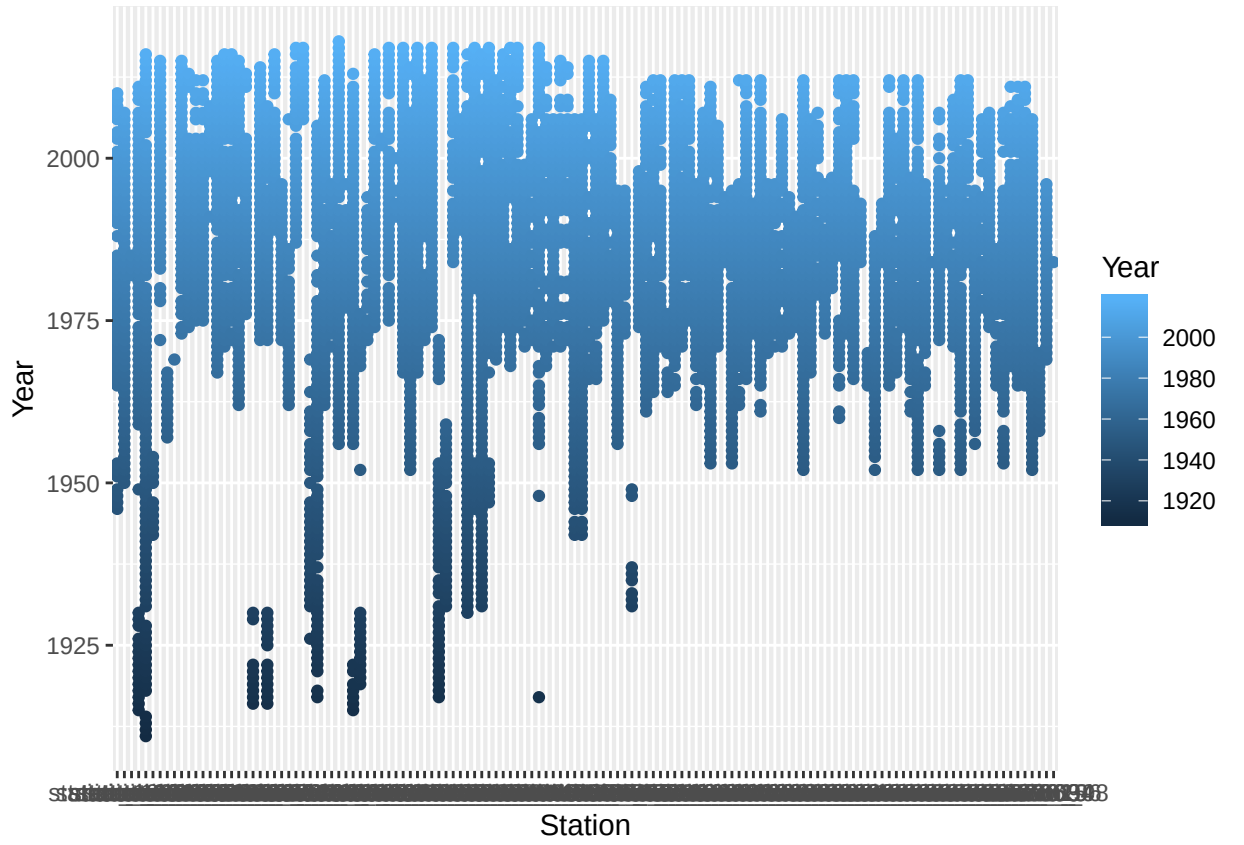
In order to quantify the effect of each variable, the non-standardized coefficient estimates (fixed effects) of the model described above can be used. If any transformations are made to any of the variables, these coefficients must be interpreted/transformed accordingly.

After removing data that have less than 10% of the missing data, there are a total of 121 stations that have more than 10 years of data. This minimize a large data set with 121 station to work with.

```
sum(t>10) #shows how many stations have more than 10 years of data
```

```
## [1] 121
```

The plot below shows the relationship between Years and Stations. This allows us to see which variables is more independent. In this graph, the plot seem to be more concentrated in the top region. Thus, it gives us a general scope that station should be the independent variables since the year can depend on the previous year and it could show a bias result if we compare year by year.



Using linear mixed-effect models to compare which varibles shows the best fit. Station is the random effect of all 6 models and the rest of the variables are fixed effect. Thus, I will only be looking at which variables shows the best fit. Model 1 is a linear mixed-effect model that fit the streamflow data with minimum temperature (T_min) and maximum temperature (T_max). Model 2 is a linear mixed-effect model that fit the streamflow data with Percipitation. Model 3 is a linear mixed-effect model that fit the streamflow data with snowmelt and maximum temperature (T_max). Model 4 is a linear mixed-effect model that fit the streamflow data with the interaction of precipitation and average temperature (T_avg). Model 5 is a linear mixed-effect model that fit the interaction of snowmelt and average temperature (T_avg). Model 6 is a linear mixed-effect model that fit the ratio of snowmelt and precipitation.

```
model1 <- lmer(streamflow.data1 ~ T_min + T_max + (1|Station),
               data = streamflow.data1, REML = FALSE)
model2 <- lmer(streamflow.data1 ~ Precipitation + (1|Station),
               data = streamflow.data, REML = FALSE)
model3 <- lmer(streamflow.data1 ~ Snowmelt + T_max + (1|Station),
               data = dat_10, REML = FALSE)
model4 <- lmer(streamflow.data1 ~ Precipitation * T_avg + (1|Station),
               data = dat_10, REML = FALSE)
model5 <- lmer(streamflow.data1 ~ Snowmelt * T_avg + (1|Station),
               data = dat_10, REML = FALSE)
```

```
model6 <- lmer(streamflow.data1 ~ Snowmelt / Precipitation + (1|Station),
               data = dat_10, REML = FALSE)
```

Results

Since 10% of the missing data was generally randomly distributed throughout the year, but larger number of missing data was attributed to continuous two-three month missing data which could cause significant bias on trend analysis between years. Moreover, the client did not want to compare a year with 365 data to a year with 300 days data with half-winter missing data so 90% threshold is standard in hydrology for the calculation of annual average (Table 2).

Initial observations using boxplot of streamflow and stations for the variability between each stations shows that each stations varies between each other with some outliers. The boxplot seems scattered, which means there are no obvious trend (Figure 1).

However, the boxplot of variability between years seem to be well distributed in the same range compared to the boxplot of variability between each stations, but still with some outliers (Figure 2). Thus, making stations independent would be a better approach. Since year can depend on the previous year, it could show a bias result if we compare year by year.

From the boxplot of precipitation variability between stations, the boxplots seem to be scattered but within the lower range of the plot. Since the outliers are also within the lower range, it would not effect the result of the plot (Figure 3).

In the boxplot of snowmelt variability between stations, the boxplot seem to be more concentrated in the lower range of the plot. This shows that all stations have the same snowmelt effect. Thus, it could possibly be an impact on the streamflow of the watershed (Figure 4).

Lastly, in the bloxplot of temperature variability between station, the boxplot seem to be very scattered apart with some outliers. This shows that there is no obvious trend in the plot, which means that the average temperature varies across stations (Figure 5).

To analyize if the data is plausible, using the QQPlot shows that the the data is roughly normal distributed, which means we can predict the model for a stardard normal distribution using the data. Since most of the points on the QQPlot fall on the black line, it shows that the data is normal (Figure 6).

Using the residual plot to see if the regressioin has any error and variance of the data. The two residual plots below shows the plots shows a bit of heteroscedasticity which means it shows a nonconstant variance in the data (Figure 7 and Figure 8). However, the residual plots seems to fit evenly on the line. This means that the regression model seem to be a good fit for the data.

Fitting model 1 to model 6, the interaction of percipitation and T_avg is the main driver since model 4 is most fitted compared to other models. Hence, percipitation and T_ave seem

to be the impact towards the streamflow of the watershed.

Conclusions

The variability between station could be impacted by temperature the most since we are taking the average of the temperature in the boxplot. From the boxplot of variability between stations and the boxplot of varaibility between years, making stations independent would be a better approach. Since year can depend on the previous year and it could show a bias result if we compare year by year.

The difference between boxplots of the other variables such as temperature, precipitation and snowmelt is that the boxplot for precipitation variability between stations and the boxplot for snowmelt variability between stations seems to be more concentrated in the lower range. However, the boxplot for snowmelt seems to be more concentrated in the lower region compared to the boxplot for precipitation. Thus, we can conclude that snowmelt is the factor that creates more impact on a watershed's streamflow and quantify each factor's contribution to streamflow change.

However, when comparing model 1 to model 6, I found that the interaction of percipitation and T_avg is the main driver since model 4 is most fitted compared to other models rather than the impact of the snowmelt. Hence, percipitation and T_ave seem to be the impact towards the streamflow of the watershed.

In this study there are several limitations such as missing values and outliers. Since we removed data that have less than 10 years of missing data, the years are not consecutive which means that some data have no overlaps. It can be hard to compare the data between stations since some stations does not have the same years compare to other stations.

Appendix

```
T_min <- streamflow.data1 %>% select(T_min)
T_max <- streamflow.data1 %>% select(T_max)
Precipitation <- streamflow.data1 %>% select(Precipitation)
Snowmelt <- streamflow.data1 %>% select(Snowmelt)
T_avg <- ((T_max+T_min)/2)
Station <- streamflow.data1 %>% select(Station)
```

```
table(streamflow.data1$Station)
summary(unique(streamflow.data1$Station))

t <- table(streamflow.data1$Station)
barplot(t)
sum(t>10)

y <- streamflow.data1$Year
```

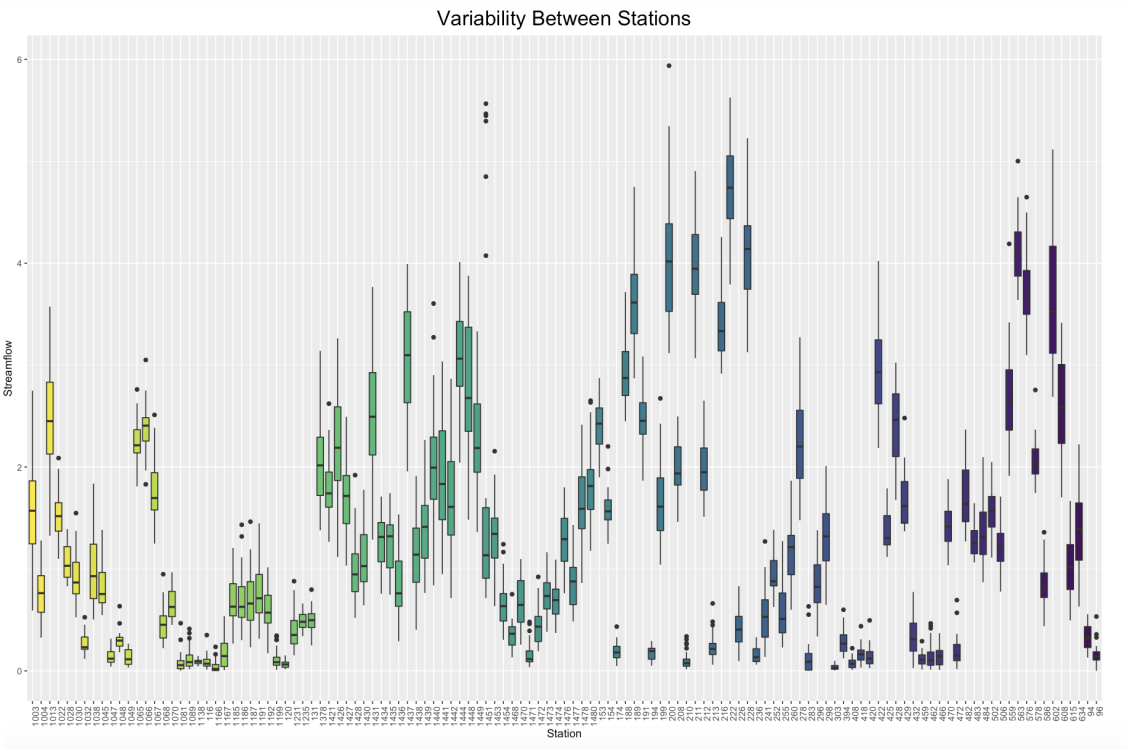


Figure 1: Variability of all stations in different region

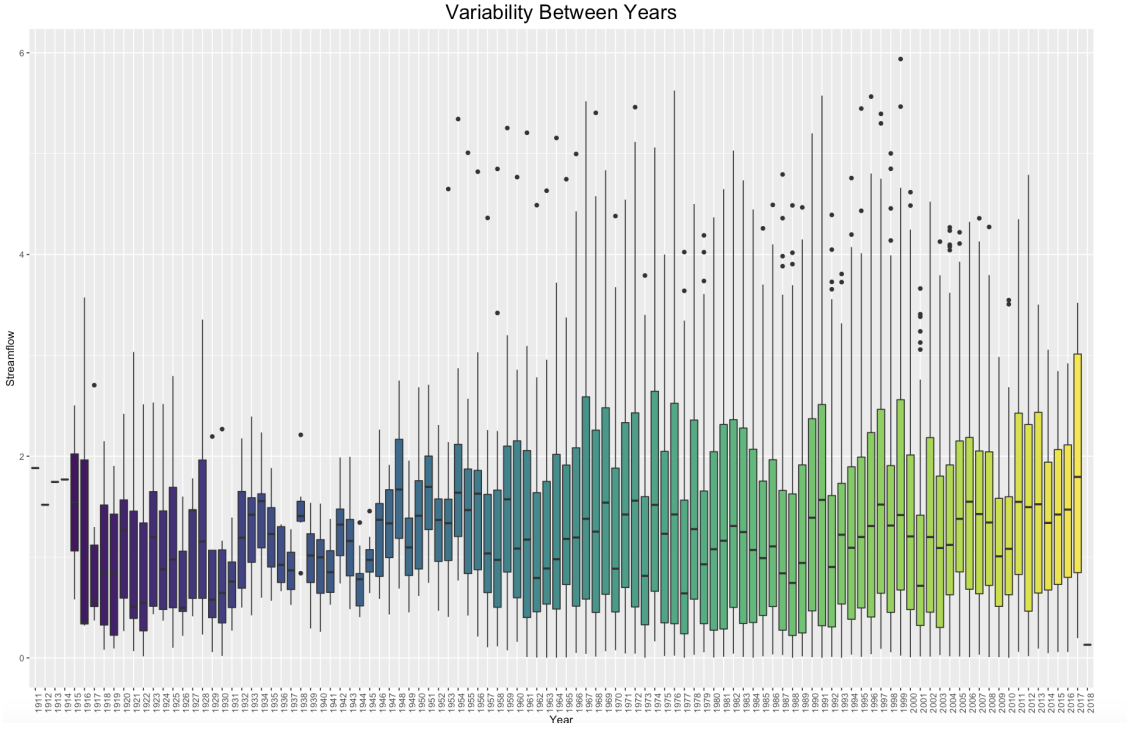


Figure 2: Variability of all stations within Prairie Pothole region from 1963 and 2012 at a natural watershed in Canada

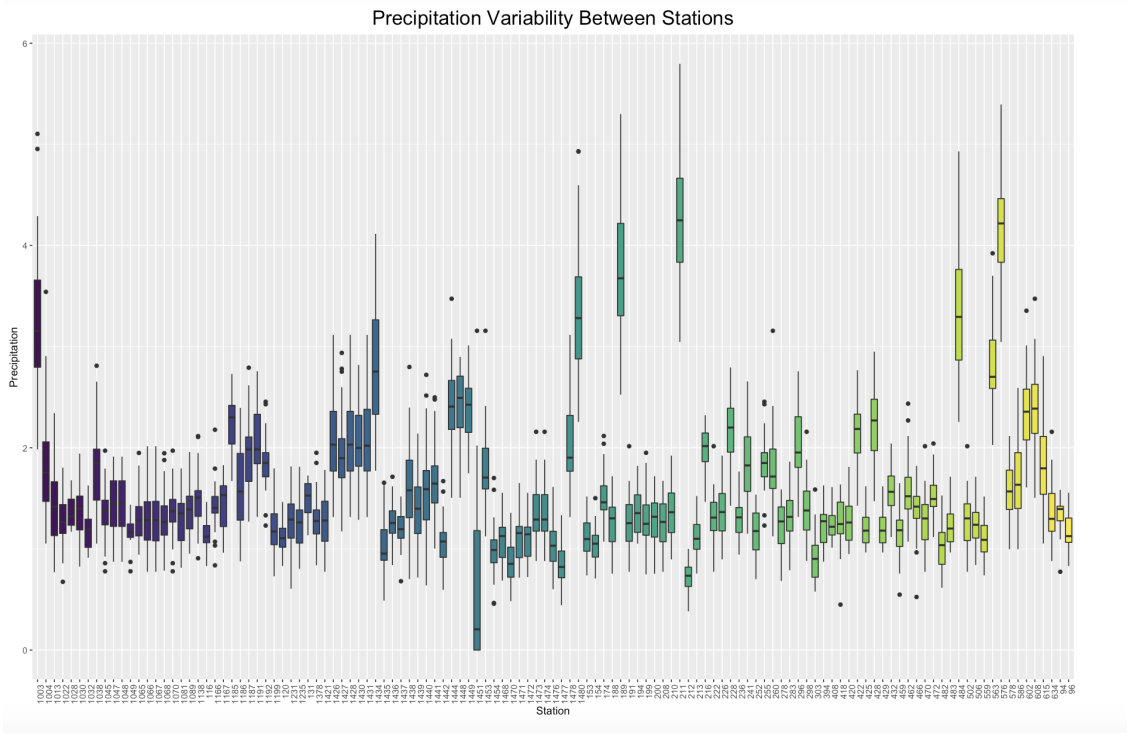


Figure 3: Variability of all stations' precipitation (mm/day) within Prairie Pothole region

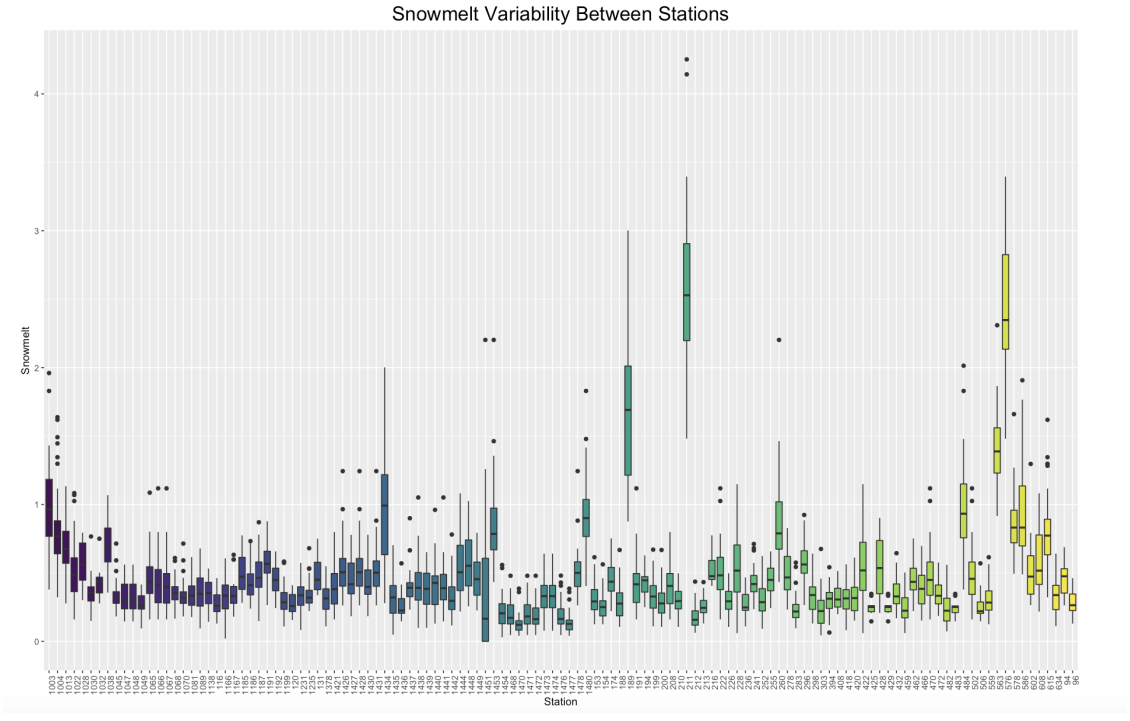


Figure 4: Variability of all stations' snowmelt (mm/day) within Prairie Pothole region

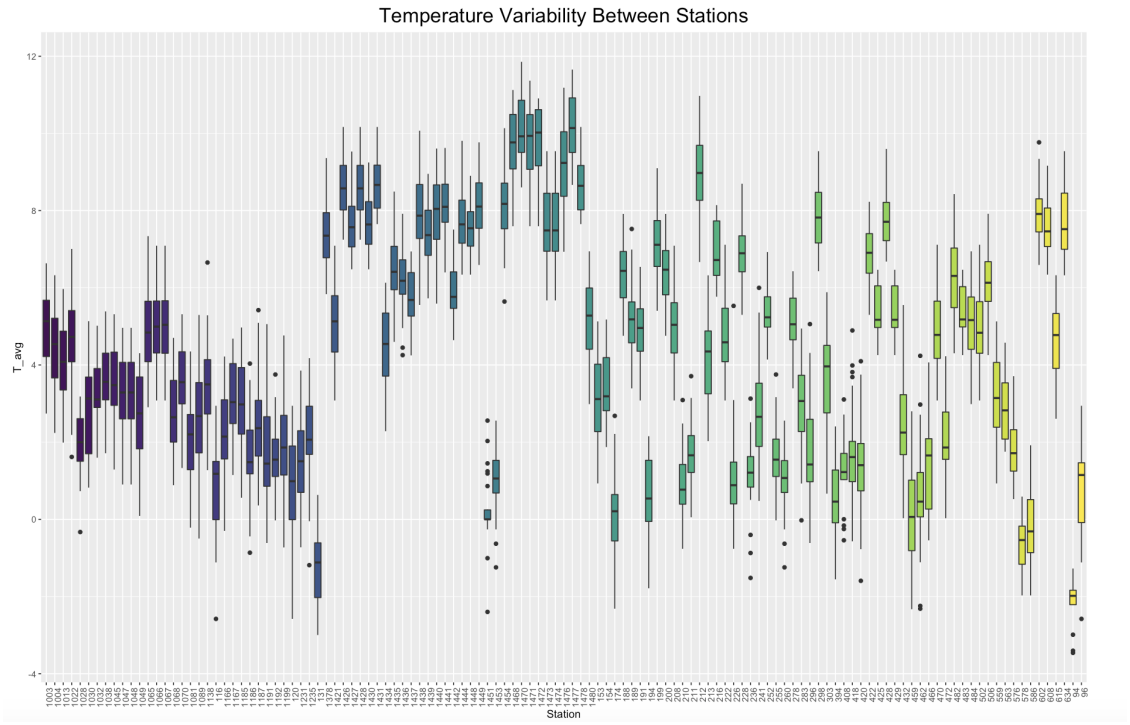


Figure 5: Variability of all stations' temperature (C) within Prairie Pothole region

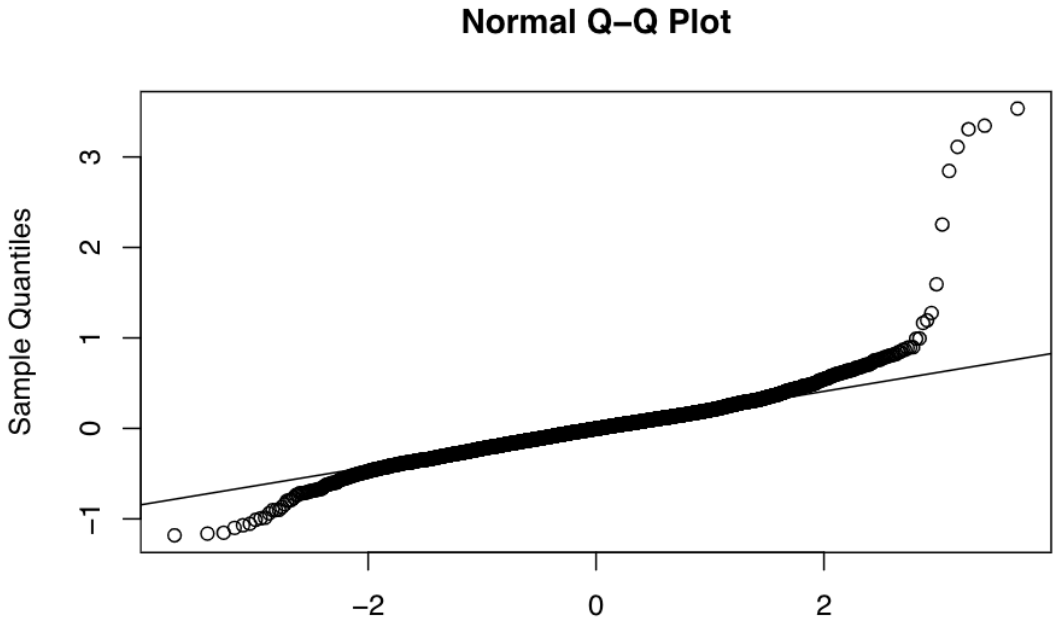


Figure 6: Normal QQ Plot of streamflow data verses standard normal

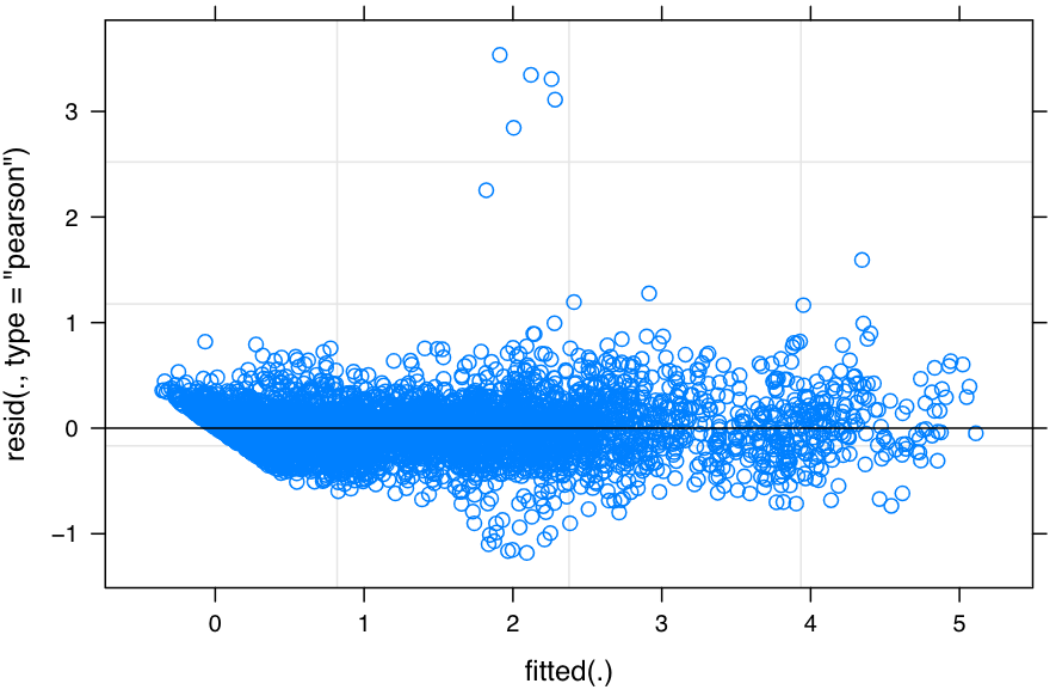


Figure 7: Residual plot of the orignial streamflow data

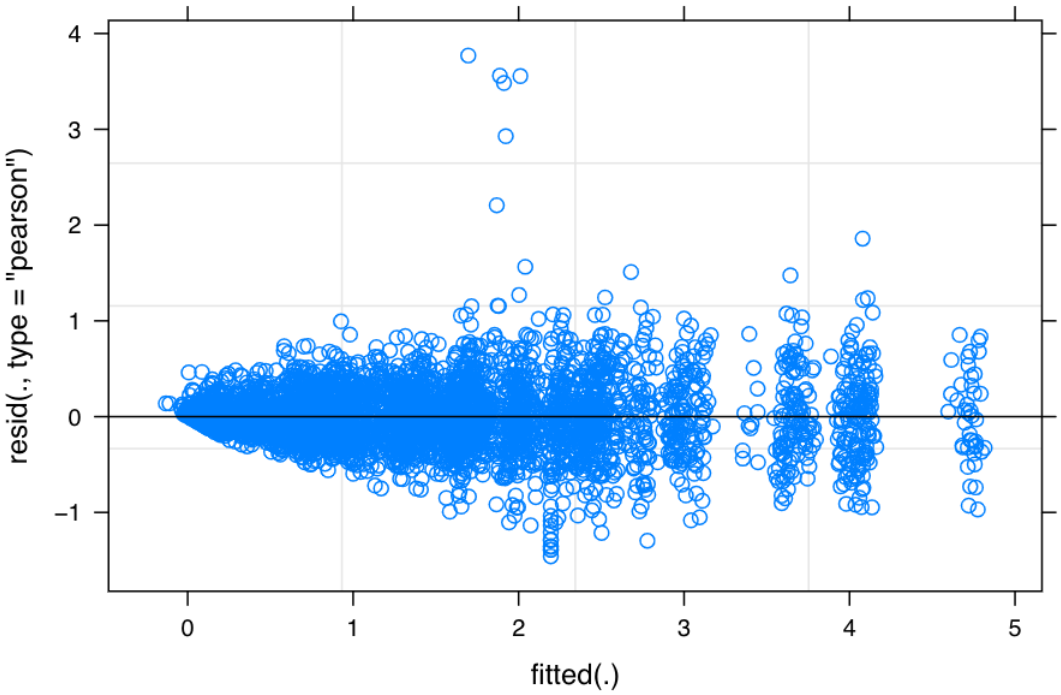


Figure 8: Residual plot of the streamflow data after aggregating and removing the missing values

```r
x <- streamflow.data1$Station

plot(x,y)

library(ggplot2)
p <- ggplot(streamflow.data1, aes(Station, Year, colour = Year))+
  geom_point()
p

summary(streamflow.data$Station)
boxplot(Year~Station, data = streamflow.data)

year.max.by.station <- tapply(streamflow.data$Year, streamflow.data$Station, max)
year.min.by.station <- tapply(streamflow.data$Year, streamflow.data$Station, min)

sum(names(year.max.by.station)!=names(year.min.by.station))

year.df <- data.frame(Station = names(year.max.by.station),
                      min = year.min.by.station,
                      max = year.max.by.station)

matplot(year.df[,2:3], xaxt='n', pch=1, ylab='range')
```