Time Series Analysis of Monthly Precipitation in Portland, OR

Christopher J. Nesseth, Sean M. Pitman, Vedavalli N. Pokala

March 2019

1 Introduction

The National Weather Service's unique local climate data of average monthly precipitation in Portland, OR from 1940 to 2017, measured at Portland International Airport is plotted below in Figure 1 [2]. Missing values are interpolated, and values marked as trace amounts of precipitation are rounded to 0. The goal of this analysis is to identify pertinent seasonal and cyclic trends in Oregon's clear volatile rainfall patterns. As a rainforest environment, we expect the data to contain clear monthly seasonality.

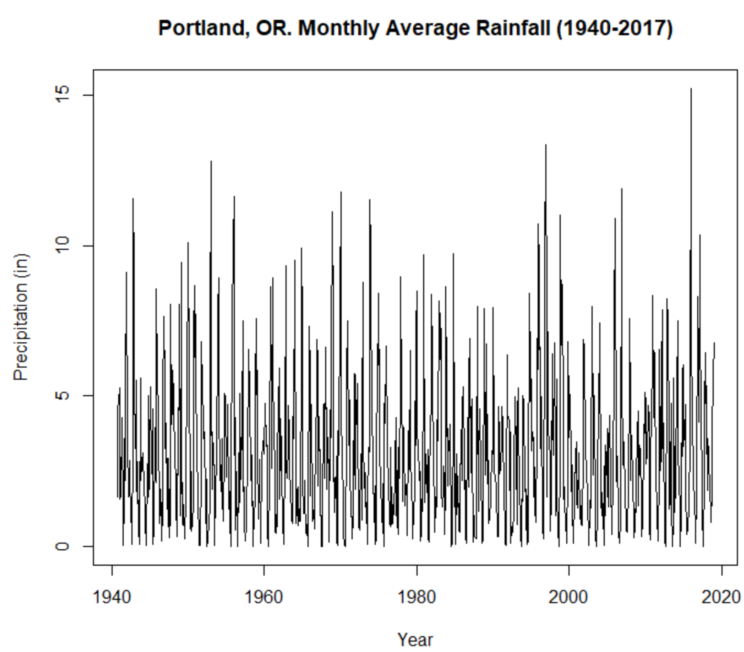

Figure 1: The National Weather Service's unique local climate data of average monthly precipitation in Portland, OR from 1940 to 2017, measured at Portland International Airport.

In the Analysis section we will describe our methods and process. We will begin by examining the series via a time domain analysis, however it would be prudent in the presence of seasonally dominated data to investigate from a frequency analysis perspective as well [1]. Our results should agree, and we will compare and discuss them in the Results section, finally concluding with interpretations about forecasting potential, seasonal trends, and cyclic trends.

2 Analysis

2.1 Time Domain Analysis

The data appear to have little to no trend; we proceed directly to considering seasonal correlation. We begin by considering the yearly seasonal differenced series seen in Figure 2 below.

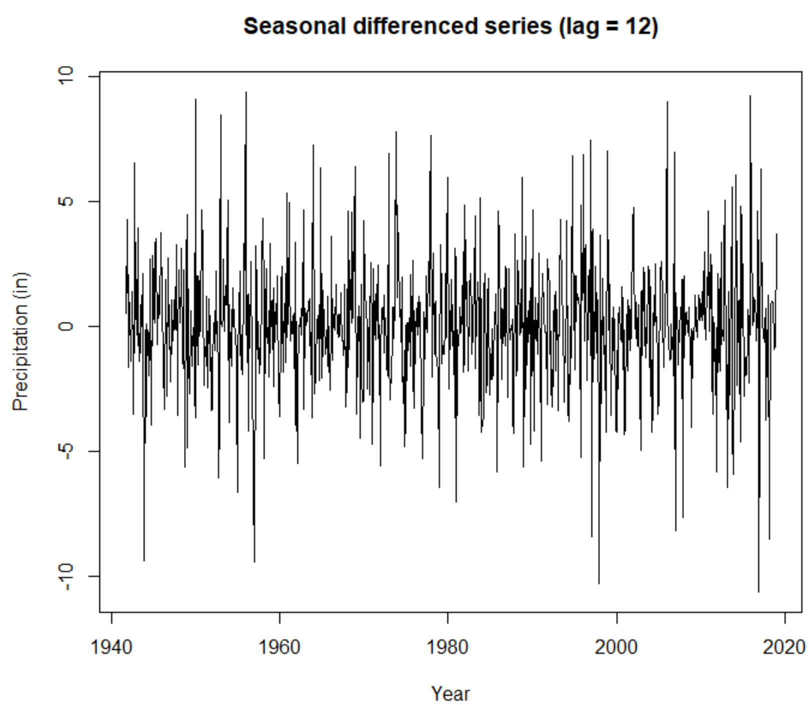

Figure 2: Seasonally differenced series at lag 12.

The data appear to be weakly stationary with mean around 0. The auto-correlation and partial auto-correlation functions for the differenced series are shown in Figure 3 below.

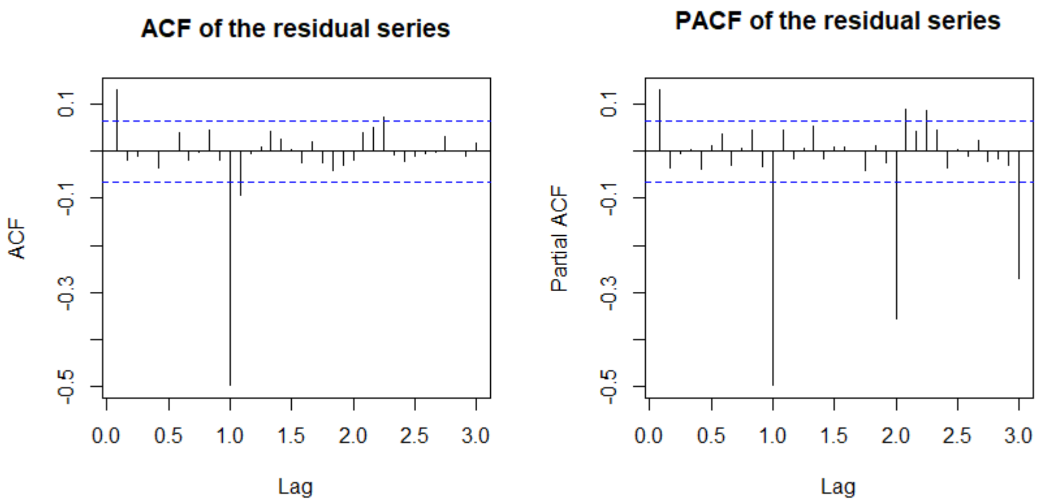

Figure 3: ACF and PACF plots. One lag corresponds to 1 year.

There is very clear yearly auto-correlation, as well as nonzero partial auto-correlation at yearly lags. Since we have differenced seasonally without differencing for trend, and the PACF is highly reminiscent of an $SMA(1)$ model, we considered model candidates $SARIMA(0,0,1)\times(0,1,1)_{12}$, $SARIMA(1,0,1)\times(0,1,1)_{12}$, $SARIMA(1,0,0)\times(0,1,1)_{12}$, as well as the R forecasting package's automatic selection method's model, $SARIMA(5,0,1)\times(2,0,0)_{12}$ via `auto_arima()`. Of these models, $SARIMA(1,0,0)\times(0,1,1)_{12}$ possesses the lowest AIC, and so we proceed to the diagnostic stage of this model fitting.

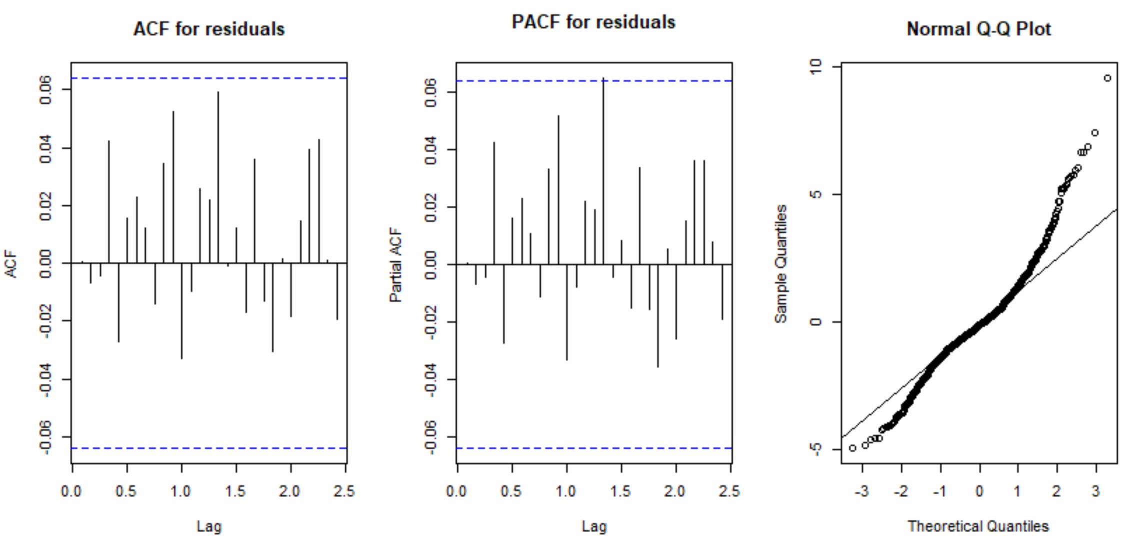

Figure 4: Residual ACF and PACF plots. One lag corresponds to 1 year. QQ-Plot of fitted residuals.

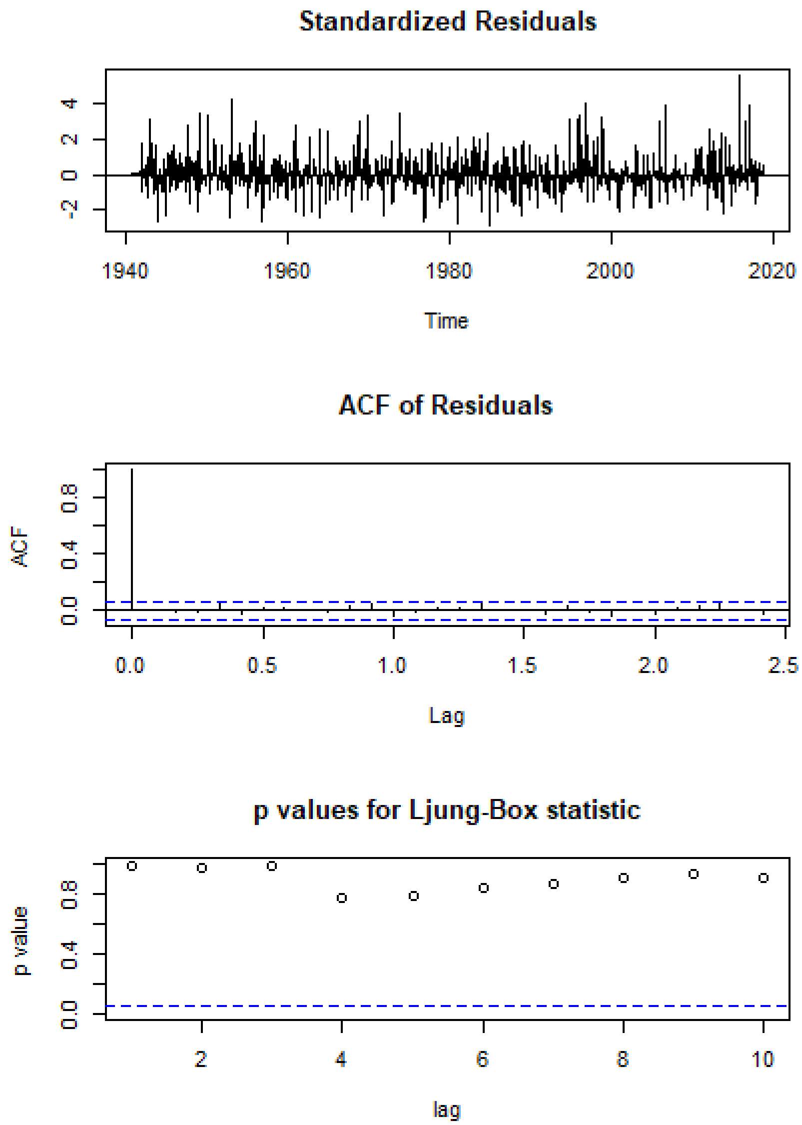

Figure 5: Time series diagnostic function output.

Despite the presence of some outliers influencing the shape of our QQ-Plot, the diagnostics indicate overall that the $SARIMA(1,0,0) \times (0,1,1)_{12}$ is a great fit for the data, especially reinforced by the Ljung-Box statistics seen in Figure 5. We proceed to forecasting 10 years using this model, comparing its performance to Holt-Winters' automatic forecasting method.

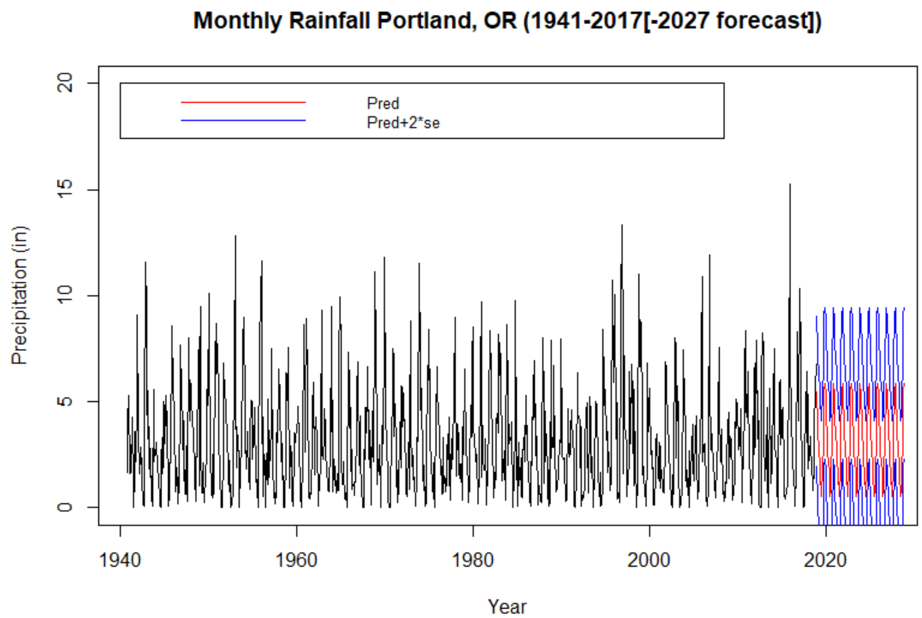

Figure 6: 10 year forecast of Portland, OR average monthly precipitation, via model-based methods.

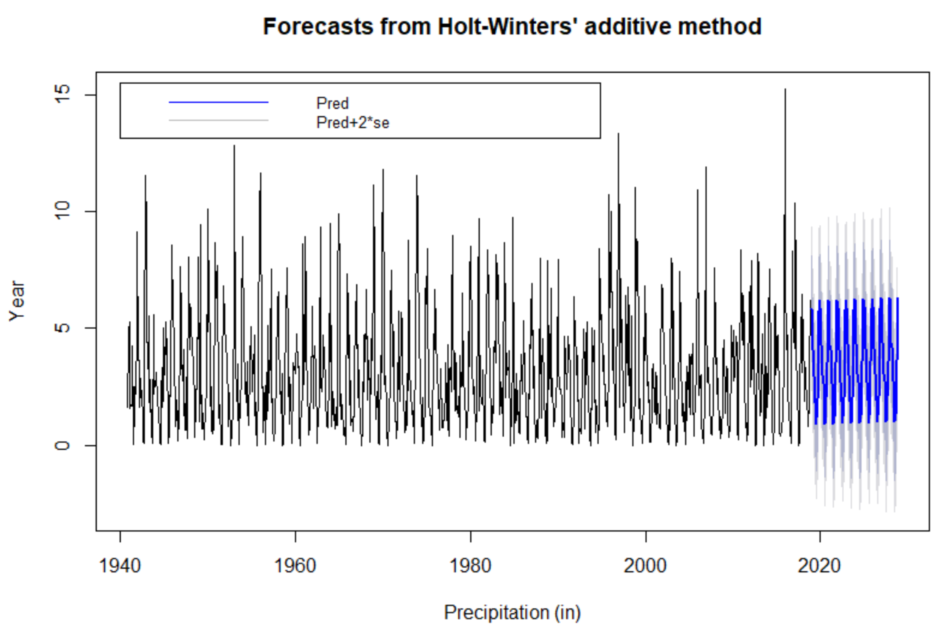

Figure 7: 10 year forecast of Portland, OR average monthly precipitation, via automatic forecasting methods.

We see our model based forecasting methods that consider no trend forecast similarly to automatic methods that consider a linear trend (<< 1 in). With forecasting accomplished, we take a closer look at seasonality through the frequency domain.

2.2 Spectral Analysis

We begin by considering the periodogram of the time series data.

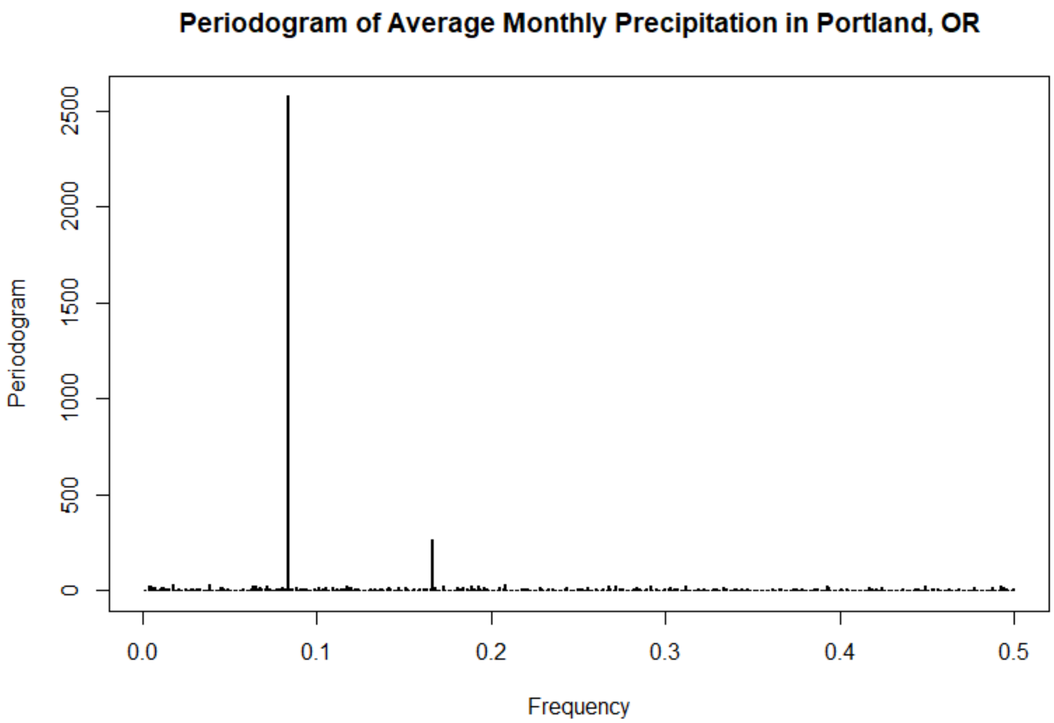

Figure 8: Periodogram of Portland, OR average monthly precipitation.

It is clear from Figure 8 that two frequencies dominate this data. Additionally, the larger weights on lower frequencies agrees with our time domain model selection of an $AR(1)$ model in nonseasonal components. After trying spans of $\sqrt{939 = \text{series length}}$, 1, 5, 10, 12, 25, we settle on a span of 12 which highlights the values of interest.

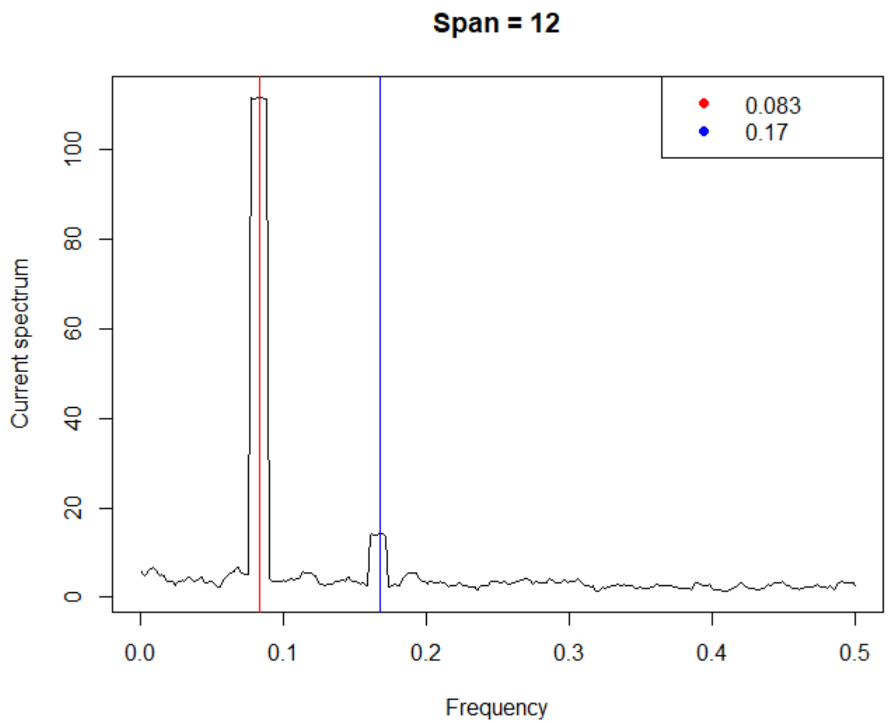

Figure 9: Smoothed Periodogram of Portland, OR average monthly precipitation.

As can be seen in Figure 9 the periodogram suggests a dominating frequency of about 0.083, corresponding to a period of about 1 year (12 months). Additionally, the sub-dominating frequency of about 0.17 corresponds to a period of about 5.8 months, or half a year. The dominating frequency agrees well with our time domain analysis of yearly seasonality, reinforcing our results.

3 Results

Though there is no apparent trend in Portland's rainfall, time domain and spectral analysis both show there is pronounced seasonality.

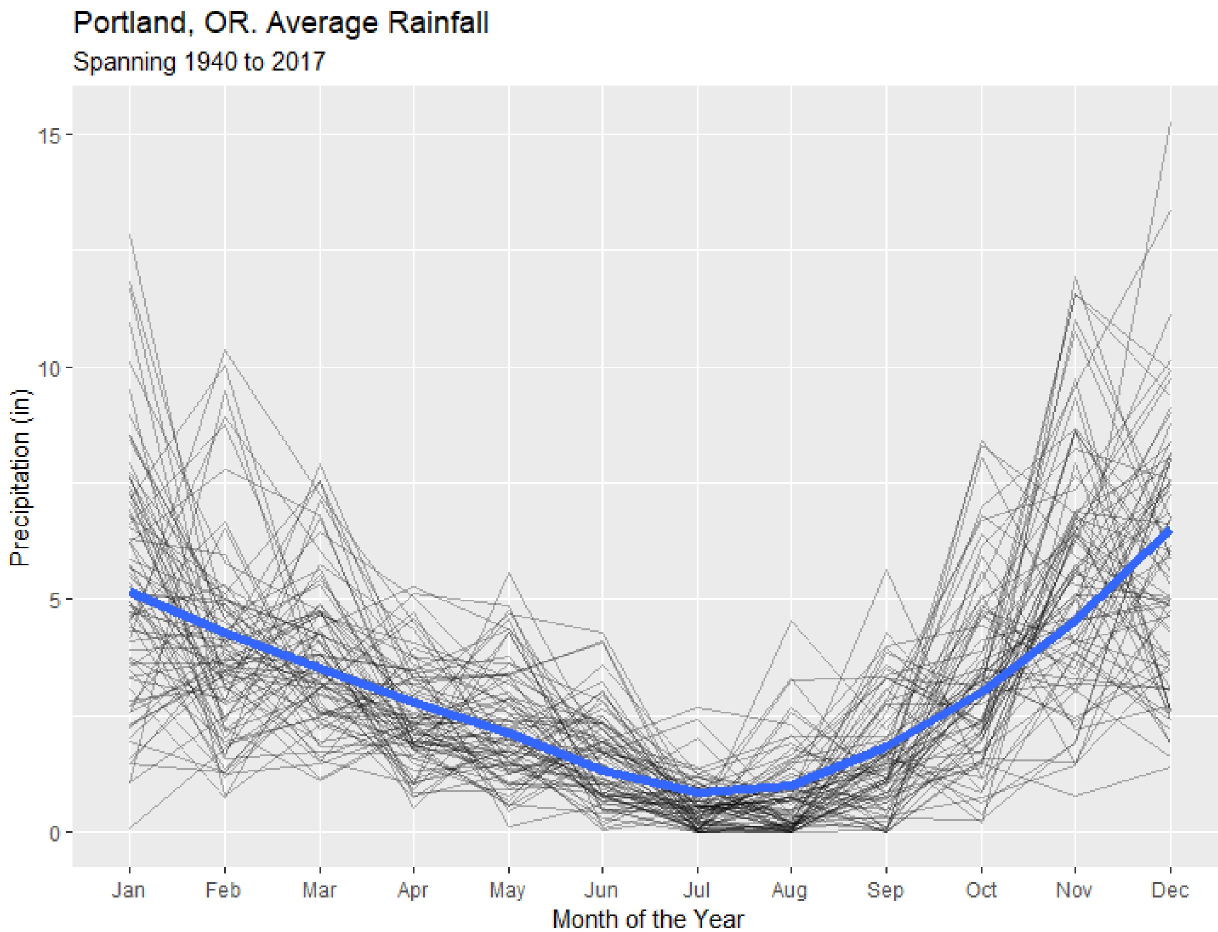

Figure 10: A visualization of yearly rain patterns in Portland, OR from 1940-2017. The blue line describes the trend with a loess smoothing function (size = 2).

Additionally, it is possible that the off-year cyclic effect seen in the sub-dominating frequency in Figure 9 could characterize the meteorological effects of irregular weather systems such as El Niño and La Niña on Portland's rainfall. A brief comparison with weather data on these phenomena suggests that this is a five to six year cycle of variation rather than a five to six month process [3].

Furthermore, inspection reveals that the secondary peak is coming from the cycle that exists in the range of rainfall per month. To highlight the cycle better we removed the average from the rainfall data, and computed the average absolute value of the residual. That data when plotted for every month shows an approximate 6 month cycle. The physical source of this variation would be an interesting extension to this analysis.

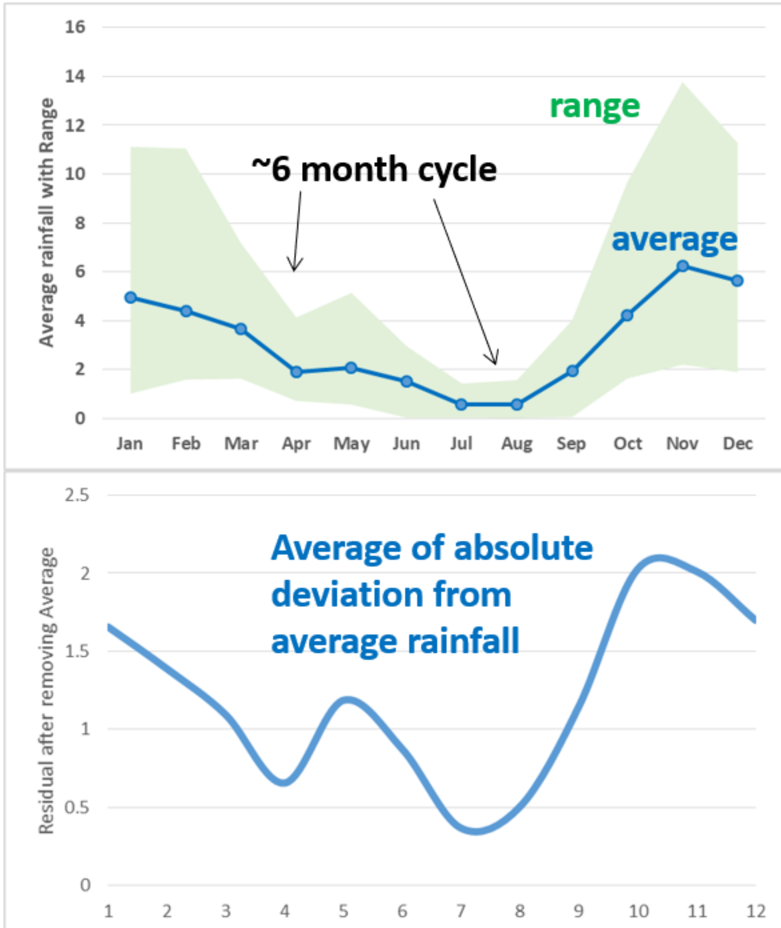

Figure 11: The 6 month cyclic effect, visualized through the average absolute deviations from the mean rainfall in one year.

In summary,

- Portland, Oregon rainfall's greatest source of variation is its annual pattern of rainfall.
- There appears to be no trend in Portland, Oregon rainfall aside from months of heavy rain that pull the overall average up negligibly (<< 1 in).
- Following this pattern Portland, OR is wettest in December and driest in July as seen in Figure 10. December shows higher variation than July but the long term trend does not seem to change.
- Irregular weather systems could influence Portland, Oregon's rainfall in a sub-dominant cycle of about half a year.

References

[1] Jonathan D Cryer and Natalie Kellet. *Time series analysis*. Springer, 1991.

[2] National Weather Service Corporate Image Web Team. Unique local climate data: Portland, or, Oct 2005.

[3] Kevin Trenberth. Nino sst indices (nino 1 2, 3, 3.4, 4; oni and tni), Feb 2018.

4 R Appendix

```
#install.packages("forecast")
#install.packages("ggplot2")
#install.packages("TSA")
#install.packages("lubridate")
#install.packages("plyr")
library(forecast)
library(ggplot2)
library(TSA)
library(readr)
library(lubridate)
library(plyr)

pdc_rains <- read.csv("http://www.wrh.noaa.gov/pqr/
                    climate/Portland_dailyclimatedata.csv",
                      skip = 6)
pdc_rains <- pdc_rains[pdc_rains$X == 'PR',] # consider only rain data

# pull out and clean up the data we want to use
rainfall_month <- pdc_rains$AVG.or.Total
rainfall_month <- gsub("M", NA, rainfall_month)
rainfall_month <- gsub("T", 0, rainfall_month)
rainfall_month <- as.numeric(rainfall_month)

rain.ts <- ts(rainfall_month, start = c(1940,10), deltat = 1/12)

# This is a big assumption made in order to fix the data
rain.ts <- na.interp(rain.ts) # interpolating missing values


par(mfrow = c(1,1))
```

```r
# default plot
plot(rain.ts, main = "Portland, OR. Monthly Average Rainfall (1940-2017)",
     xlab = "Year",
     ylab = "Precipitation (in)")

# ggplot for investigating trend (probably isn't much)
rain <- data.frame(Year = time(rain.ts), Prec = rainfall_month)
ggplot(rain, aes(Year,Prec))+
  geom_line()+
  ggtitle("Portland, OR. Monthly Average Rainfall (1940-2017)")+
  xlab("Year")+
  ylab("Precipitation (in)")+
  geom_smooth(method = "loess")
# no trend to consider, it's all seasonality

# let's investigate seasonality
rain_seas <- data.frame(Month = month(pdc_rains$MO),
                        Year = pdc_rains$YR,
                        Rain = rainfall_month)
ggplot(rain_seas, aes(Month, Rain))+
  geom_line(aes(group = Year), alpha = I(0.3))+
  geom_smooth(method = "loess", se = FALSE, size = 2) +
  scale_x_discrete(name = "Month of the Year",
                   limits = c("Jan", "Feb", "Mar", "Apr",
                              "May", "Jun", "Jul", "Aug",
                              "Sep", "Oct", "Nov", "Dec"))+
  ylab("Precipitation (in)")+
  ggtitle(label = "Portland, OR. Average Rainfall",
                    subtitle = "Spanning 1940 to 2017")

# spectral analysis might be a good approach, will try after
# time domain approach

# since this is monthly, take the year's seasonal difference
diff12 <- diff(rain.ts, lag = 12)
plot(diff12, main = "Seasonal differenced series (lag = 12)",
     xlab = "Year",
     ylab = "Precipitation (in)")

par(mfrow = c(1,2))
acf(diff12, main = "ACF of the residual series", lag.max = 36,
    na.action = na.pass)
```

```r
pacf(diff12, main = "PACF of the residual series", lag.max = 36,
    na.action = na.pass)

# there is clear yearly correlation even after differencing by year.
# one seasonal difference, no trend difference; d = 0, D = 1
# this is heavily reminiscent of an SMA(1) model; Q = 1
# for non seasonal, nonzero lags at 1, 11 ACF and 1 PACF
# might consider...
# SARIMA(0,0,1)x(0,1,1)12
# SARIMA(1,0,1)x(0,1,1)12
# SARIMA(1,0,0)x(0,1,1)12
# also compare with results from auto.arima()

(fit.sarima.001 <- arima(rain.ts, order = c(0,0,1),
                        seasonal = list(order = c(0,1,1), period = 12),
                        method = "ML", include.mean = F))
(fit.sarima.101 <- arima(rain.ts, order = c(1,0,1),
                        seasonal = list(order = c(0,1,1), period = 12),
                        method = "ML", include.mean = F))
(fit.sarima.100 <- arima(rain.ts, order = c(1,0,0),
                        seasonal = list(order = c(0,1,1), period = 12),
                        method = "ML", include.mean = F))
auto.arima(rain.ts)

# we beat auto.arima! We proceed with SARIMA(1,0,0)x(0,1,1)12 since it has
# the lowest AIC value

par(mfrow = c(1,2))
res <- fit.sarima.100$residuals
acf(res, main = "ACF for residuals", na.action = na.pass)
pacf(res, main = "PACF for residuals", na.action = na.pass)

# well that does a darn good job, pure white noise residuals

par(mfrow = c(1,1))
tsdiag(fit.sarima.100)

# pretty good looking diagnostics too, some nice Ljung-Box stats

qqnorm(res)
qqline(res)
```

```r
# that could be a problem though...
# It doesn't look like our normality assumption
# is well met. We could check with something
# like a Bonferroni outlier criterion... we proceed with caution

# forecasting time!

pred <- predict(fit.sarima.100, n.ahead = 120)
plot(rain.ts, xlim = c(1941, 2027), ylim = c(0,20),
     main = "Monthly Rainfall Portland, OR (1941-2017[-2027 forecast])",
     ylab = "Precipitation (in)",
     xlab = "Year")
lines(pred$pred, col = "red")
lines(pred$pred-2*pred$se,col='blue')
lines(pred$pred+2*pred$se,col='blue')
legend(1940, 20, legend=c("Pred", "Pred+2*se"),
        col=c("red", "blue"), lty=1:1, cex=0.8)

# This should agree with automatic methods
plot(hw(rain.ts, seasonal = "additive", h = 120),
     ylab = "Year",
     xlab = "Precipitation (in)")
legend(1940, 15.5, legend=c("Pred", "Pred+2*se"),
        col=c("blue", "grey"), lty=1:1, cex=0.8)

# This works
# Spectral analysis might be more helpful
# since this is almost purely seasonal
periodogram(rain.ts)

# that has a very clear dominant frequency or two
# lets smooth and investigate
#per.smooth <- TSA::spec(rain.ts, plot = F, span = sqrt(length(rain.ts)))
per.smooth <- TSA::spec(rain.ts, plot = F, span = 12)
freq.rain <- per.smooth$freq

plot(freq.rain, per.smooth$spec, main = "span = sqrt(939)",
     xlab = "Frequency", ylab = "Current spectrum", type="l")

dom_freq <- freq.rain[per.smooth$spec == max(per.smooth$spec)]

# dominant frequency about 0.08 12.5 months (just over a year)
```

```
# secondary frequency about 0.18 or 5.55 months (just under half a year)
# reminiscent of positive coefficient AR model spectrum,
# so it supports decisions made in the time domain
```