Division of Computer Science

International Technological University

PROJECT SUPERVISOR: DR. HELEN HUYNH

CS 960 INTRODUCTION TO DATA SCIENCE

PRECIPITATION PROJECT

SPRING 2014

Team IX Participants

- **ZHI ZHANG (86164)**
- **MITAL JANI (81299)**
- **AMRITHA BHAT (83601)**
- **RAJ KIRAN (84176)**
- **SHRAVAN KUMAR (83356)**

Declaration

We hereby declare that this project has been composed by team 9 and has not been submitted for any other degree elsewhere. The work presented herein is our own and all referenced work and assistance given to us is duly acknowledged.

TEAM IX

SPRING 2014

Acknowledgement

We would like to express our deep sense of gratitude to ***Prof. Helen*** *for her valuable suggestions and guidance that enabled us to complete this research project. She was a great source of help and encouragement throughout our project. Working under her supervision was a great experience.*

We would like to extend my thankfulness to all our team members, and teaching staff for making this project a memorable one.

Thank you all

.

TABLE OF CONTENTS

DESCRIPTION **PAGE NUMBER**

DECLARATION..... 3

ACKNOWLEDGEMENT..... 4

ABBREVIATIONS..... 7

PROJECT SCHEDULE..... 8

1. ABSTRACT..... 9

2. INTRODUCTION..... 10

2.1 PROJECT DEFINITION..... 10

2.2 BACKGROUND..... 11

2.3 OBJECTIVE..... 14

2.4 ADVANTAGES AND DISADVANTAGES OF PROJECT RELATED RESEARCHES..... 14

2.4 EXPECTED OUTCOME..... 15

3. FLOWCHART..... 17

4. PROJECT SCHEDULE..... 17

5. HYPOTHESIS/GOALS..... 18

5.1 POSITIVE HYPOTHESIS..... 18

5.2 NEGATIVE HYPOTHESIS..... 19

6. METHODOLODY..... 18

6.1 SOFTWARE REQUIREMENTS..... 21

7. IMPLEMENTATION..... 22

7.1 TOOL..... 22

7.2 ANALYSIS & DESIGN DOCUMENT..... 23

8. CONCLUSION..... 33

9. GLOSSARY..... 34

10. BIBILOGRAPHY..... 35

LIST OF FIGURES

FIGURE 1 AVERAGE PERCENTAGE OF RAINFALL IN 2013…….. 14

FIGURE 2 SAMPLES OF DATA…….. 23

FIGURE 3 NORMALIZED OUTPUT DATA…….. 24

FIGURE 4 NUMBER OF DAYS OF NON-MISSING DATA…….. 25

FIGURE 5 SEASON RAINFALL FOR 1903-2013 (INCHES)…….. 25

FIGURE 6 PLOT FOR DAILY RAINFALL…….. 26

FIGURE 7 DAILY RAINFALL (in) DISTRIBUTION BY WATER YEAR……..26

FIGURE 8&9 PLOT YEARLY ACCUMILATED RAINFALL …….. 27

FIGURE 9 MONTHLY RAINY DAYS…… 28

FIGURE 10 SEASONAL RAINFALL 190-2013 WITH GAM FIT…….. 29

FIGURE 11 MOVING AVERAGE MODEL…….. 29

FIGURE 12 SEASONAL RAINFALL MOVING AVERAGE FOR 10 YEARS ……..30

FIGURE 13 SEASONAL RAINFALL MOVING AVERAGE FOR 10 YEARS ……..30

FIGURE 14 SEASONAL RAINFALL MOVING AVERAGE FOR 10 YEARS …….. 31

FIGURE 15 SEASONAL RAINFALL MOVING AVERAGE FOR 30 YEARS …….. 31

FIGURE 16 SERIES

ABBREVIATIONS

DS – Data Science

USHCN - United States Historical Climatology Network

GHCN - Global Historical Climatology Network

POP - Probability of precipitation

OLR - Outgoing Long wave radiation

QPE - Quantitative precipitation Estimate

WEQ – Water Equivalent

DM – Data – Mining

HRWR – High resolution weather Radars

MRR – Micro rain Radar

MA – Moving Average

ACF – Auto Correlation function

PROJECT SCHEDULE

Date	Time	Location	Agenda
22/15/2014	6:00-6:45 pm	Class room 101	Brainstorming session
29/15/2014	6:15-6:45 pm	Class room 101	Requirements Gathering, Purpose, Scope.
02/5/2014	6:15-6:45 pm	Class room 101	Decision on tool to be used
02/12/2014	6:00-6:45 pm	Class room 101	Task assignment, data gathering
02/19/2014	6:15-6:45 pm	Class room 101	Data massaging, updates on assigned tasks
02/26/2014	6:00-6:45 pm	Class room 101	Updates on task assignment
03/05/2014	6:00-6:45 pm	Class room 101	Updates on task assignment
03/12/2014	6:00-6:45 pm	Class room 101	Updates on task assignment
03/19/2014	6:00-6:45 pm	Class room 101	Discussion regarding documentation, template
03/26/2014	6:00-6:45 pm	Class room 101	Discussion on presentation, report submission

1.ABSTRACT

It is important to accurately estimate rainfall for effective use of water resources and optimal planning of water structures. For this purpose, the models were developed to estimate rainfall in San Francisco using the data-mining process. Analysis so far proves 2012-2013 has been the driest year.

Important characteristics of local precipitation (including global and regional means, extremes, variations and trends) will be determined by extending, improving and analyzing the 100+ years merged precipitation data. 2013 was the driest year in California since having record. On January 17, 2014, California governor Jerry Brown officially declared a drought emergency. We will look into the current and historical precipitation data for California (mainly for bay area locations).

The specific problems to be solved are mentioned below:

Usually, 80% of a data analysis task is to do data preparation. A large amount of our project is also doing data massaging on the raw data.

a. Understanding data and getting the data:

There are multiple sources of precipitation data. Find the data needed and extracting the right data from different sites is the first step, and not a trivial task.

b. Data preparation:

The current and historical data are from different sources, with different formats. We need to find the connection between the datasets, combine, clean, and normalize the data.

c. Looking into the data:

Summarize the statistics of data, with charts and graphs. Show the big picture.

d. Sample the data, Deep analysis and comparison on certain locations: (e.g. San Francisco)

e. Try to find patterns or insights in the data could be trends, certain periodic patterns, or clusters, anything could be etc. It is quite open-end here.

f. Learn statistics, data visualization, new tools (R, Python, and JavaScript's)

2.INTRODUCTION

2.1 PROJECT DEFINITION

Precipitation: Precipitation is a product of condensation of atmospheric water vapor that falls under gravity. The main forms can be said as drizzle, rain, sleet, snow, graupel and hail.

Probability of precipitation (POP): This has been a part of weather forecast since late 1960's and this is the only option or forecast element that includes a probability. POP is defined in most commonly among meteorologists, this is confidence probability that at least one hundredth inch of liquid equivalent precipitation will fall in a single spot. One common way to understand a POP of 60 percent is; if we had ten identical weather conditions, rain would fall on six of them at a given point, or 60 percent of days. And rain will not fall on four of those days.

But this PoP is a forecast for ten potential tomorrows, not a forecast for the next ten days. One thing can be taken to the consideration is as POP increases, precipitation becomes more likely.

Mathematically PoP is define as follows:

PoP = C x A

Where "C" = the confidence that precipitation will occur somewhere in the forecast area

Where "A" = the percent of the area that will receive measurable precipitation.

Outgoing Long wave radiation (OLR): Is the energy leaving the earth as infrared radiation at low energy to space. OLR is a critical component of the earth's radiation budget and represents the total radiation going to the space emitted by the atmosphere. The OLR is

affected by clouds and dust in the atmosphere, which then tend to reduce it below clear sky values.

OLR is dependent on the temperature of the radiating body. It is affected by the earth's skin temperature, skin surface emissivity, atmospheric temperature, water vapor profile, and cloud cover.

Quantitative precipitation Estimate (QPE): Is a method of approximating the amount of precipitation that has fallen at a location or across a region. Maps of estimated amount of precipitation to have fallen over a certain areas and time span are compiled using several different data sources including manual, radar and satellite data. Different algorithms can be used to estimate precipitation amounts from the data collected by radar.

The main goal of this project is to predict the rainfall by using the data available from various resources like, meteorological, state and federal government websites mentioned in the references column.

In this process we choose a particular station and compare data for the past 100 years of historical data to check and see what the percentage of rainfall was over past years.

2.2 BACKGROUND

The meteorological events affect permanently human life. Considering the meteorological phenomena, which have no possibility of intervention, they cause the important results in human life, accurate estimation and analysis of these variables is also very important. Precipitation, which is generating flow, is an important parameter. The occurrence of extreme rainfall in a short time causes significant events that affect human life such as flood. However, in the event of insufficient rainfall in long period occurs drought. Thus, rainfall estimation is very important in terms of effects on human life, water resources, and water usage areas. However, rainfall affected by the geographical and regional variations and features is very difficult to estimate.

There are several reasons for the cause of drought such as, water vapor is not brought by air currents to the right areas at the right times or if the mountains prevent wind from blowing moisture to needed regions. As air is moving past a mountain range, it is forced to rise in order to pass over the peaks.

California is enduring its driest calendar year on record, with no signs of relief coming anytime soon. In San Francisco, the city is seeing its driest year since records began during the Gold Rush year of 1849. Soil moisture is depleted, reservoir storage is down and even if we had average rainfall statewide, we probably wouldn't see average runoff just because soil moisture is so depleted. The catastrophic wildfires remain the greatest risk the state would likely face next year after a third-straight dry winter.

Meteorologists say the reason behind the low precipitation is a massive zone of high pressure nearly four miles high and 2,000 miles long that has been blocking storms for more than a year.

On January 31, the Department of Water Resources announced several actions to protect Californians' health and safety from more severe water shortages. Those actions include dropping the anticipated allocation of water to customers of the State Water Project from 5 percent to zero; notifying long-time water rights holders in the Sacramento Valley that they may cut be 50 percent, depending upon future snow survey results; and asking the State Water Resources Control Board to adjust requirements that hinder conservation of currently stored water.

In normal years, the snowpack stores water during the winter months and releases it through melting in the spring and summer to replenish rivers and reservoirs. However, relatively dry weather conditions this year have reduced the amount of snowpack in California's mountains. Each of this season's first three snow surveys – conducted in early January, late January and late February found a statewide snowpack water equivalent (WEQ) far below average for the dates of the surveys.

One of the aims of storing this data in databases and receiving data from many sources is to convert raw data into information at present. This process is called as data-mining (DM)

process of converting data into information. In recent years, the use of data-mining process in the field of hydrology is increasing.

One of the aims of storing this data in databases and receiving data from many sources is to convert raw data into information at present. This process is called as data-mining (DM) process of converting data into information. In recent years, the use of data-mining process in the field of hydrology is increasing.

Up to now, available summaries were based on relatively crude analyses of rainfall data collected through the 1960s. This project has updated precipitation intensities based on the compilation of hundreds of years of rainfall monitoring locations in and around san Francisco with continuous data collected through 2012.

We may have to follow the steps to do some precipitation analysis:

i) **Data cleaning:** remove noise, inconsistent data and unnecessary data which may slow down the queries

ii) **Data integration:** data from multiple sources are integrated and reconfigured to a consistent format, if necessary

iii) **Data selection:** data relevant for the analysis are retrieved from the database

iv) Data transformation: data are transformed into forms appropriate for mining

v) **Data mining:** A crucial step where intelligent methods are applied to extract patterns from the data.

vi) **Pattern evaluation**: Unique/Useful/interesting patterns are identified from the extracted patterns.

Using 4km rainfall data from the PRISM Climate Group for the time period 1895-2013, a study compares the annual total for 2013 with the long-term and near-term historical averages from PRISM.

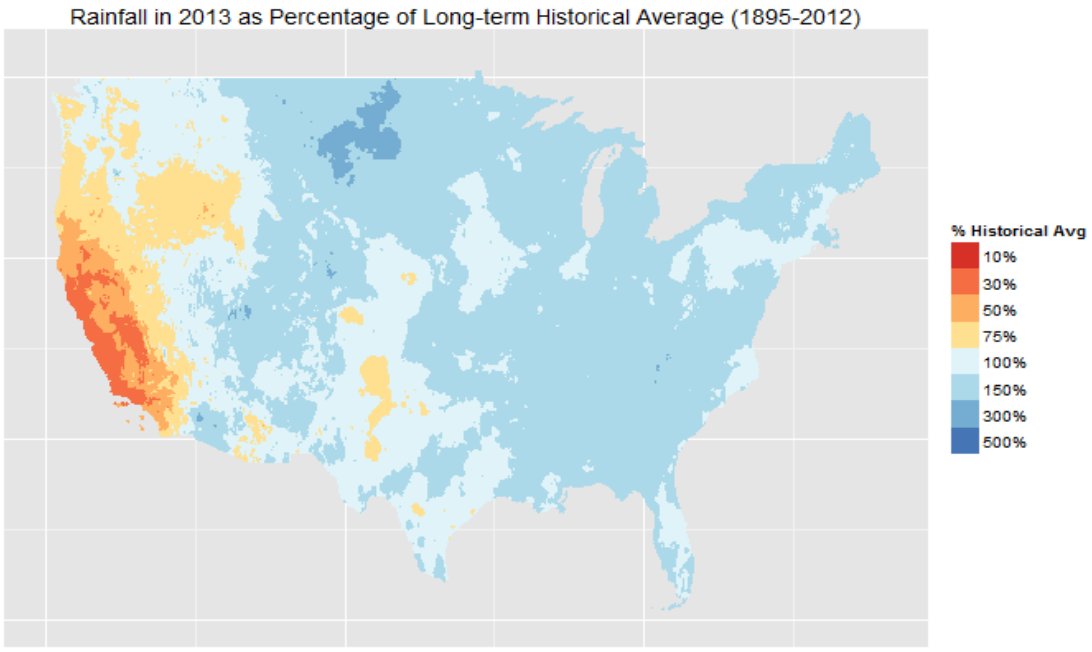


Figure 1 Average percentage of Rainfall in 2013

2.3 OBJECTIVE

The objective of this paper is to explain a new technique for the prediction of rainfall using open source data available on government sites of previous 100 years historic data and also focusing on current year data. The proposed system can predict onset 10-30 days in advance. There are lots of other factors that get affected or affect the rainfall onset. The efficiency of the technique can be improved by adding those additional parameters like Outgoing long wave radiation (OLR), Quantitative Precipitation Estimate (QPE) and atmospheric pressure. For this paper, we are focusing on gathering raw data, manipulating raw data, and using techniques to solve the missing data and massaging the data and finally will plot graphs based on different input to predict the rainfall for coming 10-30 days in advance. Also more regions for comparison can be added.

2.4 ADVANTAGES AND DISADVANTAGES OF PROJECT RELATED RESEARCHES

Precipitation observations with radars operating in the X-band frequency range are essential for meeting present and future requirements for flood forecasting, water

management, and other hydro-meteorological applications. Besides having higher resolution, these systems are cost-effective compared to S- or C-band radars because of smaller antenna size. Disadvantages of single X-band radars are the large influence of attenuation by liquid water and a relatively short range.

The project Precipitation and Attenuation Estimates from a High-Resolution Weather Radar Network (PATTERN) is designed to demonstrate that a network of high-resolution weather radars (HRWRs) can overcome the apparent drawbacks of single X-band radars. The University of Hamburg and the Max Planck Institute for Meteorology have set up a network of four modified ship navigation radars near Hamburg, Germany. The network has been operational since January 2012.

Each radar has a maximum range of 20 km with 60 m spatial and 30 s temporal resolutions. A large area in the network is covered by at least two radars at the borders and up to four radars in the center. Several rain stations consisting of micro-rain radar (MRR) and a rain gauge complement the network. These stations are used to calibrate and evaluate the quality of the X-band radars.

In addition to identifying advantages and disadvantages of the HRWR network and single X-band radar, Dr. Lengfeld will briefly describe the algorithms used to derive precipitation from reflectivity measurements — particularly algorithms exploiting benefits of the network, such as clutter removal and replacement of disturbed pixels with measurements from other radars rather than interpolation. A comparison with measurements of the German Weather Service weather radar operating in C-band will focus on the ability of high-resolution observations to generate information about-small scale structures of rain events. Dr. Lengfeld will also describe the specifications of the modified HRWR systems and the design of the network.

2.5 EXPECTED OUTCOME

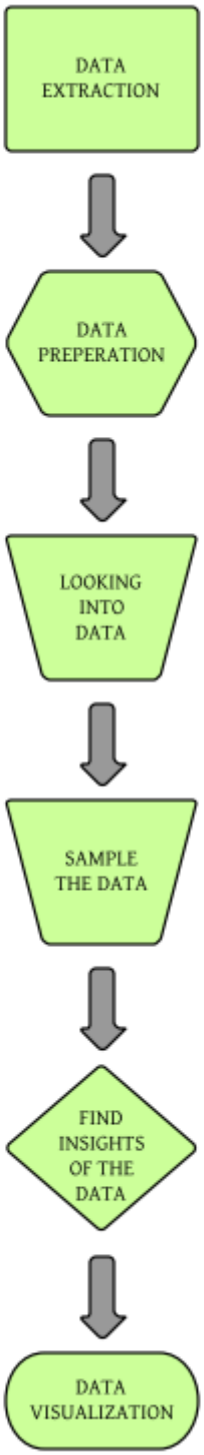
Rainfall is most essential for our life. We predict that rainfall in the certain period. Therefore, we avoid flood, cyclone, forest fire detection, global warming etc. We computed values for rainfall using 100 years input data by using R language and predicted for future years.

One of the aims of this project is to convert raw data into information at present. In recent years, the use of data science and data -mining process in the field of hydrology is increasing. The studies have been performed using DS (Data Science) process in many areas. The main aim of study is to evaluate the use of data –science process to estimate rainfall of San Francisco in California. This study is performed using rainfall data of particular station and station ID.

Rainfall falls in ground level by multiple linear regressions. In future we predict the rainfall forecasting and other applications done by using the artificial intelligence, neural network and fuzzy sets etc. We do the research on public sectors and save the world.

Finally predicted values lie below the computed values. So, it does not show an accurate but shows an approximate value.

3. FLOWCHART


DATA EXTRACTION
DATA PREPERATION
LOOKING INTO DATA
SAMPLE THE DATA
FIND INSIGHTS OF THE DATA
DATA VISUALIZATION

4. HYPOTHESIS/GOALS

4.1 POSITIVE HYPOTHESIS

Precipitation is a major component of the water cycle, and is responsible for depositing most of the fresh water on the planet. Rainfall should be adequate to fill all the reservoirs. For significant amounts of precipitation to fall, there must be an adequate amount of moisture in the atmosphere. There also must be a way to lift this moisture so that it can turn into precipitation. The most common ways for adequate precipitation are mentioned below:

Rain

All precipitation starts out as ice or snow crystals at cloud level. When this frozen precipitation falls into a layer of sufficiently warmer air (with temperatures above freezing) it melts into rain. If this warm air extends all the way to the surface of the earth, rain will fall at ground level.

Freezing rain

Rain droplets that fall into a shallow layer of cold air near the earth's surface can freeze upon contact with the ground, leaving a coating of glaze. This is known as freezing rain.

Snow:

Snow is an aggregate of ice crystals that form into flakes. Snow forms at temperatures below freezing. For snow to reach the earth's surface the entire temperature profile in the troposphere needs to be at or below freezing. It can be slightly above freezing in some layers if the layer is not warm or deep enough the melt the snowflakes much. The intensity of snow is determined by the accumulation over a given time. Categories of snow are light, moderate and heavy.

Hail

Hail is a form of solid precipitation. It forms due to ice crystals and super cooled water that freeze or stick to the embryo hail stone.

4.2 NEGATIVE HYPOTHESIS:

There is a possibility of flood. Flood can be caused due to following reasons.

Winter rain season

Heavy rains are common in many parts of the United States during the winter months. These storms often lead to flooding. Several large wildfires in the United States have dramatically changed the landscape and ground conditions, resulting in fire-scorched land that can lead to flash floods and mudflows under heavy rain. Experts say that it might take years for vegetation, which will help stabilize these areas, to return.

In cold climates across the nation, heavy rain sometimes follows heavy snow. The frozen ground cannot absorb the rain-soaked snow pack, so flooding occurs.

Spring Thaw

During the spring, frozen land prevents melting snow or rainfall from seeping into the ground. Each cubic foot of compacted snow contains gallons of water and once the snow melts, it can result in the overflow of streams, rivers and lakes. Add spring storms to that and the result is often serious, spring flooding.

Ice Jams

Long cold spells can cause the surface of rivers to freeze, leading to ice jams. When a rise in the water level or a thaw breaks the ice into large chunks, these chunks can become jammed at man-made and natural obstructions, resulting in severe flooding.

Heavy Rains

Several areas of the country are at heightened risk for flooding due to heavy rains. This excessive amount of rainfall can happen throughout the year, putting your property at risk.

Snow melts

A midwinter or early spring thaw can produce large amounts of runoff in a short period of time. Because the ground is hard and frozen, water cannot penetrate and be reabsorbed. The water then runs off the surface and flows into lakes, streams and rivers, causing excess water to spill over their banks.

Drought

Droughts are caused by a depletion of precipitation over time. One of the scariest parts of a drought is the onset time. Unlike other forms of severe weather or natural disasters, droughts often develop slowly. There are mainly three different types of drought

Hydrological Drought: Many watersheds experience depleted amounts of available water. Lack of water in river systems and reservoirs can impact hydroelectric power companies, farmers, wildlife, and communities.

Meteorological Drought: A lack of precipitation is the most common definition of drought and is usually the type of drought referred to in news reports and the media. Most locations around the world have their own meteorological definition of drought based on the climate normal in the area. A normally rainy area that gets less rain than usual can be considered in a drought.

Agricultural Drought: When soil moisture becomes a problem, the agricultural industry is in trouble with drought. Shortages in precipitation, changes in evapo-transpiration, and reduced ground water levels can create stress and problems for crops.

Acid Rain

Acid rain is a serious environmental problem that affects large parts of the United States. Acid rain is particularly damaging to lakes, streams, and forests and the plants and animals that live in these ecosystems. Acid rain causes acidification of lakes and streams and contributes to the damage of trees at high elevations and many sensitive forest soils.

Global warming

Air pollution from vehicles, industry and the burning of plant material can choke off the formation of precipitation in some semi-arid mountainous areas, threatening critical water sources. Global warming will result in more frequent hot days and fewer cool days, with the greatest warming occurring over land. Longer, more intense heat waves will become more common. Storms, floods, and droughts will generally be more severe as precipitation patterns change.

5. METHODOLODY

5.1 SOFTWARE REQUIREMENTS

- Microsoft Office (Word, Excel)
- R language

6.IMPLEMENTATION

6.1 TOOL

R is a language and environment for statistical computing and graphics. It is similar to the S language and environment that was developed at Bell Laboratories. R can be considered as a different implementation of S. There are some important differences, but much code written for S runs unaltered under R.

R provides a wide variety of statistical like linear and nonlinear modeling, classical statistical tests, time-series analysis, classification, clustering etc. and graphical techniques, and is highly extensible. R provides an Open Source route to participation vehicle of choice for research in statistical methodology.

One of R's strengths is the ease with which well-designed publication-quality plots can be produced, including mathematical symbols and formulae where needed. Main advantage is its design choices in graphics that are very easy to plot, and the user retains full control.

R is available as Free Software. It compiles and runs on a wide variety of UNIX platforms and similar systems (including FreeBSD and Linux), Windows and MacOS.

R is an integrated suite of software facilities for data manipulation, calculation and graphical display. It includes

- an effective data handling and storage facility,
- a suite of operators for calculations on arrays, in particular matrices,
- a large, coherent, integrated collection of intermediate tools for data analysis,
- graphical facilities for data analysis and display either on-screen or on hardcopy, and
- a well-developed, simple and effective programming language which includes conditionals, loops, user-defined recursive functions and input and output facilities.

6.2 ANALYSIS AND DESIGN DOCUMENT

We will look into the current and historical precipitation data for California through 2013.

- **Data Extraction**:

The long-term daily precipitation (rain/snow, and also temperature) records for the United States are available from the United States Historical Climatology Network - USHCN. These are observations of precipitation for typically about 100 years or longer. The USHCN stations are a subset of the larger/denser (but shorter) observation network from the Global Historical Climatology Network - GHCN. The data for California is available through 2012. The format of each record is in an ASCII data file, be it a state-level file (e.g., state04_CA.txt is California). The sample of data looks like this:

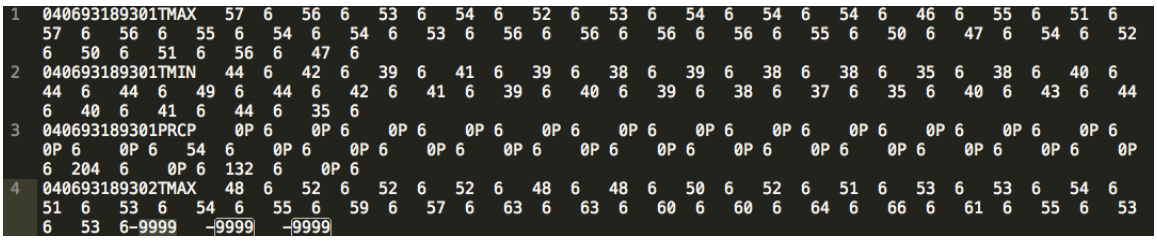
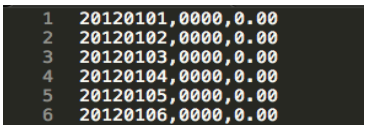


Figure.2 Samples of data

Each record in a file contains one month of daily data. The first field of each record contains Station ID, Year, month, and type of element. For our project, we only need the precipitation (hundredths of inches) data as 'PRCP'. '-9999' stands for missing data.

On the other hand, we can obtain the real time and recent precipitation data from Department of Water Resources California Data Exchange Center website. Through the query tools provided, we can extract the precipitation data for a certain station from 2008 till current in csv file. The precipitation data is in inches. The sample of data is:



We need to compare the rainfall data in 2013 to historical data for a certain area. We choose ITU's address (37.3315073N, 121.895659W) as the target area.

Ideally, we require all the data from the same station. Unfortunately, USHCN stations use the different monitor system from that of California Data Exchange Center. The nearest station from historical data is at Livermore (37.6922N, 121.7692W). As to current data, the nearest station to the Livermore station is called DUBLIN-SAN RAMON FIRE HOUSE (37.732000°N, 121.927000°W).

Because the two locations are only less than 30 miles apart, and share the similar geographic attributes, grouping the two stations' data from history and current and stitching them together is still a valid approximation for this area. So we choose extract the historical data (1903-2012) of Livermore station from CA data and the year 2013 data from Dublin station, then group them as a complete record of rainfall data from 1903 till 2013.

- **Data Normalization and Cleaning**:

The data from two sources are in totally different formats, as we have seen. We load them into R, and treat them separately, then join them together. We only extract the necessary information, i.e., year, month, day, and rainfall. Need to point out here, that the historical precipitation data is in hundredths of inches, however, the current data is in inches. We normalize them into inches. The output data is like this:

```
> head(full_data)
  year month day PRCP
1 1903     1   1 0.03
2 1903     1   2 0.00
3 1903     1   3 0.00
4 1903     1   4 0.00
5 1903     1   5 0.00
6 1903     1   6 0.00
>
```

Figure. 3 Normalized output Data

Before we explore the data, we need to make sure the quality of data. Therefore, we do a complete data sanity check. We check the gaps, duplicate records, missing data, data inaccuracy, and special characters, etc. Except certain missing data, the data is relatively clean. The reason could be that the two sources are already pre-processed for public use. Due to the time span of precipitation data collection, the missing data is unavoidable. The missing labels as '-9999' in historical data, and as 'm' in current data.

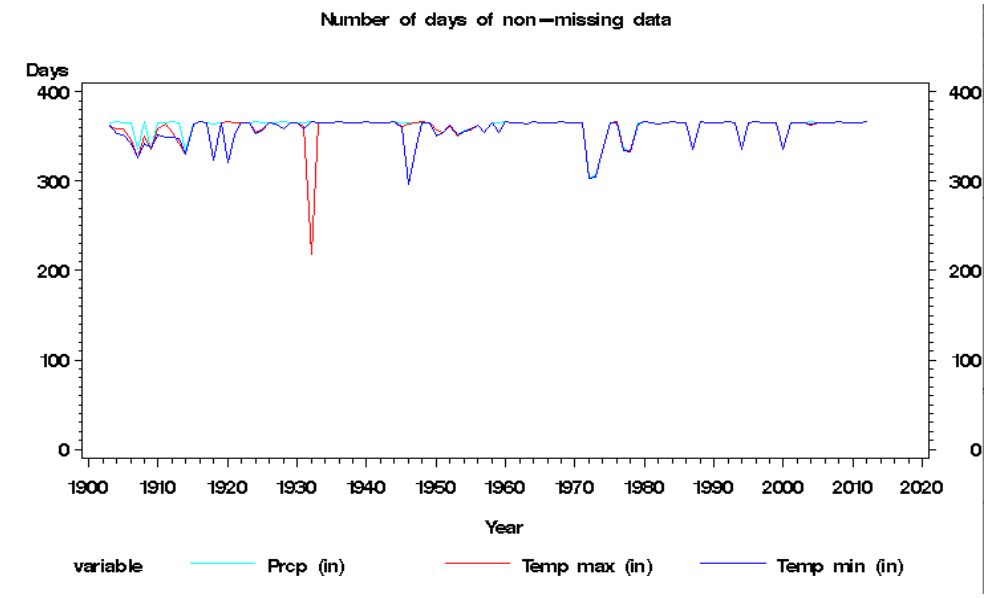


Figure.4 Number of days of Non-Missing Data

We analyze the yearly distribution of number of days with non-missing data (see figure above). The days of missing precipitation (PRCP) data stand only a very small percentage of our data, and will have minimal impact to our analysis (mostly in month/year granularity). Therefore, we choose to remove the missing data records, instead of estimating the values.

- **Data Exploration**:

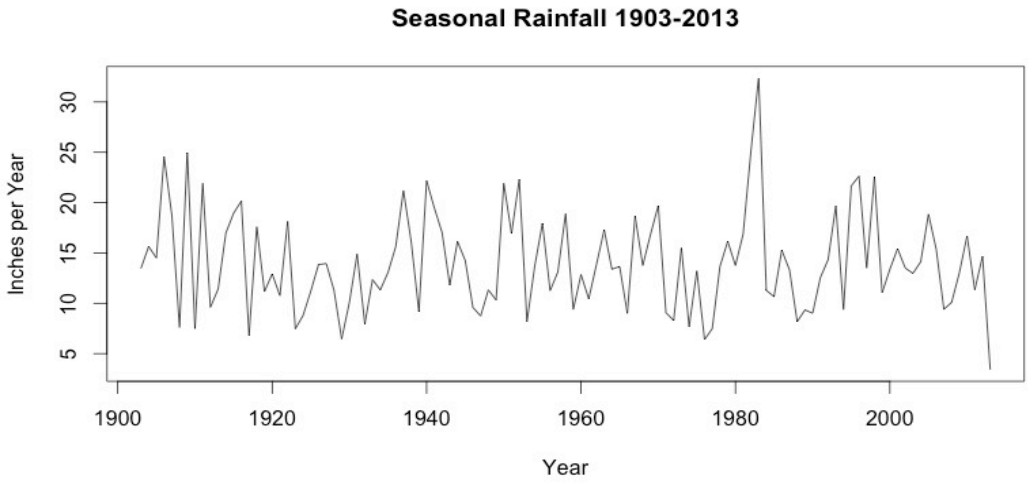


Figure. 5 Season Rainfall for 1903-2013 (inches)

This is the plot of yearly rainfall total (in) from 1903 to 2013 (Fig 1). The average yearly rainfall is 14 inches. The max rainfall happens at 1983 with 32.37 inches, and the min happens at 2013 with only 3.41 inches, which is less than a quarter of the year average.

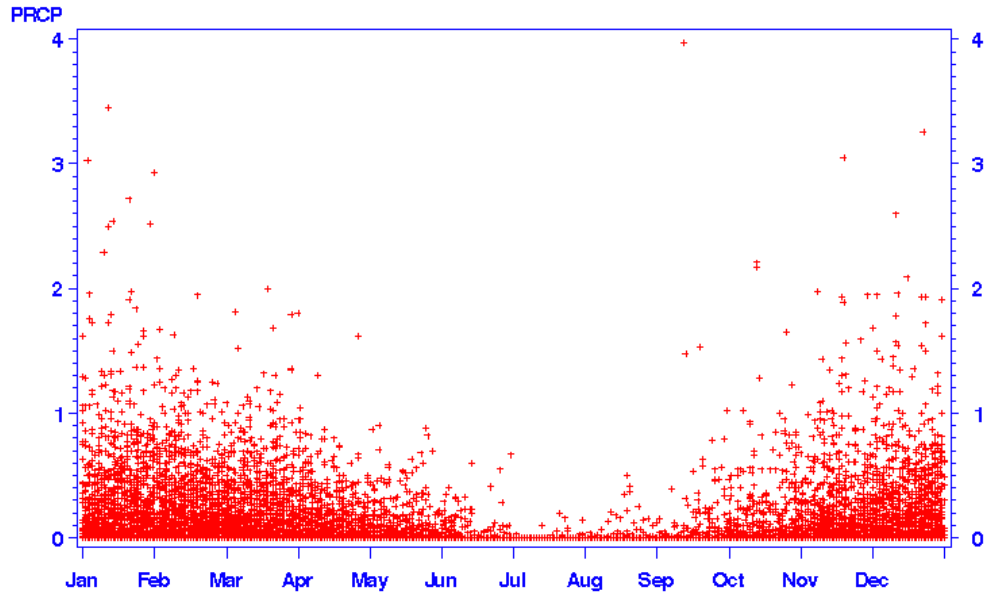


Figure.6 Plot for daily rainfall

We can easily tell that there is a clear rain (Oct-June) and dry (June-Sept) season pattern. Therefore, the United States Geological Survey defines a **water year** as the period between October 1st of one year and September 30th of the next. Figure 7 shows the daily rainfall (in) distribution by water year.

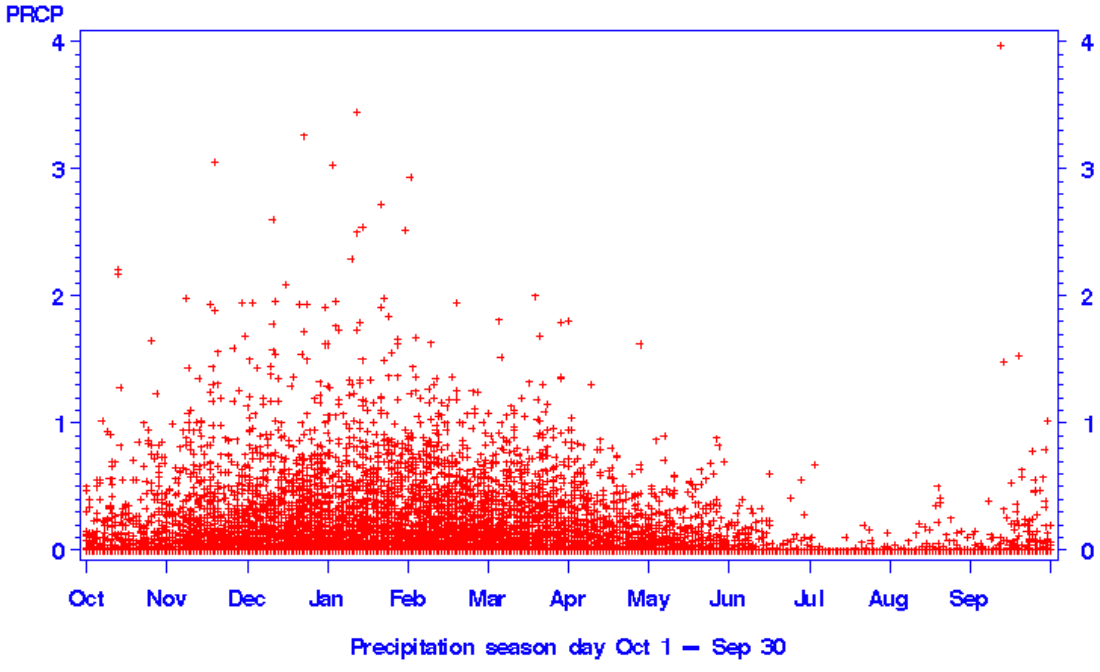


Figure.7 Daily rainfall (in) distribution by water year.

We also plot the yearly-accumulated rainfall from 1903-2012 by a calendar year and a water year (Figure 8 and Figure 9).

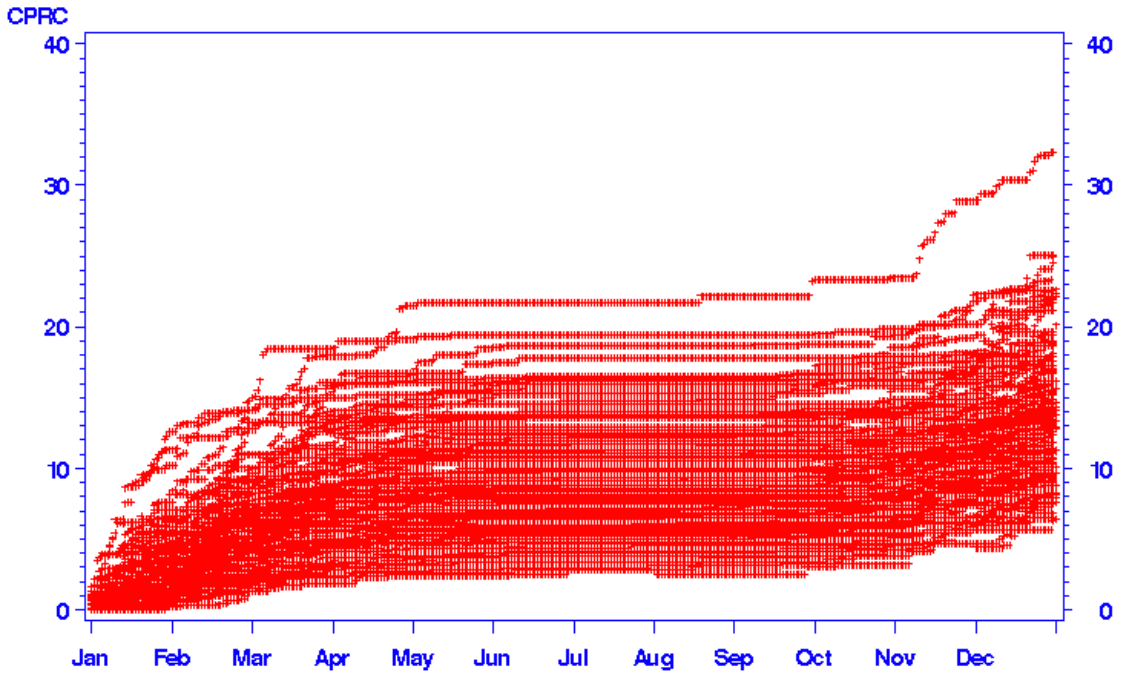


Figure.8 Scatterplot for yearly-accumulated rainfall from 1903-2012

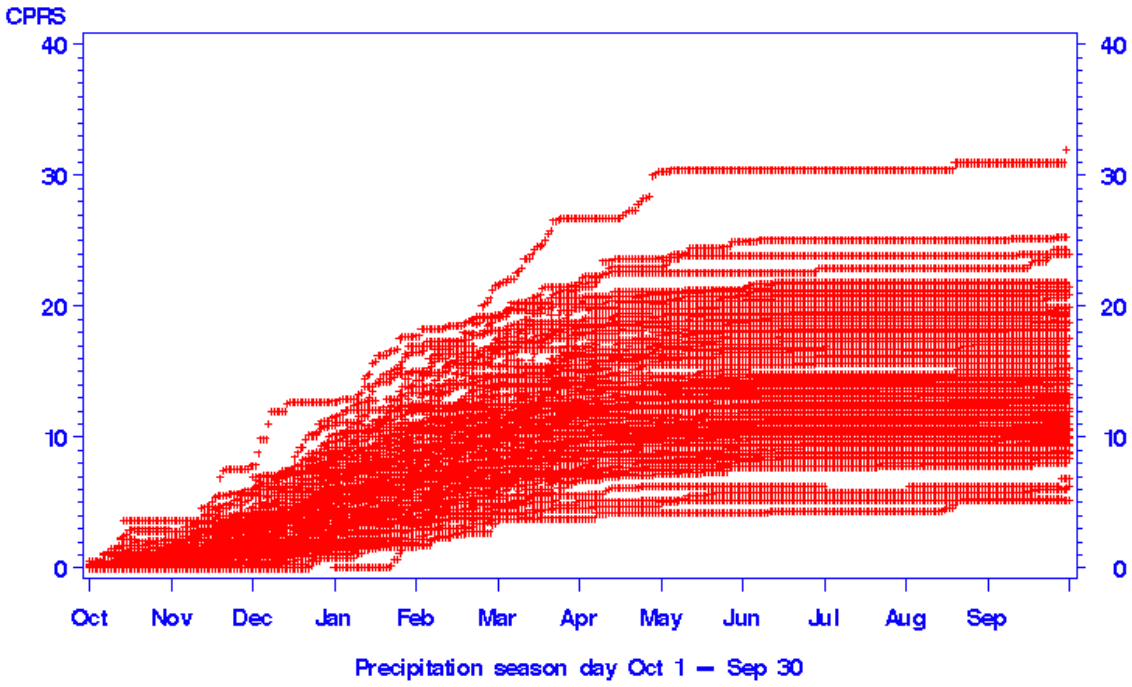


Figure.9 Scatterplot for yearly-accumulated rainfall from 1903-2012

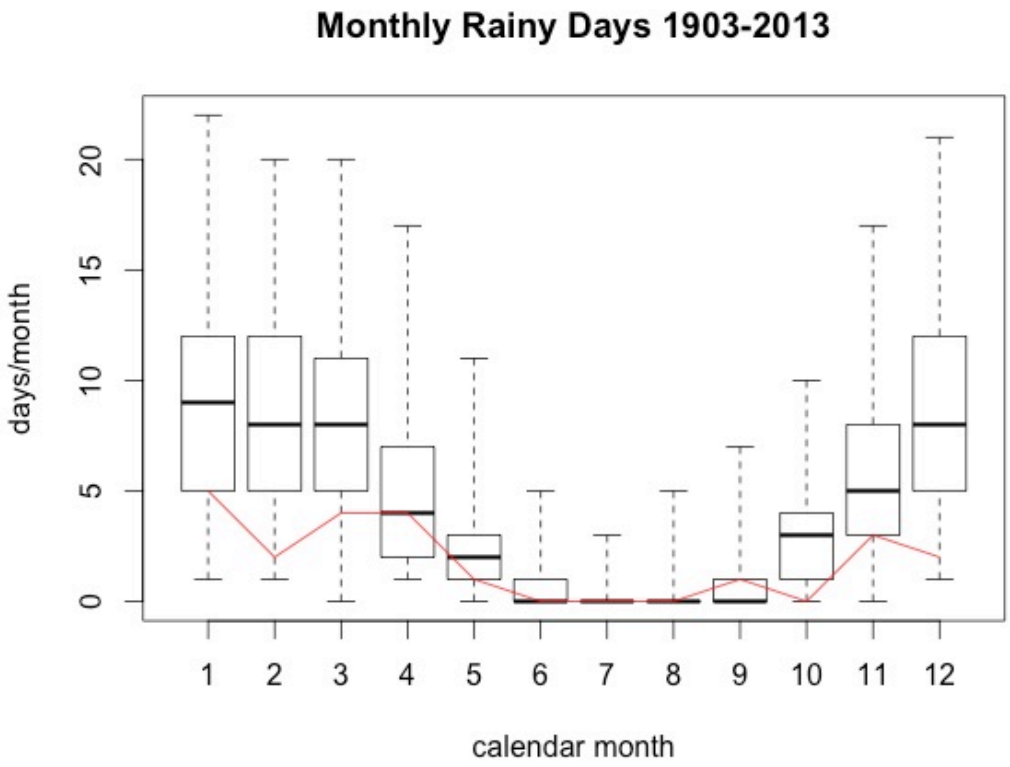


Figure.10 Monthly rainy days

We also look into historical number of rainy days per month and how it compares to that of 2013 (Figure 6). The boxplot is based on data from 1903-2012. The red line shows the rainy days per month in 2013. The thick horizontal line is median, the top and bottom lines stand for max and min rainy days. The box shows the range between the first and third quartiles. From 1903-2012, the number of average rainy days is 52.21, and max is 95 days. **That number for 2013 is 22.**

- **Predictive Modeling**

Based on the historical rainfall data from 1903-2012, could we possibly predict the rainfall for 2013?

We will look into this problem by working on yearly rainfall total data (see Fig 1). First, we try to fit the data with a regression model GAM (Figure 7). GAM (generalized additive model) is a generalized linear model in which the linear predictor depends linearly on unknown smooth functions of some predictor variables, and interest focuses on inference

about these smooth functions. Due to the highly variation of data, the output GAM fitting is basically a linear regression, which doesn't help much.

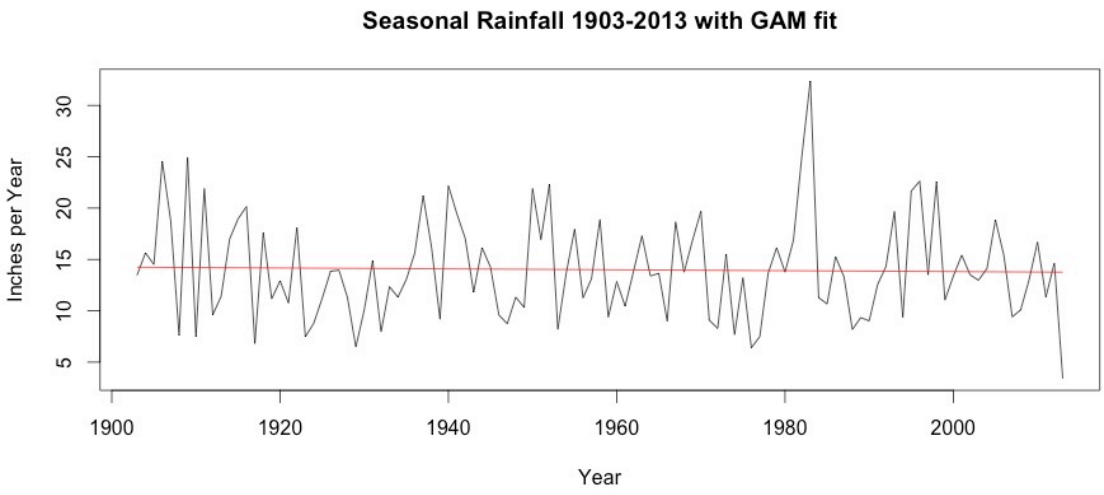


Figure. 11 Seasonal Rainfall 1903-2013 with GAM fit

The yearly rainfall data is a typical kind of time series data, like Apple (AAPL) stock price

Figure. 12 Moving average model

The moving-average (MA) model is a common approach for modeling univariate time series models. For instance, *at time t, a "centered moving average of order 3" with equal weights would be the average of values at times t -1, t, and t+1*.

We will use the basic MA model to smooth the yearly rainfall data to discover the possible pattern or trend.

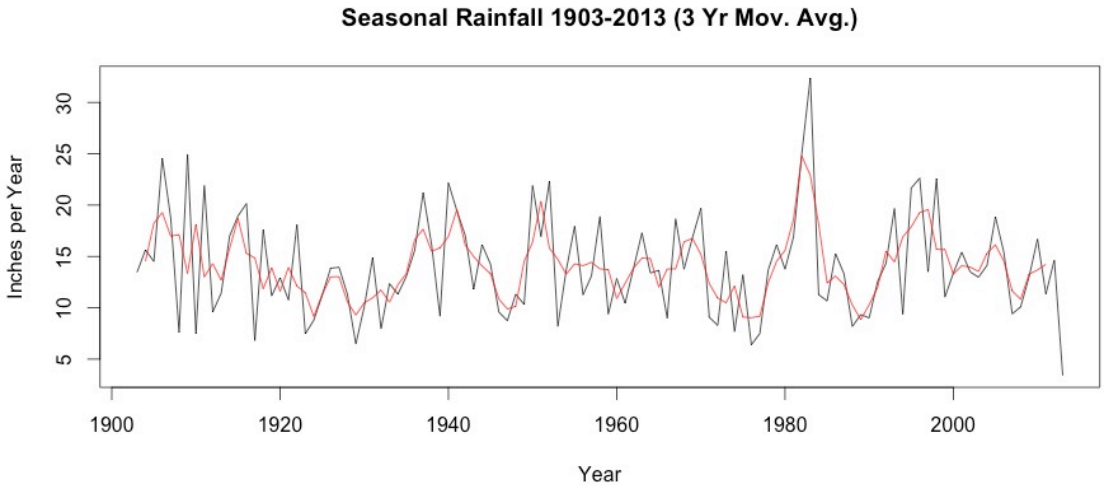


Figure 13 Seasonal rainfall 3 yr moving average

We start out with 3 Year moving average (Fig 11). The 3-Yr MA model does smooth the original data a little bit, but still left with much white noise. Then we increase the order of MA to 5, 10, and 30 Years, (Figures 12, 13, 14) respectively.

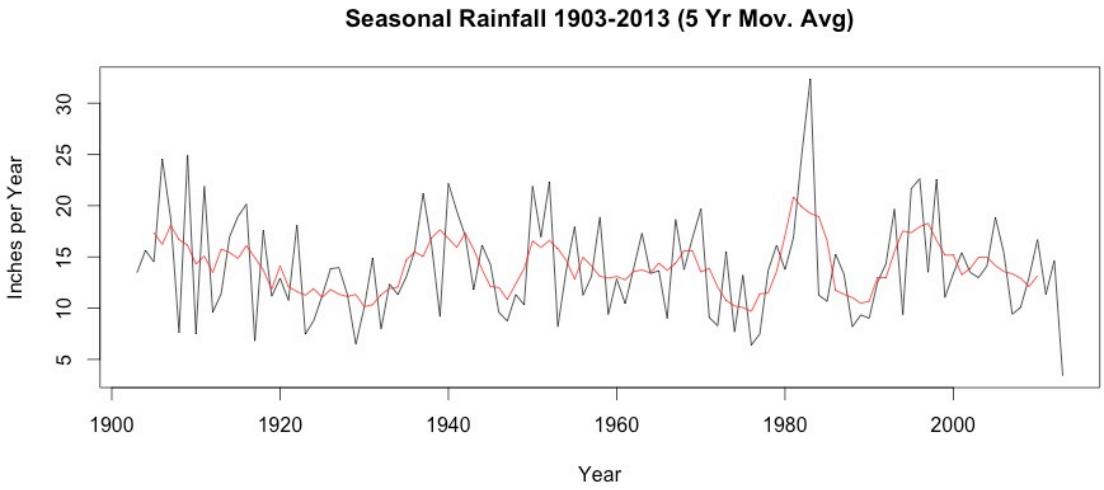


Figure. 14 Seasonal rainfall 5 year moving average

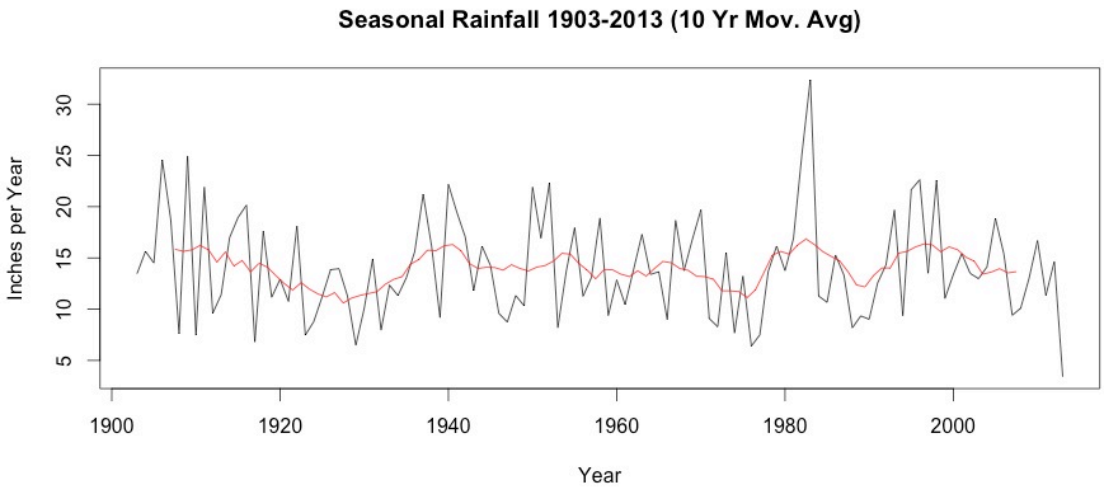


Figure. 15 Seasonal Rainfall moving average for 10 years

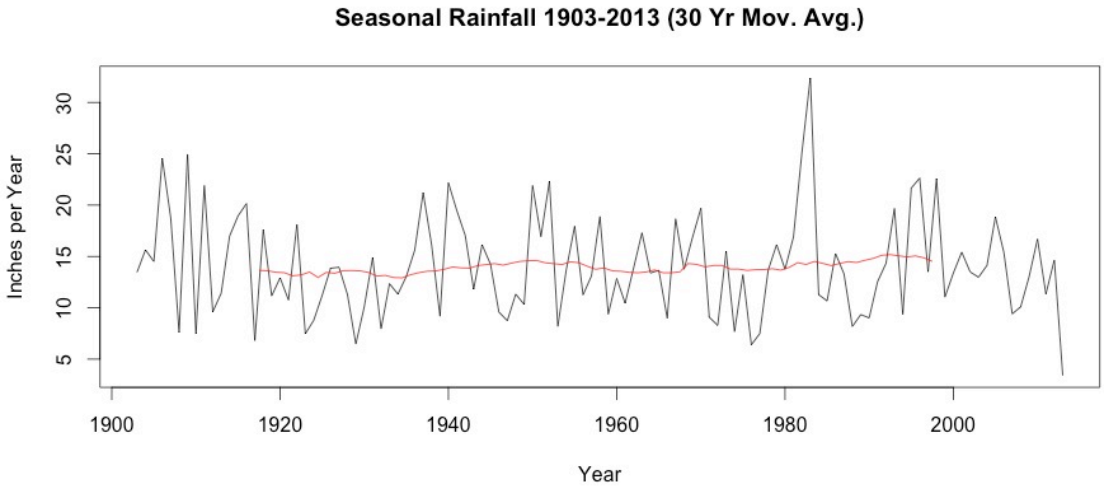


Figure 15 Seasonal Rainfall moving average for 30 years

Out of different of orders of MA, which MA model is a better choice to see the actual trend? We can answer that question by using *autocorrelation function* (ACF) to the original data.

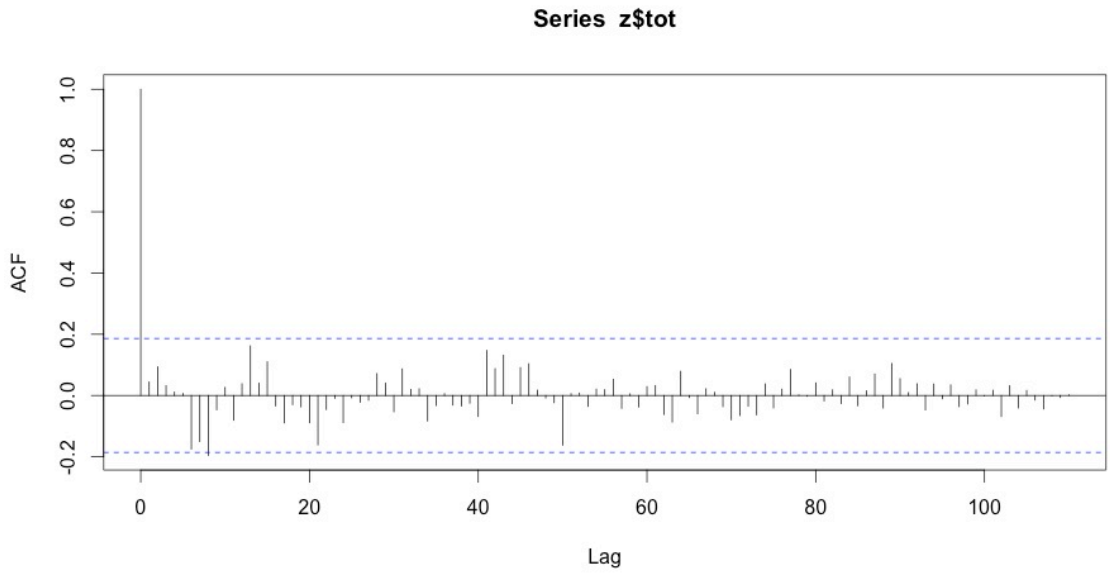


Figure 16 series

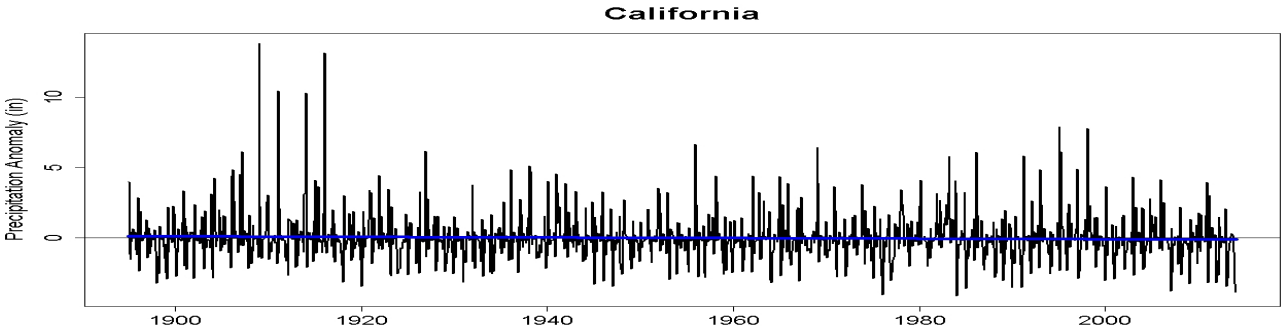
From Figure 16, we can see that most variations fall in the two-sided blue dotted zone. That means the variation is close to white noise, and no conclusions can be drawn. However, we can see a weak 10-year cycle among the data. Therefore, the best MA model is using 10-Year Moving Average. In Figure 11, we found that the actual rainfall data in 2013 is way off the predicted value, based on the trend of 10-Year MA.

7.CONCLUSION

The main goal of this experiment is to estimate rainfall for effective use of water resources and optimal

Planning of water sources.

Rainfall is measured most simply by noting periodically how much has been collected in an exposed vessel since the time of the last observation. Care must be taken to avoid underestimating rainfall due to evaporation of the collected water and the effects of wind Time series can be constructed and analysis performed in a similar manner to those of temperature.



We computed values for rainfall using 100 years input data by using R language and predicted for future years. Climate variability is a reality that is affecting rural livelihood today and presenting a growing challenge in the region, as in many other parts of the continent and elsewhere.

Precipitation has been included Precipitation is a product of condensation of atmospheric water vapor that falls under gravity also considered The meteorological events affect permanently human life. Considering the meteorological phenomena, which have no possibility of intervention, calculations has been made considering hypothesis ,rain snow ,freezing rain, hail and negative hypothesis with all the data extractions set accordingly with the years looking in to the data

Each record in a file contains one month of daily data. The first field of each record contains

Station ID, Year, month, and type of element

8. GLOSSARY

Accurate

Accurately estimate the given data

Authentication

To establish the authenticity of is to prove genuine.

Aggregation

Several things grouped together or considered as a whole

Background

Reasons for the cause

Biometric:

Refers with the historical data

Context

Specific problems, understanding data, data preparation

Constraint

A limitation or Restriction

Definition

Product of condensation

Explanation

Expected outcome of the process

Extraction

Scientific images with forecast

9.BIBLIOGRAPHY

- http://www.enggjournals.com/ijet/docs/IJET10-02-06-28.pdf
- http://www.slideshare.net/csandit/rainfall-prediction-using-data
- http://cdec.water.ca.gov/
- http://cdiac.ornl.gov/ftp/ushcn_daily/
- https://github.com/RationShop/rain_prism