CHAPTER 1

INTRODUCTION

1.1 Introduction

Statistics and geography are closely related disciplines that often work together to analyze spatial data and understand patterns and relationships in the world around us. Statistics plays a crucial role in weather prediction by enabling meteorologists to analyze vast amounts of meteorological data and use statistical methods to model, forecast, and interpret weather patterns. Statistical techniques are employed in data collection and quality control to ensure the accuracy of data from weather stations, satellites, and radars. Time series analysis helps identify trends and cycles in the data for future predictions. Statistical methods also aid in model development, data assimilation, and ensemble forecasting, which are vital for improving forecast accuracy and accounting for uncertainty. Overall, statistics is essential for refining weather prediction models and evaluating their performance, ultimately contributing to more reliable and accurate forecasts.

1.2 Weather and Climate

Earth's atmosphere has no definite boundary, gradually becoming thinner and fading into outer space. Three-quarters of the atmosphere's mass is contained within the first 11 km (6.8 mi) of the surface; this lowest layer is called the troposphere. Energy from the Sun heats this layer, and the surface below, causing expansion of the air. This lower-density air then rises and is replaced by cooler, higher-density air. The result is atmospheric circulation that drives the weather and climate through redistribution of thermal energy**.**

The primary atmospheric circulation bands consist of the trade winds in the equatorial region below 30° latitude and the westerlies in the mid-latitudes between 30° and 60°. Ocean heat content and currents are also important factors in determining climate, particularly the thermohaline circulation that distributes thermal energy from the equatorial oceans to the polar regions. Further factors that affect a location's climates are its proximity to oceans, the oceanic and atmospheric

circulation, and topology. Places close to oceans typically have colder summers and warmer winters, due to the fact that oceans can store large amounts of heat. The wind transports the cold or the heat of the ocean to the land. Atmospheric circulation also plays an important role: San Francisco and Washington are both coastal cities at about the same latitude. San Francisco's climate is significantly more moderate as the prevailing wind direction is from sea to land. Finally, temperatures decrease with height causing mountainous areas to be colder than low-lying areas.

Water vapor generated through surface evaporation is transported by circulatory patterns in the atmosphere. When atmospheric conditions permit an uplift of warm, humid air, this water condenses and falls to the surface as precipitation. Most of the water is then transported to lower elevations by river systems and usually returned to the oceans or deposited into lakes. This water cycle is a vital mechanism for supporting life on land and is a primary factor in the erosion of surface features over geological periods. Precipitation patterns vary widely, ranging from several meters of water per year to less than a millimeter. Atmospheric circulation, topographic features, and temperature differences determine the average precipitation that falls in each region.

1.3 Global Climatology

There are different types of landforms which are connected with climate throughout the year. That are described as follows,

Deserts: Northern half of Africa is dominated by the world's most extensive hot, dry region, the Sahara Desert. Some deserts also occupy much of southern Africa: the Namib and the Kalahari. Across Asia, a large annual rainfall minimum, composed primarily of deserts, stretches from the Gobi Desert in Mongolia west-southwest through western Pakistan and Iran into the Arabian Desert in Saudi Arabia. Most of Australia is semi-arid or desert, making it the world's driest inhabited continent. In South America, the Andes Mountain range blocks Pacific moisture that arrives in that continent, resulting in a desert-like climate just downwind across western Argentina. The drier areas of the United States are regions where the Sonoran Desert overspreads the Desert Southwest, the Great Basin and central Wyoming.

Polar deserts: Since rain only falls as liquid, it rarely falls when surface temperatures are below freezing, unless there is a layer of warm air aloft, in which

case it becomes freezing rain. Due to the entire atmosphere being below freezing most of the time, very cold climates see very little rainfall and are often known as polar deserts. A common biome in this area is the tundra which has a short summer thaw and a long frozen winter. Ice caps see no rain at all, making Antartica the world's driest continent.

Rain forests: Rain forests are areas of the world with very high rainfall. Both tropical and temperate rainforests exist. Tropical rainforests occupy a large band of the planet mostly along the equator. Most temperate rainforests are located on mountainous west coasts between 45 and 55 degree latitude, but they are often found in other areas. Around 40–75% of all biotic life is found in rainforests. Rainforests are also responsible for 28% of the world's oxygen turnover.

Monsoons: The equatorial region near the Inter tropical Convergence Zone (ITCZ), or monsoon trough, is the wettest portion of the world's continents. Annually, the rain belt within the tropics marches northward by August, then moves back southward into the Southern Hemisphere by February and March. Within Asia, rainfall is favored across its southern portion from India east and northeast across the Philippines and southern China into Japan due to the monsoon advecting moisture primarily from the Indian Ocean into the region. The monsoon trough can reach as far north as the 40th parallel in East Asia during August before moving southward thereafter. Its poleward progression is accelerated by the onset of the summer monsoon which is characterized by the development of lower air pressure (a thermal low) over the warmest part of Asia. Similar, but weaker, monsoon circulations are present over North America and Australia.

1.4 Rain

Rain is water droplets that have condensed from atmospheric water vapor and then fall under gravity. Rain is a major component of the water cycle and is responsible for depositing most of the fresh water on the Earth. It provides water for hydroelectric power plants, crop irrigation, and suitable conditions for many types of ecosystems.

The major cause of rain production is moisture moving along three-dimensional zones of temperature and moisture contrasts known as weather fronts. If enough moisture and upward motion is present, precipitation falls from convective

clouds (those with strong upward vertical motion) such as cumulonimbus (thunder clouds) which can organize into narrow rainbands. In mountainous areas, heavy precipitation is possible where upslope flow is maximized within windward sides of the terrain at elevation which forces moist air to condense and fall out as rainfall along the sides of mountains. On the leeward side of mountains, desert climates can exist due to the dry air caused by downslope flow which causes heating and drying of the air mass. The movement of the monsoon trough, or intertropical convergence zone, brings rainy seasons to savannah climes.

The urban heat island effect leads to increased rainfall, both in amounts and intensity, downwind of cities. Global warming is also causing changes in the precipitation pattern globally, including wetter conditions across eastern North America and drier conditions in the tropics. Antarctica is the driest continent. The globally averaged annual precipitation over land is 715 mm (28.1 in), but over the whole Earth, it is much higher at 990 mm (39 in). Climate classification systems such as the Köppen classification system use average annual rainfall to help differentiate between differing climate regimes. Rainfall is measured using rain gauges. Rainfall amounts can be estimated by weather radar.

1.5 Measurement and Uses

Rain is measured in units of length per unit time, typically in millimeters per hour, or in countries where imperial units are more common, inches per hour. The "length", or more accurately, "depth" being measured is the depth of rain water that would accumulate on a flat, horizontal and impermeable surface during a given amount of time, typically an hour. One millimeter of rainfall is the equivalent of one liter of water per square meter. Commonly, rain gauge is the instrument which is used to measure the amount of rain. A rain gauge (also known as udometer, pluviometer, pluviameter, ombrometer, and hyetometer) is an instrument used by meteorologists and hydrologists to gather and measure the amount of liquid precipitation over a predefined area, over a period of time. It is used to determine the depth of precipitation (usually in mm) that occurs over a unit area and measure rainfall amount.

The significance of rainfall prediction is to use science and technology to forecast the amount of rainfall over a specific region and also for

- ***Water Resource Management***: Precise rainfall forecasts help in managing water resources effectively. By knowing when and how much rain is expected, authorities can plan water storage, distribution, and usage efficiently.
- ***Crop Productivity***: Farmers rely on rainfall for crop irrigation. Accurate predictions allow them to plan planting and harvesting schedules, optimize irrigation, and enhance crop productivity.
- ***Flood Avoidance***: Timely and accurate rainfall forecasts aid in flood prevention. By monitoring rainfall patterns, authorities can take preventive measures to mitigate flood risks, safeguard lives, and protect property.
- ***Urban Planning***: Urban areas need to prepare for heavy rainfall events. Predictions help city planners design drainage systems, flood control measures, and emergency response plans.
- ***Disaster Management****:* Reliable rainfall forecasts assist disaster management agencies in preparing for natural calamities. Early warnings enable evacuation plans and resource allocation during extreme weather condition.

CHAPTER 2

RAINFALL DISTRIBUTION

2.1 Introduction

Forecasting of weather is the application of science and technology to predict the conditions of the atmosphere for a given location and time. People have attempted to predict the weather informally for millennia and formally since the 19th century. Weather forecasts are made by collecting quantitative data about the current state of the atmosphere, land, and ocean and using meteorology to project how the atmosphere will change at a given place.

Once calculated manually based mainly upon changes in barometric pressure, current weather conditions, and sky conditions or cloud cover, weather forecasting now relies on computer-based models that take many atmospheric factors into account. Human input is still required to pick the best possible model to base the forecast upon, which involves pattern recognition skills, teleconnections, knowledge of model performance, and knowledge of model biases.

The inaccuracy of forecasting is due to the chaotic nature of the atmosphere, the massive computational power required to solve the equations that describe the atmosphere, the land, and the ocean, the error involved in measuring the initial conditions, and an incomplete understanding of atmospheric and related processes. Hence, forecasts become less accurate as the difference between the current time and the time for which the forecast is being made (the range of the forecast) increases. The use of ensembles and model consensus helps narrow the error and provide confidence in the forecast.

There is a vast variety of end uses for weather forecasts. Weather warnings are important because they are used to protect lives and property. Forecasts based on temperature and precipitation are important to agriculture, and therefore to traders within commodity markets. Temperature forecasts are used by utility companies to estimate demand over coming days. On an everyday basis, many people use weather forecasts to determine what to wear on a given day. Since outdoor activities are severely curtailed by heavy rain, snow and wind chill, forecasts can be used to plan activities around these events, and to plan ahead and survive them.

2.2 Rainfall distribution in India

India is a vast country in geographical terms, with various regions experiences very different climatic conditions. This is also reflected in the distribution of rainfall in India. Some regions experience very high rainfall and others receive very scanty rainfall. The difference between the recorded highest and lowest rainfall in India is approximately 1178 cm. Precipitation in India is irregular over the course of a year, with a well-defined rainy season over most of the country starting in about June and ending in September. According to the Koppen climate classification, it has seven different climatic regions:

- Tropical semi-arid
- Sub-tropical arid desert
- Sub-tropical semi-arid
- Tropical rainforest
- Tropical Savannah
- Sub-tropical humid
- Alpine

The average rainfall in India is 118 cm according to annual data from the Meteorological Department. The following is the distribution of rainfall in India:

Extreme Precipitation regions*:* North-eastern regions and the windward side of the Western ghats experience an average of 400 cm of annual rainfall. Areas like Assam, Meghalaya, Arunachal Pradesh and hilly tracts of the Western Ghats are host to tropical rainforests. The highest rainfall in India and the world is recorded at Mawsynram village of Meghalaya.

Heavy Precipitation regions: The regions experiencing 200-300 cm rainfall belong to this zone. Most of Eastern India is covered under this zone. These regions are also home to tropical rainforests. States such as West Bengal, Tripura, Nagaland, Manipur, Odisha and Bihar are included in this zone. Most of the areas in the sub-Himalayan belt also fall under this zone.

Moderate Precipitation regions: Areas which experience 100 to 200 cm of rainfall include parts of West Bengal, Bihar, Odisha, Madhya Pradesh, Andhra Pradesh, and

the leeward side of the Western Ghats. Wet Deciduous forests comprise the most common natural vegetation of these regions.

Scanty Precipitation regions: Areas having 50 to 100 cm of rainfall consisting of parts of Maharashtra, Gujarat, Karnataka, Tamil Nadu, Andhra Pradesh, Madhya Pradesh, Punjab, Haryana and Western Uttar Pradesh. Tropical Grasslands, Savannah and Dry Deciduous forests are commonly found in these areas.

Desert and Semi-desert Regions: These are the areas that receive below 50 cm of rainfall. The states of Rajasthan, Gujarat and adjacent areas are classified as desert or semi-desert based on the amount of rainfall they receive. Some parts of Jammu & Kashmir such as the Ladakh plateau are also included in this zone as cold deserts. The vegetation consists of hardy species which can withstand extended droughts. Some areas like parts of Gujarat have Savannah vegetation in the wetter regions. The lowest rainfall in India has been recorded in Ruyli village, Rajasthan.

2.3 Climate in Southern India

The region has a tropical climate and depends on monsoons for rainfall. According to the Köppen climate classification, it has a non-arid climate with minimum mean temperatures of 18 °C (64 °F). The most humid is the tropical monsoon climate characterized by moderate to high year-round temperatures and seasonally heavy rainfall above 2,000 mm (79 in) per year. The tropical climate is experienced in a strip of south-western lowlands abutting the Malabar Coast, the Western Ghats and the Lakshadweep islands.

A tropical wet and dry climate, drier than areas with a tropical monsoon climate, prevails over most of the inland peninsular region except for a semi-arid rain shadow east of the Western Ghats. Winter and early summer are long dry periods with temperatures averaging above 18 °C (64 °F); summer is exceedingly hot with temperatures in low-lying areas exceeding 50 °C (122 °F); and the rainy season lasts from June to September, with annual rainfall averaging between 750 and 1,500 mm (30 and 59 in) across the region.

Only the southwest monsoon and the northeast monsoon determines the weather condition mainly southern states of India whereas three sides are surrounded by water bodies.

2.3.1 Southwest Monsoon

The southwestern summer monsoons occur from July through September. The Thar Desert and adjoining areas of the northern and central Indian subcontinent heat up considerably during the hot summers. This causes a low pressure area over the northern and central Indian subcontinent. To fill this void, the moisture-laden winds from the Indian Ocean rush into the subcontinent. These winds, rich in moisture, are drawn towards the Himalayas. The Himalayas act like a high wall, blocking the winds from passing into Central Asia, and forcing them to rise. As the clouds rise, their temperature drops, and precipitation occurs. Some areas of the subcontinent receive up to 10,000 mm (390 in) of rain annually. The southwest monsoon is generally expected to begin around the beginning of June and fade away by the end of September. The moisture-laden winds on reaching the southernmost point of the Indian Peninsula, due to its topography, become divided into two parts: the Arabian Sea Branch and the Bay of Bengal Branch.

*The **Arabian Sea Branch*** of the Southwest Monsoon first hits the Western Ghats of the coastal state of Kerala, India, thus making this area the first state in India to receive rain from the Southwest Monsoon. This branch of the monsoon moves northwards along the Western Ghats (Konkan and Goa) with precipitation on coastal areas, west of the Western Ghats. The eastern areas of the Western Ghats do not receive much rain from this monsoon as the wind does not cross the Western Ghats.

*The **Bay of Bengal Branch*** of Southwest Monsoon flows over the Bay of Bengal heading towards north-east India and Bengal, picking up more moisture from the Bay of Bengal. The winds arrive at the Eastern Himalayas with large amounts of rain. Mawsynram, situated on the southern slopes of the Khasi Hills in Meghalaya, India, is one of the wettest places on Earth. After the arrival at the Eastern Himalayas, the winds turns towards the west, travelling over the Indo-Gangetic Plain at a rate of roughly 1–2 weeks per state, pouring rain all along its way. June 1 is regarded as the date of onset of the monsoon in India, as indicated by the arrival of the monsoon in the southernmost state of Kerala.

2.3.2 Northeast Monsoon

Around September, with the sun retreating south, the northern landmass of the Indian subcontinent begins to cool off rapidly, and air pressure begins to build

over northern India. The Indian Ocean and its surrounding atmosphere still hold their heat, causing cold wind to sweep down from the Himalayas and Indo-Gangetic Plain towards the vast spans of the Indian Ocean south of the Deccan peninsula. This is known as the Northeast Monsoon or Retreating Monsoon.

While travelling towards the Indian Ocean, the cold dry wind picks up some moisture from the Bay of Bengal and pours it over peninsular India and parts of Sri Lanka. Cities like Chennai, which get less rain from the Southwest Monsoon, receive rain from this Monsoon. About 50% to 60% of the rain received by the state of Tamil Nadu is from the Northeast Monsoon. In Southern Asia, the northeastern monsoons take place from October to December when the surface high-pressure system is strongest. The jet stream in this region splits into the southern subtropical jet and the polar jet. The subtropical flow directs northeasterly winds to blow across southern Asia, creating dry air streams which produce clear skies over India. Meanwhile, a low pressure system known as a monsoon trough develops over South-East Asia and Australasia and winds are directed toward Australia. In the Philippines, northeast monsoon is called Amihan

2.4 Objectives of the study

The main objectives of this study is to analyse the rainfall distribution in southern states of India from 1951 to 2017 and

- To present the data in the form of tables
- To visualize the data using line graph
- To test the significance difference between the means of two states
- To check the control limits of the data
- To deduce the trend for future
- To test stationarity
- To use different forecasting methods
- To find the best method.

CHAPTER 3

METHODOLOGY

3.1 Introduction

Methodology is the systematic process of gathering and analysing data to advance knowledge. It involves defining research problems, formulating hypotheses, collecting and organizing data, using deductive reasoning, and testing conclusions. Researchers employ various methods such as theoretical procedures, experiments, and statistical analyses. The objectives of research include exploration, description, explanation, application, and theory development. In essence, research methodology serves as the compass guiding the pursuit of knowledge.

3.2 Functions of Statistics

Prof R. A. Fisher defines statistic as, "The science of statistics is essentially a branch of applied mathematics and may be regarded as a mathematics applied to observational data". This section deals with the function of statistics. These are mainly classified into four categories.

- Collection of data
- Presentation of data
- Analysis of data
- Interpretation of data

3.2.1 Collection of data:

Data collection is one of the most important aspects of research. There are two types one is primary data and other one is secondary data.

Primary data: Primary data are those which are collected from the units or individuals directly and those data have never been used for any purpose earlier.

Secondary data: The data which had been collected by some individuals or agency and statistically treated to draw certain conclusions. Again, the same data are used to draw certain conclusions and analyzed to extract some other information are termed as secondary data.

3.2.2 Presentation of data

After the data has been systematically collected and edited, the next step is classification of data. Classification is a process of arranging data into different classes according to their resemblances and affinities. The objectives of classification of data are Simplifying the raw data, facilitating comparison, Duplicating the silent features of data, Making the data more intelligent, eliminating unnecessary details, Facilitating statistical interpretation.

3.2.3 Classification of data

Classification is the process of arranging things or item in group or class. According to their resemblance and affinities and give expression to the units of attributes that may subsist amongst the diversity of individuals. The classification mainly deals with

Geographical classification, i.e., in relation to place

Chronological classification, i.e., on the basis of time

Qualitative classification, i.e., according to some attributes

Quantitative classification, i.e., based on figures or characteristics

3.2.4 Tabulation of data

Tabulation is the process of presenting data collected through survey, experiment or record in rows and columns. So that it can more easily understood and it can be used for further statistical analysis. There are five parts of tables,

Title: This is a brief description of the contents and is shown at the top of the table.

Stubs: The extreme left part of the table are descriptions of rows are shown is called stubs.

Caption and Box-head: The upper part of table which the description of columns is called caption. Units of measurements and column-numbers, if any, are called Box-head.

Body: It is a part of the table which shows the figure.

Source note: This is the part below the Body, where the source of data and any explanations are shown.

3.2.5 Analysis of data

Analysis of data also known as data analysis, is a process of inspecting, cleansing, transforming and modelling data with the goal of discovering useful information, suggesting, conclusions and supporting decision-making. Data analysis has multiple facts and approaches. In the data analysis for this study has carried out based statistical software like Excel, SPSS, Minitab and R software.

I. Statistical Package for the Social Sciences (SPSS)

The Statistical Package for the Social Sciences is a widely used program for statistical analysis in social sciences, particularly in education and research. However, because of its potential, it is also widely used by market researchers, health-care researchers, survey organizations, governments and, most notably, data miners and big data professionals, which allows the user to do case selection, create derived data and perform file reshaping. Another feature is data documentation, which stores a metadata dictionary along with the data file.

II. Minitab

Minitab is a statistics package developed at the Pennsylvania State University by researchers Barbara F. Ryan, Thomas A. Ryan, Jr., and Brian L. Joiner in 1972. Minitab is a statistical software package used for data analysis, statistical testing, and quality improvement. It is commonly used in various industries, including manufacturing, healthcare, finance, and education, for tasks such as statistical process control, hypothesis testing, regression analysis, and Six Sigma projects

III. R (Programming Language)

R is a programming language and free software environment for statistical computing and graphics that is supported by the R Foundation for Statistical computing. The R language is widely used among statisticians and data miners for developing statistical software and data analysis. Polls, surveys of data miners, and studies of scholarly literature database show that R's popularity has increased substantially in recent years. As of January 2018, R ranks 13th in the TIOBE index. The source code for the R software environment is written primarily in C, Fortan, and R. It is freely available under the GNU General Public License, and pre-

compiled binary versions are provided for various operating systems. While R has a command line interface, there are several graphical front-ends available.

IV. Microsoft Excel

Microsoft Excel is a powerful spreadsheet software developed by Microsoft, commonly utilized for data analysis, financial modelling, statistical analysis, and more. It allows users to organize data into worksheets within workbooks, manipulate data using functions and formulas, create various charts and graphs for visualization, and perform tasks such as sorting, filtering, and pivot tables. Excel offers a wide range of formatting options for customization and supports data import/export from external sources. Its versatility and ease of use make it a popular choice in businesses, educational institutions, and personal finance management.

3.2.6 Interpretation of data

Interpretation of data refers to the process of analyzing and making sense of the information collected from various sources. This process involves examining the data to identify patterns, trends, relationships, and insights that can provide valuable information or answer specific questions. Interpretation of data often involves statistical analysis, visualization techniques such as charts and graphs, and contextual understanding of the subject matter. In essence, data interpretation transforms raw data into meaningful insights that can inform decision-making, problem-solving, or further research. It requires critical thinking, domain knowledge, and an understanding of the context in which the data was collected. Effective data interpretation can help uncover hidden patterns, validate hypotheses, identify outliers or anomalies, and ultimately extract actionable insights that drive informed decisions and actions.

3.3 Diagrams

A graphical is a pictorial presentation of the relationship between variables. Many types of graphs are employed in statistics, depending on the data is involved and the purpose of graphs is intended. Sometimes referred to as charts or diagrams.

Bar diagram*:* A bar diagram represents the magnitude of a single factor according to time period, places, items etc. But when the magnitude of the factors, each bar is

further sub-divided into components in proportion to the magnitude of the sub-factors. Such a diagram is known as a sub divided bar diagram.

Multiple bar diagram: A multiple bar diagram is used to denote more than one phenomenon, e.g., for import and export trend. Multiple bars are useful for direct comparison between two values. The bars are drawn side by side. In order to distinguish the bars, different colours, shades, etc., may be used and a key index to this effect be given to understand the different bar.

Line diagram: A line diagram is a one-dimensional diagram in which the highest of line represents the frequency corresponding to the value of the items or a factor. This is the simplest of all the diagrams. On the basis of size of the figures, heights of bars or lines are drawn. The distance between lines is kept uniform. It makes comparison easy. This diagram is not attractive; hence it is less important.

Pie diagram: A pie diagram is a circular diagram which is usually used for depicting the components of a single factor. The circle is divided into segments which are in proportion to the size of the components. They are shown by different patterns or colors to make them attractive.

3.4 Descriptive Statistics

Descriptive statistics refers to the branch of statistics that focuses on summarizing and describing the characteristics of a dataset. It provides simple numerical summaries or visual representations that help in understanding the main features of the data. Descriptive statistics do not involve making inferences or drawing conclusions beyond the data at hand; instead, they aim to describe and summarize the data in a meaningful way. Common measures of descriptive statistics include:

Measures of Central Tendency**:** These statistics indicate the central or average value of a dataset. The most common measures of central tendency are the mean, median, and mode. The *mean* is the arithmetic average of all the values in the dataset. The *median* is the middle value in a sorted dataset. The *mode* is the value that appears most frequently in the dataset.

Measures of Variability or Dispersion: These statistics quantify the spread or variability of the data points around the central tendency. Common measures of variability include the range, variance, and standard deviation. The *range* is the difference between the maximum and minimum values in the dataset. The *variance* measures the average squared deviation of each data point from the mean. The *standard deviation* is the square root of the variance and provides a measure of the dispersion of data points around the mean.

***Measures of Distribution Shape*:** These statistics describe the shape of the distribution of the data. They include skewness and kurtosis. *Skewness* measures the asymmetry of the distribution. A positive skew indicates that the distribution is skewed to the right, while a negative skew indicates a skew to the left. *Kurtosis* measures the peakedness or flatness of the distribution. It indicates whether the distribution has heavy tails or is more concentrated around the means.

***Frequency Distributions*:** Descriptive statistics also include visual representations such as histograms, frequency polygons, and bar charts, which show the frequency of different values or intervals in the dataset.

3.5 t-test:

The t-test is a statistical method used to determine if there is a significant difference between the means of two groups. It's commonly employed when you have two sets of data points and want to assess whether the difference between their means is likely due to chance or if it's statistically significant. There are different types of t-tests, and the choice of which one to use depends on the nature of your data and the hypothesis you want to test.

Independent Samples t-test:

- Used when comparing the means of two independent groups.
- Assumptions include normality of data and homogeneity of variances between groups.
- The formula for the t-statistic in this case is: $t = \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{\frac{s_1^2}{n_1^2} - \frac{s_2^2}{n_2^2}}}$

Where n_1 and n_2 are the sample sizes, $\bar{x}_1$ and $\bar{x}_2$ are means of sample sizes and S_1 and S_2 are standard deviations.

3.6 Control charts:

Control charts, also known as Shewhart charts or process behaviour charts, are a statistical tool used in quality control to monitor processes and detect any changes or abnormalities in the process. They were developed by Walter A. Shewhart in the 1920s and have since become a fundamental tool in quality management. The main purpose of control charts is to distinguish between natural process variation and variations that are indicative of a problem or a change in the process. By monitoring key process parameters over time, control charts help in identifying when a process is operating within acceptable limits (in control) and when it has moved outside of these limits (out of control).

X-bar and R charts*:* These charts are used when monitoring the central tendency (mean) and variability (range) of a process. The X-bar chart tracks the average value of a sample over time, while the R chart monitors the range of variation within each sample.

3.7 Forecasting

Forecasting is the process of making predictions of the future based on past and present data and most commonly by analysis of trends. Prediction is a similar, but more general term. Both might refer to formal statistical methods employing time series, cross- sectional or longitudinal data, or alternatively to less formal judgmental methods. Risk and uncertainty are central to forecasting and prediction; it is generally considered good practice to indicate the degree of uncertainty attaching to forecasts. Time series methods use historical data as the basis of estimating future outcomes. They are based on the assumption that past demand history is a good indicator of future demand.

3.8 Time series analysis

Time series analysis comprises methods for analysing time series data in order to extract meaningful statistics and other characteristics of the data. Time series forecasting is the use of a model to predict future values based on previously observed values. While regression analysis is often employed in such a way as to test theories that the current values of one or more independent time series affect the current value of another time series, this type of analysis of time series is not called

"time series analysis", which focuses on comparing values of a single time series or multiple dependent time series at different points in time. Interrupted time series analysis is the analysis of interventions on a single time series.

Time series data have a natural temporal ordering. This makes time series analysis distinct from cross-sectional studies, in which there is no natural ordering of the observations. Time series analysis is also distinct from spatial data analysis where the observations typically relate to geographical locations. A stochastic model for a time series will generally reflect the fact that observations close together in time will be more closely related than observations further apart. In addition, time series models will often make use of the natural one-way ordering of time so that values for a given period will be expressed as deriving in some way from past values, rather than from future values.

Time series analysis constitutes of four major components which are given below,

- Trend
- Seasonal variations
- Cyclic variations
- Irregular or Random movements

The above components can be deduced by decomposition. Here, trend is the long-term movement of the data. Seasonal and cyclic variations are the short-term movements of data. Irregular or random movements are those fluctuations in data that cannot be foreseen and are erratic in nature. The functional relationship between these components can take two forms: additive and multiplicative. These are written as,

$$s_t = \mu_t + \phi_t + \gamma_t + \varepsilon_t \quad (1)$$

$$s_t = \mu_t \times \phi_t \times \gamma_t \times \varepsilon_t \quad (2)$$

Here, is the trend, is the cyclic variation, is the seasonal component, is the random component.

Trend is the general tendency of the data to increase or decrease during a long period of time. It is not necessary that the increase or decrease should be in the same direction throughout the given period. However, the overall tendency may be upward or downward or stable. It operates in an evolutionary manner and do not

reflect sudden changes. In general trend can be linear or non-linear. If the time series values plotted on graph cluster more, or less, round a straight line, then the trend exhibited is termed as linear trend otherwise it is termed as non-linear trend. In practice linear trend is mostly used as the rate of growth (or decline) is constant.

3.9 Stationary process

The theory of time series data analysis is mostly based on two conditions namely, stationary process and ergodic process. Initially the data is tested to find its underlying process so as to provide the right forecasting methods and models.

A stationary process (or a strict/strictly stationary process or strong/strongly stationary process) is a stochastic process whose unconditional joint probability distribution does not change when shifted in time. Consequently, parameters such as mean and variance also do not change over time. Since stationarity is an assumption underlying many statistical procedures used in time series analysis, non-stationary data are often transformed to become stationary. The most common cause of violation of stationarity is a trend in the mean, which can be due either to the presence of a unit root or of a deterministic trend. In the former case of a unit root, stochastic shocks have permanent effects, and the process is not mean-reverting. In the latter case of a deterministic trend, the process is called a trend stationary process, and stochastic shocks have only transitory effects after which the variable tends toward a deterministically evolving (non-constant) mean.

A trend stationary process is not strictly stationary, but can easily be transformed into a stationary process by removing the underlying trend, which is solely a function of time. Similarly, processes with one or more unit roots can be made stationary through differencing. An important type of non-stationary process that does not include a trend- like behaviour is a cyclostationary process, which is a stochastic process that varies cyclically with time.

3.10 Moving average method

A moving average is a common technique used in time series analysis to smooth out short-term fluctuations and highlight longer-term trends or cycles in the data. It is calculated by taking the average of a specified number of consecutive data points in the time series. There are different types of moving averages, including:

Simple Moving Average (SMA)*:* A single moving average forecast is based on finding the average for a specified number of observations and using it as a forecast for the next period in time. The calculation of the moving average is based on a constant number of observations, and this is done by adding a new observation and dropping the oldest observation. To compute the moving average Mk of order k to forecast at time k + 1 for the observations, S_1,..., S_k, the following form is used.

$$M_k = \frac{1}{k}\sum_{i=1}^{k} s_i \quad (3)$$

Weighted Moving Average (WMA): In a weighted moving average, different weights are assigned to each data point in the averaging period. Typically, more recent data points are assigned higher weights, while older data points are assigned lower weights.

Exponential Moving Average (EMA)*:* The exponential moving average gives more weight to recent data points and less weight to older ones. It is calculated recursively, giving exponentially decreasing weights to older observation.

Moving averages are used for various purposes in time series analysis, such as identifying trends, detecting seasonality, and smoothing out noise. However, it's important to choose the appropriate type of moving average and the right parameters (e.g., the number of data points to include in the average) based on the characteristics of the data and the specific objectives of the analysis.

3.11 Exponential Smoothing

The Exponential Smoothing time series method works by assigning exponentially decreasing weights for past observations. It is called so because the weight assigned to each demand observation is exponentially decreased. It is a broadly accurate forecasting method for short-term forecasts. The technique assigns larger weights to more recent observations while assigning exponentially decreasing weights as the observations get increasingly distant. This method produces slightly unreliable long-term forecasts. Exponential smoothing can be most effective when the time series parameters vary slowly over time.

Double Exponential Smoothing: This method is known as Holt's trend model or second-order exponential smoothing. Double exponential smoothing is used in timeseries forecasting when the data has a linear trend but no seasonal pattern. The

basic idea here is to introduce a term that can consider the possibility of the series exhibiting some trend. In addition to the alpha parameter, Double exponential smoothing needs another smoothing factor called beta (b), which controls the decay of the influence of change in trend. The method supports trends that change in additive ways (smoothing with linear trend) and trends that change in multiplicative ways (smoothing with exponential trend). The Double exponential smoothing formulas are:

$$S_1 = x_1 \quad (4)$$

$$B_1 = x_1 - x_0 \quad (5)$$

For t >1, $S_t = \alpha x_t + (1 - \alpha)(S_t\text{-}1 + B_t\text{-}1)$ (6)

$$\beta t = \beta(S_t - S_{t-1}) + (1 - \beta)B_t\text{-}1 \quad (7)$$

here,

B_t = best estimate of the trend at time t

β = trend smoothing factor; $0 < \beta < 1$

3.12 ARIMA model

ARIMA models or autoregressive integrated moving average models produce forecasts based on detecting patterns in historical data. It is appropriate if the observations of time series are statistically dependent on each other. The model consists of an autoregressive term, a moving average term and differences. This can be used when the data is non-stationary. The autoregressive model of order p, or AR(p) can be computed by

$$s_t = \phi_0 + \phi_1 s_{t-1} + \phi s_{t-2} + \cdots + \phi s_{t-p} + e_t \quad (8)$$

Where, e is the error term ϕ and are coefficients.

The moving average model of order q, or MA(q) is computed by,

$$s_t = \theta_0 + e_1 - \theta_1 e_{t-1} - \theta_2 e_{t-2} - \cdots - \theta_q e_{t-q} \quad (9)$$

In ARIMA (p, d, q) denotes the number of differencing required to make the time series stationary. When the given time series is stationary, then ARMA model or autoregressive moving average model can be used. It is when the difference term in ARIMA is zero. Forecasting using ARIMA model usually involves three steps. The

first step is to identify the appropriate model; the next step is to estimate and test the model. Once the model is adequate, the final step forecasting can be done.

3.13 Accuracy

The accuracy of the above stated methods is computed to infer the better method to forecast the rainfall distribution. The errors are computed to test the accuracy. The method or model with the least value of errors is termed as abetter method. The list of accuracies are

- Mean Absolute Percentage Error (MAPE): This metric calculates the average percentage difference between actual and forecasted values. It is a useful measure for understanding the relative accuracy of a forecast across different scales of data.
- Mean Absolute Deviation (MAD): This metric measures the average absolute deviation of forecasted values from actual values. It provides an indication of the forecast error in absolute terms and is useful for understanding the scale of forecast errors.
- Mean Squared Deviation (MSD): This metric, also known as Mean Squared Error (MSE), calculates the average squared difference between actual and forecasted values. It gives higher weight to larger errors due to the squaring of the differences, making it useful for assessing the impact of large forecast errors.

CHAPTER 4

ANALYSIS AND INTERPRETATION

4.1 Introduction

The data used for the analysis is the secondary data collected from the websites www.kaggle.com, the data related to the rainfall distribution in southern states of India. For our analysis we used the Microsoft Excel (MS Excel), Statistical Package for Social Science (SPSS) Software packages, Minitab, R Software. The statistical analysis has been carried out and the results are presented in this chapter.

4.2 Diagrams

The secondary data of the rainfall distribution in southern states of India from 1951 to 2017 is presented in the form of tables in annexure and for that data suitable diagrams also given in Figure 4.1.

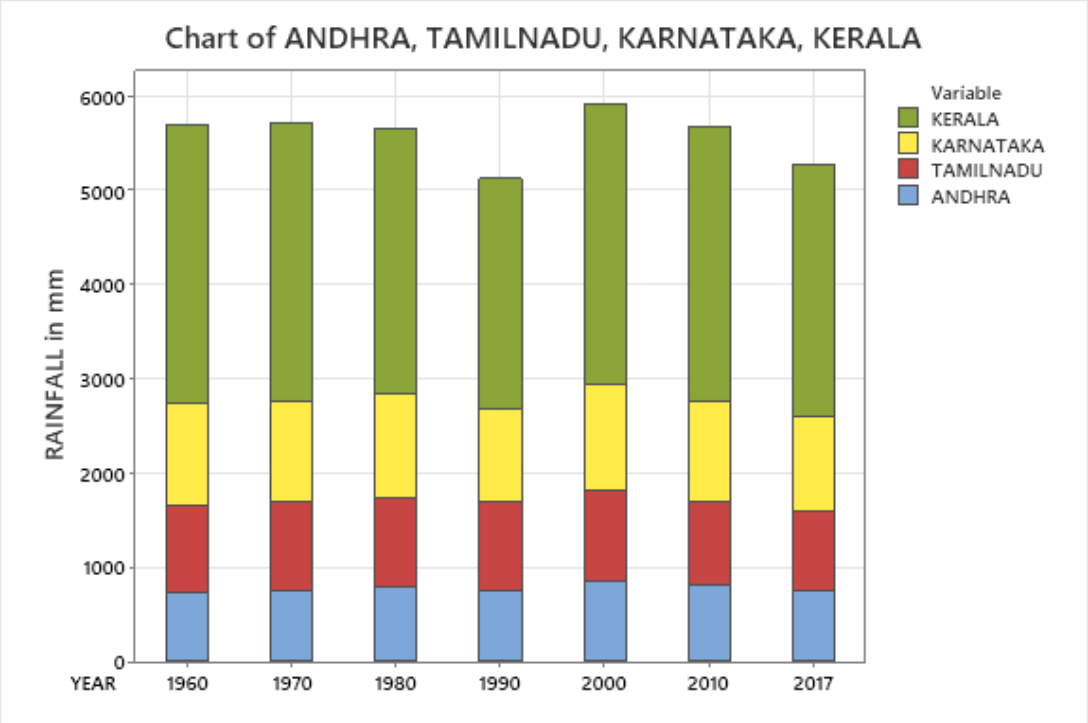


Figure 4.1 Bar graph of Rainfall Distribution in mm

The subdivided bar graph shows that rainfall in Kerala, Tamilnadu, Karnataka, Andhra for 67 years. Rainfall data is plotted on the Y-axis, ranging from 0 to 6000 mm. The years are indicated on the X-axis at intervals of approximately a decade. The total height of each bar corresponds to the cumulative rainfall from all four states for that particular year.

The following graphs shows the rainfall distribution in each state separately for each year of for the southern states which is given in Figure 4.2 to 4.5.

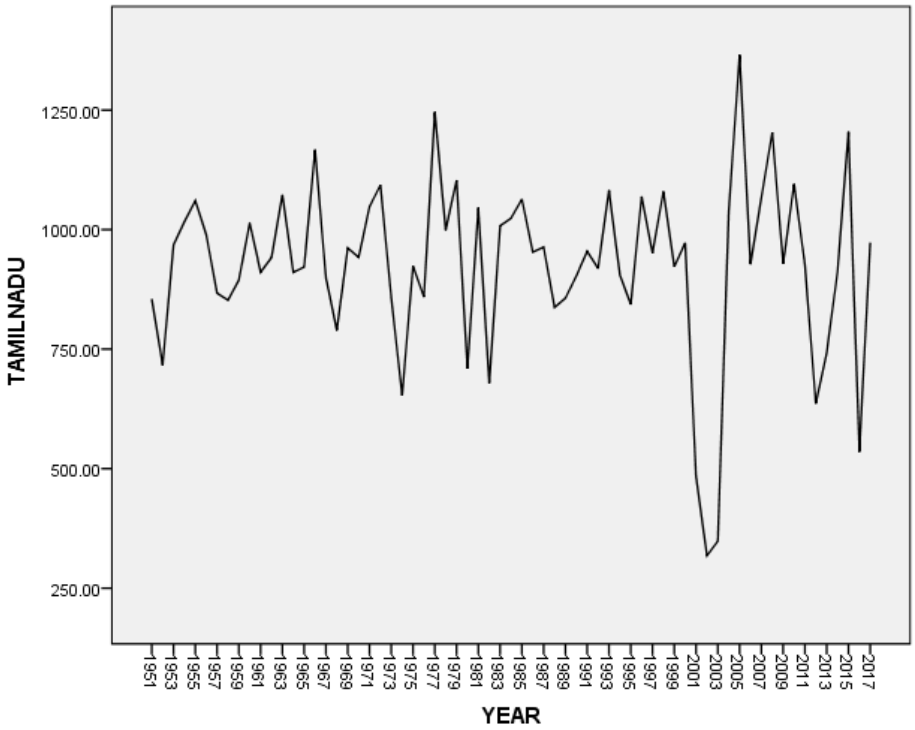


Figure 4.2 Rainfall Distribution in Tamilnadu

The Y-axis is labelled “TAMIL NADU” and ranges from 250.00 at the bottom to 1250.00 at the top. The X-axis denotes the years from 1951 to 2017. It’s almost over a decade.

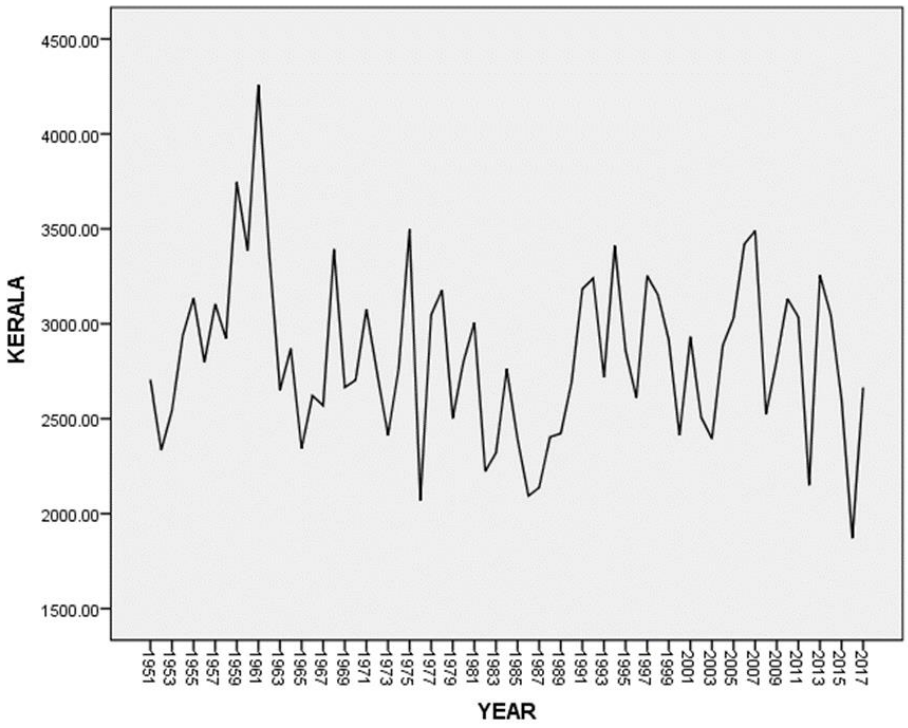


Figure 4.3 Rainfall Distribution in Kerala

The Y-axis is labelled “KERALA” and ranges from 1500.00 to 4500.00 (units not specified).

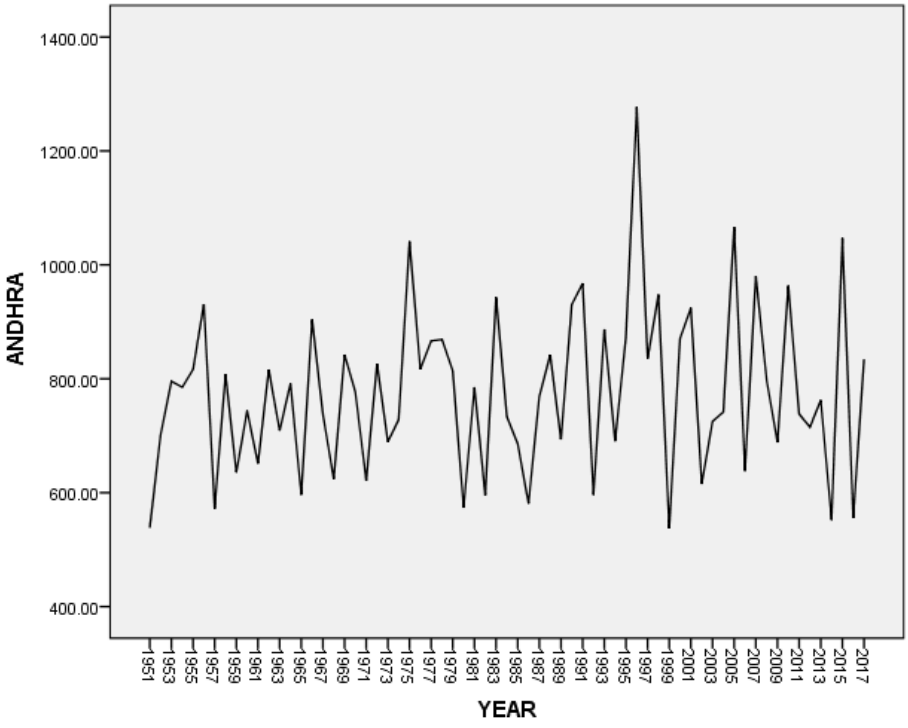


Figure 4.4 Rainfall Distribution in Andhra

The y-axis ranges from 400.00 to 1400.00. Every year there are changes in distribution.

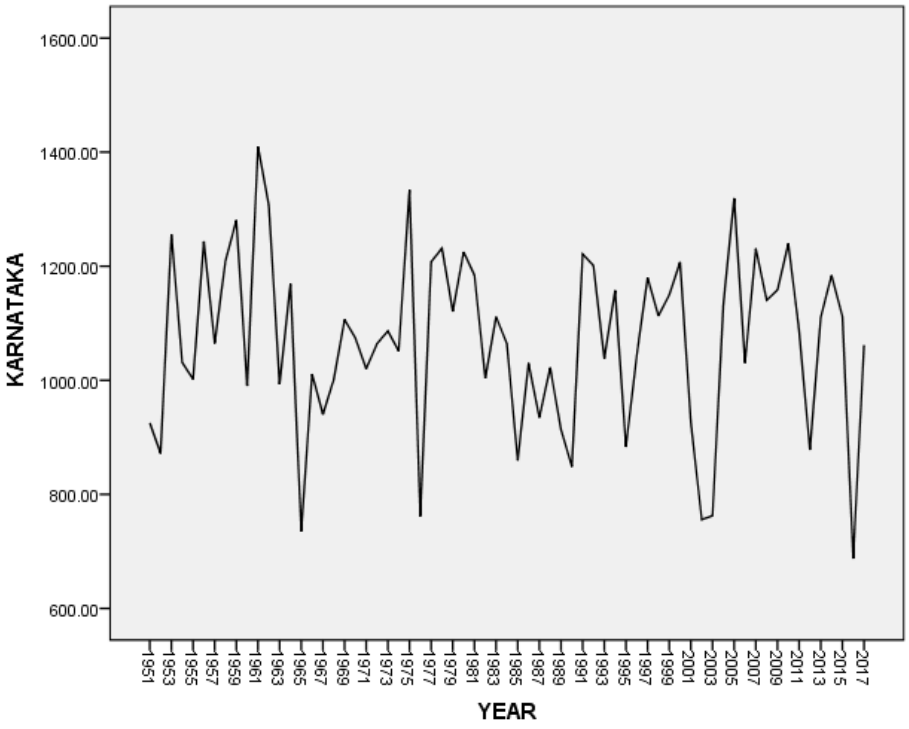


Figure 4.5 Rainfall Distribution in Karnataka

The highest value appears to be around 1600.00, while the lowest is approximately 600.00. The above figure shows the separate line graph and Andhra has more fluctuations than other 3 states which may add difficulty to predict.

4.3 Descriptive Statistics

In this section the descriptive statistics analysis is performed for the southern states of the year 1951-2017 which is consolidated in the Table 4.1.

Table 4.1 Descriptive Statistics table

	N	Range	Minimum	Maximum	Mean	Std. Deviation	Variance
KERALA	67	2387.10	1870.80	4257.90	2822.5985	439.53236	193188.697
KARNATAKA	67	722.10	687.40	1409.50	1070.3134	155.66680	24232.152
ANDHRA	67	740.50	537.20	1277.70	776.3209	147.35009	21712.049
TAMILNADU	67	1047.40	318.00	1365.40	925.0955	186.61817	34826.342
Valid N (listwise)	67						

The above table gives the mean, range separately. Variance shows how the rainfall differs for each states. Kerala has a wide range of values, with a minimum of 1870.8 and a maximum of 4257.9. Karnataka's values are more tightly clustered. Andhra also shows a narrower range. Karnataka, Andhra, Tamilnadu are less in variance shows similarity in rainfall distribution. These statistics provide insights into the distribution of the data for each state.

4.4 t- test

Comparing the means of two states for the year 2011 to 2017 with one standard state Tamilnadu and Andhra is presented in Table 4.2.

Table 4.2 Rainfall Distribution in Tamilnadu and Andhra

Year	Tamilnadu	Andhra
2011	926.7	738.2
2012	636.2	715.1
2013	742.0	762.6
2014	913.0	551.9
2015	1204.6	1047.1
2016	535.2	555.4
2017	973.0	834.5

Null hypothesis: There is no significance difference between Andhra and Tamilnadu annual rainfall.

Alternative hypothesis: There is significance difference between Andhra and Tamilnadu annual rainfall.

T-Value : -0.97 DF : 11 P-Value : 0.823

Hence the above value gives the result of the test that p value is greater than level of significance (0.05). Therefore, there is no evidence to reject the null hypothesis. There is no significance difference between Andhra and Tamilnadu annual rainfall. In short, the rainfall distribution of Andhra and Tamilnadu from 2011 to 2017.

Comparison of two states rainfall from 2011 to 2017 of Tamilnadu and Karnataka is presented in Table 4.3

Table 4.3 Rainfall Distribution in Tamilnadu and Karnataka

Year	Tamilnadu	Karnataka
2011	926.7	1087.3
2012	636.2	878.0
2013	742.0	1110.6
2014	913.0	1184.1
2015	1204.6	1112.5
2016	535.2	687.4
2017	973.0	834.5

Null Hypothesis: There is no significance difference between Tamilnadu and Karnataka annual rainfall.

Alternative Hypothesis: There is significance difference between Tamilnadu and Karnataka annual rainfall.

T-Value : 1.58 DF : 11 P-Value : 0.071

Hence the above table gives the result of the test that p value is greater than level of significance (0.05). Therefore, there is no evidence to reject the null hypothesis. There is no significance difference between Karnataka and Tamilnadu annual rainfall. In short, the rainfall distribution of Karnataka and Tamilnadu from 2011 to 2017.

Comparison of two states rainfall from 2011 to 2017 of Tamilnadu and Kerala is presented in Table 4.4

Table 4.4 Rainfall Distribution in Tamilnadu and Kerala

Years	Tamilnadu	Kerala
2011	926.7	3035.3
2012	636.2	2151.2
2013	742.0	3255.5
2014	913.0	3046.6
2015	1204.6	2600.7
2016	535.2	1870.8
2017	973.0	2664.8

Null Hypothesis : There is significance difference between Tamilnadu and Kerala annual rainfall.

Alternative Hypothesis : There is significance difference between Tamilnadu and Kerala annual rainfall.

T-Value : 8.67 ***DF : 8*** ***P-Value : 0.000***

Hence the above table gives the result of the test that p value is less than level of significance (0.05). Therefore, there is evidence to reject the null hypothesis. There is significance difference between Karnataka and Tamilnadu annual rainfall. In short, the rainfall distribution of Karnataka and Tamilnadu for the given years differs very badly.

4.5 Control Charts

Control charts are typically used for time-series data (continuous data or variable data), but can also be adapted for data with logical comparability. Control charts shows the limit of the given data in both upper as well as lower. If the analysis of a control chart indicates that the process is currently under control, no immediate corrections or changes to process control parameters are necessary. If the chart is not in control, there needs alertness in the process. Mean control chart for southern states are presented in Figure 4.6 to 4.9:

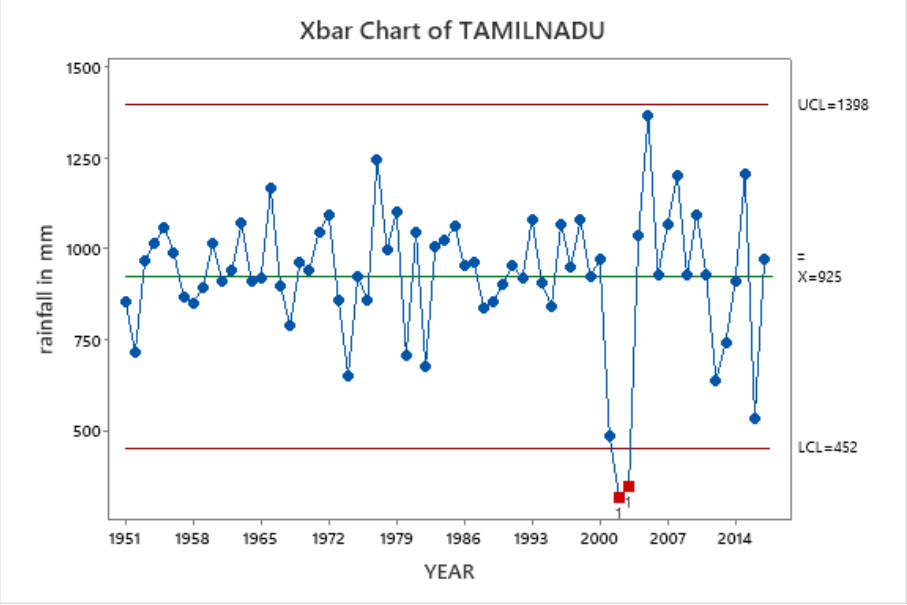


Figure 4.6 Mean chart of Tamilnadu

If the data points fall within the UCL and LCL, the process is stable. Notably, there is a significant drop in rainfall around the year 2000 and 2001.

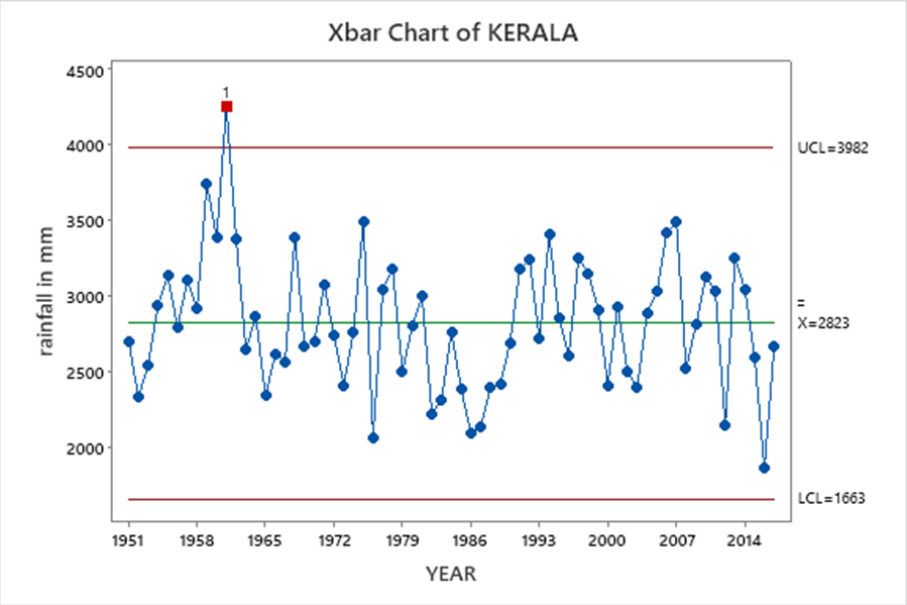


Figure 4.7 Mean chart of Kerala

The peak in rainfall around 1961 suggests a temporary increase in rainfall during that year. Already, comparing to other states Kerala has highest rainfall though the limit was not under control in a year.

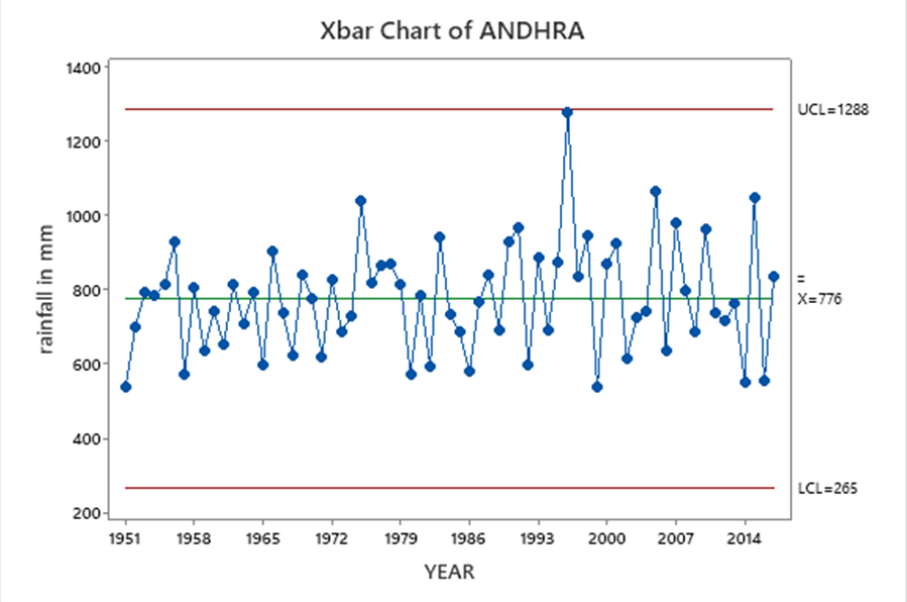


Figure 4.8 Mean chart of Andhra

Variability beyond the control limits may indicate special causes affecting rainfall (e.g., extreme weather events). There is variability in rainfall over the years, with some points exceeding the UCL and others falling below the LCL.

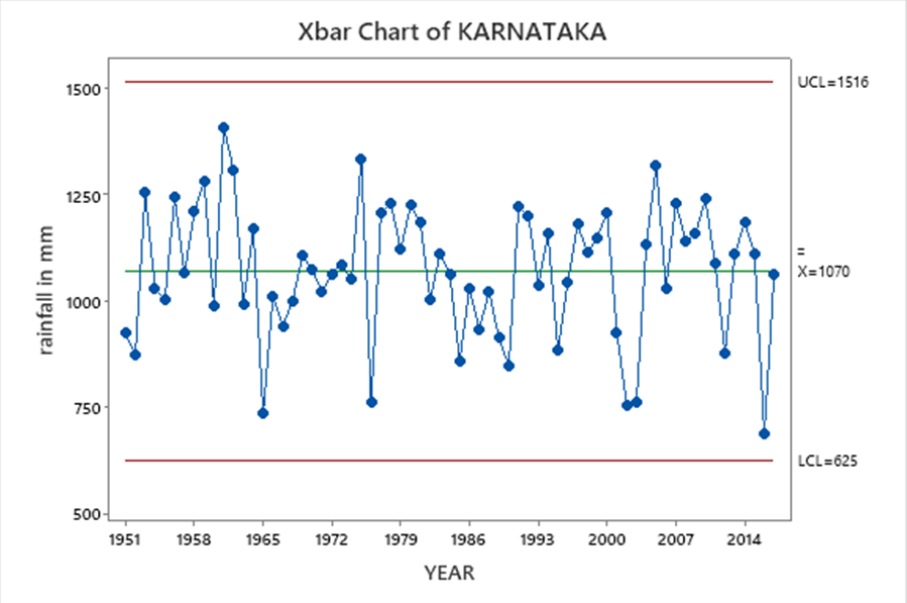
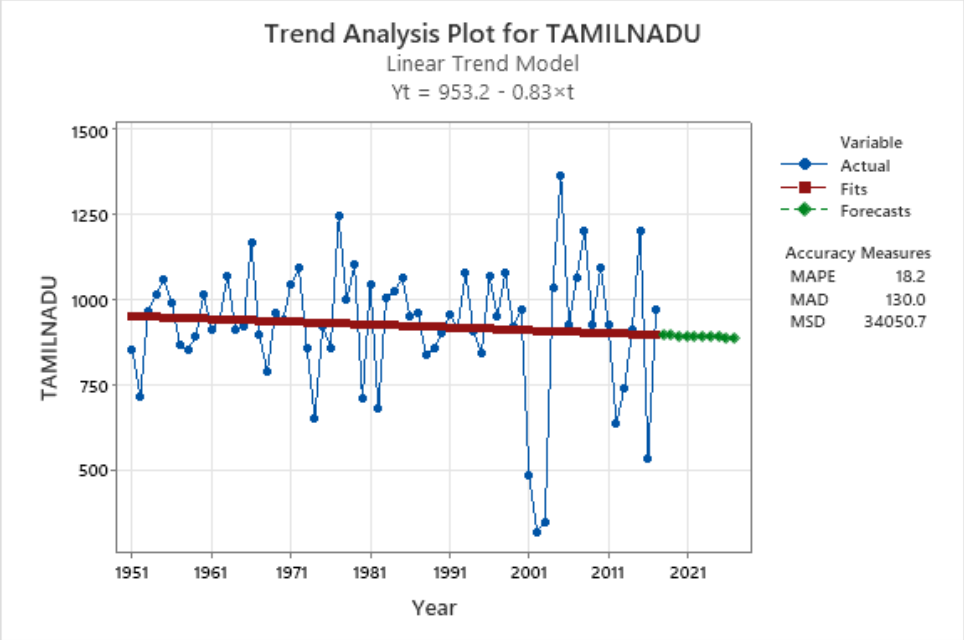


Figure 4.9 Mean chart of Karnataka

The chart displays the mean rainfall (labelled as "X=1070") along with the Upper Control Limit (UCL) and Lower Control Limit (LCL) lines. These control limits indicate the range within which the rainfall data points are considered normal.

4.6 Trend Analysis

Trend analysis relies on time series data, which is a sequence of observations or measurements collected and recorded over successive intervals of time (e.g., daily, monthly, yearly). Analysts examine the data to identify recurring patterns, trends, or cycles. These patterns could be upward (indicating growth), downward (indicating decline), or cyclical. The suitable trend analysis carried out for the given data, forecasted for the upcoming years and presented in Figure 4.10

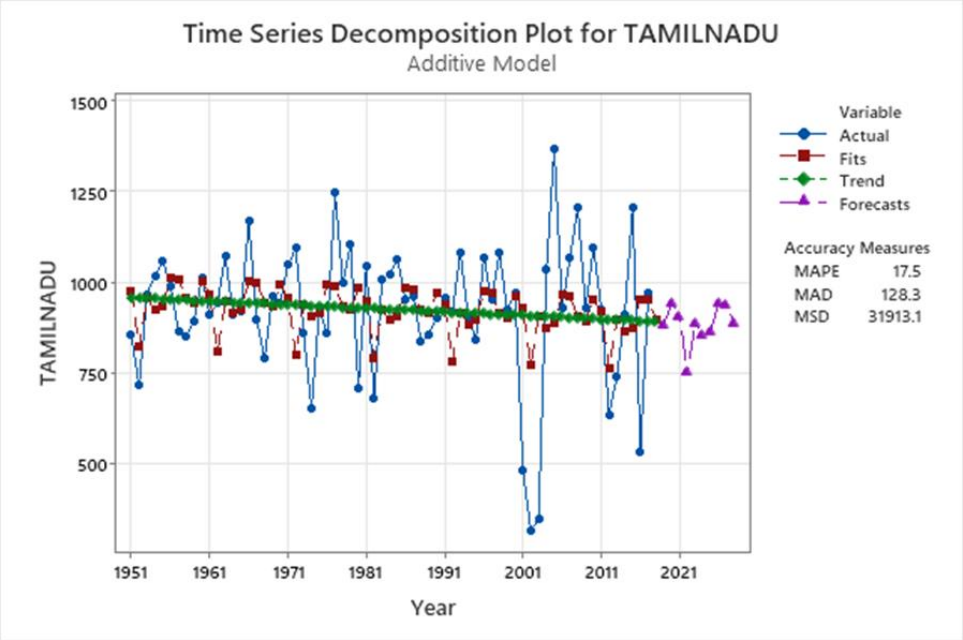


4.10 Trend Analysis for Tamilnadu

Tamilnadu's variable has been on a declining trend over the years, as indicated by the negative slope of the trend line. The forecasts suggest this trend will continue into the future. From this, linear trend model is good in fitting the data rather than quadratic or exponential. After 2000, there is a sudden decrease in the rainfall which is less than 500 mm. There are fluctuations within 750-1250 before 2000.

4.7 Rainfall decomposition

This section deals with the decomposition of time series data. rainfall decomposition analysis helps researchers understand changes in rainfall patterns over time, which is important for studying climate change and its effects on ecosystems and water resources. Seasonal patterns deals with the decomposition of data and is presented in the Figure 4.11



4.11 Decomposition Plot for Tamilnadu

The above figure shows the actual, fits, trend and forecasts of the data. It shows the slight decreasing trend. It has the error value of less comparing to trend analysis. From this, it can be declared that decomposition is better than trend analysis.

4.8 Stationary test

In this section the stationarity of the data is determined using the tests augmented Dickey Filler test. The null hypothesis of the ADF test states that the data is not stationary.

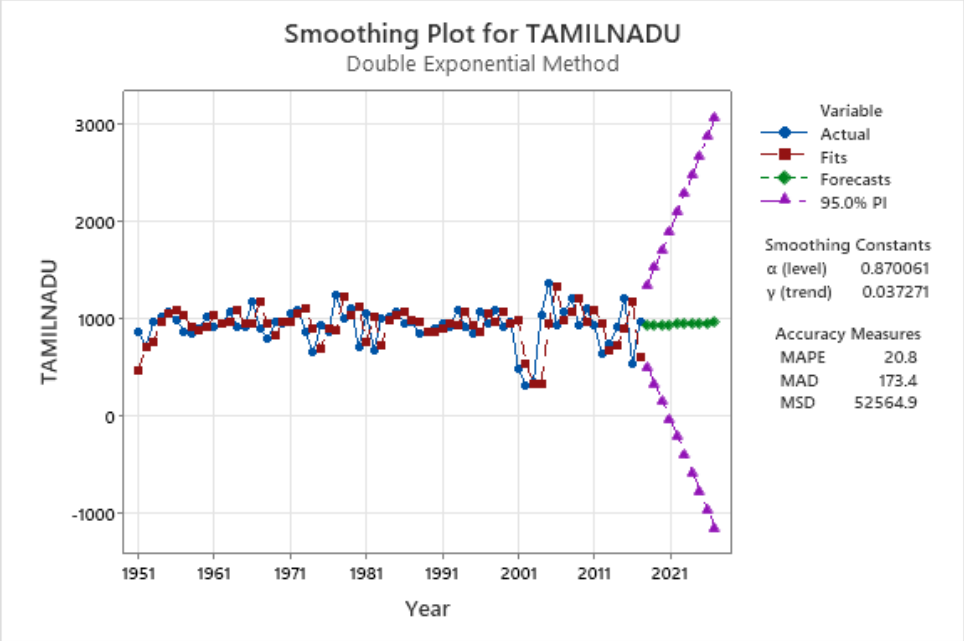
The results of the adf test gave the value for statistic as –6.34559 which is less than the critical value of –2.90644.

Hence the null hypothesis is not accepted. Therefore, the data appears to be stationary. When the data is stationary, arima model can be used without differencing.

4.9 Exponential Smoothing

Exponential smoothing is of three types namely simple exponential smoothing, double exponential smoothing, triple exponential smoothing.

Simple or single exponential smoothing is the method of time series forecasting used with univariate data with no trend and no seasonal pattern whereas *Double exponential smoothing* is used in time-series forecasting when the data has a linear trend but no seasonal pattern. Triple exponential smoothing is used for time series forecasting when the data has linear trends and seasonal patterns. Double exponential smoothing is used and presented in the Figure 4.12

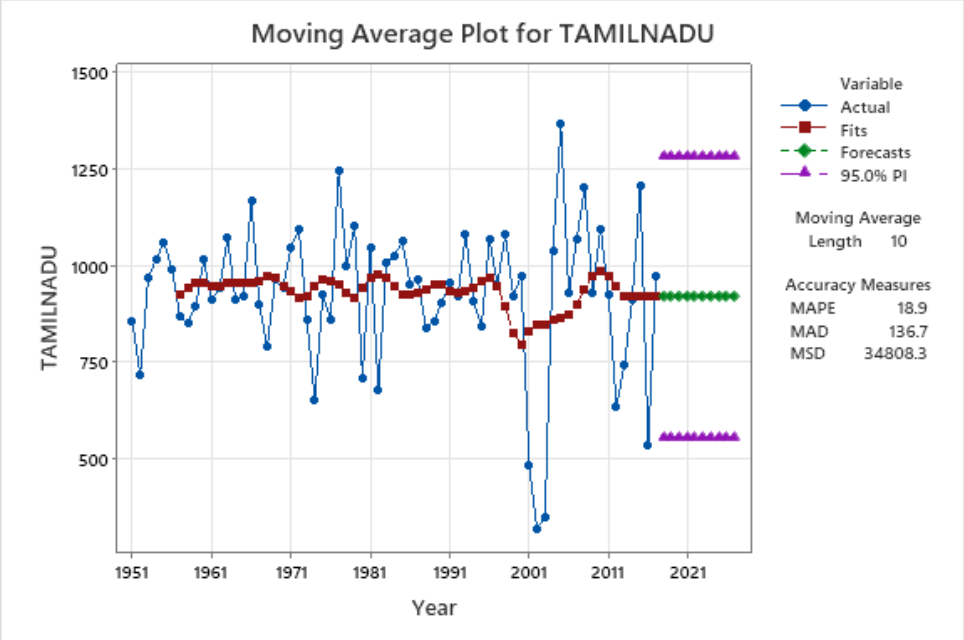


4.12 Smoothing Plot for Tamilnadu

The graph shows fluctuating actual data until around the year 2010; after this point, only forecasted data is shown, indicating a sharp increase. By adjusting α and γ constants, the method balances the influence of past observations and recent trends to create accurate forecasts.

4.10 Moving average method

A simple moving average is a basic method of smoothing and analysing data in time series forecasting and financial analysis. It is a lagging indicator that helps identify trends by calculating the average of a fixed number of data points over a specified period. Simpe moving average fits the data is presented in Figure 4.13

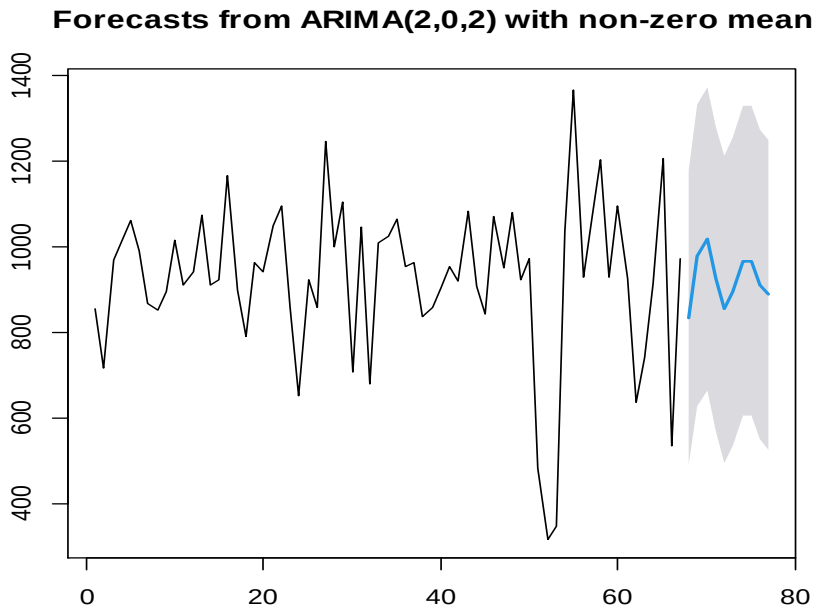


4.13 Simple Moving Average for Tamilnadu

The moving average helps smooth out fluctuations in the data. The fitted line likely represents a model fit to historical data. The forecast line extends into the future, indicating predictions. This interval captures the uncertainty in the forecast. The upward or downward trend in the moving average can provide insights into long term changes.

4.11 ARIMA

ARIMA is a powerful time series modelling approach that combines autoregressive and moving average components with differencing to produce reliable forecasts. Rainfall distribution is forecasted using Autoregressive Integrated Moving Average method. Selecting the appropriate values of p, d, and q is essential for building a good ARIMA model. This is often done using statistical techniques such as the Akaike Information Criterion (AIC) or Bayesian Information Criterion (BIC). Once the model is selected, parameters are estimated using methods such as maximum likelihood estimation. ARIMA model with fit of least error is presented in Figure 4.14



4.14 ARIMA model for Tamilnadu

The ARIMA model (2,0,2) indicates the order of the ARIMA model: 2 represents the number of autoregressive terms (lags), 0 signifies that the data has been differenced (integrated) to make it stationary (i.e., remove trends and seasonality), 1 denotes the number of moving average terms. The X-axis represents the year from 1951 to 2017 and the Y-axis denotes the rainfall in mm. It gives the accuracies of MAPE = 20.2, MAD = 162.2, MSD = 42337.7

4.12 Accuracy

This section compares the accuracy of all the models used for forecasting the rainfall distribution in Table 4.5

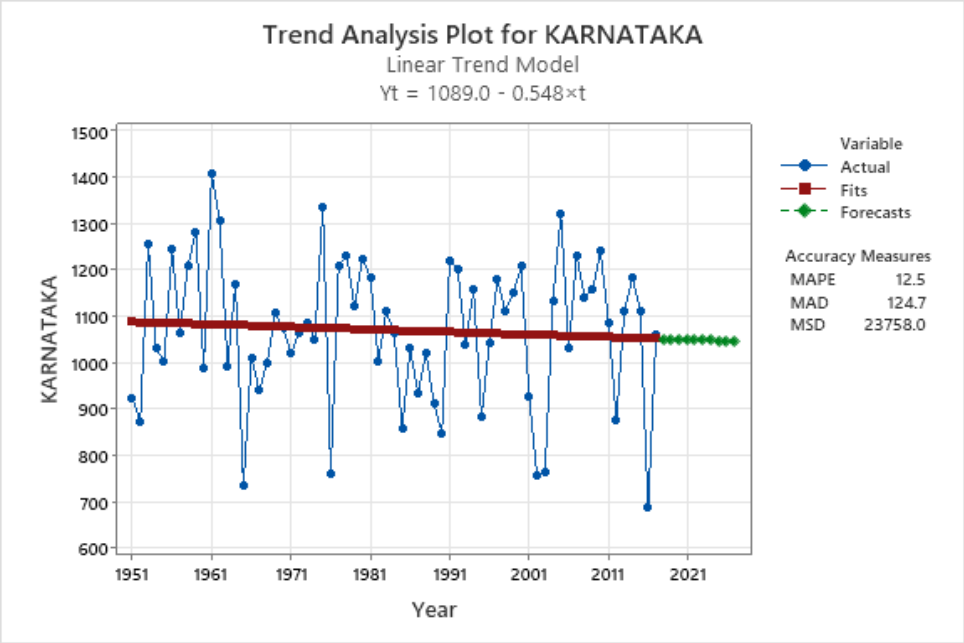
Table 4.5 Accuracy of forecasting methods

Method	MAPE	MAD	MSD
Trend	18.2	130.0	34050.7
Exponential Smoothing	20.8	173.4	52564.9
Moving Average	18.9	136.7	34808.3
ARIMA	20.2	162.2	42337.7

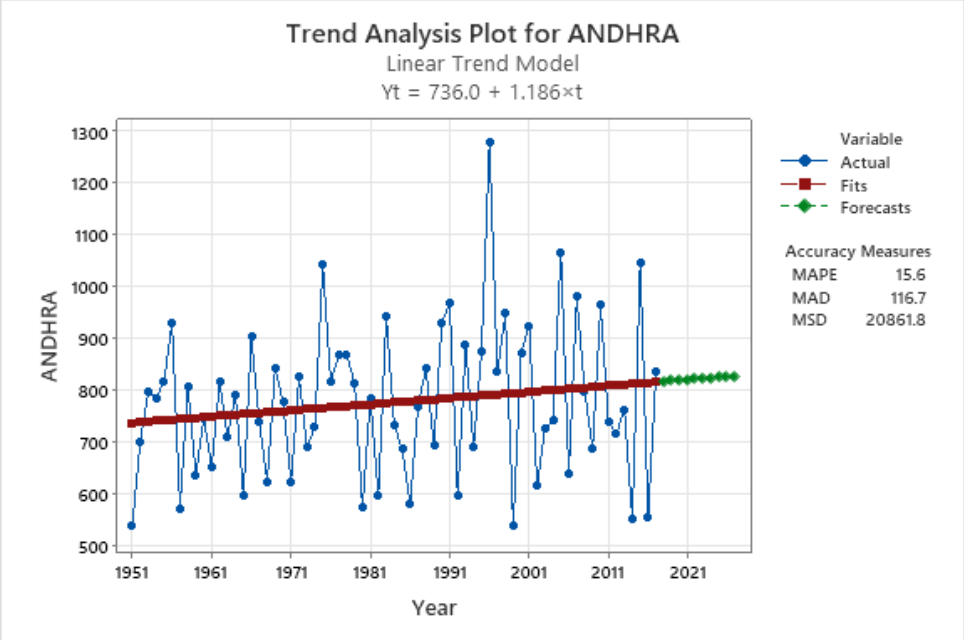
Table 4.5 shows the accuracy of the forecasting methods. It can be seen that exponential smoothing and ARIMA model holds the large values when compare with other methods. Hence it can be concluded that the best method for computing rainfall distribution is trend analysis.

4.13 FORECASTING

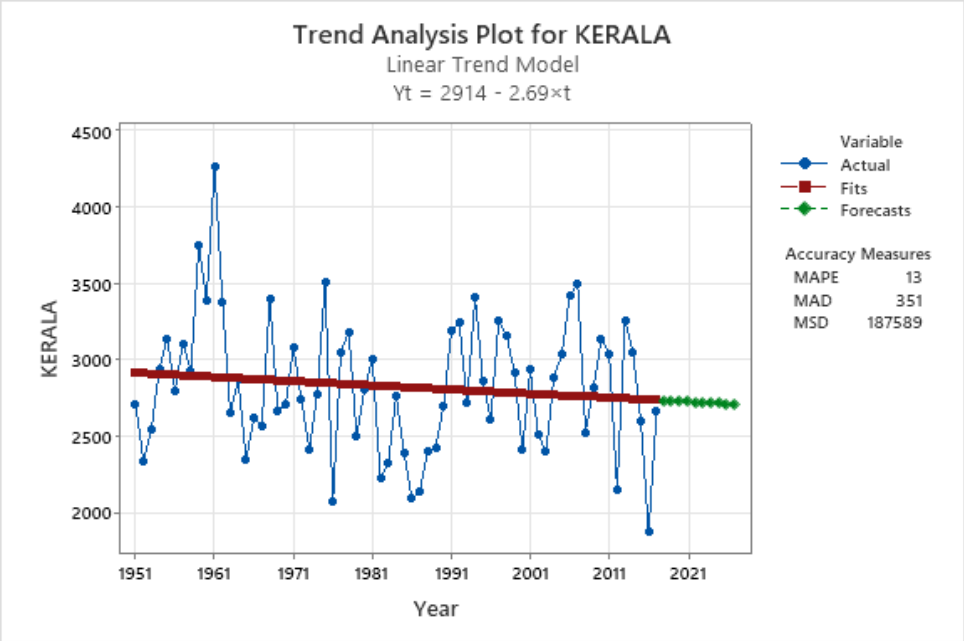
From the above accuracy, trend analysis will suit for this data to forecast and it has less error value. So, forecasting for the upcoming years is to be done by trend analysis for all the states. Forecasting in weather condition will always be an approximation and there may be changes also. From 2018, forecasting has been started for Karnataka, Kerala, Andhra and is presented in Figure 4.15 to 4.17



4.15 Trend Analysis for Karnataka



4.16 Trend Analysis for Andhra



4.17 Trend Analysis for Kerala

The trend analysis for the southern states is presented in Figure 4.10, 4.15, 4.16, 4.17. Only Andhra has increasing trend. All other shows decreasing trend. It will be useful for forecasting the values.

Forecasting values for all the 4 states from 2018 to 2025 are presented in Table 4.6

Table 4.6 Forecasting values of Rainfall for the Southern states

YEAR	TAMILNADU	KERALA	KARNATAKA	ANDHRA
2018	896.9	2731.0	1051.7	816.6
2019	896.1	2728.3	1051.1	817.8
2020	895.3	2725.6	1050.6	819.0
2021	894.5	2722.9	1050.0	820.2
2022	893.7	2720.2	1049.5	821.4
2023	892.8	2717.5	1048.9	822.6
2024	892.0	2714.8	1048.4	823.8
2025	891.2	2712.1	1047.8	825.0

This table shows the predicted values for the southern states where Tamilnadu and Andhra has the low rainfall and with low variance between them.

CHAPTER 5

SUMMARY AND CONCLUSION

Rainfall prediction offers a valuable glimpse into the future, impacting decisions across numerous sectors. From strategically planning agricultural activities to optimizing water resource management and even mitigating flood risks, knowing how much rain to expect is crucial. However, predicting rainfall accurately remains a complex task.

The main objective of the study is to forecast the future rainfall and to analyse the past data. From the data, southern states of India's rainfall can be seen. The graphical representation shows the entire data in the subdivided bar chart where Kerala has the highest rainfall in south. The separate line graphs shows the rainfall amount in four states in each year from 1951 to 2017. There is no graph without fluctuation that is for every year there are changes. Descriptive Statistics shows the minimum and maximum rainfall, and the difference between the annual rainfall received in each state.

Comparison of Tamilnadu and other states in t test gives that Karnataka and Andhra has no significance difference between them. For Tamilnadu and Andhra, both receives northeast monsoon at same time and if the cyclone occurs, two states gets affected at same time. Due to land area, cyclone direction, speed defines the affecting areas. And for tamilnadu and Karnataka, both receives almost same amount of rainfall annually but differs in monsoon.

Control charts is useful in finding when the limit has been crossed. Karnataka and Andhra are within their control. Even though, Andhra has touched its highest limit in 1996. Tamilnadu has suffered by drought in 2000 and 2001. Kerala

has suffered by flood in 1961. Lower limit denotes the condition of any state which is affected by drought or famine. Upper limit shows the flood condition or the extreme rainfall distribution.

The time series analysis is used to find the pattern of the data to check whether it has cyclic, or seasonal, or irregular in it. Trend analysis shows the data with very less error and moving average shows the data with less error in forecasting. ARIMA is the advanced techniques in forecasting but it doesn't suit for the data. Forecasted values are displayed for next 8 years from 2018 to 2025. In conclusion, rainfall prediction is a powerful tool with inherent limitations due to the complexities of weather systems.

REFERENCES

Textbooks

- Andrew Metcalfe V and Paul Cowpertwait S.P. (2009). **Introductory Time Series with R**. Springer.
- Gupta S.C. and Kapoor V.K. (2015). **Fundamentals of Applied Statistics**. Sultan Chand & Sons.
- Gupta S.P. (2015). **Statistical Methods**. Sultan Chand & Sons
- Hanke, J.E and Wichern, D.W. (2009), **Business Forecasting**, PHI Learning Pvt Limited, 8th edition, New Delhi.

Website

https://www.geeksforgeeks.org/time-series-and-forecasting-using-r/

https://mausam.imd.gov.in/imd_latest/contents/rainfall_statistics_3.php

https://en.wikipedia.org/wiki/Weather

https://a-little-book-of-r-for-time-series.readthedocs.io/en/latest/src/timeseries.html

ANNEXURE

The data gives annual rainfall distribution in southern states of India from 1951 to 2017 as follows:

YEAR	KERALA	KARNATAKA	ANDHRA	TAMILNADU
1951	2705.6	925.3	538.4	855
1952	2334.8	871.8	701.4	716.2
1953	2544.8	1256	795.6	968
1954	2937.9	1031.1	785.2	1016.3
1955	3134.6	1002	816.4	1060.6
1956	2798.3	1243.6	930.8	989.8
1957	3103.2	1064.8	571.8	867
1958	2923.1	1209.9	808.2	852.5
1959	3746.2	1280.8	635.7	894.3
1960	3385.6	990.1	744.2	1013.8
1961	4257.9	1409.5	651.4	910.8
1962	3375.8	1306.9	816.2	941.6
1963	2651	992.9	709	1072.7
1964	2869.2	1169.4	791.6	910.6
1965	2342.5	735.1	596.2	921.7
1966	2621.8	1010.9	904.1	1166.9
1967	2569.1	940.2	737.8	899.2
1968	3392.7	1001	624	789.4
1969	2664.9	1106.5	842.3	961.8
1970	2703.3	1073.9	776.4	942
1971	3076.7	1020.2	621.2	1047.6
1972	2739.2	1064.1	826.1	1093.3
1973	2412.4	1086.5	689.4	858.7
1974	2767.5	1051.4	728.3	652.8
1975	3498.4	1334.3	1041.5	923.7
1976	2068.8	761.3	817	859
1977	3047.6	1207.8	866.5	1246
1978	3176.7	1231.5	869.1	998.9
1979	2503.1	1120.7	813.5	1103.3
1980	2803.4	1225	573.6	709.2
1981	3005.8	1184.5	785.3	1046
1982	2223.3	1003.5	595.4	678.8
1983	2320.5	1111.1	943	1007.9
1984	2762.2	1063.9	733	1023.7
1985	2390.8	859.4	685.9	1063.4
1986	2093.4	1030.2	580.9	953
1987	2137.6	934.6	769.4	963.4
1988	2403.4	1022.2	842.2	837.2
1989	2422.6	913.4	693.6	856.7
1990	2693.3	848.5	930.7	903.2

1991	3184.5	1221.2	967.6	954.8
1992	3239.4	1201.1	596.1	918.8
1993	2717.7	1038.1	886.9	1082.3
1994	3410.8	1157.7	690.5	905.9
1995	2858.8	883.2	874.3	843.4
1996	2609.9	1044.1	1277.7	1069.3
1997	3252.5	1180.1	835.1	950.5
1998	3151.4	1113	947.9	1079.8
1999	2914.7	1149.7	537.2	922.6
2000	2412.5	1207.3	871.3	972.2
2001	2931	926.4	924.6	483.5
2002	2507.5	755.8	615.3	318
2003	2394.8	762.7	725.4	348.5
2004	2886	1133.5	741.7	1037.6
2005	3031.2	1319.3	1066.4	1365.4
2006	3420.6	1030.5	638.1	927.8
2007	3489.6	1231	980.6	1067.1
2008	2524.7	1140.6	797.9	1203.4
2009	2810.7	1158.4	688.3	928.5
2010	3131.9	1239.8	963.5	1095.3
2011	3035.3	1087.3	738.2	926.7
2012	2151.2	878	715.1	636.2
2013	3255.5	1110.6	762.6	742
2014	3046.6	1184.1	551.9	913
2015	2600.7	1112.5	1047.1	1204.6
2016	1870.8	687.4	555.4	535.2
2017	2664.8	1061.8	834.5	973