Analysis of Weather Parameters Using Machine Learning

Ramdas D. Gore(✉) and Bharti W. Gawali

Department of Computer Science and Information Technology, Dr. Babasaheb Ambedkar,
Marathwada University, Aurangabad 43104 (MS), India
ramdasgore1888@gmail.com

Abstract. Rainfall forecasts at various time and space scales have been one of the essential ingredients for not only industries, businesses, and politicians but also farmers to minimize losses. For agriculture, forecasts of other atmospheric parameters are also relevant. Atmospheric techniques must be known, measurements improved and ongoing study and improvement to ensure reliable weather predictions. In India, the Ministry of Earth Science (MoES) works to develop forecasting in its separate programs and divisions such as India Meteorological Department (IMD). The efforts have culminated in fairly reliable predictions of rainfall patterns. This research aims to enhance prediction accuracy in different geographical areas ranging from weather subdivisions to agro-climate areas. The machine learning techniques have introduced detailed experimental position forecasts for the Marathwada region of Maharashtra state, India. We have developed prediction model for Marathwada region using machine learning techniques. We have used Autocorrelation and seven machine learning techniques for the prediction of weather model such as Linear, Exponential, Quadratic, Additive seasonality, Additive Seasonality Quadratic Trend, Multiplicative Seasonality, and Multiplicative Seasonality Linear Trend. Linear, Exponential, Quadratic and Additive seasonality are not given good result for weather parameter.

Additive Seasonality Quadratic Trend is best fit model for the highest maximum temperature (1.42), lowest minimum temperature (1.87), wind speed (1.06), relative humidity (5.07), mean station (1.19), and mean sea level pressure (1.43). Multiplicative Seasonality model is the best model for mean minimum temperature (1.12) total rainfall in the month (24.13), heavy rainfall (8.74), and number of rainy days (1.2). Multiplicative Seasonality Linear Trend is given good accuracy for mean maximum temperature (1.2). The linear, exponential, quadratic and additive seasonality are not given good result for weather parameter. Rainfall is not the same every year. Some areas get more rain and some areas get less rain and its effect falls on all Marathwada region. The low rainfall and high temperature in the Marathwada region in most of the year due to this comes under the drought condition. So there is a need to change the crop pattern in this region like temperature tolerant crops.

Keywords: Rainfall · Temperature · Weather · Prediction model · Forecasting Model · Correlation · Autocorrelation · Linear Regression · Data Science · Machine Learning · Marathwada Region

R. Manza et al. (Eds.): ACVAIT 2022, AISR 176, pp. 569–589, 2023.
https://doi.org/10.2991/978-94-6463-196-8_44

1 Introduction

In rapid events, climate research relies entirely on long-run traits. Climate data has risen significantly in the last 30 years due to the rapid expansion of information technology, and the rate of increase will accelerate in the future. [1]. The long-term average of weather patterns, usually assessed over a period of 30–40 years, is referred to as climate. Some commonly measured meteorological variables include temperature, humidity, atmospheric pressure, rainfall, wind, and precipitation. In a broader sense, the climate is the state of the climate system components which includes the Earth's ocean and ice. The climate of a location is influenced by factors such as its latitude, geography, and altitude, as well as adjacent bodies of water and currents. A region's climate is the general state of the present-day climate system at that location [2]. Meteorological parameters such as temperature, humidity, wind speed, and the current climate all influence rainfall. When the temperature is high, rainfall and humidity are both low. Wind speed is high then rainfall is also high and wind speed is low, then average rainfall. Climate parameters (Rainfall, Temperature, Humidity, and Wind) affected living and non-living things. Climate is affected directly or indirectly in various sectors such as Human Health, Agriculture, Forest, Animal, Birds, Business, Environment, Educations, polities, National and International organizations (Private and Government). We have found the number of forecasting techniques in the literature survey. The data science techniques are wildly used for forecasting analysis. The machine learning is the one of the technique in data science [3].

The objective of this research, the machine learning techniques is helped forecasting, it is established prediction models five homogeneous monsoon regions of the district of Marathwada. The analysis's goal is to compare forecasts and assess the model's usefulness. It is possible to compare the findings of both studies to draw acceptable conclusions that show the value and usefulness of models.

1.1 Weather and Atmospheric

The state of the atmosphere is referred to as weather. Most weather occurs in the troposphere or the lowest layer of the atmosphere. Weather is influenced by a variety of elements for example, air temperature, atmospheric pressure, humidity, precipitation, solar radiation, and wind. Each of these factors is monitored in order to assess the quality of local atmospheric conditions and identify common weather patterns [4]. According to recent studies on fluctuations in rainfall over India, the global average yearly rainfall does not follow a clear pattern. Despite the fact that there was no clear pattern in monsoon rainfall in India over time, particularly on a national scale, multiple studies revealed areas with substantial long-term fluctuations in rainfall [5].

1.1.1 Rainfall (in mm)

Rain is an important part of the water cycle since it is responsible for depositing the majority of the world's fresh water. It provides habitat for a variety of habitats as well as water for hydroelectric power facilities and farmland irrigation. In many places of the globe, rainfall or precipitation is the primary supply of water for agricultural production.

Water is required for all crops to grow and provide yields. Rainfall is the most essential source of water for agricultural growth. Rainfall is defined by the amount, intensity, and timing of its occurrence [6].

1.1.2 Temperature (in Degree Celsius)

Temperature, an essential determinant, influences the rate of plant growth. Temperatures predicted by climate change, as well as the potential of more severe temperature occurrences, will have an impact on plant productivity, or agricultural production [7]. The temperature categories are as follows: mean and highest maximum temperature, mean and lowest minimum temperature, and mean and lowest minimum temperature. One of the earliest studies looked at trends in global yearly maximum and minimum temperatures and determined that there was no regular tendency for these temperatures to increase/decrease. Seasonal and monthly air temperature sequences from 1881 to 1997 revealed a significant growing tendency of 0.57 degrees Celsius every 100 years. The magnitude of heat was larger during the post-monsoon and winter seasons. With the exception of a severe negative pattern in northwest India, the monsoon temperature did not show a discernible trend in any significant portion of the world [8].

1.1.3 Wind Speed (in *Kmps)*

It is a fundamental atmospheric quantity created by the movement of air from high to low pressure, mostly as a result of temperature variations. Because of the rotation of the Earth, wind direction is nearly parallel to isobars (rather than perpendicular, as one might anticipate). Wind speed influences weather forecasts, aviation and maritime operations, construction projects, plant growth and metabolic rates, and a variety of other factors. The use of an anemometer to determine wind speed is becoming more widespread [9].

1.1.4 Humidity (in %)

Humidity has a big influence on the weather. It is a natural characteristic of our environment that relates to the amount of water vapour in the air. Humidity levels that are too high or too low can harm people and crops. The combination of high humidity and hot temperatures can be dangerous to one’s health, especially for the young and elderly. It indicates whether precipitation, dew, or fog are likely [10].

2 Related work

The following are the most important outcomes of this enormous collection of interconnected works. (1) Between 1901 and 2007, the annual temperatures (mean, maximum, and minimum) all increased at significant rates of 0.51, 0.72, and 0.27 degrees Celsius, respectively. The temperature has gradually and steadily climbed over this time (a hundred years). The warming was mostly caused by higher temperatures throughout the winter and post-monsoon seasons. The annual average temperature climbed by 0.20 degrees Celsius per decade between 1971 and 2007, resulting in substantial swings in both hot and low temperatures. In the same way, the lowest temperature rise was significantly higher than the median. Despite the fact that post-monsoon temperatures climbed drastically in a small number of places, total winter and summer monsoon temperatures rose significantly throughout almost the entire world. (2) Temperature patterns in India were found to be very stable and in line with global and hemispheric trends on a broader geographically aggregated scale. Patterns on smaller regional sizes and for specific sub-periods, on the other hand, have not consistently paralleled Indian aggregate temperatures. Rainfall variability has also altered patterns throughout the monsoon and post-monsoon seasons. Recent temperature increases in some regions of the world might be attributed to the proportional influence of aerosols and greenhouse gases. (3) Rapid warming occurred between 1971 and 2007, with considerable warming occurring between 1998 and 2007. Maximum temperatures in India were somewhat higher than the long-term average (1901–2007) throughout that time period, with a consistent pattern, whereas minimum temperatures exhibited a rising trend, nearly equivalent to that recorded between 1971 and 2007. It's worth remembering that, according to World Meteorological Organization (WMO) data, 2010 was one of the top three warmest years on record since the start of observational climate records in 1850. (4) The average temperature in each season grew dramatically between 1901 and 2007 [11–13].

The research takes use of monthly, seasonal, and annual data. Daily temperature scale increases were also seen at three additional locations during the monsoon and post-monsoon seasons. Temperatures in NE India were fairly steady throughout the winter and pre-monsoon seasons, but they rose dramatically during the monsoon and post-monsoon seasons. Over the annual, seasonal (winter and pre-monsoon), and monthly time periods, there were decreasing trends in sunshine duration. The annual maximum temperature in Central Northeast India rose by 0.008 °C during the monsoon season, 0.014 °C during the post-monsoon season, and 0.008 °C during the post-monsoon season between 1914 and 2003 [14]. During the post-monsoon season, the lowest temperature climbed by 0.012 °C each year, while the maximum temperature declined by 0.002 °C each year. Minimum temperature swings have been demonstrated to be more difficult than maximum temperature swings, both temporally and geographically, and to have less implications [15].

Rainfall patterns during the monsoon season were studied over various divisions, sub-divisions, and the whole nation of India. In India, there has been a declining trend in monsoon rainfall, rainy days, and yearly rainfall since the second half of the 1960s. Growing trends in mean annual and monsoon rainfall were observed over the meteorological subdivisions of Punjab, Haryana, western Rajasthan, eastern Rajasthan, and western Madhya Pradesh from 1901 to 1982. Monsoon rainfall has been decreasing in the northeast peninsula, northeast India, and the northwest peninsula, according to climatic statistics (ranging from -6 to -8 percent of average per 100 years). Rainfall during the monsoon season, on the other hand, rose in the west coast, in the central peninsula, and in northwest India (by around 10 to 12 percent of normal every 100 years) [16].

From 1941 to 2002, the monsoon rainfall in northwest and central India decreased. There was no discernable pattern in yearly, seasonal, or monthly rainfall in India, according to an examination of 135 years of rainfall data (1871 to 2005). Annual and monsoon rainfall have decreased throughout time, however rainfall during the pre-monsoon, post-monsoon, and winter seasons has increased, with the pre-monsoon season highest [17]. During the monsoon months of June, July, and September, precipitation patterns in India decreased, but surged in August. Despite the fact that just three sub-divisions were statistically significant, Haryana, Punjab, and coastal Karnataka, additional investigation on a sub-divisional basis (30 sub-divisions) revealed that half of them had a rising trend in yearly precipitation. Only the Chhattisgarh sub-division had a substantial decline in annual precipitation, showing that annual precipitation is decreasing. For the majority of the months, there was no discernible pattern in annual, seasonal, or monthly precipitation in any of the five zones. Rainfall shows small increases in both the winter and summer seasons [18].

2.1 Background of Marathwada Region *(Maharashtra State)*

Agriculture in Maharashtra's Marathwada region is dependent on rainfall, and the region is currently experiencing the worst drought due to poor rainfall. In this region, agricultural production is also lower than the rate of investment. Agricultural yields are affected by a variety of factors, including erratic weather, low or excessive rainfall, climate change, the lack of soil monitoring, the absence of soil-based crops, or associated fertilizers, the application of too little or too much fertilizer, and planting the same crops every year.

3 Study Area

Maharashtra is India's third biggest state (307,713 km^2) and the second most populous subdivision (118,809 sq mi). Maharashtra is divided into 36 districts, 355 talukas, 535 towns, 63,663 villages, and five regional areas (Konkan, Pune, Marathwada, Vidarbha, and Nashik). Some of the significant rivers are presented in the State (Krishna, Bhima, Godavari, Tapi-Purna, and Wardha-Wainganga). Low rainfall in the state's central region, with several dams on the majority of the rivers in the area.

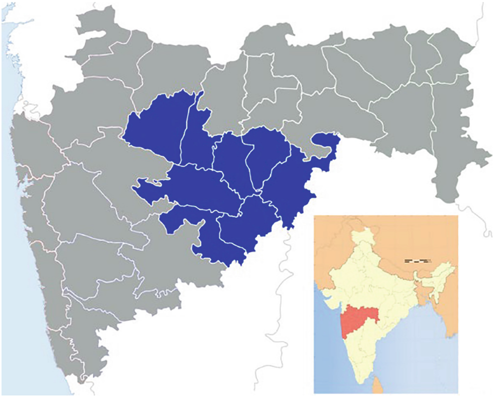


Fig. 1. Study Area (Blue colour shows Marathwada Region)

The Marathwada region occupies 64590 km^2 (24,940 sq mi) and has a population of 18,731,872 people, according to the 2011 census. (See Fig. 1). The region is divided into eight districts (Aurangabad, Beed, Hingoli, Jalna, Parbhani, Latur, Osmanabad, and Nanded). The main city is Aurangabad, which is located in the district of Aurangabad.

3.1 Database

The temperature (in degrees Celsius), rainfall (in millimeters), relative humidity (in percent), pressure (in hpa), and wind speed (in kilometers per hour) are described in the dataset. There are 120 sub-parameters observations of weather parameters (mean and highest maximum temperature, mean and lowest minimum temperature, total rainfall in the month, heavy rainfall in 24 h, number of rainy, mean wind speed, relative humidity, mean station and sea level pressure) in the Marathwada area of Maharashtra State (India) spanning a ten-year period (1976–1985).

IMD is the most widely used climate database. We have taken five districts of Marathwada as Aurangabad, Beed, Parbhani, Nanded, and Osmanabad. The source of the data is credited as the Indian Metrological Department (IMD), Ministry of Earth Science, Pune, India.

4 Methods and Techniques

In order to produce educated estimates that are predictive in identifying the direction of future trends, historical data is used as inputs in the forecasting process. The application of science and technology to anticipate atmospheric conditions for a given area and time is known as weather forecasting. People have been attempting to predict the weather informally and systematically since the nineteenth century. Weather predictions are based on the collection of quantitative data on the present position of the atmosphere, land, and water, as well as the use of meteorology to predict how the atmosphere will change in a particular region. It might be challenging to determine which model will best fit the data from a scatter plot. To identify which model best matches the data, we need first find several models for it, then compare the y-values of each model to the true y-values [19].

4.1 Pre-processing

The raw data is translated into the excel format. We have divided database as district wise and year wise and find the missing of year, month and days.

4.1.1 Missing Values

When no data value is saved for a variable in an observation, it is referred to as missing data or values in statistics. Missing data is a typical occurrence that has a major impact on the conclusions formed from the data. For missing numeric data, the mean and median imputers are utilized (int and float).

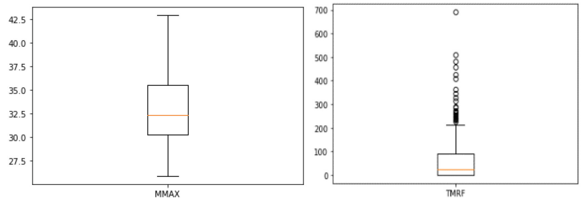


Fig. 2. Data Visualization-Outliers

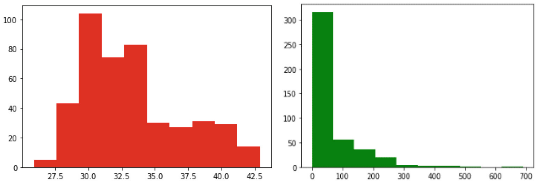


Fig. 3. Data Visualization-Histrogram

4.1.2 Exploratory Data Analysis (EDA)

An outlier is a value in a random sample from a Weather that is abnormally far apart from other values. It is up to the analyst to determine what constitutes aberrant behavior. It is crucial to classify typical observations before identifying anomalous observations. When the distribution is normal, two graphical techniques are employed to discover outliers: Box plots and scatter plots are two types of plots. The box plot is an effective graphical representation for understanding data behavior in the middle and tails of a distribution. The median, as well as the lower and higher quartiles, are used in the box plot (25th and 75th percentiles). The interquartile range, or IQ, is the gap between the lowest and upper quartiles of a population (Q3 - Q1). A moderate outlier is defined as a point on each side of an inner fence, whilst a severe outlier is defined as a point on either side of an outer barrier [20] (Fig. 2).

4.1.3 Histogram

A histogram is a graphical representation of data in the form of groups. It is an accurate method for graphically depicting numerical data distribution. It's a bar plot with the X-axis representing bin ranges and the Y-axis representing frequency [21] (Fig. 3).

4.1.4 Time Series Plot

Visualisation is essential in time series research and forecasting. Raw sample data plots provide important diagnostics for spotting temporal patterns like as trends, cycles, and seasonality, all of which impact model selection. The first and possibly most common time series visualisation is the line plot. The x-axis represents the month, while the y-axis represents the observation data [22] (Fig. 4).

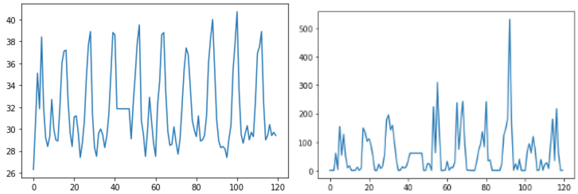


Fig. 4. Data Visualization-Time Series Plot

4.2 Correlation

The mean, standard deviation, variance, and covariance are all statistical numbers that are closely related to correlation. The associations between two or more variables (or features) of a dataset are frequently studied in statistics and data science [23]. The features are the traits or attributes of the observations, and each data point in the dataset is an observation. Variables and observations were used in every dataset we worked with. The features or variables are presented in the dataset (highest and mean maximum temperature, lowest and mean minimum temperature, total monthly rainfall, heavy rainfall in 24 h, number of rainy days, mean wind speed, relative humidity, mean station and mean sea level pressure).

When data is displayed in the form of a table, the observations are usually represented by the rows, while the features are represented by the columns. It is seen in a database table sample (Table 1).

Each entry in this table represents a single observation, or data regarding a single weather characteristic (either month of Jan, Feb, Mar, etc.). For all of the weather data, each column displays one feature (MN). We've looked at any two features in a dataset and discovered some sort of relationship between them. The following pair diagram shows the pairwise correlation between these highly associated qualities (Figs. 5, 6 and Tables 2, 3, 4, 5, 6 and 7).

Each of these plots shows one of three different forms of correlation:

a. **Negative correlation:** The y values tend to drop as the x values increase in the plot on the negative correlation. It was discovered that there is a strong negative association when large values of one trait correspond to small values of the other, and vice versa. The mean and lowest minimum temperatures and as well as the mean wind speed have a substantial negative association with the mean station and the mean sea level pressure.
b. **Weak or no correlation:** The centre plot displays no discernible trend. This is a type of weak correlation that happens when there is no evident or hardly discernible link between two features. The mean and highest maximum temperature, mean and lowest minimum temperature, mean wind speed, mean station and mean sea level pressure have no link with relative humidity.
c. **Positive correlation:** The y values tend to grow when the x values increase in the strong positive correlation. Strong positive correlation arises when large values of one property correspond to large values of the other, and vice versa [24]. Because a greater maximum temperature corresponds to a mean maximum temperature, and mean station level pressure relates to mean sea level pressure and vice versa, the

Table 1. Sample of one year (1985) weather database

MN	MMAX	HMAX	MMIN	LMIN	TMRF	HVYRF	RD	MWS	SLP	MSLP	RH
Jan	30.6	33.9	15	8.4	1	1	0	2	953.3	1009.9	38
Feb	33.1	35.2	13.3	9.6	1	1	0	2.5	949	1004.8	24
Mar	38.3	39.6	20.9	17.8	8.4	8.4	1	2.9	949.8	1004.8	19
Apr	38.1	42.4	22.5	18	8	8	1	4.1	947.2	1002.8	25
May	39.8	43	25.5	18.6	42.4	36.4	2	6	944.2	998.8	23
Jun	34.4	37.5	24.2	22.4	18	10.6	2	7.3	944.2	999.9	50
Jul	29.5	33.6	23	21.4	111.2	35.6	8	5.3	945.4	1001.7	60
Aug	30.1	35.4	22.3	21	11.3	4.4	2	5.7	945.2	1001.4	54
Sep	31.4	35	22.2	20.6	134.4	70.2	8	4.3	947.3	1003.3	51
Oct	30.2	32.8	18.4	13.2	59	28.4	4	4.2	950.1	1006.6	45
Nov	29.5	31.7	13.8	11.4	1	1	0	3.4	953.2	1009.8	31
Dec	29.8	32	13.7	9	1	1	0	3	953.6	1010.5	35

Table 2. Lag sample of temperature (1976 year)

Time (t)	Original data (Y_t)	1 step lagged (Y_{t-1})	2 step lagged (Y_{t-2})
1	29.1		
2	32.4	29.1	
3	36.8	32.4	29.1
4	38.4	36.8	32.4
5	40.1	38.4	36.8
6	34	40.1	38.4
7	31.2	34	40.1
8	30.2	31.2	34
9	31.1	30.2	31.2
10	35	31.1	30.2
11	29.1	35	31.1

connection between mean and highest maximum temperature, mean and lowest minimum temperature, mean wind speed, mean station and sea level pressure are positive correlation.

d. There is a medium (0.71) and negative (-0.55) association between relative humidity and total monthly rainfall.

4.2.1 Heat Map of Weather Dataset

A correlation matrix is a table that shows how two variables in a dataset are related to one another. Each value in this matrix reflects the correlation coefficient between

Table 3. Overall accuracy of weather parameter

Sr. no.	Methods	MMAX	HMAX	MMIN	LMIN	TMRF	HVYRF	RD	MWS	SLP	MSLP	RH
1.	LINEAR	2.59	2.79	2.86	3.26	30.77	13.82	3.36	2.25	1.76	1.84	15.5
2.	Exponential	2.72	2.92	2.85	3.22	37.39	13.8	3.2	2.44	1.74	1.84	13.7
3.	Quadratic	2.85	2.1	2.93	3.32	32.92	13.97	3.21	2.18	1.68	1.84	13.8
4.	Additive seasonality	1.14	1.5	1.14	2.3	28.14	11.46	1.84	1.42	1.21	1.45	5.63
5.	Additive Seasonality Quadratic Trend	1.14	1.42	1.19	1.87	29.3	11.43	1.51	1.06	1.19	1.43	5.07
6.	Multiplicative Seasonality	1.15	1.5	1.12	2.34	24.13	8.74	1.2	1.27	1.21	1.45	5.38
7.	Multiplicative Seasonality Linear Trend	1.2	1.44	1.15	2.32	27.59	10.13	1.35	1.23	1.23	1.48	6.48
Overall Accuracy		**1.83**	**1.95**	**1.89**	**2.66**	**32.16**	**11.91**	**2.24**	**1.69**	**1.43**	**1.62**	**9.37**

Table 4. Best fit model for weather parameter

Name of Models	MMAX	HMAX	MMIN	LMIN	TMRF	HVYRF	RD	MWS	SLP	MSLP	RH
Additive Seasonality Quadratic Trend		✓		✓				✓	✓	✓	✓
Multiplicative Seasonality			✓		✓	✓	✓				
Multiplicative Seasonality Linear Trend	✓										

the variables represented by the relevant row and column, and each row and column represents a variable. The correlation matrix is a vital data analysis metric, it is used to summarize data in order to better understand the link between different variables and make informed decisions. When dimensionality reduction on huge quantities of data is necessary, computing and evaluating the correlation matrix is a crucial pre-processing step in Machine Learning pipelines. Each cell in the correlation matrix represents a correlation coefficient between the two variables denoted by the cell's row and column [25] (Fig. 7).

4.3 Linear Method

Figure 8 depicts a basic technique of imputation that assumes a linear connection between missing and non-missing variables. This method fits a distinct linear polynomial between

Table 5. Original values of Weather parameters

Month	MMAX	HMAX	MMIN	LMIN	TMRF	HVYRF	RD	MWS	SLP	MSLP	RH
Jan	28.7	31.6	12.6	6	0	0	0	9.4	946.6	1010.1	23
Feb	31.2	34	15.6	11.9	3.3	2.5	1	10.8	944.6	1007.4	22
Mar	36	40.2	18.9	13.7	0.1	0.1	0	12.7	942.8	1004.5	14
Apr	39.3	41.2	22.7	12	0	0	0	10.9	940.7	1001.6	15
May	39.9	42	23.6	20.5	0	0	0	18.4	938.9	999.7	18
Jun	33.8	39.4	22.5	18.5	202.3	87.4	9	17.2	936.5	998.7	53
Jul	30.3	34.6	21.8	20.2	102	64	7	22.1	938.5	1001.5	59
Aug	28.7	31.7	20.7	18.7	127.7	45.8	6	17.6	939.1	1002.5	65
Sep	32.4	35	21.1	19	38.3	13.4	4	11.6	941.4	1004	45
Oct	34.5	37.6	18.6	12.7	13.1	13.1	1	6.9	944.3	1006.7	25
Nov	31.1	34	15	9.7	8	7.8	1	5.8	946.1	1009.2	32
Dec	28.2	31.5	11.2	7.1	30.1	18.6	1	6.1	948	1011.9	37

Table 6. Predicted values of Weather parameters

Month	MMAX	HMAX	MMIN	LMIN	TMRF	HVYRF	RD	MWS	SLP	MSLP	RH
Jan	29.33	31.85	11.81	7.01	0	0	0	8.8	946.53	1010.08	27
Feb	32.03	35.17	13.34	8.39	0	0	0	10.09	944.79	1007.48	18
Mar	36.09	39.14	17.58	11.55	0	0	0	11.74	943.5	1005.37	13
Apr	39.07	41.55	21.28	16.46	0	0	0	13.9	940.73	1001.85	12
May	40.29	42.21	23.44	19.05	0	0	0	18.9	938.77	999.53	16
Jun	35.35	40.27	21.92	19.18	165.2	31.25	5	21.29	936.23	998.05	44
Jul	31.12	34.58	21.17	19.5	119.5	37.74	9	20.4	936.67	999.32	61
Aug	29.62	32.56	20.32	18.55	86.89	26.16	8	21.48	937.6	1000.53	63
Sep	31.23	33.79	19.86	17.57	46.58	43.2	6	14.51	940.49	1003.36	55
Oct	32.84	35.09	17.77	13.51	17	14.3	2	9.22	943.92	1006.65	32
Nov	30.59	32.74	14.86	9.74	35.96	23.57	1	9.16	945.61	1008.85	32
Dec	29.13	31.04	11.53	6.11	20.67	14.4	1	7.96	946.92	1010.51	27

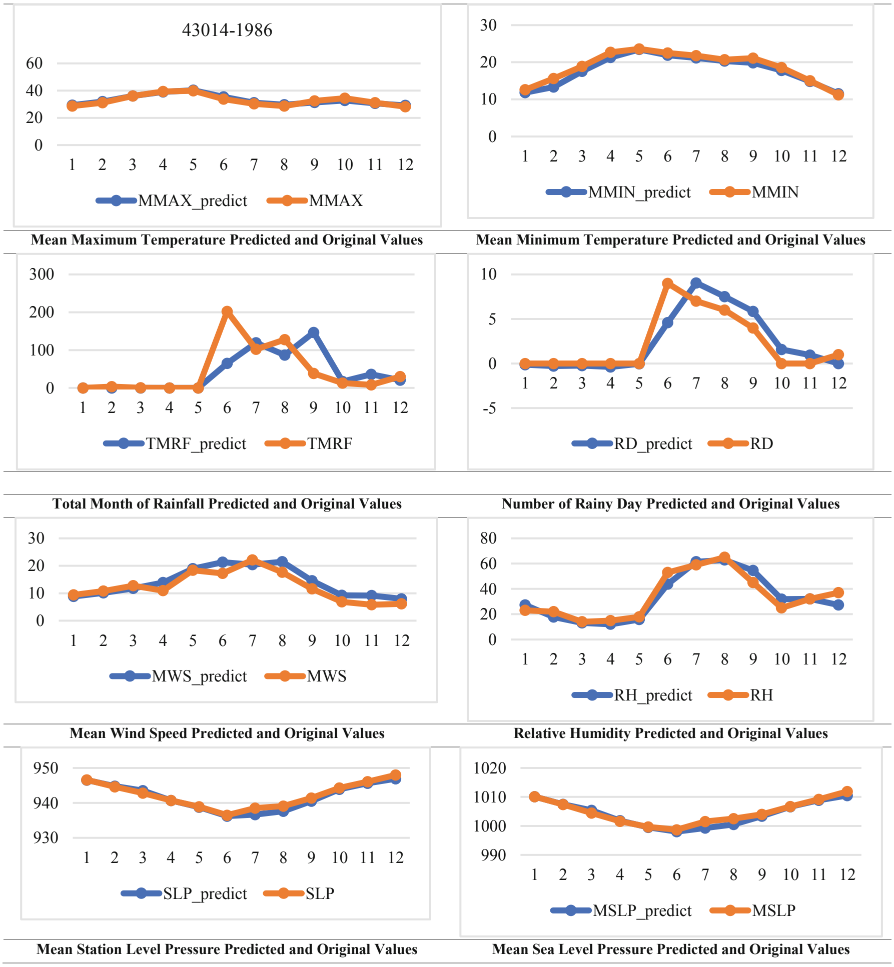
each pair of data points for curves, and a different linear polynomial between each set of three points for surfaces.

$$Y_t = \beta_0 + \beta_1 t + \varepsilon \tag{1}$$

$$y = y_1 + (x - x_1)\frac{(y_2 - y_1)}{(x_2 - x_1)} \tag{2}$$

The known value is x, while the unknown value is y. The coordinates x1 and y1 are below the known x value, whereas the coordinates x2 and y2 are above the known x value (Fig. 8). If the value at point Z is absent, the value at point Z will be computed using both the last real assessment performed before to point Z, which is point X, and the first actual assessment conducted after point Z, which is point Y [26].

Table 7. Original and Predicted values of Weather parameters



Mean Maximum Temperature Predicted and Original Values

Mean Minimum Temperature Predicted and Original Values

Total Month of Rainfall Predicted and Original Values

Number of Rainy Day Predicted and Original Values

Mean Wind Speed Predicted and Original Values

Relative Humidity Predicted and Original Values

Mean Station Level Pressure Predicted and Original Values

Mean Sea Level Pressure Predicted and Original Values

4.4 Exponential Model

Exponential functions are used to describe many physical phenomena, including populations, interest rates, radioactive decay, and the quantity of drugs in the circulation.

$$\mathrm{Log}(Y_t) = \beta_0 + \beta_1 t + \varepsilon \tag{3}$$

4.5 Quadratic Model

A mathematical model represented by a quadratic equation such as,

$$Y_t = \beta_0 + \beta_1 t + \beta_2 t^2 + \varepsilon \tag{4}$$

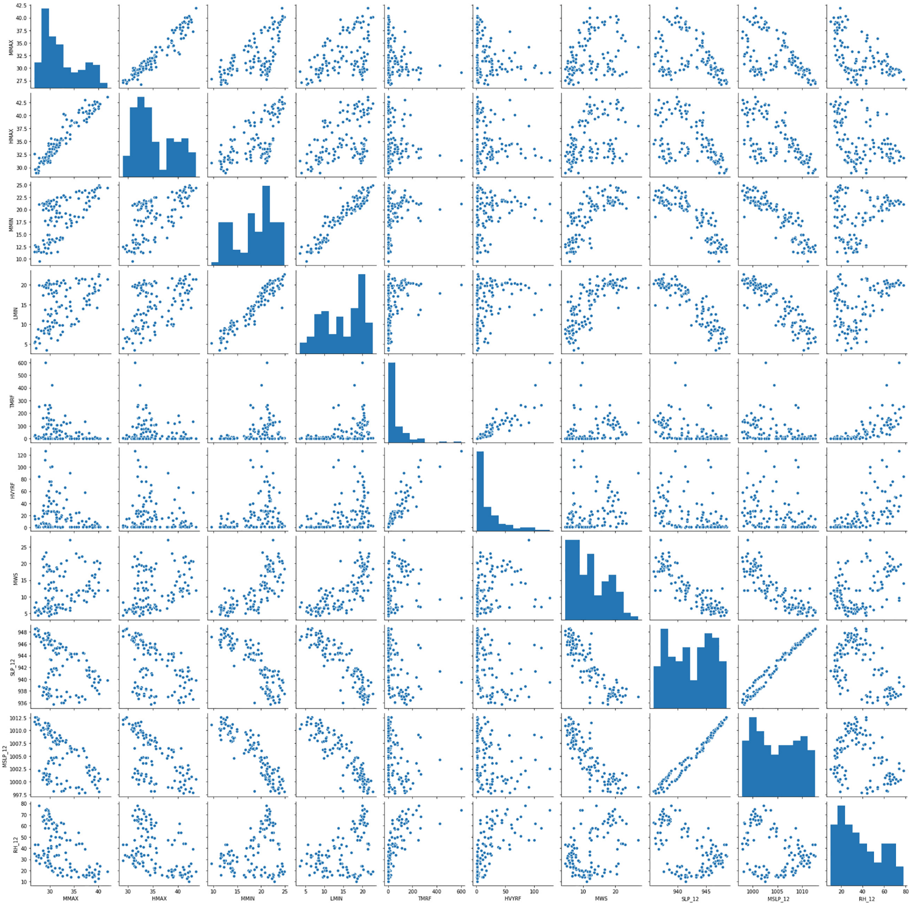


Fig. 5. Correlation of weather parameters

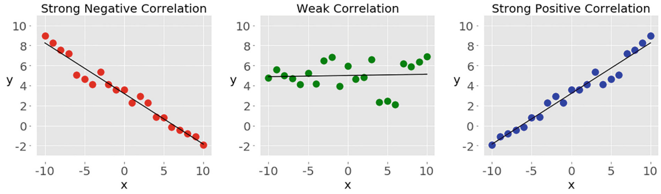


Fig. 6. Three different forms of correlation

A parabola is the graphed link between the variables in a quadratic equation. We plotted the data in scatter plots and used a graphing application to get the least squares regression lines. A similar approach might be used to find a model for nonlinear data. There are several methods to identify the model after utilizing a scatter plot to determine

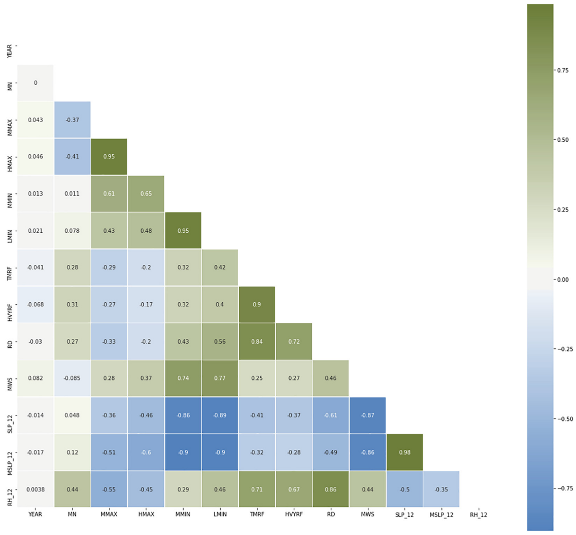


Fig. 7. A correlation matrix of data

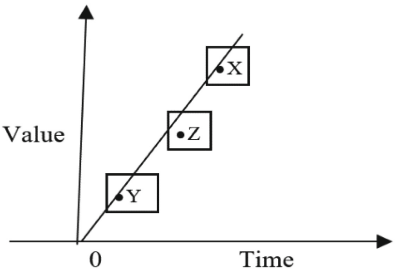


Fig. 8. Linear Relationship between data and missing values.

the type of model that would best suit a collection of data. Each method works effectively when utilizing a computer or calculator rather than hand calculations [27].

4.6 Forecasting Strategy

Seasonality is an important quality to consider when analyzing a Time Series. This property is said to be retained by a Time Series with a default behavior throughout a specified time period. We will have a sample that demonstrates seasonal activity if the pattern repeats itself within the same time range. This study is made easier by the Stats models package, but first, let's define a time series and its analytical properties. The average value of the series is called level. The growing or decreasing value is the series trend. The series' short-term cycle is repeated, which is referred to as seasonality. A noise is a term used to describe the random fluctuation of a series. There are two methods for studying a time series seasonality (additive and multiplicative) [28].

4.6.1 The Additive Model

The additive model is a form of data model that distinguishes and adds the effects of various components to synthetically represent the data. It is a systematic component of the forecasting model. It is represented by:

$$Y_t = L + T + S + N \tag{5}$$

where, L is level, T is the trend, S is seasonality and N is noise. The behavior is linear in the additive model, where adjustments are regularly made by the same amount over time, similar to a linear trend. Following equation is shown the Additive Seasonality (6) and Additive Seasonality with Quadratic Trend Eq. (7).

$$Y_t = \beta_0 + \beta_1 D_{Jan} + \beta_2 D_{Feb} + \beta_3 D_{Mar} + \ldots\ldots + \beta_{11} D_{Nov} + \varepsilon \tag{6}$$

$$Y_t = \beta_0 + \beta_1 t + \beta_2 t^2 + \beta_3 D_{Jan} + \beta_4 D_{Feb} + \beta_5 D_{Mar} + \ldots\ldots + \beta_{13} D_{Nov} + \varepsilon \tag{7}$$

4.6.2 The Multiplicative Model

The error component is compounded and added to the trend and seasonal components. It's a part of a systematic forecasting model [29]. A curving line represents the multiplicative model, which is not linear but might be quadratic or exponential.

$$Y_t = L \times T \times S \times N \tag{8}$$

The multiplicative model, in contrast to the additive model, has an amplitude and frequency that increases or decreases with time. Following equation is shown the Multiplicative Seasonality (9) and Multiplicative Seasonality Linear Trend Eq. (10).

$$\mathrm{Log}(Y_t) = \beta_0 + \beta_1 D_{Jan} + \beta_2 D_{Feb} + \beta_3 D_{Mar} + \ldots\ldots \beta_{11} D_{Nov} + \varepsilon \tag{9}$$

$$\mathrm{Log}(Y_t) = \beta_0 + \beta_1 t + \beta_2 D_{Jan} + \beta_3 D_{Feb} + \beta_4 D_{Mar} + \ldots\ldots + \beta_{12} D_{Nov} + \varepsilon \tag{10}$$

4.7 Forecasting Error

The discrepancy between the actual and anticipated value of a time series or any other event of interest is known as the forecast error [30]. Because the forecast error is *calculated from the same scale of data, it is only possible* to compare the forecast errors of various series when they are on the same scale. Forecast error (e_t) equation is observed values – predicted values.

$$e_t = Y_t - \hat{Y}_t \tag{11}$$

Where, Y_t is observed values and $\hat{Y}_t$ is forecasted values.

4.7.1 Lag Plot

Plots a variable against its own lagged sample to reveal probable associations between consecutive samples. For example, monthly rainfall/temp/wind/humidity in a year (Y_t = rainfall/temp in time period t, Y_{t+k} = rainfall/temp in time period t-k).

4.7.2 Evaluating Predictive Accuracy Equations

Mean Error

$$ME = \frac{1}{T}\sum\nolimits_{t-1}^{n} e_t \tag{12}$$

Mean Absolute Deviation

$$MAD = \frac{1}{n}\sum\nolimits_{t=1}^{n} |e_t| \tag{13}$$

Mean Squared Error.

$$MSE = \frac{1}{n}\sum\nolimits_{t=1}^{n} e_t^2 \tag{14}$$

Mean Root Squared Error

$$RMSE = \sqrt{\frac{1}{n}\sum\nolimits_{t=1}^{n} e_t^2} \tag{15}$$

Mean Percentage Error.

$$MPE\frac{1}{n}\sum\nolimits_{t=1}^{n} \frac{e_t}{y_t} \tag{16}$$

Mean Absolute Percentage Error.

$$MAPE = \frac{1}{n}\sum\nolimits_{t=1}^{n} \left|\frac{e_t}{y_t}\right| \tag{17}$$

4.7.3 Root Mean Squared Error Values (RMSE)

The square root of the mean of all errors squared is used to calculate the RMSE. Because it is scale-dependent and not across variables, it is only helpful for comparing prediction errors of different models or model configurations for a single variable. RMSE is the residuals' standard deviation (prediction errors). The residuals show how far the data points are from the regression line, and the RMSE shows how widely these residuals are distributed. It represents how tightly the data clusters around the best-fit line. To validate experimental data, RMSE is frequently used in climatology, forecasting, and regression analysis [31].

Additive Seasonality Quadratic Trend is best fit model for the highest maximum temperature (1.42), lowest minimum temperature (1.87), wind speed (1.06), relative humidity (5.07), mean station (1.19), and mean sea level pressure (1.43). Multiplicative Seasonality model is the best model for mean minimum temperature (1.12) total rainfall in the month (24.13), heavy rainfall (8.74), and number of rainy days (1.2). Multiplicative Seasonality Linear Trend is given good accuracy for mean maximum temperature (1.2). The linear, exponential, quadratic and additive seasonality are not given good result for weather parameter.

The temperature (mean and high maximum, mean and lowest minimum), number of rainy day, mean wind speed, mean station and sea level pressure are given good RMSE values such as 1.83, 1.95, 1.89, 2.66, 2.24,1.69, 1.43, 1.62 respectively. Relative humidity is given 9.37 RMSE values. Total month of rainfall and heavy rainfall is given 32.16 and 11.9 RMSE values. Total month of rainfall accuracy is not good because rainfall is deepened location, time, climate and geographical region. Due to this rainfall is high, medium and low in the rainy season. The mean maximum temperature prediction model gives 98.98% accuracy using Multiplicative Seasonality Linear Trend and rainfall prediction model is given 75.87% accuracy using Multiplicative Seasonality machine learning techniques.

We have used seven machine learning techniques for the prediction of weather model such as linear, exponential, quadratic, additive seasonality, additive seasonality quadratic trend, multiplicative seasonality, and multiplicative seasonality linear trend. The linear, exponential, quadratic and additive seasonality are not given good result for any weather parameter.

The additive seasonality quadratic trend is best fit model for the HMAX, LMIN, MWS, RH, SLP and MSLP. The multiplicative seasonality model is the best model for the MMIN, TMRF, HVYRF and RD. The multiplicative seasonality linear trend is given good accuracy for MMAX.

4.7.4 Time Series Partitioning

We have used one decade historical data for forecasting. We have used 70% data for training. It is fitted the model only to training period and 30% data is used for validation. It assess performance on validation period.

4.8 Autocorrelation

The correlation between a variable and its lagged form is known as autocorrelation (one step, two step or multiple time step).

$$r_k = \frac{\Sigma_{t=k=1}^{n}\left(Y_t - \overline{Y}\right)\left(Y_{t-k^{-Y}}\right)}{\Sigma_{i=1}^{n}\left(Y_t - \overline{Y}\right)^2},\ \mathrm{K} = 0, 1, 2, \ldots \tag{18}$$

Where,

(Y_t) is observation in time period (t).

($Y_{t\text{-}k}$) is observation in time period (t-k).

($\bar{Y}$) is mean of the values of the series.

(r_k) is autocorrelation coefficient for k-step lag.

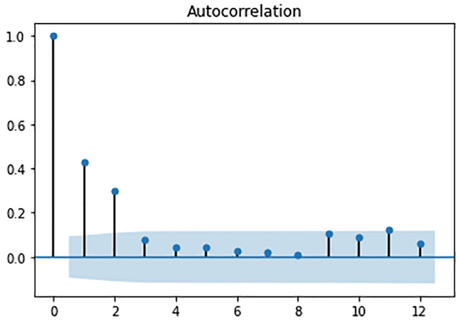


Fig. 9. ACF plot of data

4.8.1 Correlogram or ACF Plot

The ACF is plotted versus the lag. Limits to be exceeded for statistical significance are indicated as plus and minus two standard errors. Identifies lagged variables that is beneficial for predicting [32].

4.8.2 ACF Plot on Residuals

ACF is a (complete) auto-correlation function that returns auto-correlation values for any time series with lagged values (Fig. 9).

4.8.3 Partial Auto-correlation Function (PACF)

It finds correlations of present with lags of the residuals (prediction errors) of the time series (Fig. 10).

We have got the slightly changes in the original and predicted values for mean maximum temperature, mean station and sea level pressure. The rainfall, wind speed and humidity are increased in the monsoon season. Rainfall is highest in the months of July and September, and lowest in the months of June and August. The RMSE value in rainfall is high since rainfall is deepened based on current meteorological knowledge. Sometime rainfall is high and some time it is low or medium. Rainfall is higher in the

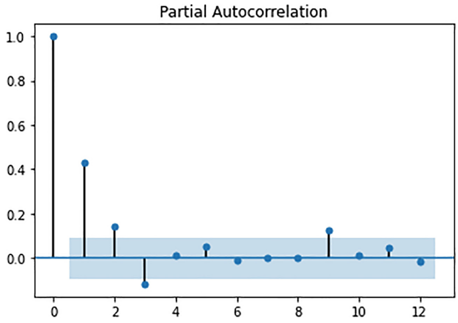


Fig. 10. PACF plot of data

months of June, July, and September, and lower in the month of August. The number of rainy day is depended on how much day rainfall is happened in the month. Number of rainy day is highest in the month of July. The wind speed and relative humidity are increased in the monsoon season. It correlation between rainfall. The station and sea level pressure are decreased in the rainy season. It is increased in the summer and winter season.

5 Conclusions

The weather parameters are playing vital role. The weather parameters are interdependent and current situation. If the extreme events are happened in the year such as flooding, earthquake, cyclone, etc. it will effect on the weather condition. We have taken database form the Indian Metrological Department (IMD), Pune, Maharashtra, India. We have covered the five Marathwada region districts such as Aurangabad, Beed, Parbhani, Osmanabad and Nanded. We have implemented missing imputation, outliers, histogram, scatter plot, barplot and time series plot for preprocessing dataset. It is helped to improve the result accuracy. We have used autocorrelation and seven models for furcating such as linear, exponential, quadratic, additive seasonality, additive seasonality quadratic trend, multiplicative seasonality, and multiplicative seasonality linear trend.

Additive seasonality quadratic trend is the best model for the highest maximum temperature, lowest minimum temperature, wind speed, relative humidity, mean station and sea level pressure. Multiplicative seasonality model is the best model for mean minimum temperature, total rainfall in the month, heavy rainfall, and number of rainy days. Multiplicative Seasonality Linear Trend is given good accuracy for mean maximum temperature. The linear, exponential, quadratic and additive seasonality are not given good result for weather parameter. We have created forecasting system for Marathwada region by using machine learning tools. We have got overall accuracy 1.83, 1.95, 1.89, 2.66, 32.16, 11.91, 2.24, 1.69, 1.43, 1.62, 9.37 for mean and highest maximum temperature, mean and lowest minimum temperature, total month of rainfall, heavy rainfall in the month, number of rainy day, mean wind speed, mean station and sea level pressure and relative humidity respectively.

References

1. Semenov MA, Barrow EM. Use of a stochastic weather generator in the development of climate change scenarios. Climatic Change 1997; 35: 397-414.
2. Wilks DS. Adapting stochastic weather generation algorithms for climate change studies. Climatic Change 1992; 22: 67-84.
3. Pruski FF, Nearing MA. Climate-induced changes in erosion during the 21st century for eight U.S. locations. Water Resour Res 2002; 38: 341–3411.
4. Zhang XC, Nearing MA, Garbrecht JD, Steiner JL. Downscaling monthly forecasts to simulate impacts of climate change on soil erosion and wheat production. Soil Sci Soc Am J 2004; 68: 1376–85.
5. Zhang XC. Spatial downscaling of global climate model output for site-specific assessment of crop production and soil erosion. Agr Forest Meteorol 2005; 135: 215–29.

6. Zhang XC, Liu WZ. Simulating potential response of hydrology, soil erosion, and crop productivity to climate change in Changwu tableland region on the Loess Plateau of China. Agr Forest Meteorol 2005; 131: 127-42.
7. Kilsby CG, Jones PD, Burton A, Ford AC, Fowler HJ, Harpham C, James P, Smith A, Wilby RL. A daily weather generator for use in climate change studies. Environ Modell Softw 2007; 22: 1705-19.
8. Richardson CW. Stochastic simulation of daily precipitation, temperature, and solar radiation. Water Resour Res 1981; 17: 182-90.
9. Richardson CW, Wright DA. WGEN: A model for generating daily weather variables. U.S. Depart. Agr, Agricultural Research Service. Publ. ARS-8; 1984, p. 1–86.
10. Stockle CO, Campbell GS, Nelson R. ClimGen Manual. Biological Systems Engineering Department, Washington State University, Pullman, WA; 1999.
11. Semenov MA, Barrow EM. LARS-WG, A Stochastic Weather Generator for Use in Climate Impact Studies, User Manual; 2002.
12. Dubrovsky M, Buchteke J, Zalud Z. High-frequency and low-frequency variability in stochastic daily weather generator and its effect on agricultural and hydrologic modeling. Climatic Change 2004; 63: 145-79.
13. Hansen JW, Mavromatis T. Correcting low-frequency variability bias in stochastic weather generators. Agr Forest Meteorol 2001; 109: 297-310.
14. Wang QJ, Nathan RJ. A method for coupling daily and monthly time scales in stochastic generation of rainfall series. J Hydrol 2007; 346: 122-30.
15. Chen J, Brissette PF, Leconte R. A daily stochastic weather generator for preserving low-frequency of climate variability. J Hydrol 2010; 388: 480-90.
16. Mehrotra, R., Sensitivity of runoff, soil moisture and reservoir design to climate change in central Indian river basins. Climatic Change, 1999, 42, 725-757.
17. IPCC, Summary for policymakers. In Climate Change 2007: The Physical Science Basis (eds Solomon, S. D. et al.), Cambridge University Press, Cambridge, UK, 2007.
18. Kundzewicz, Z. W., Change detection in hydrological records - a review of the methodology. Hydrol. Sei., J., 2004, 49(1), 7–19.
19. Sen, P. K., Estimates of the regression coefficient based on Kend all's tau. J. Am. Stat. Assoc., 1968, 63, 1379-1389.
20. Lettenmaier, D. P., Wood, E. F. and Wallis, J. R., Hydro climatological trends in the continental United States, 1948-88. J. Climate, 1994, 7, 586-607.
21. Yue, S. and Hashino, M., Temperature trends in Japan: 1900 1990. Theor. Appl. Climatol., 2003, 75, 15-27.
22. Partal, T. and Kahya, E., Trend analysis in Turkish precipitation data. Hydrol. Process, 2006, 20, 2011-2026.
23. Yu, Y. S., Zou, S. and Whittemore, D., Non-parametric trend analysis of water quality data of rivers in Kansas. J. Hydrol., 1993, 150, 61-80.
24. Douglas, E. M., Vogel, R. M. and Knoll, C. N., Trends in flood and low flows in the United States: impact of spatial correlation. J. Hydrol., 2000, 240, 90-105.
25. Yue, S., Pilon, P. and Phinney, B., Canadian streamflow trend detection: impacts of serial and cross-correlation. Hydrol. Sei. J., 2003,48, 51-63.
26. Burn, D. H., Cunderlik, J. M. and Pietroniro, A., Hydrological trends and variability in the Liard river basin. Hydrol. Sei. J., 2004, 49, 53-67.
27. Singh, P., Kumar, V., Thomas, T. and Arora, M., Changes in rain fall and relative humidity in different river basins in the northwest and central India. Hydrol. Process., 2008, 22, 2982-2992.
28. Singh, P., Kumar, V., Thomas, T. and Arora, M., Basin-wide assessment of temperature trends in the north-west and central India. Hydrol. Sei. J., 2008, 53, 421-433.
29. Salas, J. D., Analysis and modeling of hydrologie time series. In Handbook of Hydrology (ed. Maidment, D. R.), McGraw-Hill, New York, 1993, p. 19.1–19.72.

30. Helsel, D. R. and Hirsch, R. M., Statistical Methods in Water Resources, Elsevier, New York, 1992.
31. Hirsch, R. M., Helsel, D. R., Cohn, T. A. and Gilroy, E. J., Statistical treatment of hydrologie data. In Handbook of Hydrology (ed. Maidment, D. R.), McGraw-Hill, New York, 1993, p. 17.1–17.52.
32. Srivastava, H. N., Sinha Ray, K. C., Dikshit, S. K. and Muk hopadhaya, R. K., Trends in rainfall and radiation over India. Vayu Mandai, 1998, 41–45.