Journal Homepage: - www.journalijar.com

INTERNATIONAL JOURNAL OF ADVANCED RESEARCH (IJAR)

Article DOI: 10.21474/IJAR01/20652
DOI URL: http://dx.doi.org/10.21474/IJAR01/20652

RESEARCH ARTICLE

UNRAVELING PATTERN AND FORECASTING URBAN RAINFALL USING TIME SERIES ANALYSIS

M. Manoprabha and Joel Jossy
Department of Statistics and Applied Mathematics, Central University of Tamil Nadu, Thiruvarur.

Manuscript Info

Manuscript History
Received: 21 January 2025
Final Accepted: 24 February 2025
Published: March 2025

Key words:-
Pattern Classification, Clustering, Sarima, STL Decomposition, Seasonal Naïve Forecasting, Performance Metrics

Abstract

The primary objective of this urban study is to identify the most effective forecasting method for highly seasonal time series data, using monthly rainfall records for Chennai from 1901 to 2021. The analysis begins with data visualization to uncover long-term trends and seasonal variations. We apply clustering techniques specifically to seasonal components of the rainfall data to group similar seasonal behaviours and reveal distinct rainfall regimes across different periods. The structure and distribution of data within each cluster are analyzed to better understand rainfall variability and recurring seasonal patterns. Following this, three forecasting models-ARIMA, STL decomposition, and seasonal naïve forecasting-are implemented. The performances of these methods are evaluated using the standard metrics of Mean Absolute Error (MAE), Root Mean Squared Error (RMSE), and Mean Absolute Scaled Error (MASE). Among the models tested, STL decomposition performs the best, achieving the lowest MAE (67.99), RMSE (125.03), and MASE (0.67). Its ability to isolate trend, seasonality, and residuals allows for more accurate forecasting of complex and highly seasonal rainfall patterns. These findings demonstrate the value of integrating clustering with seasonal analysis and underscore the robustness of STL decomposition in environmental time series forecasting. Leveraging this finding, STL decomposition is utilized to forecast rainfall for the entire dataset. Forecasted values are merged with the original data to reapply K-means clustering and validate consistency in rainfall regimes. The analysis reveals a remarkable similarity in the distribution of data across the new clusters, indicated by an Adjusted Rand Index of 0.95.This shows that STL decomposition has effectively captured the underlying trends and patterns in this highly seasonal data.

Introduction:-

Motivation

Extreme rainfall events have become increasingly common due to global climate change, transforming urban floods from rare disasters into frequent crises [3]. Cities like Chennai, located on the southeast coast of India, are particularly vulnerable. Characterized by a tropical wet and dry climate, Chennai relies heavily on the monsoon season for agriculture, water management, and urban sustainability. However, rainfall variability— from intense

Corresponding Author:- M. Manoprabha
Address:- Department of Statistics and Applied Mathematics, Central University of Tamil Nadu, Thiruvarur.

flooding to prolonged droughts—poses significant challenges for disaster preparedness, infrastructure planning, and resource allocation [4].

The uncertainty in flood forecasting stems from various sources: unpredictable rainfall, complex model structures, and variability in parameters. Machine learning and statistical approaches have been increasingly employed to address this complexity, yet forecast accuracy remains a key challenge, particularly for highly seasonal and non-linear systems [1]. As highlighted by recent studies, improving forecast reliability is a pressing need for urban planning and disaster mitigation [9].

Objectives:-

This study aims to develop a data-driven approach for understanding and forecasting rainfall in Chennai by:

1. Analyzing over a century's worth of monthly rainfall data (1901–2021) to identify long-term trends, cycles, and anomalies.
2. Employing clustering techniques to group similar seasonal patterns, thereby uncovering latent structures within the data.
3. Comparing multiple time series forecasting models—specifically SARIMA, STL decomposition, and seasonal naïve forecasting—for their ability to predict future rainfall patterns with high accuracy.
4. Evaluating forecasting performance using metrics such as Mean Absolute Error (MAE), Root Mean Squared Error (RMSE), and Mean Absolute Scaled Error (MASE) [2].

Contribution

The primary contribution of this study lies in the integration of clustering with time series forecasting to enhance rainfall prediction for a highly seasonal dataset. By identifying homogenous clusters of rainfall behaviour, this approach aids in capturing distinct seasonal regimes, which improves the granularity and interpretability of forecasting models [5].

By combining statistical forecasting with unsupervised learning, this research advances methodologies for rainfall prediction in monsoon-dependent urban areas and provides a replicable framework for similar climate-vulnerable regions [1]. The use of time series modelling in climate forecasting, as supported by previous literature, shows promise for informed decision-making and sustainable development planning [6].

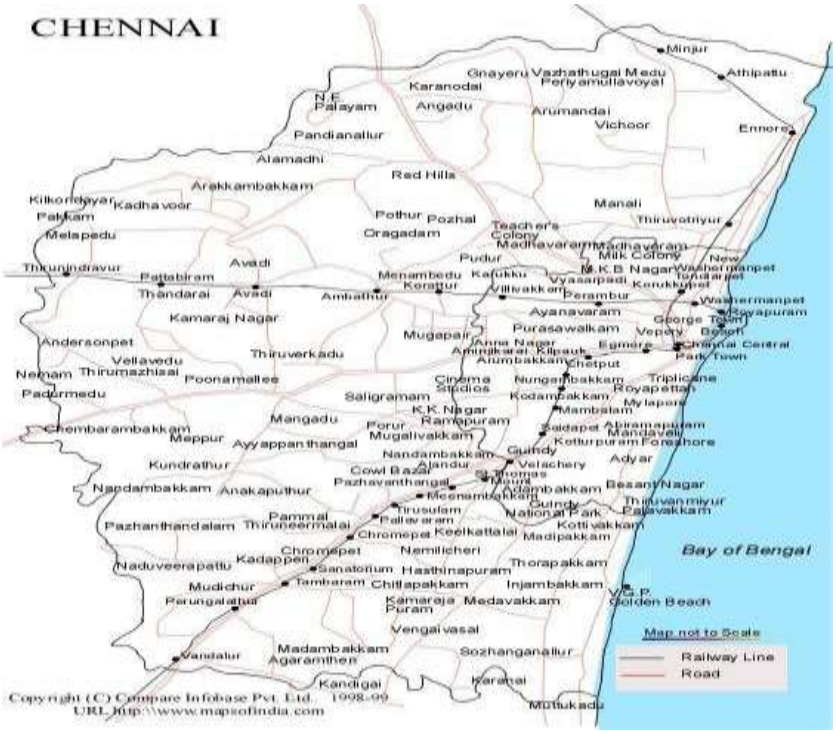

Figure 1.1:- Geographical area of the study.

Methodology:-

Sarima

Seasonal Auto-Regressive Integrated Moving Average (SARIMA) is an extension of ARIMA model that incorporates seasonality in addition to non-seasonal components. [10]

STL decomposition

Seasonal-Trend decomposition using LOESS (Locally weighted regression and scatter plot smoothing) is a method used to decompose the time series into 3 components seasonal, trend and remainder components.[11]

Seasonal Naïve Forecasting

The idea behind seasonal naïve forecasting is to use the observation from the previous season as forecast for the corresponding season in the future. [12]

Performance Metrics of forecasting models

In this paper, performance of the forecasting models is assessed using 3 metrics MAE, RMSE, MASE. The lower values of this metrics indicate better performance of the model.

Mean Absolute Error (MAE)

MAE measures the mean of the errors between actual value and forecasted value. It can be mathematically expressed as,

$$\mathbf{M.A.E} = \frac{\mathbf{1}}{\mathbf{n}}\sum_{\mathbf{i=1}}^{\mathbf{n}} |\ \widehat{\mathbf{Y}}_\mathbf{i} - \mathbf{Y_i}| \qquad \ldots\ldots\ldots\ldots\ldots (2.1)$$

Here, n is the number of observations, $\widehat{Y}_i$ is the forecasted value and Y_i is the actual value.[12]

2.4.1. Root Mean Squared Error (RMSE)

RMSE is similar to MAE but it penalises large errors more heavily because of the squaring of errors. It is expressed mathematically as,

$$\mathbf{R.M.S.E} = \sqrt{\frac{\mathbf{1}}{\mathbf{n}}\sum_{\mathbf{i=1}}^{\mathbf{n}}(\widehat{\mathbf{Y}}_\mathbf{i} - \mathbf{Y_i})^\mathbf{2}} \qquad \ldots\ldots\ldots\ldots\ldots (\mathbf{2.2})$$

Here, n is the number of observations, $\widehat{Y}_i$ is the forecasted value and Y_i is the actual value.[12]

Mean Absolute Scaled Error (MASE)

MASE is a normalized version of MAE that compares the performance of a forecasting model to that of a naive model. It's particularly useful when dealing with data that has seasonality. It is expressed mathematically as,

$$\mathbf{M.A.S.E} = \frac{\mathbf{M.A.E}}{\frac{\mathbf{1}}{\mathbf{n-m}}\sum_{\mathbf{i=m+1}}^{\mathbf{n}} |\mathbf{Y_i} - \mathbf{Y_{i-m}}|} \qquad \ldots\ldots\ldots\ldots\ldots (\mathbf{2.3})$$

Here, n is the number of observations, m is the seasonal period, Y_i is the actual value at i, Y_{i-m} is the actual value at (i-m) and M.A.E is mean absolute error.[12]

K-Means Clustering for Seasonal Rainfall Analysis

K-means clustering was chosen over other unsupervised learning methods, such as DBSCAN and hierarchical clustering, due to its efficiency, scalability, and suitability for continuous numerical data like monthly rainfall records. Unlike DBSCAN, which is sensitive to the selection of parameters like epsilon and can struggle with varying densities, K-means performs well when clusters are relatively spherical and similar in size, conditions that align with seasonal rainfall groupings. Hierarchical clustering, while useful for capturing nested relationships, becomes computationally intensive with large datasets and lacks scalability compared to K-means. Additionally, K-means provides clear, non-overlapping cluster assignments and centroids that are interpretable in terms of average seasonal patterns. These characteristics make it a pragmatic and effective choice for identifying distinct rainfall regimes across the long-term dataset used in this study.[13,14]

Adjusted Rand Index (ARI)
ARI is a measure of similarity between two data clusters. It accounts for chance agreement between the clusters.ARI returns a value between -1 and 1, where 1 indicates perfect agreement between clusters,0 indicates that the clustering is no better than random and negative values indicate disagreement worse than random[15]

Results and Discussion:-

Monthly rainfall data for Chennai city for the years 1901 to 2021 have been obtained from https://data.opencity.in. This extensive dataset offers a valuable resource for understanding the historical precipitation patterns in one of India's major urban centers. Over the course of more than a century, these records encapsulate the fluctuations and trends in rainfall that have impacted Chennai's environment.

The data from January 1901 to December 2021 is plotted in the figure 3.1. This figure depicts the fluctuation in rainfall for the city. This visualization provides valuable insights into the temporal variability of rainfall patterns, highlighting periods of abundance and scarcity.

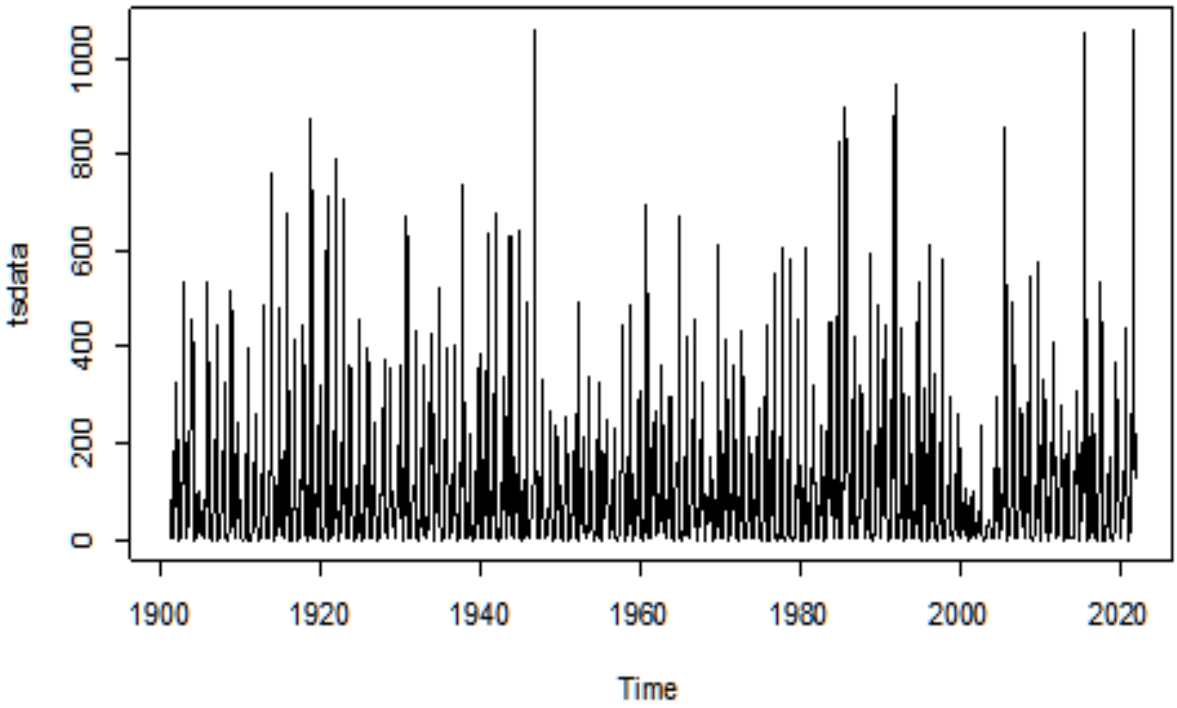

Figure 3.1:- Graph of Rainfall.

The rainfall data is subjected to a K-means clustering algorithm, resulting in the classification of the dataset into seven distinct clusters. These clusters represent groups of data points with similar patterns or characteristics in terms of rainfall variability[13]. **Figure 3.2 – Figure 3.8** visually presents these clusters, providing a graphical representation of how the data points are grouped together based on their respective rainfall patterns. These 7 patterns are then used to investigate the rainfall data in a very comprehensive way. This graph clearly shows that pattern 2 is associated with low rainfall and pattern 6 is associated with high levels of rainfall.

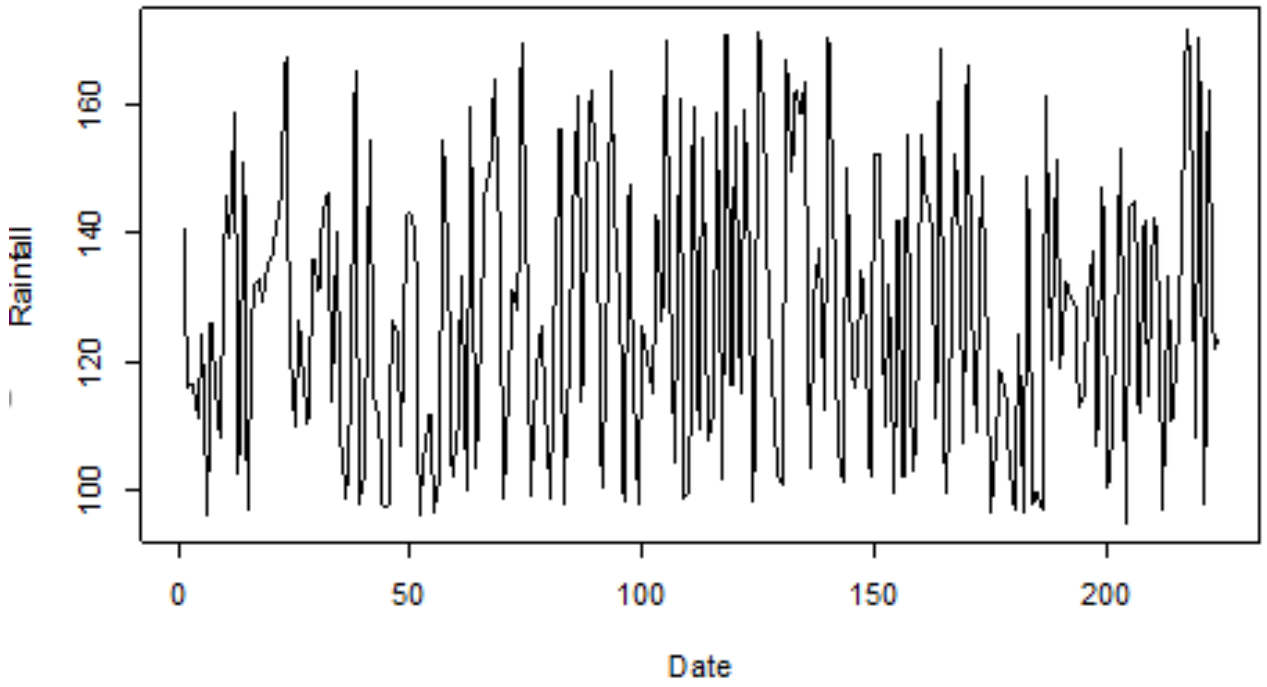

Fig. 3.2:- Diagrammatic representation of pattern 1.

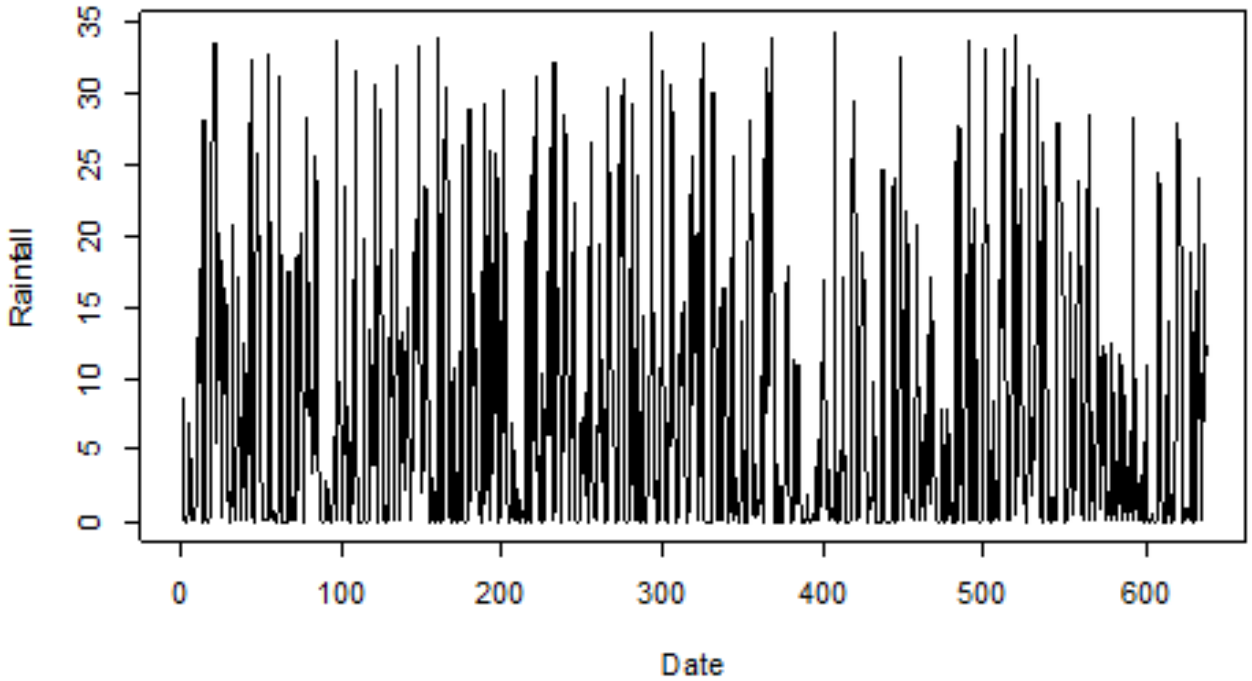

Fig. 3.3:- Diagrammatic representation of pattern 2.

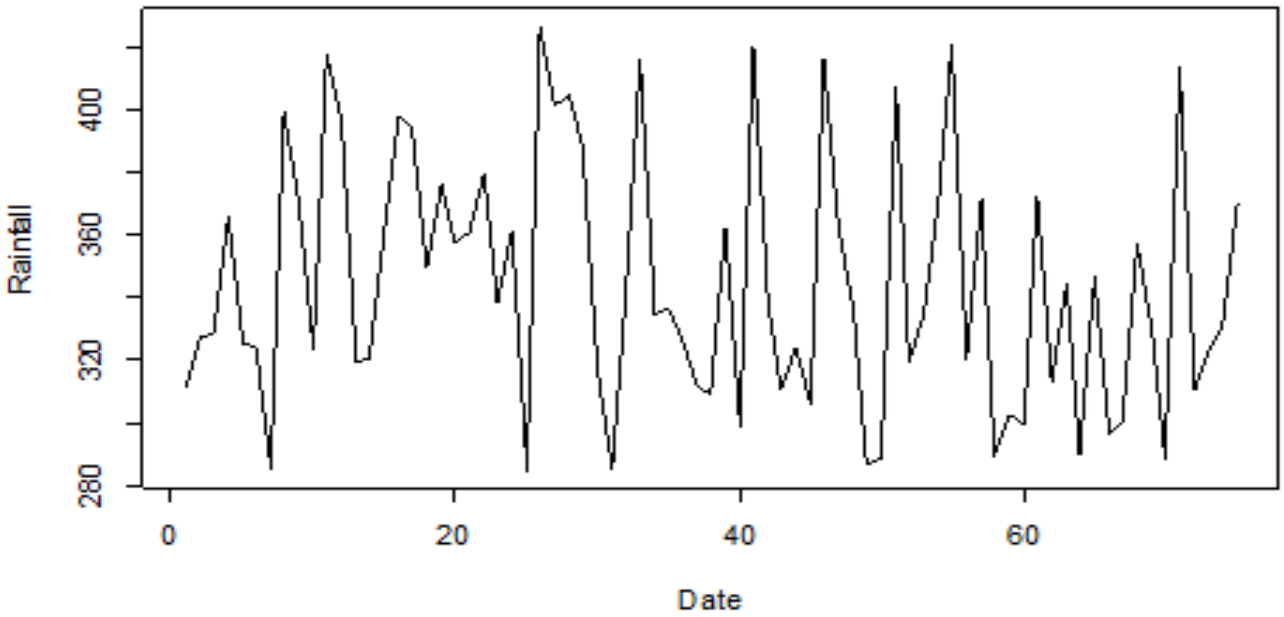

Fig. 3.4:- Diagrammatic representation of pattern 3.

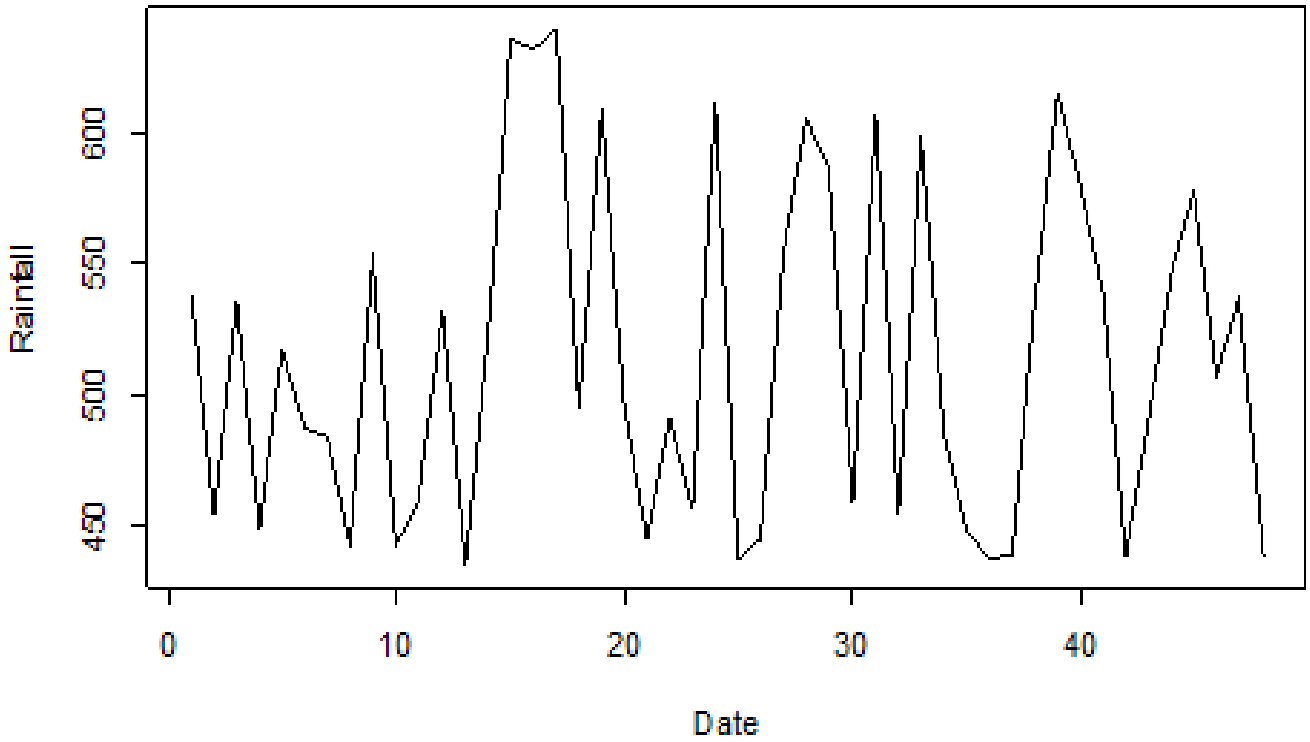

Fig. 3.5:- Diagrammatic representation of pattern 4.

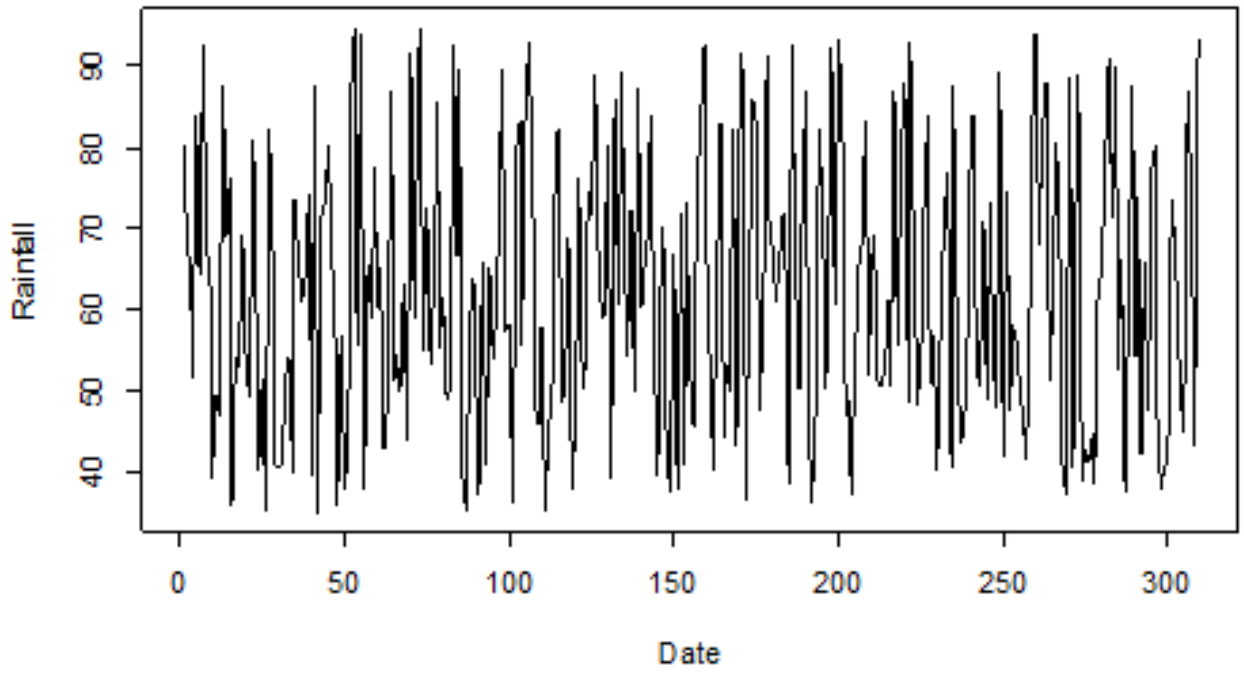

Fig. 3.6:- Diagrammatic representation of pattern 5.

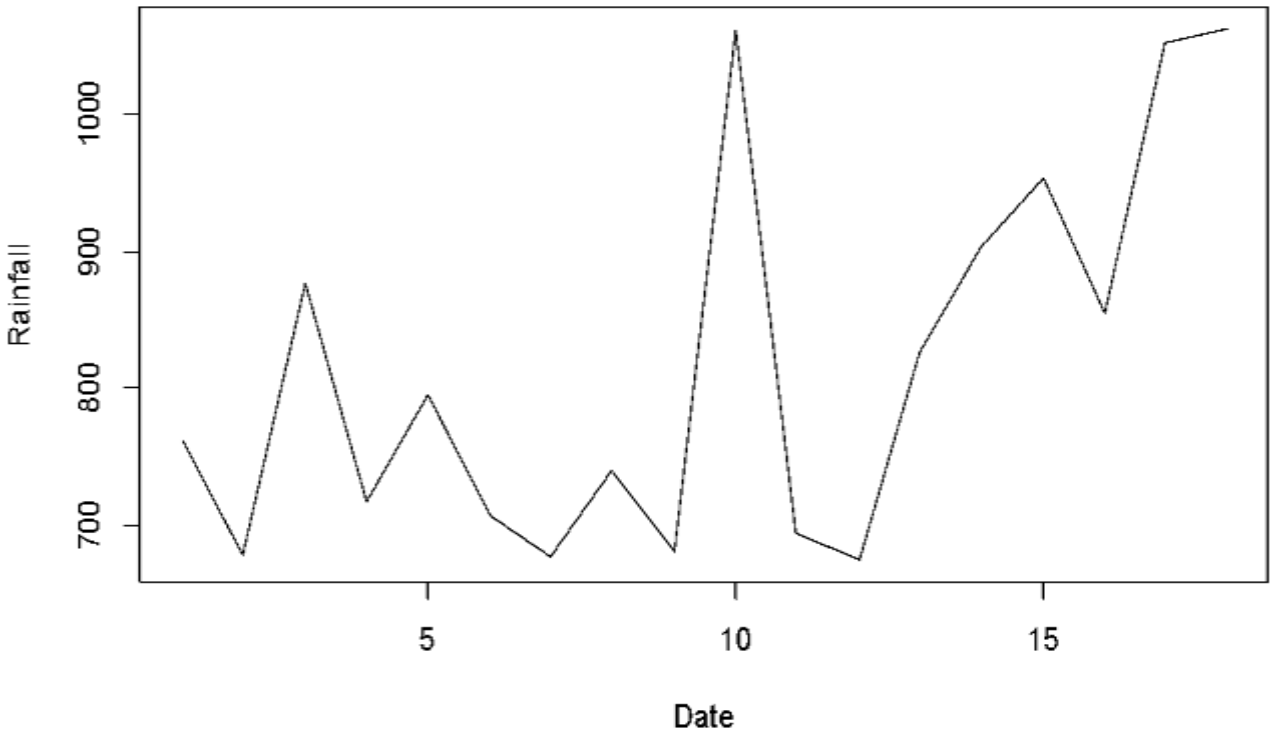

Fig. 3.7:- Diagrammatic representation of pattern 6.

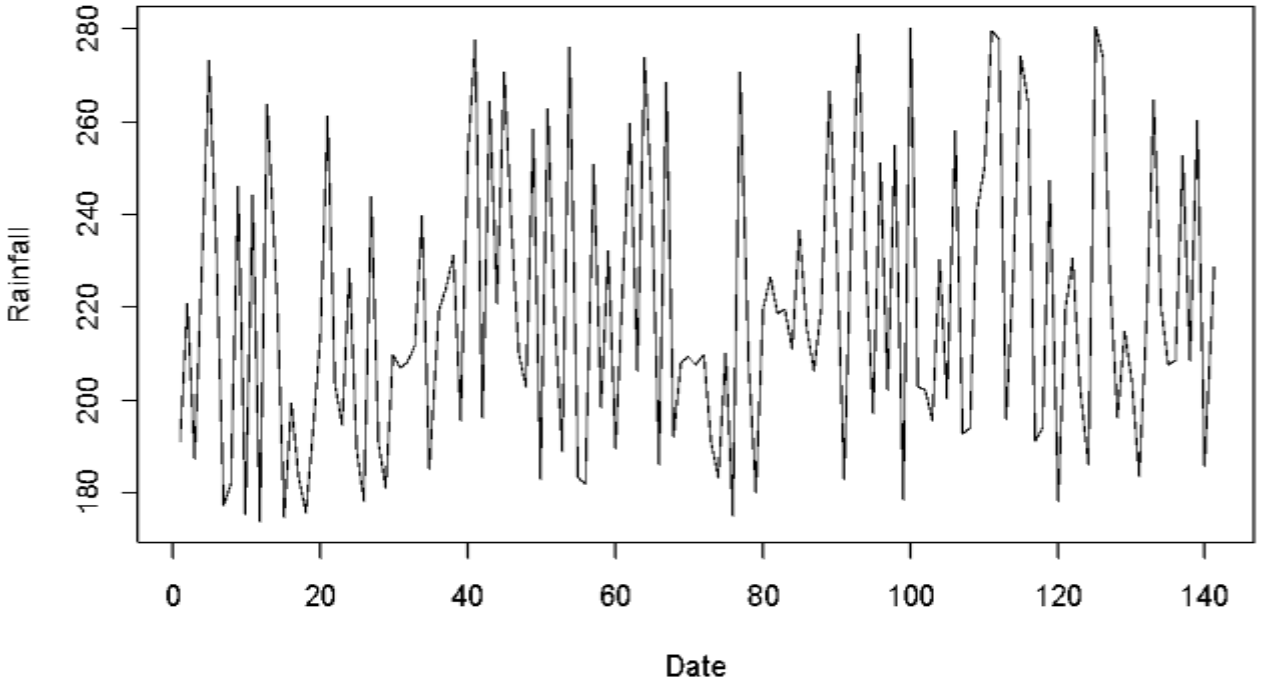

Fig. 3.8:- Diagrammatic representation of pattern 7.

Table 3.1:- Characteristics of each cluster.

Descriptive Statistics	C1	C2	C3	C4	C5	C6	C7
Data count	224	637	75	48	309	18	141
Data percent	15.43	43.87	5.17	3.31	21.28	1.24	9.71
Min	95.22	0	285.19	435.03	35.21	674.19	173.93
Max	171.82	34.31	426.35	639.64	94.69	1061.64	280.46
Mean	127.61	7.64	344.68	515.36	62.63	817.22	219.55
Median	126.04	2.21	336.48	501.13	61.12	778.08	214.22
Mode	95.22	0	285.19	435.03	35.21	674.19	173.93
Standard Deviation	22.21	9.97	40.53	65.63	17.05	139.35	30.67
Variance	493.25	99.42	1642.57	4307.47	290.71	19417.4	940.6
IQR	37.28	13.53	59.16	106.79	27.38	198.20	48.21
Skewness	0.3	1.18	0.40	0.39	0.22	0.66	0.45
Kurtosis	1.9	3.14	2.08	1.84	1.90	2.08	2.08

In table 3.1, C1, C2, ….., C6 corresponds to cluster 1, cluster 2,….., cluster 6.The inference drawn from Figure 3.2 is true since heavy rainfall is associated with C6 with a mean rainfall of 817.22 while C2 is associated with low rainfall with a mean rainfall of 7.64.

We can comment about the distribution of each cluster using the table. For example, Cluster 1 represents a moderate rainfall level, with values between 95.22 and 171.82. The rainfall distribution in this cluster is approximately symmetric, as indicated by the skewness and kurtosis values close to zero. The value of mode in this cluster is 95.22. The mean rainfall in this cluster is approximately 127.61, close to the median value of 126.04, indicating a relatively balanced distribution. Given a standard deviation of 22.21, the data points are relatively close to the mean, there is only a moderate level of variability in the rainfall.

From both **Figure 3.2** to **Figure 3.8**, as well as the data presented in **Table 3.1**, stakeholders can draw meaningful conclusions and gain a general understanding of the rainfall behaviour during each time period associated with the respective clusters predicted using the K-means clustering algorithm. This enables data-driven insights into seasonal trends and variability, which can be useful for planning and decision-making.

ARIMA, STL Decomposition and Seasonal Naïve Forecasting methods are used to predict the rainfall. The best predictive model has been obtained by splitting the data into two sets. The one set of data will contain the rainfall data from January 1901 to December 2010, while the second set will contain the rest of the data from January 2011 to December 2021.We used the first set of data to forecast the rainfall during Jan 2011 – Dec 2021 using the three methods and find the actual vs. forecasted value plot as well as the residual to find the metrics of Mean Absolute Error (MAE),Root Mean Squared Error (RMSE) and Mean Absolute Scaled Error (MASE).The method with least values of this metrics is the best method for this case.

Table 3.2:- Best method for forecasting.

	ARIMA	**STL**	**Seasonal naive**
MAE	90.73846	67.99901	95.83648
RMSE	150.7801	125.0251	154.3388
MASE	0.8951691	0.6708358	0.945463

It is observed from table 3.5 that because of the low values of MAE, RMSE and MASE for STL decomposition that STL decomposition is the best method for forecasting in this case.

Now, STL decomposition is used to forecast for the entire data. The STL decomposition graph for the data is given in figure 3.9.

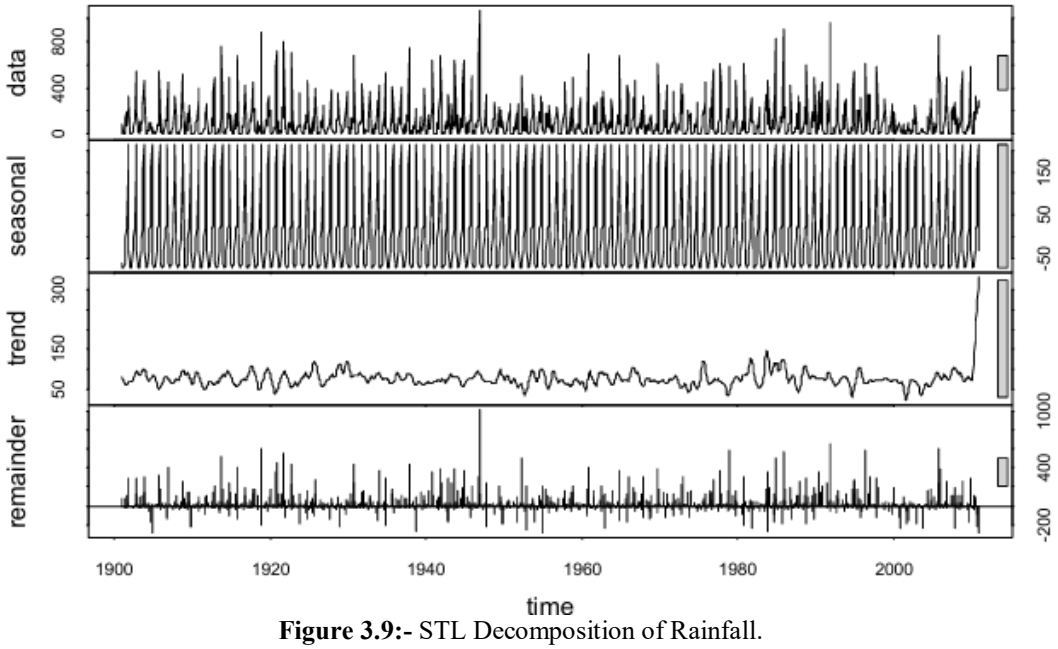

Figure 3.9:- STL Decomposition of Rainfall.

Figure 3.10 gives the plot of the forecast for the entire data using STL decomposition for the time period of January 2022 to December 2026 (ie,5 years)

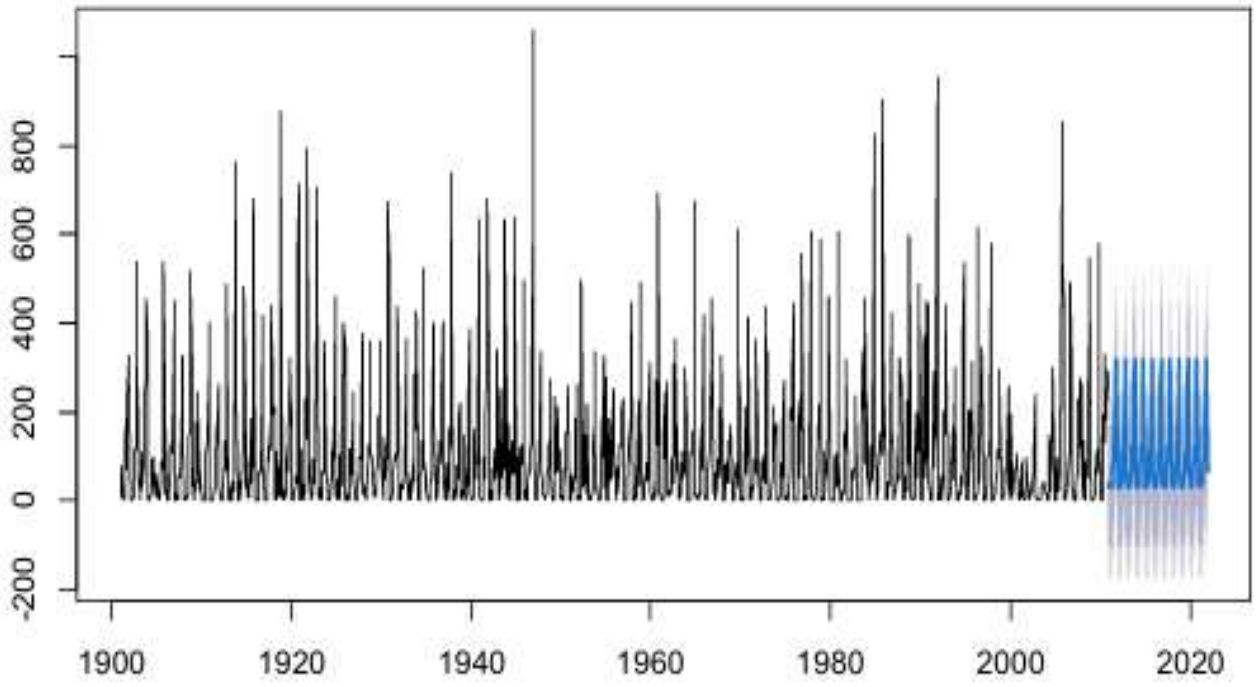

Figure 3.10:- Forecast using STL Decomposition method.

The final forecasted value is combined with the original data to form a combined new data and K-means clustering is done in this combined data to form 7 patterns and investigation is done on this and the original 7 clusters. The Adjusted Rand Index (ARI) between the existing and the forecasted pattern is 0.950. This indicates a high level of agreement between the actual and predicted patterns. This indicates that STL decomposition effectively captured the underlying patterns in the historical data.[15]

Conclusion:-

In this case study, we attempted to predict rainfall patterns in an urban city context, with a focus on Chennai. The K-means clustering algorithm was employed to classify the city's monthly rainfall from 1901 to 2021 into seven distinct clusters. Each cluster represented a unique pattern of rainfall behaviour, and the distribution and characteristics of rainfall within each were analyzed in depth.

An extensive comparative study of forecasting methods—SARIMA, STL decomposition, and seasonal naïve forecasting—was conducted to determine the most effective model for highly seasonal data. Based on the performance metrics MAE, RMSE, and MASE, STL decomposition emerged as the best-performing technique. This method was then used to forecast future rainfall, and the resulting predictions were merged with the original data. Upon reapplying clustering to this enriched dataset, it was observed that the properties of the clusters, particularly measures of variability, showed minimal divergence, confirming that the STL decomposition effectively preserved the underlying seasonal patterns of the historical data.

However, the study is not without limitations. The reliance on monthly data limits the granularity and responsiveness of the analysis. To conduct a high-resolution and more actionable study, daily or hourly rainfall data is essential—it would allow for more precise detection of short-term anomalies, better modelling of extreme events, and stronger support for real-time urban infrastructure planning. Furthermore, the analysis does not explicitly account for external changes such as urban development.

The practical implications of identifying seven distinct rainfall clusters are significant. For urban planners, this classification supports the development of tailored flood mitigation infrastructure, water storage planning, and improved drainage design based on specific rainfall regimes. For meteorologists and disaster management

authorities, the cluster-based insights provide a data-driven foundation to refine seasonal forecasts, identify extreme rainfall patterns early, and enhance overall preparedness for climate-related disasters.

Supplementary Material

1.SARIMA

SARIMA model is represented as

ARIMA (p, d, q) (P, D, Q) [m]

Where the 1st bracket represents the non-seasonal components and the 2nd bracket represents the seasonal components and 'm' is the number of observations per year or the period of the model.

p = Number of Auto-regressive terms of the non-seasonal component

d = Number of the differencing of raw observations to allow the time series to become stationary

q = Number of Moving Average terms of the non-seasonal component

P = Number of seasonal AR terms. This component captures the relationship between the current value of the series and its past values, specifically at seasonal lags.

D = Number of seasonal differences. Similar to the non-seasonal differencing, this component accounts for the differencing required to remove seasonality from the series.

Q = Number of Seasonal Moving Average terms this component models the dependency between the current value and the residual errors of the previous predictions at seasonal lags.

For example, the ARIMA (0,0,0) (2,0,0) [12] can be expressed mathematically as,

$$Y_t = c + \Phi_1 Y_{t-12} + \Phi_2 Y_{t-24} + \in \quad \ldots\ldots\ldots\ldots (1)$$

Where Y_t is the time series data at time t

Intercept is represented by c.

The auto-regressive parameters at lags 12 and 24 are represented by Φ_1 and Φ_2 respectively.[10]

2.STL Decomposition

The formula for STL decomposition is

$$Y_t = S_t + T_t + R_t \quad \ldots\ldots\ldots\ldots (2)$$

Here the trend component T_t is calculated using the formula,

$$T_t = LOESS(Y_t) \quad \ldots\ldots\ldots\ldots (3)$$

LOESS stands for locally weighted regression and scatter plot smoothing.

Now,

$$Y_t - T_t = S_t + R_t \quad \ldots\ldots\ldots\ldots (4)$$

Use moving average method to the RHS of the above equation i.e., $S_t + R_t$ to obtain the seasonal component of the time series

$$S_t = Moving\ Average(S_t + R_t) \quad \ldots\ldots\ldots\ldots (5)$$

Use this seasonal component to obtain the remainder component.

$$R_t = S_t + R_t - S_t \quad \ldots\ldots\ldots\ldots (6)$$

After having obtained trend, seasonal and remainder component forecast for each component using appropriate forecasting method. [11]Then combine this forecast to obtain the required forecast,

$$\hat{Y}_t = \hat{T}_t + \hat{S}_t + \hat{R}_t \quad \ldots\ldots\ldots\ldots (7)$$

3. Seasonal naive forecasting

Mathematically, this can be given as,

$$\hat{Y}_{t+h} = Y_{t+h-k(m\lfloor \frac{h}{m} \rfloor)} \quad \ldots\ldots\ldots\ldots (8)$$

Here, k is the number of seasons ago, is the length of the seasonal cycle (here since the data is monthly, m = 12),$\hat{Y}_{t+h}$ is the forecast for time period t+h [12].

References:-

1. Bentivoglio, R., Isufi, E., Jonkman, S. N., & Taormina, R. (2022). Deep learning methods for flood mapping: a review of existing applications and future research directions. Hydrology and earth system sciences, 26(16), 4345-4378.
2. Bora, S., & Hazarika, A. (2023, April). Rainfall time series forecasting using ARIMA model. In 2023 International Conference on Artificial Intelligence and Applications (ICAIA) Alliance Technology Conference (ATCON-1) (pp. 1-5). IEEE.

3. Box, G. E. P., Jenkins, G. M., Reinsel, G. C., & Ljung, G. M. (2015). Time series analysis: Forecasting and control (5th ed.). Wiley.
4. Chan, F. K. S., Yang, L. E., Scheffran, J., Mitchell, G., Adekola, O., Griffiths, J., ... & McDonald, A. (2021). Urban flood risks and emerging challenges in a Chinese delta: The case of the Pearl River Delta. Environmental Science & Policy, 122, 101-115.
5. Cleveland, R. B., Cleveland, W. S., McRae, J. E., & Terpenning, I. (1990). STL: A seasonal-trend decomposition procedure based on loess. Journal of Official Statistics, 6(1), 3–73
6. Hubert, L., & Arabie, P. (1985). Comparing partitions. Journal of Classification, 2(1), 193–218
7. Hyndman, R. J., & Athanasopoulos, G. (2018). Forecasting: Principles and practice (2nd ed.). OTexts.
8. Kamath, R. S., & Kamat, R. K. (2018). Time-series analysis and forecasting of rainfall at Idukki district, Kerala: Machine learning approach. Disaster Adv, 11(11), 27-33.
9. MacQueen, J. (1967, June). Some methods for classification and analysis of multivariate observations. In Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability, Volume 1: Statistics (Vol. 5, No. 1, pp. 281-297). University of California Press
10. Precipitation Time Series Analysis and Forecasting for Italian Regions Ebrahim Ghaderpour, Hanieh Dadkhah, Hamed Dabiri, Francesca Bozzano, Gabriele Scarascia Mugnozza and Paolo Mazzanti
11. Stedinger, J. R. (2012). Statistical Methods for Assessing Flood Risk and the Climate Change Challenge. Revista de Ingeniería, (36), 48-53.
12. Tang, Y., Sun, Y., Han, Z., Wu, Q., Tan, B., & Hu, C. (2023). Flood forecasting based on machine learning pattern recognition and dynamic migration of parameters. Journal of Hydrology: Regional Studies, 47, 101406.
13. Wang, X., Zhang, T., Wang, Y., Huang, X., Gong, H., & Chen, B. (2023). Temporal and Spatial Distribution Characteristics of Flood Disasters with Different Intensities in Arid-Semiarid Region in Northern Xinjiang, China. In E3S Web of Conferences (Vol. 394, p. 01009). EDP Sciences.
14. Xu, D., & Tian, Y. (2015). A comprehensive survey of clustering algorithms. Annals of Data Science, 2(2), 165–193.
15. Zhao, J., Wang, J., Abbas, Z., Yang, Y., & Zhao, Y. (2023). Ensemble learning analysis of influencing factors on the distribution of urban flood risk points: a case study of Guangzhou, China. Frontiers in Earth Science, 11, 1042088.