International Journal of Engineering Trends and Technology
ISSN: 2231 – 5381 /doi:10.14445/22315381/IJETT-V69I10P212

Volume 69 Issue 10, 89-96, October, 2021

Time Series Analyzation and Prediction of Climate using Enhanced Multivariate Prophet

J. Jagannathan[1], Dr. C. Divya[2]

[1]Research Scholar, [2]Assistant Professor
Centre for Information Technology and Engineering, Manonmaniam Sundaranar University,
Abishekapatti, Tirunelveli, Tamil Nadu - 627012, India.

[1]jagannathan161091@gmail.com, [2]cdivyame@gmail.com

Abstract — *In a development of a country, a huge challenge faced was Climatic change. For an agriculture-based country like India, it affects the quality and quantity of the products. Increase in the temperature and decrease in the rainfall. The main factor behind this was the urbanization of many cities. Based on the increase in the industries and density of population, the heat island was created, and it affects the climate dynamically. But in the last two years, the climate seems to be improving, and it was due to the lockdown effect. The Prediction of temperature with more accuracy is a great challenge. Here the temperature is a factor that can be affected by various other factors, mainly humidity, wind speed, precipitation. In this paper, a multivariate prediction of temperature was proposed. Here the historical climatic data has been taken with a dataset of temperature, wind speed, humidity, precipitation, and pressure. The data has been trained and tested with various time series algorithms like Auto-ARIMA, LSTM, Prophet models. And a proposed Enhanced Multivariate Prophet (EMP) algorithm has been employed to find the seasonality and trend. Based on the analyzation the temperature has been forecasted for the future 365 days. In comparison with the other algorithms. The testing of EMP provides a very good accuracy of 99.9%. MAE and RMSE of 0.02.*

Keywords — *Auto-ARIMA, LSTM, Prophet, Temperature, Wind Speed, Precipitation, Humidity***.**

I. INTRODUCTION

In recent years, a dramatic change in climatic factors as a result of industrialization and technological growth has been seen. According to a report published by India's According to the Ministry of Earth Sciences, the country's average surface air temperature increased by 0.7 degrees Celsius between 1901 and 2018, accompanied by a rise in atmospheric moisture content. Sea surface temperatures in the tropical Indian Ocean rose by around 1°C between 1951 and 2015. Clear indications of human-induced climate change have emerged across the Indian subcontinent, leading to an increase in climatic extremes as a result of anthropogenic GHG and aerosol forcing, as well as changes in land use and land cover. The complex interactions between earth system components in response to a warming environment and regional anthropogenic influences have resulted in an increase in the frequency of localized heavy rainfall events, drought and flood occurrences, and the intensity of tropical cyclones, among other things, in the last few decades. Sean J. Taylor et al. [1] present a modular regression model with decipherable parameters that analysts with domain knowledge of the time series can adjust intuitively. It entails comparing and evaluating predicting performance studies. techniques, as well as projections that are automatically flagged for manual inspection and revision. Reliable, realistic forecasting of business time series is made possible by tools that assist analysts in making the most of their expertise. YUNJUN YU et al. [2] suggested a technique to increase cloudy day prediction accuracy, using the clearness-index as input data for the LSTM model and k-means to During data processing, determine the kind of weather, with cloudy days being categorized as cloudy and mixed (partially cloudy). To compare the accuracy of different approaches, the cross-regional study establishes if the strategy can be generalized by using NN models. Toni Toharudin et al. [3] proposed a model for data that includes trend, seasonality, holidays, missing data, and outliers. The results reveal that Prophet outperforms LSTM. On max air temperature, Prophet outperforms LSTM, whereas, on min air temperature, Prophet outperforms LSTM. The RMSE value difference, on the other hand, is not significant. very significant. Tim Palmer and colleagues [4] suggested a numerical study for weather forecasting in 2030. They described the constraints using a variety of numerical models. The prediction model is presented as a means of mitigating climate change. Jonathan A. Weyn et al. [5] suggested an updated model that delivers indefinitely stable weather forecasts and realistic weather patterns with advance times of many weeks and longer. But the projections are worse than those produced by a high-resolution NWP system. AlexBihlo et al. [6] forecast the geopotential height of the 500hPa pressure level, the total precipitation, and two-meter temperature for the next 24 hours over Europe using a conditional deep convolutional procreant adversarial network. The suggested models are trained on four years of ERA5 reanalysis data in order to forecast the linked meteorological fields in 2019. Cho et al. [7] used support vector regression, random forest, a multi-model ensemble (MME), and an artificial neural network. Input variables included 14 LDAPS model forecast data, daily minimum and maximum air temperatures from in situ observations, and various assisting data. To determine the hydroclimatic variance in rice growing, Emmanuel Nyadzi et al. [8]

utilized a time series model to evaluate the seasonal climatic forecast. They employed time-series models such as LSTM and Prophet. Sarthak Gupta [9] proposed a method for improving prediction accuracy by utilizing the time aspects of a stock dataset. The prophet method was utilized to forecast the stock price in this study. Forecasting of univariate time-series datasets. It was meant to automatically select a good combination of model hyperparameters, and It's simple to use so that credible forecasts can be generated for data with patterns and seasonal structure by default. An approach was proposed by Victor Aquiles Alencar [11]. uses climatic attributes in addition to historical data. As a result, adding meteorological data enhanced the model's performance. An average Mean Absolute Error (MAE) of approximately 61.13 travels was obtained with the demand data on Evo, whereas MAE equals 32.72 travels was observed when adding the climatic data. Taghreed Alghamd et al. [12] uses the extensively used ARIMA model to assess and forecast the flow of traffic annotations in a defined study region in California, USA, on an hourly basis. To equivalence with the model recommended by the auto.Arima function, which employs arbitrary walk with drift, many ARIMA models are settled using PACF, ACF analyses of the time series. The model residual shows good performance in predicting future traffic conditions. Shakir Khan[13] used auto ARIMA and compared
the results (Auto-Regressive Integrated Moving average model). ARIMA (1,1,33) outpaced the further two models in terms of computing the MAPE and holdout testing, demonstrating the ARIMA model's potential for accurate stock forecasting. Krishnamurthy, V et al. [14] considers slowly shifting components like soil moisture, sea surface temperature, sea ice, and snow cover may give a root for long-term climate prediction. Monsoon intraseasonal alternation can be better awaited at long leads, according to a prediction model based on stage space renovation. Gouda, K. C et al. [15] ANN in the supercomputer platform was used to suggest a back-propagation technique to predict rainfall over Bangalore, India. The back-propagation was trained and confirmed using actual rainfall data from the region. The results revealed that ANNs could accurately predict long-term rainfall with acceptable accuracy. Sindhu, P et al. [16] used Theoretical analysis, field observations, and numerical modeling are all used in this project. After discretizing the domain, the temporal change of temperature at each nodal point was assessed using the fundamental heat transfer equation.

II. DATASET AND PRE-PROCESSING

The climatic data of Chennai was purchased from the weather data provider world weather and ISRO. The raw data consists of various parameters like temperature, wind direction, sun rice, etc., for every one hour. The data set was available from the year 2008 to 2021. Thus, for the prediction of the temperature of the urban heat island, the humidity, pressure, wind speed, rainfall data's to corollate were taken. So initially, the required data was filtered alone. The filtered data set is shown in table 1. The dataset contains the five major features as follows: Rainfall indicates the quantity of rain falling (mm), Temperature indicates the temperature (°F), Humidity indicates the amount of water vapor present in the air, Wind Speed indicates the movement of air from high to low pressure (Kmph), Pressure indicates the pressure within the atmosphere of the earth (MB).

Table 1 – Sample Climatic Dataset

Date	**Rain**	**Hum**	**Temp**	**Wind**	**Pres**
01-07-2008	1.5	60	91	15	1006
02-07-2008	0.2	58	94	16	1006
03-07-2008	1.7	61	94	13	1006
04-07-2008	5.2	67	93	14	1006
05-07-2008	1.3	64	95	15	1005

Table 1 shows the sample dataset used. This dataset has been extracted from the raw data with core features. Here some important data are extracted for the model. Table 2.2 shows the overall climatic data of the cyclone days. Based on the data, Chennai got affected by the various cyclone. Here the amount of precipitation is very extreme in the time of cyclone Jal, with 375.5 MM on average. But the wind speed is 42 Kmph. In the time of Cyclone vardha, Chennai got a very big impact. Here the precipitation is 212.3 MM which is low when compared to Cyclone Jal but the average wind speed to very high with 72 Kmph.

Table 2 –Climatic Dataset on Cyclone Days

date	max temp F	min temp F	avg temp F	total precip MM	wind speed Kmph	hum idity	pressure MB	Heat Index F	Dew Point F	Wind Chill F	Wind Gust Kmph	Cyclone
26-11-2008	78	76	77	175	41	93	1008	82	75	77	62	Nisha
10-03-2009	81	73	77	194	27	87	1013	83	73	77	39	Laila
07-11-2010	76	72	73	375.5	42	94	999	78	72	73	89	Jal
29-12-2011	80	76	77	32.8	73	78	1006	81	70	77	61	Thane
31-10-2012	81	76	78	64.7	53	86	1000	84	74	78	56	Nilam
09-11-2015	82	78	80	237.5	61	85	1004	86	75	80	64	Tropical Depression
12-12-2016	77	70	73	212.3	72	82	1002	77	67	73	76	Vardah
15-11-2018	84	78	81	17.8	33	76	1009	87	73	81	35	Gaja
16-12-2018	79	72	75	2.3	37	70	1011	79	65	75	49	Phethai
24-11-2020	80	70	77	43	41	81	1008	82	73	77	70	Nivar
15-05-2021	94	82	88	0	36	60	1002	96	73	88	41	Tauktae

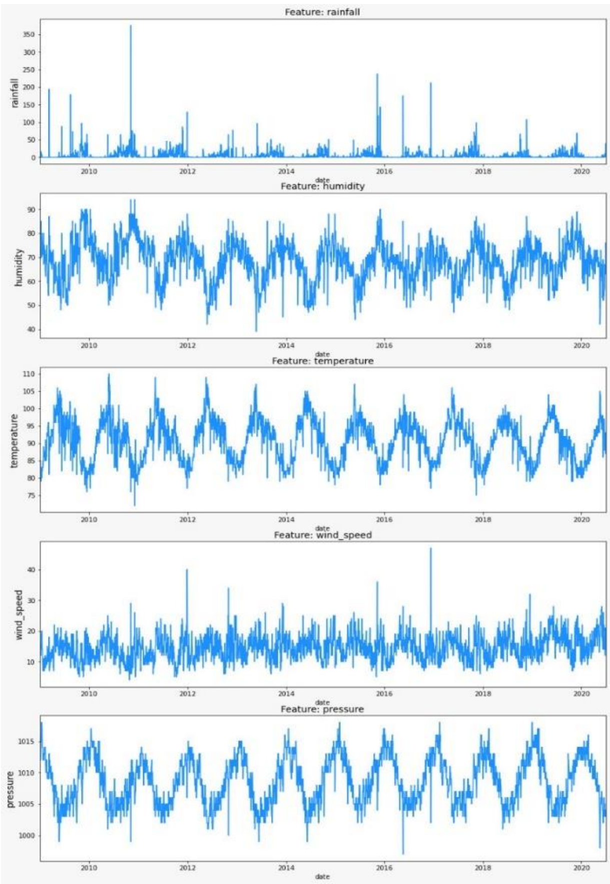

Fig. 1. : Visualization of Climatic data – Rainfall, Humidity, Temperature, Wind Speed, Pressure

In figure 1 shows the visualization of the climatic data as stated in table 1. Figure 1 shows the various climatic features for the entire data duration. Figure 1 shows the relationship between each climatic data feature. While humidity and wind speed increase, the temperature and pressure decrease, and the probability of rainfall occur. As the temperature increases gradually, humidity and Rainfall decreases which can be observed from figure 1.

A. Data Pre-Processing

In a time series, the data should be in chronological sequence, and the timestamps should be evenly spaced. The time period in the data is one day, and the information is already in chronological order. The missing data must then be dealt with. There are some null values in this dataset. The data can be tidy them up or can fill NaN with Linearly Interpolated Value using interpolate() and the neighboring value's knowledge. The processed data must then be resampled in order to provide more information. So the resample() function was used to downsample.

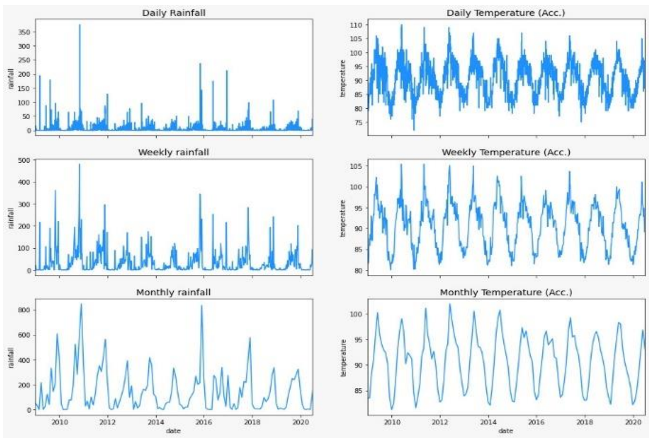

Fig. 2.: Seasonal temperature and rainfall from Sampled data.

Figure 2 shows the seasonal temperature and rainfall relationship for daily, weekly, and monthly views. Here the relationship between the data's very clearly observed. After the sampling procedure, the stationarity must be verified. Some time-series models are based on the assumption that the data is fixed. To create the stationary test.

The ADF test is part of the unit root test.

$$Yt=\alpha Yt\text{-}1+\beta Xe+\epsilon \quad (1)$$

Where the value of the time series is denoted as Yt, and at the time 't' and exogenous variable is mentioned as Xe.

A unit root specifies that the time series is non-stationary. Non-stationarity is caused by unit-roots. Thus, it can be concluded that if the null hypothesis is rejected, the time series is stationary. The null hypothesis can be rejected in two ways: if the p-value is less than a certain level of significance or if the p-value is larger than a certain level of significance. The test is less than the critical value. If it is less than the critical value default significance threshold is 5%. The p-value in the data is one percent. The time series is thus stationary.

III. FEATURE ENGINEERING

Once the data has been made stationary, the Cyclical Encoding Features has been used to avoid the confusions due to the cyclic nature of the months and date in the year. It will suddenly fall to 1 at the beginning of the month. The time series decomposition is then utilized to produce a combination of level, trend, seasonality, and noise components in order to comprehend the problem and analysis and forecast it. Seasonality is a repeating short-term cycle in the series, and Noise is a random variation in the series. The level is an average value in the series, Trend is an increasing or decreasing value in the series, Seasonality is a repeating short-term cycle in the series, and Noise is a random variation in the series. In this situation, the stats models library's seasonal decompose() function was utilized.

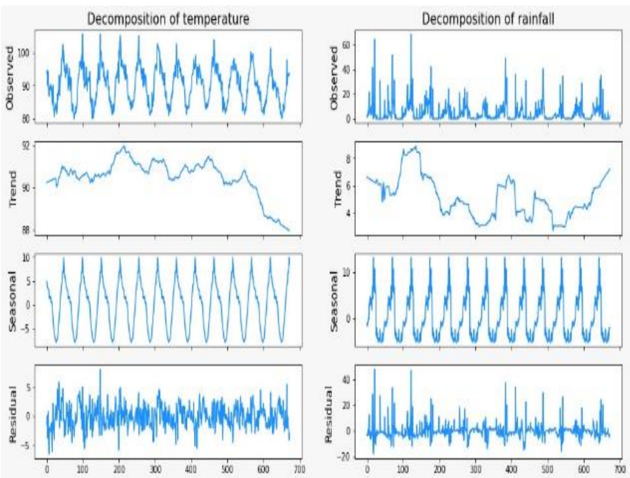

Fig. 3.: Decomposition of Features.

IV. EXPLORATORY DATA ANALYSIS

Then to plot the data and try to extract some knowledge through which the multivariate correlation can be identified. It was observed that the rainfall: reaches its maximum around October/November and its March/April temperature: reaches its maximum around April/May, and its minimum around December/January Wind Speed: reaches its maximum in June/July and minimum around October/February. Pressure: reaches its maximum around January/February and its minimum around June/July.

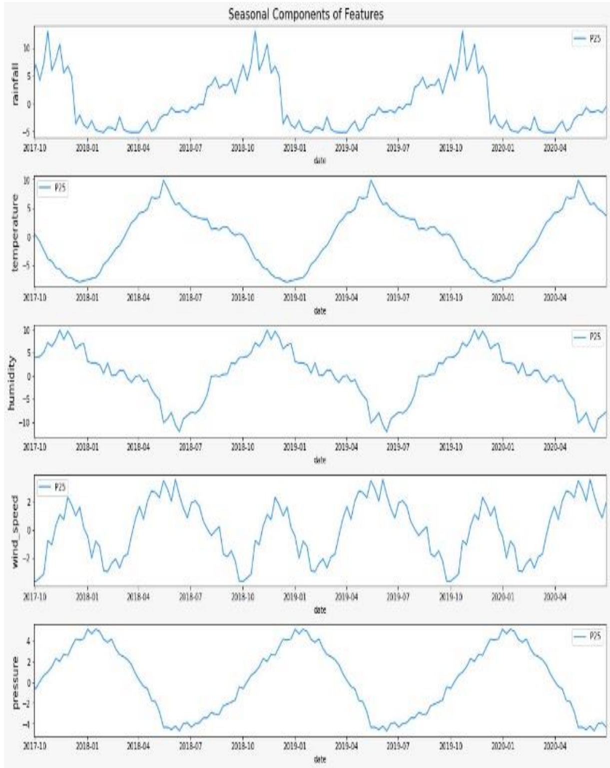

Fig. 4.: Seasonal Components of Features.

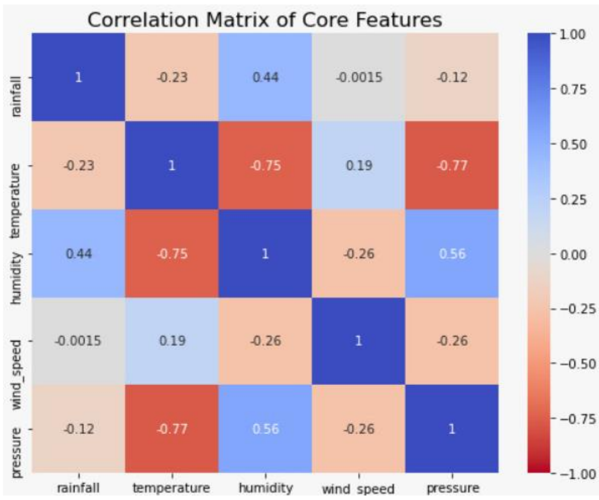

Fig. 5.: Correlation Matrix of Features

Following the use of differencing to stationarize a time series, the very next step in fitting an ARIMA model is to determine whether AR or MA terms are needed to correct any residual autocorrelation in the different fields. The partial autocorrelation (PACF) and autocorrelation function (ACF) plots of the different fields may be used to determine the number of AR and/or MA terms needed. Autocorrelation graphs may be used to determine seasonality.

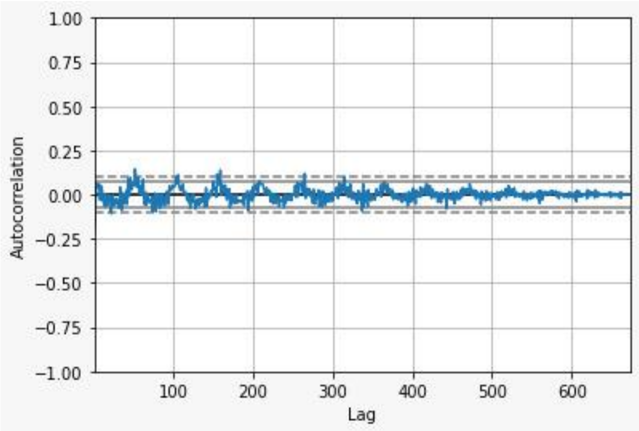

Fig. 6.: Autocorrelation of features.

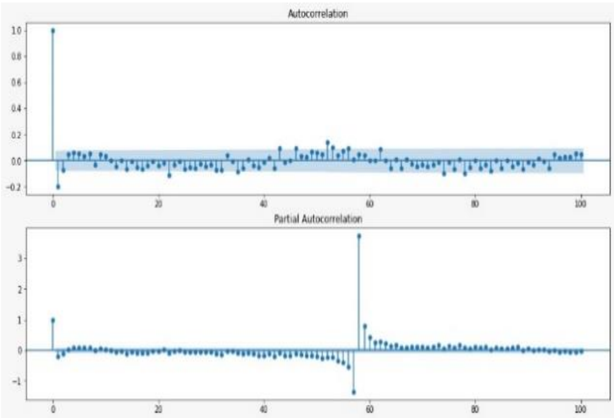

Fig. 7.: Autocorrelation and Partial-Correlation of features

A. *Modeling*

There are two types of time series: univariate and multivariate: In a univariate time series, there is only one time-dependent variable. In multivariate time series, Multiple time-dependent variables are present. First, the cross-validation technique was used, as illustrated in figure 8. In figure 8 shows in each iteration which train and test sets were used to fit the model.

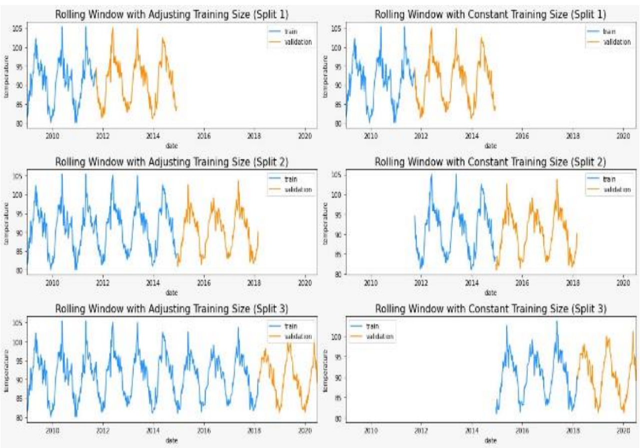

Fig. 8.: Cross-Validation of Dataset.

B. *Univariate Time Series*

Prophet: Prophet uses a time series forecasting model known as additive time series forecasting. Prophet divides time data into three categories: trend, seasonality, and holidays. It features user-friendly hyper settings that are simple to adjust. Trend + Seasonality + Holiday + Error = Prophet time series. The trend model simulates non-periodic variations in the time series value. Seasonality refers to variations that occur on a daily, weekly, or annual basis. The holiday effect occurs when schedules are disrupted for a day or a few days. The term "error words" refers to what the model does not explain.

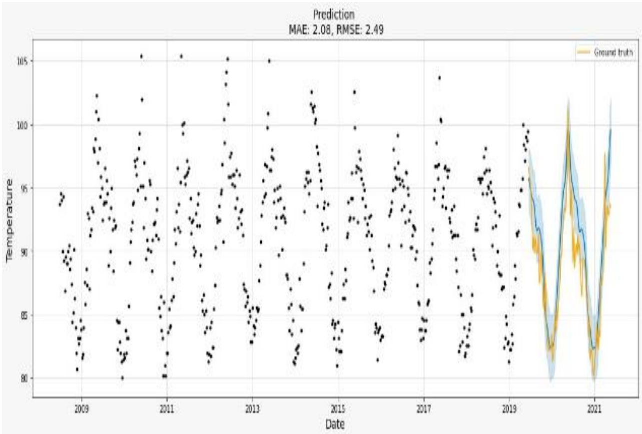

Fig. 9.: Temperature Prediction using Prophet. (MAE:2.08, RMSE: 2.49)

The result obtained was shown in figure 9 from the prophet model gives an MAE score of 2.08 and RMSE score of 2.49, which give an accuracy of 97.5%.

Auto-ARIMA: Although ARIMA is a sophisticated model for forecasting time series data, the parameter tweaking operations and data preparation take a long time. Before you can use ARIMA, first, the series was made stationary and used the plots declared earlier to find the values of p and q. Auto ARIMA makes this work much easier for us by removing the need to calculate the d value, ACF, and PACF values.

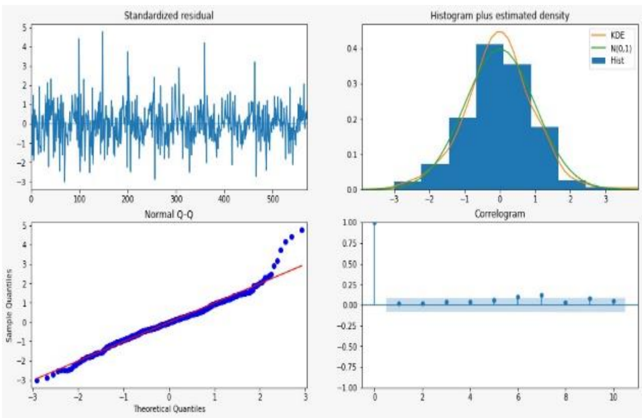

Fig. 10.: i) Standardized Residual ii) Histogram & Estimated density, iii) Normal Q-Q, iv) Correlogram

The residual errors in figure 10 appear to oscillate around a mean of zero and have a consistent variance between them (-2, 2).
The density plot in the histogram suggests a normal distribution with a mean zero. The majority of the blue dots in the normal q-q are above the red line, implying that the distribution is significantly skewed. The ACF plot in the Correlogram reveals that the residual errors are not autocorrelated.

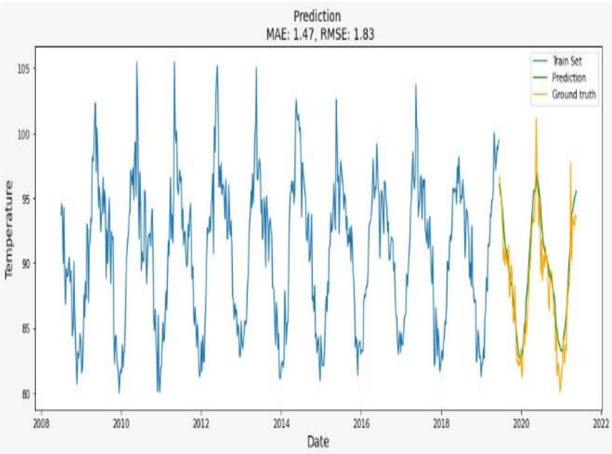

Fig. 11.: Temperature Prediction using LSTM. (MAE : 1.47, RMSE 1.83)

LSTM: To forecast the last value of a sequence of data, the multi-layered LSTM RNN is utilized. Before constructing the LSTM model, the ensuing pre-processing of data and feature engineering must be completed. Create the dataset and make sure that all of the data is float. Normalize the characteristics. Sets for training and testing have been created. Create a dataset matrix from an array of values. X=t+1 and Y=t+1 are the new shapes.
Resize the input (num samples, num timesteps, num features) to make it 3D. Reshape input to be 3D - num_timesteps, num_samples, num_features.
The result obtained was shown in figure 11 from the LSTM model gives an MAE score of 1.47 and RMSE score of 1.83, which give an accuracy of 98.1%.

V. PROPOSED METHODOLOGY

A. Enhanced Multivariate Prophet (EMP)

In the climatic dataset, the Multiple variables vary over time. The Enhanced Multivariate Prophet decomposes time series into trend, seasonality, holiday, and Regional.

$$y(t) = g(t)+s(t)+h(t)+r(t)+\epsilon t$$

Where trend function is given as g(t), which discovers changes in time series data is non-periodic, and the seasonality function is given as s(t), which may be used to find the correlation between multivariate data on a daily, weekly, monthly, or yearly basis. The holiday information provided is h(t), which is also used to determine whether or not meteorological factors are changed during the period. r(t) is a regional function that takes into account regional effects like cyclone information, industrial emissions, recent lockdown factors, and so on, providing additional information regarding microclimatic variation. The model will continue to identify a proper trend, but if it finds a large deviation in some years, the prediction will be impaired; thus, the geographical component was also taken into account. The error rate (t) represents any notable changes that the model did not account for.

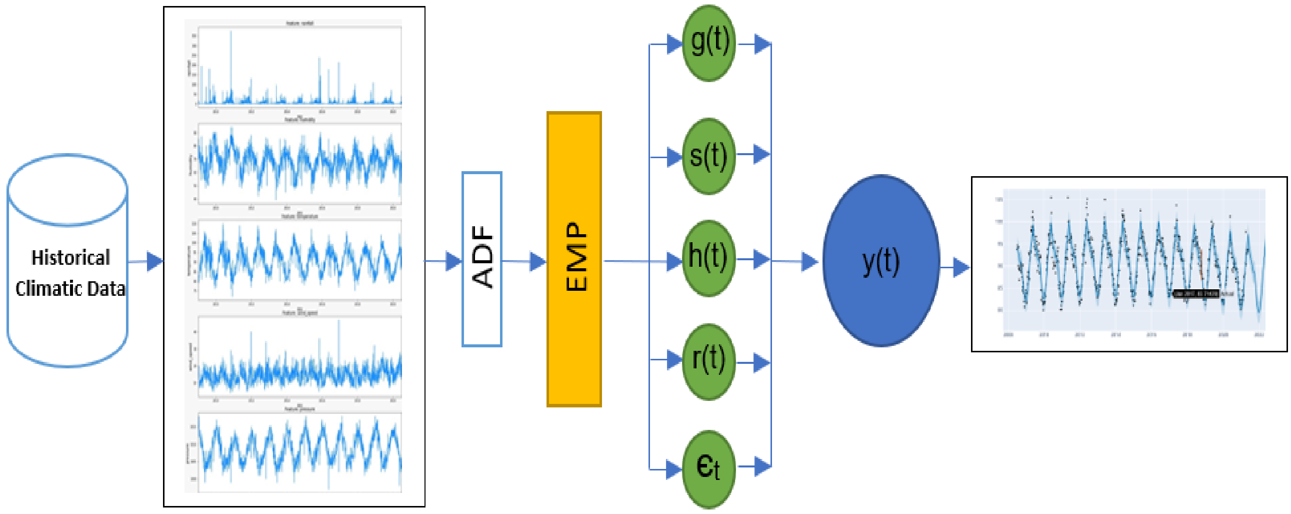

Fig. 12.: Architectural diagram of the EMP

Figure 12 shows the architecture of the Enhanced Multivariate prophet model. Where the historical data were pre-processed using the techniques mentioned previously. From the historical data was considered the main factors, temperature, precipitation, humidity, wind speed, pressure, cloud cover, heat index, dew point, wind gust was taken into consideration for the prediction. The multivariate data will provide additional information for the prediction.
From figure 1, it was observed that when the humidity increases, wind speed increases, and pressure decreases automatically, the temperature decreases, and the rainfall increases.

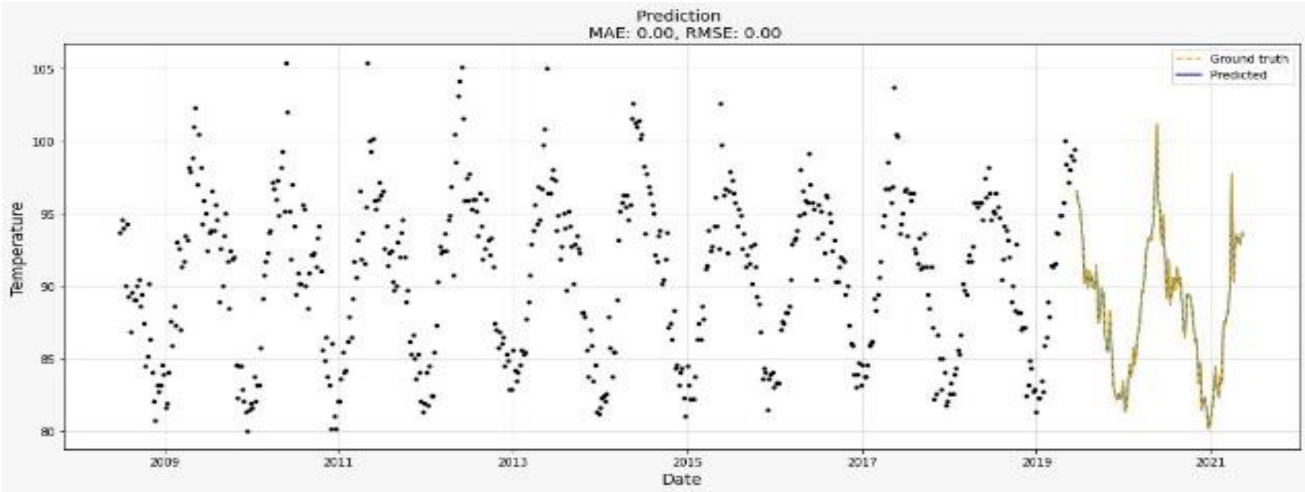

Fig. 13.: Temperature Prediction Using proposed EMP (MAE : 0.00041, RMSE : 0.00049)

Thus, each climatic factor has a very big impact on the variation in temperature and rainfall. Here all the factors, trends, and seasonality were considered. Based on the analyzed data, future values of each factor can be predicted. Using the predicted values, the temperate was predicted with more accuracy.

Here in figure 13, it was observed that the model perfectly fits the actual and predicted values. It provides an accuracy of 99.9%. The accuracy was calculated based on the Mean Absolute Error (MAE) with an error of 0.0004, which the score is linear which shows the all distinct variances are weighted equally in the average and RMSE - Root means squared error with an error of 0.0005 the average magnitude of the error was measured as the RMSE and the MAE must always be greater or equal. Thus, the EMP provides a very less error rate.

But even it is considered as an impact with minute accuracy variation.

Table 3 – MAE, RMSE, Accuracy Comparison

Model	MAE	RMSE	Accuracy
Prophet	2.08	2.49	97.5%
ARIMA	11.34	12.22	87.78%
Auto-ARIMA	4.28	4.59	95.41%
LSTM	2.78	3.05	96.95%
EMP	0.00041	0.00049	99.9%

Table 3 shows the comparative accuracy score and MAE, RMSE score for each time-series model used in this analysis. Thus based on the result, the proposed model provides the highest accuracy.

Based on the Enhanced Multivariate Prophet training and testing process, temperature for the rolling window of 1 year from the final date was forecasted. In this analysis, the dataset was available up to mid of May 2021. Thus the forecast was made up to the month of May 2022. Figure 14 shows that the overall correlated dataset from the year 2008 to 2021, and the predicted value was shown up to 2022.

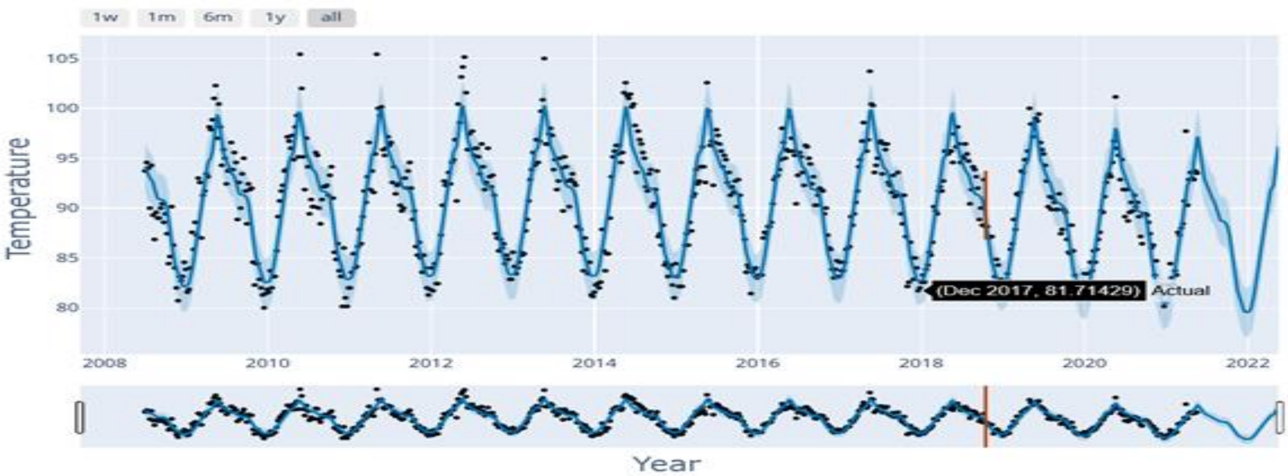

Fig. 14.: 1-year rolling window forecast of Temperature using EMP

VI. CONCLUSION

The foremost objective of the proposed research methodology is to build a most accurate prediction model to predict the temperature. Usually, the temperature prediction can be made using the historical temperature data. But the temperature cannot be predicted with historical temperature data alone. The additional features which affect the temperature, such as humidity, rainfall, pressure, the wind speed, were considered. Here an Enhanced Multivariate Prophet was proposed, which is used to predict the temperature using the multiple feature correlation with it. Here the proposed model includes an additional feature of regional value, also which gives the additional information for the model. The proposed EMP model provides an accuracy of 99.9%, which was the best score when compared with the other time-series predictive model.

VII. MOTIVATION AND CONTRIBUTION

The climatic change creates a very big impact in developing countries. Which affects agricultural production and also which also affects water resources. It leads to a demand in the water supply. It was due to the urbanization of cities. Increase in the industrial areas and increase in the emission of gas due to the more private transportation. And more number of vegetations has been converted into high rise buildings. To mitigate this, more roof gardens can be implemented in all government and high-raised buildings. And also, the unused areas of the cities can also be used for planting trees. Here this model can be used to predict the temperature, and also using the same model, the other features such as rainfall, wind speed, humidity can also be predicted as well. So, the agriculture and water storage process planning can be done for the upcoming year. The model can be linked with real-time data to get a more accurate prediction.

REFERENCES

[1] Taylor, Sean J., and Benjamin Letham. Forecasting at scale, The American Statistician 72(1)(2018) 37-45.

[2] Yu, Yunjun, Junfei Cao, and Jianyong Zhu. An LSTM short-term solar irradiance forecasting under complicated weather conditions. IEEE Access 7 (2019) 145651-145666.

[3] Toharudin, Toni, et al. Employing long short-term memory and Facebook prophet model in air temperature forecasting. Communications in Statistics-Simulation and Computation (2020) 1-24.

[4] Palmer, Tim. A vision for numerical weather prediction in 2030. arXiv preprint arXiv:2007.04830 (2020).

[5] Weyn, Jonathan A., Dale R. Durran, and Rich Caruana. Improving data - driven global weather prediction using deep convolutional neural networks on a cubed sphere. Journal of Advances in Modeling Earth Systems 12(9) (2020): e2020MS002109.

[6] Bihlo, Alex. A generative adversarial network approach to (ensemble) weather prediction. Neural Networks 139 (2021) 1-16.

[7] Cho, Dongjin, et al., Comparative assessment of various machine learning - based bias correction methods for numerical weather prediction model forecasts of extreme air temperatures in urban areas. Earth and Space Science 7(4) (2020) e2019EA000740.

[8] Nyadzi, Emmanuel, et al. Verification of seasonal climate forecast toward hydroclimatic information needs of rice farmers in Northern Ghana., Weather, Climate, and Society 11(1) (2019):127-142.

[9] Gupta, Sarthak. A Directionality & Novel Function for Stock Price Modelling Prediction Dynamics based on Multiple Learning., Turkish Journal of Computer and Mathematics Education (TURCOMAT) 12(6) (2021) 3132-3139.

[10] Rao, Jian, et al. Predictability of stratospheric sudden warmings in the Beijing Climate Center forecast system with statistical error corrections. Journal of Geophysical Research: Atmospheres 124(15) (2019) 8385-8400.

[11] Alghamdi, Taghreed, et al. Forecasting traffic congestion using ARIMA modeling. 2019 15th International Wireless Communications & Mobile Computing Conference (IWCMC). IEEE, 2019.

[12] Khan, Shakir, and Hela Alghulaiakh., ARIMA Model for Accurate Time Series Stocks Forecasting., https://thesai. org

[13] Scaife, Adam A., et al. Tropical rainfall predictions from multiple seasonal forecast systems. International Journal of Climatology 39(2)(2019) 974-988.

[14] Krishnamurthy, V. Predictability of weather and climate., Earth and Space Science 6(7)(2019) 1043-1056.

[15] Gouda, K. C., et al., An Approach for Rainfall Prediction using Soft Computing, International Journal of Engineering Trends and Technology 67(3)(2019) 158-164.

[16] Sindhu, P., and C. P. Archana., Numerical Modeling of Urban Air Temperature. International Journal of Engineering Trends and Technology 53(2) (2017) 95-99.

Authors' Profiles

J. Jagannathan received the Bachelor of Technology in Information and Communication Technology in 2012 Master of Technology in Information and Communication Technology in 2014 from Centre for Information Information Technology and Engineering, Manonmaniam Sundaranar University, Abishekapatti, Tirunelveli, Tamil Nadu, India. And pursuing Doctor of Philosophy in Centre for Information Information Technology and Engineering, Manonmaniam Sundaranar University, Abishekapatti, Tirunelveli, Tamil Nadu, India. His research interest are Deep Learning, Image Processing, Sensor Networks, Nano Devices.

C. Divya received the Master of Engineering in Communication Systems in 2010 from SSN College of Engineering and the Doctor of Philosophy in 2014 from Manonmaniam Sundaranar University, Abishekapatti, Tirunelveli, Tamil Nadu, India. She is currently working as an Assistant Professor in the department of Centre for Information Technology and Engineering, Manonmaniam Sundaranar University, Abishekapatti, Tirunelveli, Tamil Nadu, India. She has published more research papers in International / National journals / Proceedings / Books. Her current research interests includes Data Analytics, Cyber Security, Nanodevices and Low power VLSI circuits Wireless Sensor Networks, Communication Networks.