

Time Series Prediction of Temperature in Pune using Seasonal ARIMA Model

Aarati Gangshetty¹
JRF, DRDO, R&DE(Engrs)
Pune, India

Gurpreet Kaur²
Scientist E, DRDO, R&DE(Engrs)
Pune, India

Uttam Sitaram Malunje³
Technical Officer A, DRDO,
R&DE(Engrs) Pune, India

Abstract - An effort has been made to develop a SARIMA (Seasonal Autoregressive Integrated Moving Average) model for temperature prediction using historical data from Pune, Maharashtra. The historical dataset from the year 2009 to 2020 has been taken for observation. When there is a repeating cycle is present in a time series, instead of decomposing it manually to fit an ARIMA model, another very popular method is to use the seasonal autoregressive integrated moving average (SARIMA) model. Time series currently is becoming very popular, a reason for that is declining hardware's cost and capability of processing. The model can be used to set the outline for the upcoming year. Enumerate the effects of unexpected changes or disruptions in the system. The seasonal ARIMA model is designed by running Python 3.7.4 on Anaconda Jupyter Notebook and using the package matplotlib 3.2.1 for data visualization. The goodness of fit of the model was verified against standardized residuals, the autocorrelation function, and the partial autocorrelation function. We discover that SARIMA (1,1,1)(1,1,1)₁₂ can represent very well in the performance. We found MAE of 0.60850 and RMSE of 0.76233 for SARIMA model. According to the model diagnostics, the model was good for predicting temperature.

Keywords — SARIMA, Prediction, ARIMA, temperature.

I. INTRODUCTION

The primary aim of time series model is to accumulate and take a look at historic values to develop suitable models that describe the essential structure and traits of the time series. Alternatively, the time series forecasting model observes distinct values and expecting to predict future values. Regression evaluation method often tests theories that the current data of one or more time series has the impact at the present data of another time series [1]. Time series statistics occurs in lots of areas like economic analysis, sensor monitoring of community, evaluation of scientific issues, and mining of social interest. Greater recent fields focus on the subject and talk to it as time series forecasting. Forecasting contains taking models fitting on past data and using them to predict future remarks. Descriptive models can derive for the future (i.e. to smooth out or eliminate noise), they are just trying to better define the statistics. An important distinction in the forecasts is that the future is completely impossible to obtain and that one should expect only that which has already taken place.

The prediction of time series is the use of a given model to predict future values against historical values, and can therefore be understood as a method. Separate numerical forecasts of metrological forecasts and time series use a

model to predict future values based on historical values. Owing to the position of time series forecasting in uncountable practical fields, researchers should pay attention to fitting a suitable model to the time series. Over the past many years, many intelligent time series models have been developed in the literature to improve the correctness and effectiveness of time series forecasting. One of the most widely used and known time series models of statistical forecasting is the Integrated Moving Average (ARIMA)self-regulatory model. The ARIMA model is well-known for notable forecasting correctness and effectiveness in representing various types of time series [3] with simplicity as well as the associated, Box– Jenkins's methodology for optimal model construction. For seasonal time series forecasting Box Jenkins [4]

proposed a relatively effective variation of the ARIMA model called the Seasonal ARIMA (SARIMA) model. The main design objectives for this paper are set out below:

- 1) Plotting the data as a time series plot
- 2) Checking the data, if it has any trend or seasonality
- 3) Predicting values of SARIMA (p, d, q) (P, D, Q)s
- 4) Applying SARIMA (p, d, q) (P, D, Q)s to predict future values.

II. LITERATURE SURVEY

Rios-Moreno et al. [5] used outside air temperature, relative humidity, air velocity, and global solar radiation flux as outside variables to an autoregressive (AR) and an autoregressive moving average (ARMA) model. They effectively forecasted the room temperature in a university room in Mexico. The results showed that the outside variable older than 20 minutes did not recover the performance of the model. Felice et al. [6] used a non-seasonal time-series method to predict energy demand at the national and regional level in Nigeria. It was established that using temperature as an outside variable enhanced the prediction results. Mahmudur Rahman, A.H.M. Saiful Islam, Sahah Maqnoon Nadvi, Rashedur M Rahman (2013) examine the Arima and Anfis Model and outline how the Model ARIMA can more efficiently capture the dynamic concert of property say weather, Temperature, Humidity and Air pressure which must be compared by various evaluation measures, for example, with the Mean Square Error(MSE), R-Square Error and the Sum of the Square Error(SSE) [7] and the author can check whether ARIMA would yield a more precise result than other models.

In addition, [8] accepted the study to examine the trend and forecast the maximum monthly temperature in Nigeria using the SARIMA model. According to the simplest suitable SARIMA model, the predicted maximum temperature of five years is slightly stable compared to that of reference. In another review, [9] fitted SARIMA model to average temperature for the measure of 1980-2010 of Dibrugarh utilizing automatic arima function i.e., `autoarima()` in R software. Keeping these points in mind, an effort has been made to develop a SARIMA model.

III. METHODOLOGY

Temperature data recorded from 2009 to 2020 were obtained for Pune city, from the meteorology department at one-hour intervals [12]. The longitude and latitude of the automatic weather station is 73.856255 and 18.516726, respectively. The data collected has different parameters, such as date time, temperature, humidity, moonrise, wind speed, wind direction, pressure. From this, we have eliminated features that have large amounts of missing data and we have considered temperature as an input parameter. The seasonal ARIMA model is implemented by running Python 3.7.4 on Jupyter Notebook and using the package `matplotlib` 3.2.1 for data visualization. Time series plot of temperature for the year 2018 was shown in Figure 1. The hourly temperature data during 2009–2018 is used as the training set, while that during 2019–2020 is used as the testing set. To evaluate the forecast accuracy, as well as to compare the results obtained from different models, the mean-square error (MSE) is calculated.

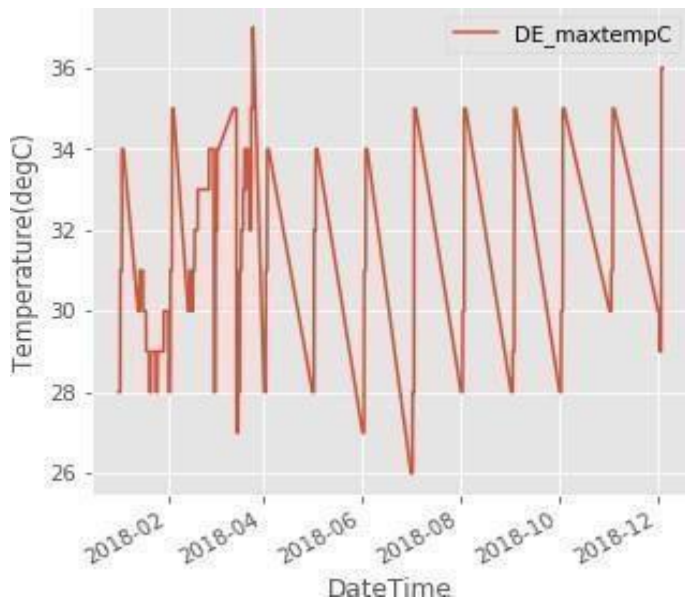


Figure 1 Time series plot of Temperature in Pune (year-2018)

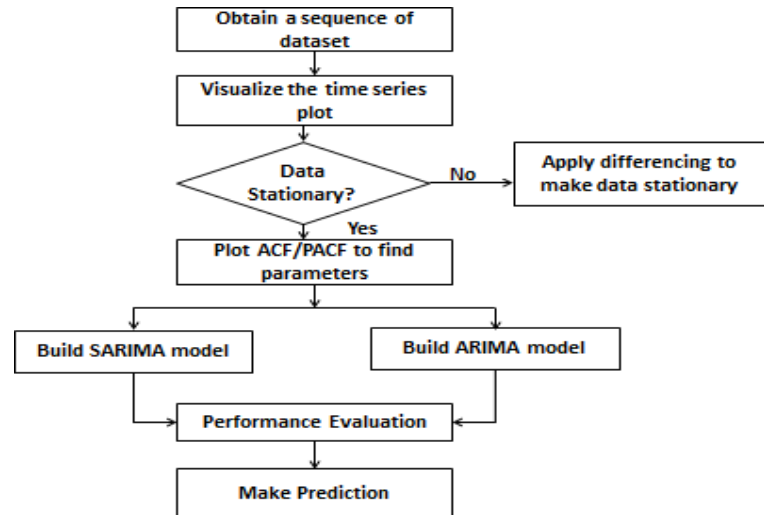


Figure 2 Flowchart of the proposed model

A. Check stationarity

If the time series is not stationary, it needs to be stationary through differencing. We tested our dataset stationarity by augmented Dickey-Fuller test and found that it is not stationary. Take the first difference then, determining stationarity with an augmented Dickey-Fuller test until P value is less than or equal to 0.05 ($P \leq 0.05$). The order of differencing d is chosen such that it minimizes the standard deviation. The previously differenced series which is now a stationary series might still have some auto-correlated faults which can be detached by adding AR terms ($p \geq 1$) and MA terms ($q \geq 1$) in the forecasting equivalence.

B. Plot ACF and PACF

In this step, the ACF and PACF of the data are plotted. Autocorrelation functions (ACF) and partial autocorrelation functions (PACF) are used to identify potential models. We can see there is gradual decreasing pattern in ACF plot while there is PACF cuts immediately after lag. Thus, the graph suggests that AR (1) would be appropriate for the model. If there's a pointy cutoff within the PACF of the differenced series and therefore the series shows mild 'under-differencing', an AR term is added to the model. If there is a high limit in the ACF of the differenced series and the series shows mild 'over-differencing', an MA term is further added to the model. Then select the optimal model based on a performance metric like AIC (Akaike Information Criteria).

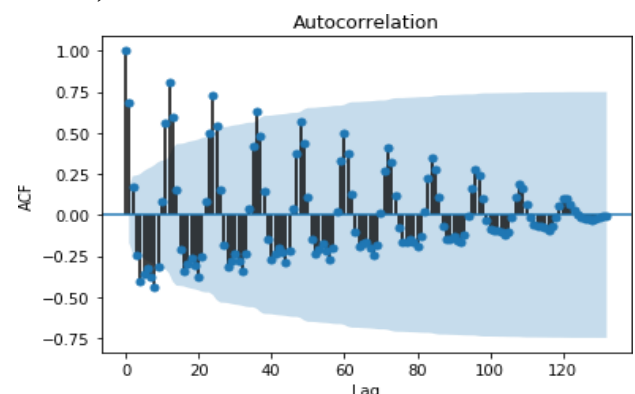


Figure 3 Autocorrelation Function

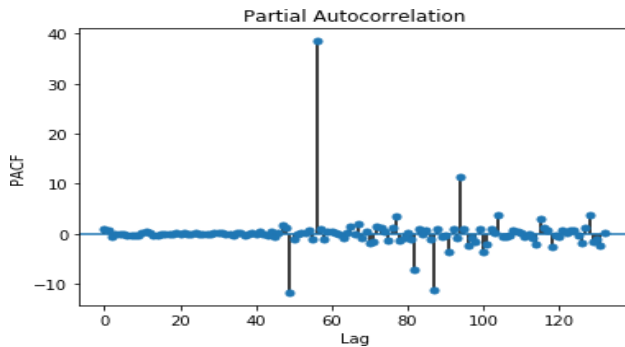


Figure 4 Partial autocorrelation Function

C. *Auto Regressive Integrated Moving Average Model*
ARIMA processes can be generally divided into two distinct processes, namely autoregressive (AR) processes, and moving average (MA) processes. The constraints can be defined as

- p: the number of interval observations in the model; also known as the interval order.
- d: differencing is needed to make the series stationary.
- q: the size of the moving average window; also known as the order of the moving average.

According to Box Jenkins methodology an ARIMA model is generally written as ARIMA (p,d,q) [10].

The **AR(p)** model is defined by the equation:

$$X_t = \alpha + \phi_1 X_{t-1} + \phi_2 X_{t-2} + \dots + \phi_p X_{t-p} + \omega_t \quad (1)$$

Where,

- X_t = response variable at time t
- $X_{t-1}, X_{t-2}, \dots, X_{t-p}$ = response variable at time t-1, t-2 and t-p respectively.
- α = constant term
- ϕ_1, ϕ_2 and ϕ_p = coefficients to be estimated
- ω_t = error term at time t

The **MA(q)** model is defined by the equation:

$$X_t = \alpha + \theta_1 w_{t-1} + \theta_2 w_{t-2} + \dots + \theta_q w_{t-q} + \omega_t \quad (2)$$

Where,

- X_t = response variable at time t
- α = constant term
- $w_{t-1}, w_{t-2}, \dots, w_{t-q}$ = forecast errors at timeseries lags t-1, t-2 and t-q
- θ_1, θ_2 and θ_q = coefficients to be estimated
- ω_t = error terms at time t

By combining equation (1) and (2) Autoregressive integrated moving average model ARIMA (p,d,q) can be written mathematically as

$$X_t = \alpha + \phi_1 X_{t-1} + \phi_2 X_{t-2} + \dots + \phi_p X_{t-p} + \theta_1 w_{t-1} + \theta_2 w_{t-2} + \dots + \theta_q w_{t-q} + \omega_t \quad (3)$$

D. *Seasonal ARIMA model*

In addition to trend, stationary series quite commonly display

seasonal behavior where a certain basic pattern tends to be repeated at regular seasonal intervals. Seasonal ARIMA model (SARIMA) is designed by addition of seasonal terms in the ARIMA models listed above. SARIMA models are written as,

$$\text{ARIMA}(p, d, q)(P, D, Q)_m \quad (4)$$

Where (p, d, q)(P, D, Q) m are the non-seasonal and seasonal part of the model, respectively. The d constraint tells how many differencing orders are going to be used to make the series stationary. The constraint m is the number of periods per season. The constraint m value is set with a period of 12.

IV. EXPERIMENTAL RESULTS

Akaike's Information Criterion (AIC) is the most commonly used model selection criterion [10]. AIC basically deals with the goodness of fit of a model. AIC is calculated as [10]:

$$\text{AIC} = -2 \ln(\text{maximum likelihood}) + 2p$$

Where, p represents the number of independent constraints estimated. Therefore, when comparing models, the one with the least AIC value is chosen. According to Table 1, SARIMA (1, 1, 1) × (1, 1, 1)₁₂ shows the lowest AIC value (AIC=196085.724). Other than this value all AIC values are larger. Thus, this model should be said that the good forecasting model.

Table1 AIC vales of SARIMA model

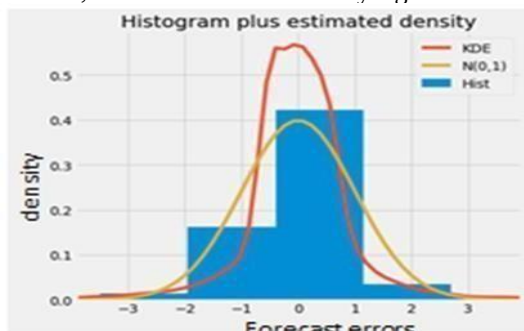
SARIMA (p, d, q) (P, D, Q) _s	AIC Values
SARIMA(0, 0, 0)x(0, 0, 0) ₁₂	AIC:204680.074
SARIMA(0, 0, 0)x(0, 0, 1) ₁₂	AIC:217984.649
SARIMA(0, 0, 0)x(0, 1, 0) ₁₂	AIC:200841.533
SARIMA(0, 0, 0)x(0, 1, 1) ₁₂	AIC:215339.952
SARIMA(0, 0, 0)x(1, 0, 0) ₁₂	AIC:197085.724
SARIMA(0, 0, 0)x(1, 0, 1) ₁₂	AIC:247760.254
SARIMA(0, 0, 0)x(1, 1, 0) ₁₂	AIC:196650.122
SARIMA(0, 0, 0)x(1, 1, 1) ₁₂	AIC:205623.637
SARIMA(0, 0, 1)x(0, 0, 0) ₁₂	AIC:245623.637
SARIMA(0, 0, 1)x(0, 0, 1) ₁₂	AIC:225623.637
SARIMA(0, 0, 1)x(0, 1, 0) ₁₂	AIC:235623.637
SARIMA(0, 0, 1)x(0, 1, 1) ₁₂	AIC:197623.637
.....	
SARIMA(1, 1, 1)x(1, 1, 1)₁₂	AIC:196085.724
.....	

Diagnostic Test: The forecast accuracy of the selected model is validated by applying a Dickey-Fuller test. According to Table2 the AIC value of SARIMA (1, 1, 1) $\times$ (1, 1, 1)₁₂ is the lowest. Table2 summarizes the results of the diagnostics test of the SARIMA (1, 1, 1) $\times$ (1, 1, 1)₁₂ model.

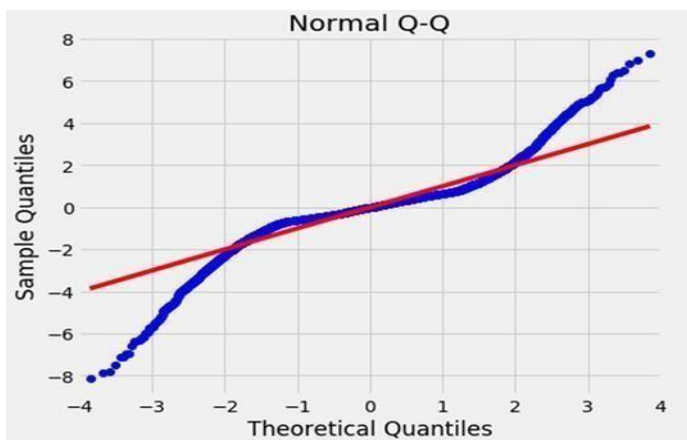
Table2 Summary of the diagnostics test of the SARIMA (1,1, 1) $\times$ (1, 1, 1)₁₂ model.

	Coef	std err	z	P> z	[0.025	0.0975]
Const	28.1213	0.461	60.976	0.000	27.217	29.025
AR.L1	0.7704	0.003	222.732	0.000	0.764	0.777
MA.L1	-1.0000	0.006	-155.452	0.000	-1.013	-0.987
AR.S.L12	-0.6651	0.005	-147.274	0.000	-0.674	-0.656
MA.S.L12	-0.8680	0.002	-381.308	0.000	-0.873	-0.864

The second column is the weight of the coefficients. The “Coef” column shows the weighting (i.e., importance) of each feature and how each one impacts the time series. The first set of AR and MA variables (AR.L1 and MA.L1, respectively) is lagged by one time step, while the second set is lagged by 12 time steps (AR.S.L12 and MA.S.L12). Since all values of $P > |z|$ are less than 0.05, the results are statistically significant.



(5a)



(5b)

Figure 5 Diagnostic tests on the residuals of the model

(5a) Distribution of standardize residuals

(5b) Normal Q-Q plot

The results of the diagnostic test on SARIMA (1,1,1) $\times$ (1,1,1)₁₂ are shown in Figure 5. According to Figure 5a, the results imply that the residual follows a normal distribution, with mean equal to 0 and standard deviation equal to 1. In Figure 5b, the Q-Q plot of the residuals implies that the residuals follow a linear trend. Thus, the residuals are normally distributed. Table3 show the comparison between actual and predicted value of temperature in C.

Table3 Actual value vs predicted values of temperature (°C)

DateTime	Actual Values	Predicted Values
2019-01-31 02:00:00	30.193548	28.765794
2019-02-28 02:00:00	32.642857	31.115332
2019-03-31 02:00:00	35.032258	35.282232
2019-04-30 02:00:00	35.466667	35.745039
2019-05-31 02:00:00	35.677419	35.375808
2019-06-30 02:00:00	30.033333	33.774278
2019-07-31 02:00:00	28.258065	27.734503
2019-08-31 02:00:00	29.129032	28.324608
2019-09-30 02:00:00	29.166667	29.434541
2019-10-31 02:00:00	28.580645	29.058096
2019-11-30 02:00:00	29.533333	28.366377
2019-12-31 02:00:00	29.516129	29.968087

Figure 6, shows the time series plot of actual value and predicted values of the temperature using SARIMA model. To evaluate the quality of the model, we will first compare the predicted values with the actual values. We can also see some kind of variations in the plot. These types of seasonal variations may cause by climate condition and any other external factors. From this figure, we can observe that the prediction results are almost equal to the actual data. We can say that the seasonal ARIMA model is performing better. Figure 7, shows the time series plot of actual value and predicted values of the temperature using ARIMA model. From the results we can say that the model is not fitted well as compare to SARIMA model. Figure 8, shows the future prediction of temperature using SARIMA model.

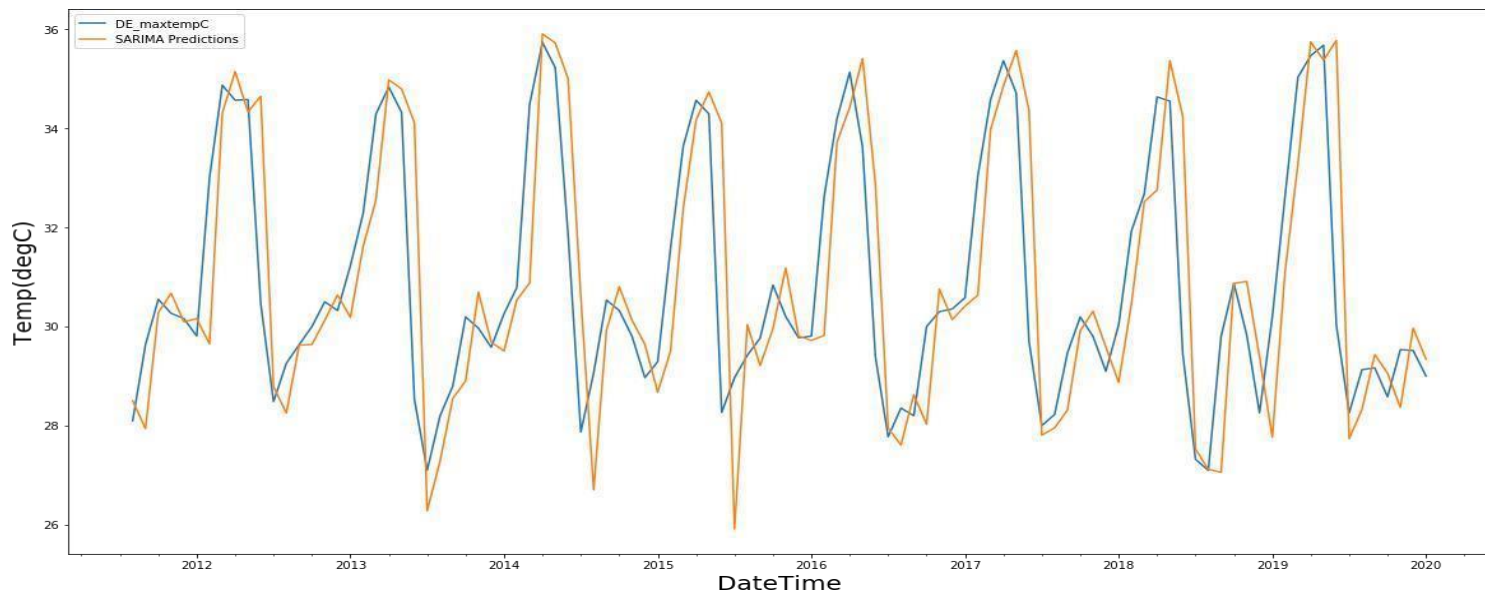


Figure 6 Actual value v/s predicted value using SARIMA model

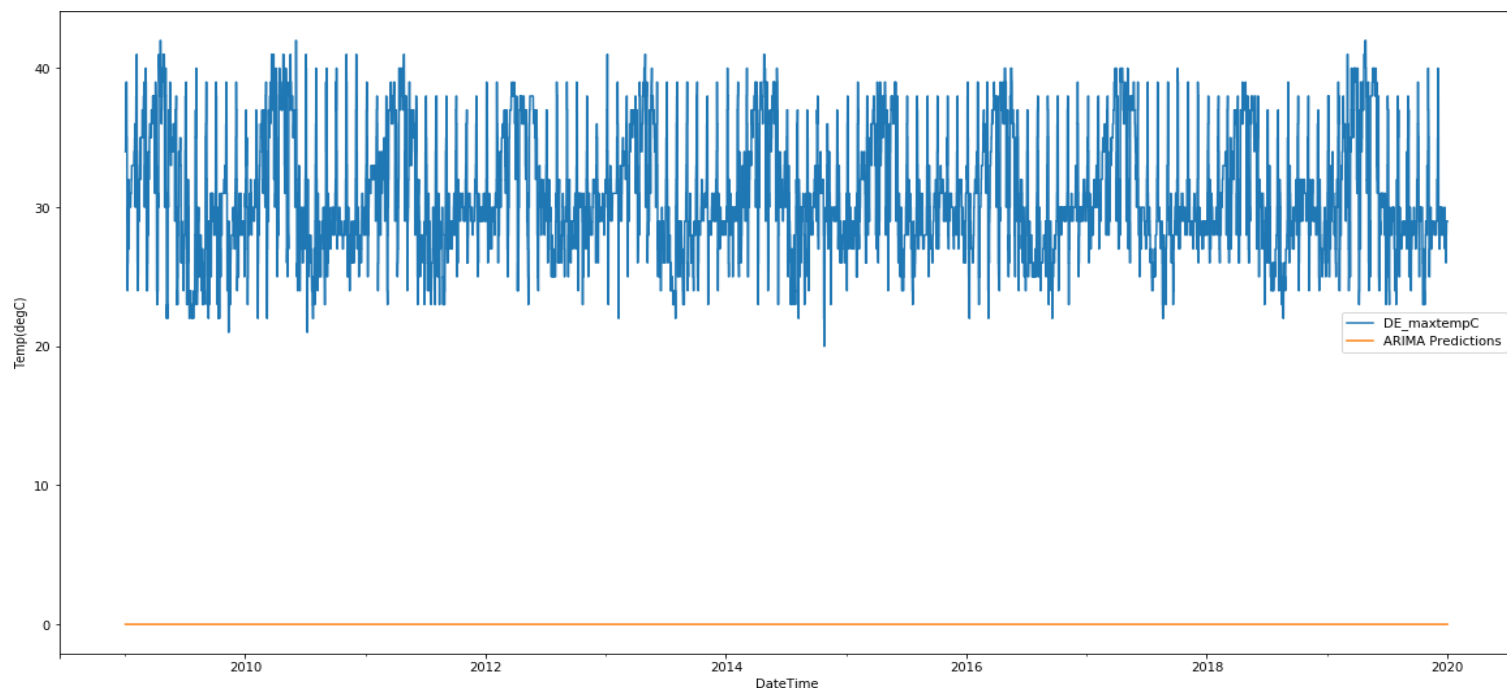


Figure 7 Actual value v/s predicted value using ARIMA model

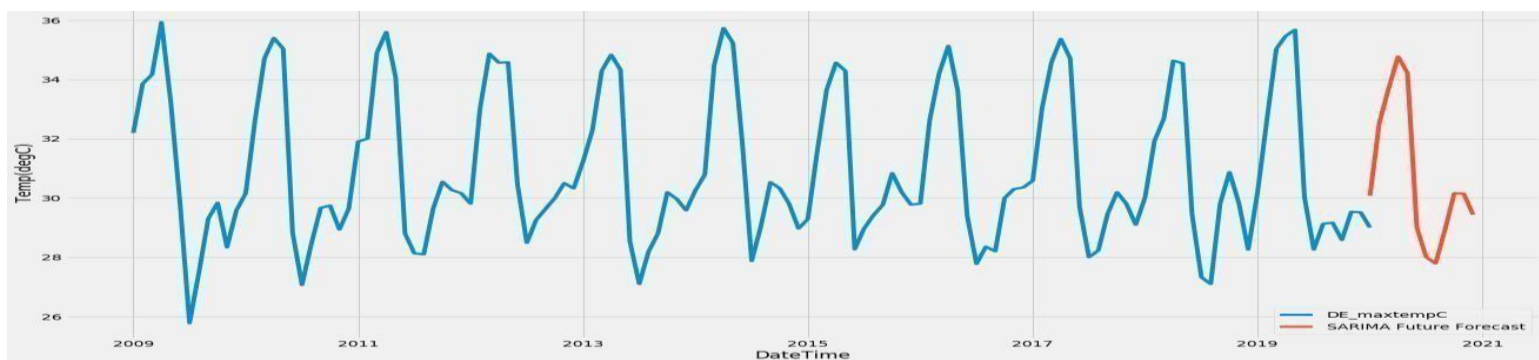


Figure 8 Future Prediction of temperature using SARIMA model

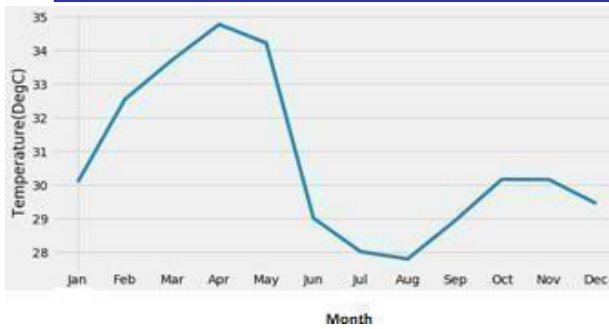


Figure 9 Time series plot of the future prediction (year-2021)

The above figure shows the time series plot of the temperature prediction. During the month from June-August we can see the sudden decrease in temperature, we can assume that this will be due to rainy season.

V. PERFORMANCE EVALUATION

Mean Square Error(MSE), Root Mean Square Error(RMSE) and Mean Absolute Error(MAE) were used as performance evaluation metrics given in Table 4. By taking the square of the errors, MSE is calculated as [11]:

$$MSE = \frac{1}{n} \sum_{t=1}^n e_t^2 \quad (5)$$

RMSE takes the root of the MSE. Thus, it has the same unit of measurement as the data. It is calculated as [11]:

$$RMSE = \sqrt{\frac{1}{n} \sum_{t=1}^n e_t^2} \quad (6)$$

Mean absolute error is the average of the absolute values of the deviation.

$$MAE = \text{mea}(|e_t|) \quad (7)$$

Table4 Results of the performance evaluation of the model

Method	MAE	MSE	RMSE
ARIMA	6.052	56.187	7.496
SARIMA	0.60850	0.58114	0.762325

The predicted temperature values are compared with actual values for accuracy based on error metrics. We obtained MAE of 0.60850 and RMSE of 0.76233for SARIMA model and MAE of 6.052 and RMSE of 7.496 for ARIMA model. From the above table, we concluded that SARIMA model forecasts yielded least error in prediction of temperature as output.

CONCLUSION

In this paper, temperature data were collected from the year 2009-2020 at one-hour intervals in Pune. The estimation and diagnostic analysis results revealed that the model adequately fitted to the historical data. Power load forecasting is the basic work of power grid control optimization and significant part of power system transmitting. In practical applications, the non-linear relationship between environmental factors and load changes defined by the improved optimization algorithms can effectively reduce the deviation between predicted results and actual results. To maintain an electrical grid's production/consumption balance, stochastic production forecasting must be implemented at multiple temporal horizons based on the level of utilization. Finally, the predicted values were compared with the actual values of both using ARIMA and SARIMA model. Forecast accuracy measures, including MAE, MSE, and RMSE were calculated.

REFERENCES

- [1] Imdadullah. "Time Series Analysis". Basic Statistics and Data Analysis. itfeature.com. Retrieved 2 January 2014.
- [2] Raicharoen, t., lursinsap, c., & sanguanbhokai, p. (2018). Application of critical support vector machine to time series prediction. International symposium on circuits and systems (vol.5, pp.v-741- v-744 vol.5). IEEE
- [3] Khandelwal, I., Adhikari, R., & Verma, G. (2015). Time series forecasting using hybrid arima and ann models based on DWT Decomposition. In Procedia Computer Science (Vol. 48, pp. 173–179) Elsevier B.V. <https://doi.org/10.1016/j.procs.2015.04.167>
- [4] Box, G. E. P., & Jenkins, G. M. (1976). Time series analysis forecasting and control - rev. ed. Oakland, California, Holden-Day, 1976, 37 (2), 238 - 242.
- [5] G. J. Rios-Moreno, M. Trejo-Perea, R. Castaneda- Miranda, V. M. Hernandez-Guzman, and G. Herrera- Ruiz, "Modelling temperature in intelligent buildings by means of autoregressive models," vol. 16, pp. 713– 722, 2014.
- [6] M. De Felice, A. Alessandri, and P. M. Ruti, "Electricity demand forecasting over Italy, Potential benefits using numerical weather prediction models," Electr. Power Syst. Res., vol. 104, pp. 71–79, 2013.
- [7] Mahmudur Rahman, A.H.M. Saiful Islam, Sahah Yaser Maqnoon Nadvi, Rashedur M Rahman (2013): Comparative Study of ANFIS and ARIMA Model for weather forecasting in Dhaka" IEEE
- [8] Chisimkwuo, J., Uchechukwu, G. And Okezie S.C. 2014 Time series analysis and forecasting of monthly maximum temperatures in south eastern Nigeria. International Journal of Innovative Research and Development. ISSN 2278 – 0211. 3(1), pp. 165- 171
- [9] Roy, T. D. and Das K. K. 2012 Time series analysis of Dibrugarh air temperature. Journal of Atmospheric and Earth Environment. 1(1), pp. 30- 34.
- [10] Box-Jenkins models, NIST handbook of statistical method
- [11] C. Chatfield, Time-series Forecasting. Chapman & Hall/CRC, 2015.
- [12] www.kaggle.com