ENTHUSIASTIC
INTERNATIONAL JOURNAL OF STATISTICS AND DATA SCIENCE
Volume 2, Issue 1, April 2022, pp. 1-9

Analysis of Changes in Atmospheric CO_2 Emissions Using Prophet Facebook

Arum Handini Primandari [a,1,*], Achmad Kurniansyah Thalib [a,2], Ayundyah Kesumawati [b,3]

[a] Statistic Study Program, Universitas Islam Indonesia, Indonesia
[b] Purwadhika Digital Technology School, Bandung, Indonesia
[1] primandari.arum@uii.ac.id*; [2] achmadkurniansyah@gmail.com; [3] ayundyah.k@uii.ac.id
* corresponding author

ARTICLE INFO

Article history
Received February 27, 2022
Revised April 25, 2022
Accepted April 27, 2022

Keywords
Carbon dioxide
Prophet
Spline interpolation
Time series

ABSTRACT

CO_2 emissions have been an environmental issue for decades. The trigger for the increasing concentration of CO2 in the atmosphere is the growth of industries related to burning fossil fuels for coal, natural gas, and petroleum. For nearly a century, several attempts have been made to suppress the rapid growth of CO_2. This study uses daily atmospheric CO_2 levels observed in Mauna Loa laboratories. The method used is a Prophet that can handle seasonality and mark the change points. Almost 20% of data was missing value, which was then imputed using spline interpolation. Based on the analysis results, CO_2 levels have an upward trend throughout the year and seasonality. There is no point of change in the last ten years that shows a decrease in CO_2 levels. Using forward chaining cross-validation evaluation and error measurement, the prophet model can follow the pattern of CO_2 levels well. The average RMSE value is less than 2.0, with an MAPE value below 0.5%.

1. Introduction

It is undeniable that CO_2 emissions are an inseparable effect of industrial developments in the use of fossil fuels, natural gas, and petroleum. The contribution of deforestation exacerbates the condition. Natural carbon sinks remove CO_2 from the atmosphere considerably slower than these emission rates. Therefore, CO_2 levels continue to grow. Reducing emissions can slow the rate of accumulation, but they will not halt it until the total amount of CO_2 in the atmosphere is zero.

Based on the Met Office report, the average carbon dioxide concentration in March 2021 was even 50% higher than the average in 1750-1800. In May 2021, the Met Office predicted that carbon dioxide concentrations would reach their highest. Although, the levels had dropped, perhaps due to the Covid-19 pandemic [1].

The pattern of CO_2 concentration has a seasonality, influenced by the cycle of seasons on earth. From June to September, CO_2 concentrations in Mauna Loa follow the usual temporary declines due to carbon sequestration by ecosystems in the Northern Hemisphere growing season. After October reaches its lowest point, CO_2 concentrations usually increase again as ecosystems release carbon during autumn and winter in the Northern Hemisphere. This natural seasonal cycle is a prominent feature of the famous Keeling Curve graph of atmospheric CO_2 at Mauna Loa, above a clear long-term upward trend. The annual average atmospheric CO_2 concentration increases yearly and is a significant driver of global warming [1].

With this explanation, the carbon dioxide data will have a seasonality, which is a pattern that repeats itself in the same time period. The following illustrates the concentration of carbon dioxide in the years 2017 to 2021. The peaks and valleys in the data appear to be repeating itself, apart from the upward trend.

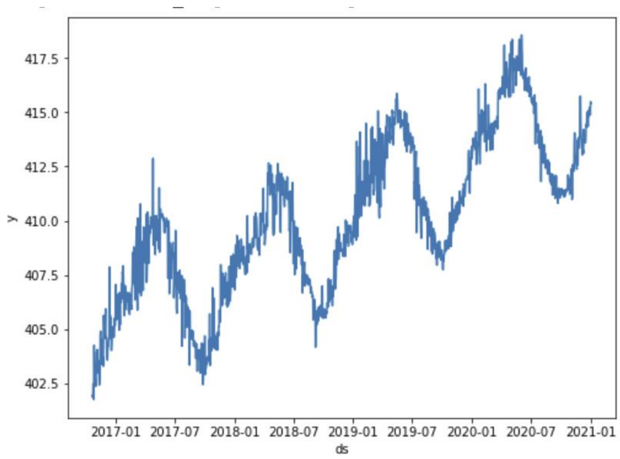


Fig. 1. Carbon dioxide concentration in 2017-2021.

The Prophet is a model created by Facebook in 2017 available in R and python language libraries. The advantage of this model is that simple changes to its parameters can significantly increase the model's accuracy. Prophet has been widely used to forecast weather temperature (NOAA data: National Oceanic and Atmospheric Administration), which can be an alternative to conventional meteorological methods [2], [3]. Research by Weytjens et al. uses Prophet to forecast the company's cash flow, along with ARIMA methods and neural networks such as MLP and LSTM. Prophet is considered a method with the second smallest MSE after LSTM [4]. For forecasting seasonal data such as the volume of train passengers in Indonesia, Prophet outperforms the Feed Forward Neural Network (FFNN). Although, the MAPE difference between the two methods is not very significant [5].

While several studies have tried to predict the numbers related to the pandemic, Lounis's research employed Prophet to predict the active, death, and recovery rates in Algeria [6]. The FB Prophet model could help adopt the best preventive measures. A similar study by Mahmud predicts daily cases of Covid-19 in Bangladesh. Although in general, the Prophet model can follow seasonal patterns. However, there are still overestimates and underestimates [7].

The next section of this paper is organized into the following way. The second part includes data and data sources used in this study and the research methodology. Discussion regarding the results of the analysis of changes in carbon dioxide over time is presented in section 4—the technical analysis of forecasting with Prophets, including parameter selection, cross-validation, and model evaluation. The last part is the conclusion.

2. Material and Method

2.1. Data and Resource

The data are in situ observations of the average daily levels of Atmospheric Carbon Dioxide Dry Air Mole Fractions categorized as greenhouse gases. Observations are quasi-continuous measurements at Mauna Loa-Hawaii, United States (MLO) laboratories [8]. Its size is Mole fraction reported in units of micromol mol-1 (10-6 mol per mole of dry air); abbreviated as ppm (parts per million). Measurement methods change according to the times. Before 2013 and 2014, CO2 measurements were based on Non-dispersive Infrared (NDIR) CO_2 analyzers. After 2017, the next-generation analytical systems entirely used laser-based spectrometers to measure CO_2 [9].

p-ISSN 2798-253X
e-ISSN 2798-3153

The data source is ESRL's Global Monitoring Laboratory (GML) of the National Oceanic and Atmospheric Administration (NOAA) website (https://gml.noaa.gov/). The institution conducts research that addresses three principal challenges: greenhouse gas and carbon cycle feedback, changes in clouds, aerosols, surface radiation, and stratospheric ozone recovery.

2.2. Method of Analysis

This research utilizes a tool provided by Facebook, Prophet, issued in 2017 as open-source software. The prophet is designed with no parameter tuning or optimization. Hence, anyone is able to tweak the model to increase performance significantly. Prophet can handle trends, seasonal, holiday, outlier, and special events. In addition, this tool provides a cutoff point of change that occurs in the trend.

The time series is a set of sequentially time observed data—the data should have the same time interval. Even a single missing value would cause the interval to be uneven. CO_2 data has many missing values, but no source explains the cause. Therefore, we need to overcome this problem first. The following **Fig. 2** is the research flowchart.

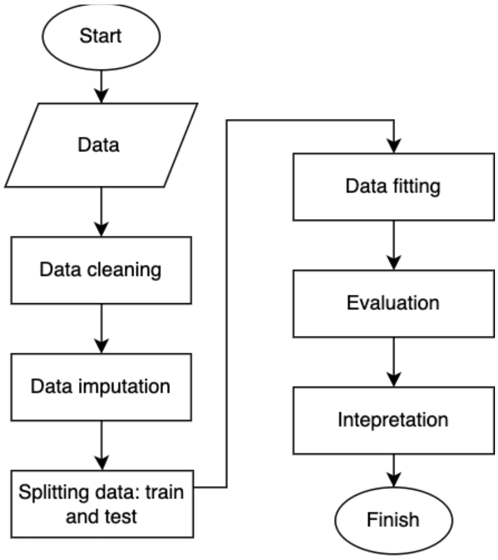


Fig. 2. Research flowchart.

After the data is downloaded, preprocessing ensures the location and number of missing values. The actual data plot that shows the trend and seasonality help to determine the imputation method. In this research, we utilized spline interpolation. After this process, the data is organized according to the input format required by Prophet. The data is divided into training and testing with a ratio of 90%: 10%. The distribution of splitting data does not use randomization but chronologically.

The data fitting leverages Prophet with an additive seasonality on weekly and yearly. Instead of k-fold, the data validation process uses the rolling-origin cross-validation method with no random shuffling data. Performance metrics measure the model's performance, including MSE, RMSE, MAE, and MAPE. The last process is interpreting the results obtained from the Prophet and comparing the CO_2 levels from the four laboratories.

2.3. Prophet

The Prophet model is an open-source library for R and python programming languages in forecasting time series. This model is based on a decomposable model that the Facebook team provided in 2017. It may achieve excellent performance, even with simple and intuitive settings, compared to other standard forecasting models. Moreover, it considers the impact of custom seasons and holidays. The Prophet model can be expressed mathematically as [2], [3], [5], [10]

$$y(t)=g(t)+s(t)+h(t)+\varepsilon_t \quad (1)$$

p-ISSN 2798-253X
e-ISSN 2798-3153

where y(t) is the model's predicted/fitted value, g(t) is the trend function, s(t) is the periodical function impacted by weekly and yearly seasonality, and h(t) is the effect of holidays with varied periods. The error term ε_t represents random changes that the model does not account.

2.3.1. Trend

There are two types of growth modes in the Prophet: the linear and the logistic. Mode selection is based on the characteristics of the data. Linear mode means that the data have straight and slope trend. Meanwhile, when the data is S-shaped curve, it is better to choose logistic mode [11]. The formulation of growth define as follows [5].

$$g(t)=(k+a(t)^T\delta)t+(m+a(t)^T\gamma) \quad (2)$$

The expression of logistic mode is represented as Eq. (3)

$$\boldsymbol{g}(t)=\frac{C(t)}{1+\exp\left(-(k+a(t)^T\delta)\left(t-(m+a(t)\gamma)\right)\right)} \quad (3)$$

where C(t) is a time-varying capacity, m is an offset parameter, k represents the rate of growth, γ is vector of change-points adjustment, and δ is vector of growth rate adjustment [3], [6], [10].

2.4. Seasonality

Periodic changes in the same time interval are called seasonal. If it is monthly data, then the Prophet can decompose the annual seasonality, which shows repeated movements in the monthly period. If it is daily data, the generated seasonality is weekly and yearly. Likewise, when the data is hourly, seasonal on the Prophet can provide multiple seasons: daily, weekly, and yearly. There also, fourier series to approximate periodic property. The seasonal model defines as

$$s(t)=\sum_{r=1}^{R}\left(a_n\cos\left(\frac{2\pi nt}{P}\right)+b_n\sin\left(\frac{2\pi nt}{P}\right)\right) \quad (4)$$

$$\beta=[a_1,b_1,a_2,b_2,\ldots,a_R,b_R]$$

the asumption of the distribution for $\beta \sim N(0,\sigma^2)$ to enforce smoothing prior to seasonality [3], [5], [6], [10] . The R is the number of parameters β and P is the expected regular period series data. Usually, it should be sufficient to take R=10 for yearly periodicity and R=3 for weekly.

2.5. Spline Interpolation

A *spline* is a particular function defined piecewise by polynomials. The general idea of the spline is to represent the graph with a simple function on each interval between data points. The simplest graph is the linear line. The quadratic function will make the graph a lot smoother, likewise with more immense polynomial powers. It is our intuitive to stop the degree of the polynomial.

The imputation process uses the imputeTS library from R. This library provides interpolation with linear, spline, and stine modes. In addition, the plot functions allow us to plot data without missing values where the missing values are marked with a red vertical line. It helps us observe the pattern of data and determine the imputation method. There is also a plot showing actual data and red dots resulting from imputation [12].

2.6. Evaluation

2.6.1. Forward-chaining Cross validation

The test set in Forward-chaining cross-validation with no randomization must be the final portion of the data. For example, if each fold is 10% data (as in 10-fold cross-validation), the test set must be at the final 10% of the data range. We choose an initial amount of data to train on, such as seven-folds, and then evaluate on the eighth fold and preserve that performance metric with the remaining data. Next, the first of eight folds are re-trained, and the ninth is evaluated. Repeat the technique until all of the folds have been exhausted. The ilustration is on the Fig. 3 where the white box is train set and the grey box is test set [11].

p-ISSN 2798-253X
e-ISSN 2798-3153

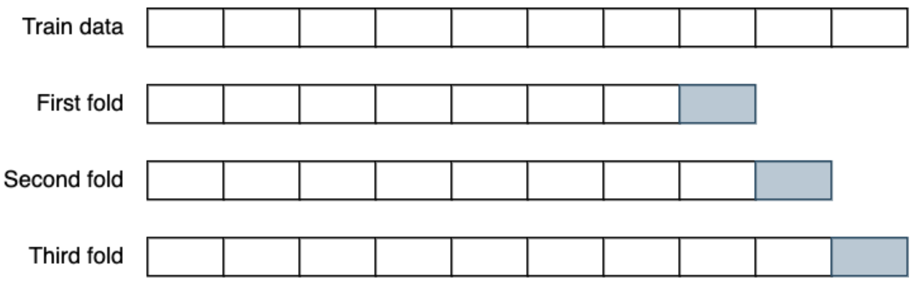


Fig. 3. Forward-chaining cross-validation with seven folds.

Forward-chaining cross validation also called rolling-origin cross-validation because the "origin" at which the forecast is based rolls forward in time. The forecast accuracy is computed by averaging over the test sets. For a given index T, k<T<n, the data $X_1, X_2, \ldots, X_n\}$ is divided into train sets $\{X_1, X_2, \ldots, X_{T+m}\}$ and test sets $\{X_{T+m+1}, X_{T+m+2}, \ldots, X_n\}$, m=0, 1, …, n-1-T respectively. The formula define as follows [13] [14].

$$ATE_T^2(k)=\sum_{m=0}^{n-1-T} \frac{1}{n-T-m} \sum_{t=T+m+1}^{n} e^2 \quad (5)$$

where $e=(x_i-\hat{x}_i)$ is the error of actual data from its fitting value.

2.6.2. Error measurement

The Prophet diagnostic package includes six metrics for assessing the model. The metrics are Mean Squared Error (MSE), Root of MSE (RMSE), Mean Absolute Error (MAE), Mean Absolute Percentage Error (MAPE), etc. The formula of these metrics is define as follows.

$$MSE=\frac{1}{n}\sum_{i=1}^{n}(x_i-\hat{x}_i)^2 \quad (6)$$

The RMSE is simply the root squared of MSE. The formulation of MAE only differ on the operation for error that is absolute.

$$MAE=\frac{1}{n}\sum_{i=1}^{n}|x_i-\hat{x}_i| \quad (7)$$

The absolute percentage error takes the percentage of the error ratio to the actual data.

$$MAPE=\frac{1}{n}\sum_{i=1}^{n}\left|\frac{x_i-\hat{x}_i}{x_i}\right| *100\% \quad (8)$$

3. Result and Discussion

NOAA GML began continuous measurements of atmospheric CO2 at Mauna Loa, Hawaii (MLO) in April, 1974. In the MLO daily average observation, 16% is the missing value. Spline interpolation is used for imputation because the data pattern has peaks and valleys due to seasonality. **Fig. 14** shows that the imputation results marked with red dots are able to follow the seasonal movement of the actual data.

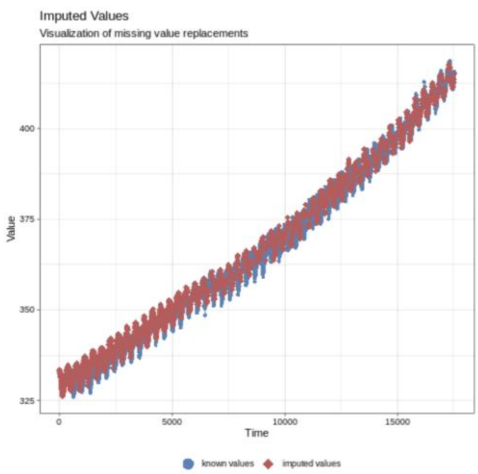


Fig. 4. Imputed values at (a) MLO; (b) BRW.

p-ISSN 2798-253X
e-ISSN 2798-3153

The following process sets the data to meet the input required by the Prophet, namely a data frame with only two columns: 'ds' is a date and 'y' is an observation. This restriction becomes the weakness of Prophet limited on univariate time series. Although Prophet also comes up with an additional regressor. Observation dates are available in raw data in separated columns year, month, and day that need to be concatenated and considered date. The y column is observations with imputed missing values.

From 1973 to 2020, there were 17,531 observations which split into a 90% training set of 15,779 observations and a 10% testing set of 1,752 observations. We built the model by utilizing a training set. The validation process was carried out after. Furthermore, the testing data evaluate the model's performance.

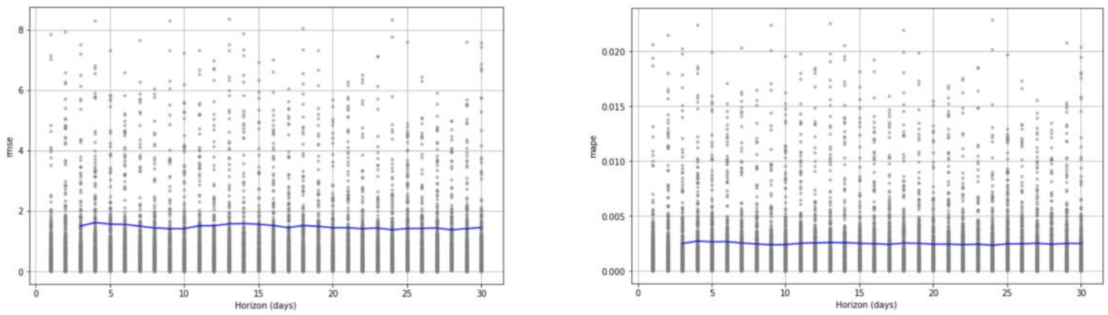
Hyperparameter tuning utilizes the grid search function to combine several parameters in the Prophet function. The seasonal mode, either additive or multiplicative, is defined using the 'seasonality_mode' input parameter. A Fourier order is employed by setting the 'yearly_seasonality' for smoothing the curvy and bumpy seasonality. The parameter 'changepoint_prior_scale' adjusts the trend flexibility by setting the strength of the sparse prior. By default, this parameter is set to 0.05. Increasing it will make the trend more flexible. There is a parameter seasonality_prior_scale which similarly adjusts the extent to which the seasonality model will fit the data.

Table 1. Hyperparameter Tuning

Index	Seasonality mode	Yearly seasonality	Changepoint prior scale	Seasonality prior scale	MSE	MAPE
0	additive	2	0.1	0.05	2.204	0.282
1	additive	2	0.1	1	2.203	0.282
2	additive	2	0.5	0.05	2.203	0.282
3	additive	2	0.5	1	2.203	0.282
4	additive	4	0.1	0.05	2.192	0.282
5	additive	4	0.1	1	2.193	0.282
6	additive	4	0.5	0.05	2.191	0.282
7	additive	4	0.5	1	2.191	0.282
8	multiplicative	2	0.1	0.05	2.181	0.280
9	multiplicative	2	0.1	1	2.180	0.280
10	multiplicative	2	0.5	0.05	2.181	0.280
11	multiplicative	2	0.5	1	2.180	0.280
12	multiplicative	4	0.1	0.05	2.170	0.279
13	multiplicative	4	0.1	1	2.170	0.280
14	multiplicative	4	0.5	0.05	2.169	0.279
15	multiplicative	4	0.5	1	2.171	0.280

Based on the fitting data in Table 1. The smallest MSE and MAPE was obtained when the seasonality mode was multiplicative with Fourier order of 4, scale of change point of 0.5, and scale of prior seasonality of 0.05. To validate the result, we run the rolling-origin cross validation by setting the horizon of 30 days as the same as the period and the initial at 20 years observation. The horizon is the period over which you want to evaluate your forecast. When we have 30 days of daily data predicting, it is a plausible length of time. Period is the amount of time between each fold. While initial is the first training period. The RMSE and MAPE of the cross-validation within 30 days harizon is presented in **Fig. 5** having value respectively less than 2 and 0.5%.

p-ISSN 2798-253X
e-ISSN 2798-3153



(a) (b)

Fig. 5. Plot of RMSE and MAPE of cross-validation.

Prophet can generate both weekly and yearly seasonality out of daily data. The **Fig. 6** shows the trend and seasonality.

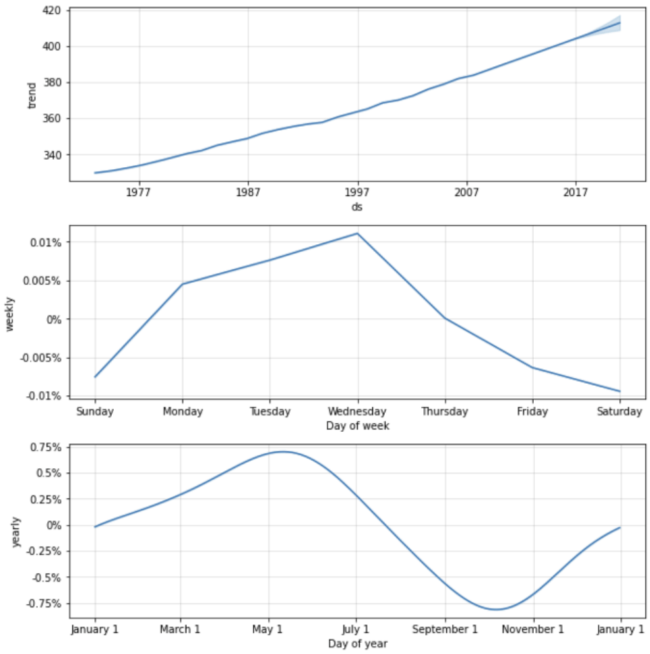


Fig. 6. Plot trend, weekly seasonality, and yearly seasonality.

The trend continues to rise, but the slope appears to be steepening as time goes on. It indicates that CO_2 concentrations in the atmosphere are increasing. Weekly seasonality indicates that day by day of the week, the value varies from -0.01% to 0.01% PPM. Intuitively, some laboratories far from residential areas do not show weekly seasonality. So, the value of the concentration movement is purely due to random chance.

Carbon dioxide shows an upward movement from January to peak in May and June in annual seasons. These months are winter in the northern hemisphere. Concentrations decrease after June to their lowest point in October when it is summertime. Based on the actual data in **Fig. 1**, it is clear that there is an annual seasonal pattern.

In the additive model, the y-axis is the absolute value. In the multiplicative model, y is the percentage. Additive seasonality is modeled as an additional factor to the trend, i.e., by adding or subtracting the value of the trend. However, multiplicative seasonality represents a deviation relative to the trend, so the magnitude of the seasonal effect depends on the value predicted by the trend at that point. The seasonal effect is the percentage of the trend.

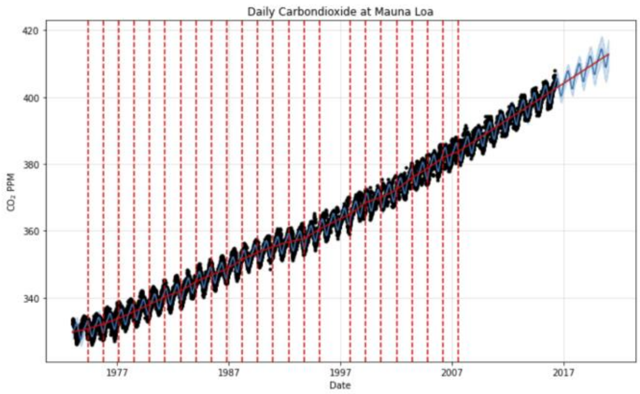


Fig. 7. Change points of trend.

The dashed red line is change point. Trend changepoints are located where the trend component of the model suddenly changes its slope. There could be many reasons. In For this case: i t is hard to tell from the chagepoints that the trends is actually bending. The magnitute before 1987 and after 1997 is sharper than the time between. It is ranged from -0.15 to 0.15.

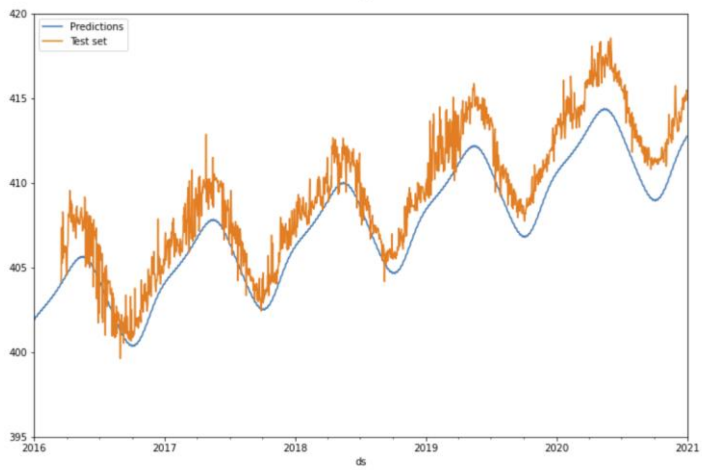


Fig. 8. Forcasting of test set.

Forecasting values follow seasonal and trend patterns well. However, most of the data testing forecasting results have underestimated values. The larger the period, the further the forecast value will deviate from the actual data. The MAPE for testing data is 0.45%.

4. Conclusion

Daily atmospheric carbon dioxide observations have many missing values imputed by the spline interpolation method. The imputed values reveal that they follow the actual data patterns well. The data clearly show seasonality and trends.

Prophet with grid search optimization technique including seasonal mode, Fourier series, and seasonal scale provides good analysis performance for carbon dioxide observation. The validation employs rolling-origin cross-validation on the data train, which does not randomize the data, but rolls the data for forecasting from its origin. Meanwhile, the testing data shows that the MSE and MAPE is relatively small, below 2 and 0.5% respectively.

The prophet model also generates weekly and annual patterns from daily carbon dioxide data. There is a tendency for CO2 to rise in winter and fall in summer in the Northern Hemisphere. A sharp increase in CO2 concentration was found from year to year based on trend analysis. Changepoints also support the interpretation of the trend, as some of the changepoint slopes down occurred in 1987-1997, whereas changepoints had a high magnitude after that.

p-ISSN 2798-253X
e-ISSN 2798-3153

References

[1] R. Betts and R. Keeling, "Atmospheric carbon dioxide at record high levels despite reduced emissions in 2020 - Met Office," 2021. https://www.metoffice.gov.uk/research/news/2021/record-co2-levels-despite-lower-emissions-in-2020 (accessed Feb. 22, 2022).

[2] Z. Z. OO and S. PHYU, "Time Series Prediction Based on Facebook Prophet: A Case Study, Temperature Forecasting in Myintkyina," *Int. J. Appl. Math. Electron. Comput.*, vol. 8, no. 4, pp. 263–267, 2020, doi: 10.18100/ijamec.816894.

[3] T. Toharudin, R. S. Pontoh, R. E. Caraka, S. Zahroh, Y. Lee, and R. C. Chen, "Employing long short-term memory and Facebook prophet model in air temperature forecasting," *Commun. Stat. Simul. Comput.*, vol. 0, no. 0, pp. 1–24, 2020, doi: 10.1080/03610918.2020.1854302.

[4] H. Weytjens, E. Lohmann, and M. Kleinsteuber, "Cash flow prediction: MLP and LSTM compared to ARIMA and Prophet," *Electron. Commer. Res.*, no. 0123456789, 2019, doi: 10.1007/s10660-019-09362-7.

[5] R. S. Pontoh, S. Zahroh, H. R. Nurahman, R. I. Aprillion, A. Ramdani, and D. I. Akmal, "Applied of feed-forward neural network and facebook prophet model for train passengers forecasting," *J. Phys. Conf. Ser.*, vol. 1776, no. 1, pp. 0–9, 2021, doi: 10.1088/1742-6596/1776/1/012057.

[6] M. Lounis, "Predicting active , death and recovery rates of COVID-19 in Al- geria using Facebook ' Prophet model," *Preprints*, vol. 1, no. March, 2021, doi: 10.20944/preprints202103.0019.v1.

[7] S. Mahmud, "Bangladesh COVID-19 Daily Cases Time Series Analysis using Facebook Prophet Model," *SSRN Electron. J.*, no. June 2020, 2020, doi: 10.2139/ssrn.3660368.

[8] K. W. Thoning, A. M. Crotwell, and J. W. Mund, "Atmospheric Carbon Dioxide Dry Air Mole Fractions from continuous measurements at Mauna Loa, Hawaii, Barrow, Alaska, American Samoa and South Pole," Boulder, Colorado, USA, 2021. doi: https://doi.org/10.15138/yaf1-bk21.

[9] C. L. Zhao and P. P. Tans, "Estimating uncertainty of the WMO mole fraction scale for carbon dioxide in air," *J. Geophys. Res. Atmos.*, vol. 111, no. 8, Apr. 2006, doi: 10.1029/2005JD006003.

[10] L. Guo, W. Fang, Q. Zhao, and X. Wang, "The hybrid PROPHET-SVR approach for forecasting product time series demand with seasonality," *Comput. Ind. Eng.*, vol. 161, no. February, p. 107598, 2021, doi: 10.1016/j.cie.2021.107598.

[11] G. Rafferty and an O. M. C. Safari, *Forecasting Time Series Data with Facebook Prophet*. Packt Publishing Ltd., 2021.

[12] S. Moritz and T. Bartz-beielstein, "imputeTS : Time Series Missing Value Imputation in R," *R J. Vol.*, vol. 9, no. 1, pp. 207–218, 2017.

[13] Z. Liu and X. Yang, "Cross validation for uncertain autoregressive model," *Commun. Stat. Simul. Comput.*, vol. 0, no. 0, pp. 1–12, 2020, doi: 10.1080/03610918.2020.1747077.

[14] M. Schnaubelt, "A comparison of machine learning model validation schemes for non-stationary time series data," Nürnberg, No. 11/2019, 2019. [Online]. Available: http://hdl.handle.net/10419/209136.

p-ISSN 2798-253X
e-ISSN 2798-3153