Predict United States CO2 emissions using numerical techniques

Judy Truong

Index Terms—CO2 emission, numerical analysis, Autoregressive, Polynomial approximation

1 Introduction

THE seriousness of global warming, which is a major environmental issue, has been drew worldwide attention for many years. Temperatures have been raised over decades and this change is accompanying the serial critical effects on environments. The lengths of the winter and summer seasons have been getting longer since the 1980s, and this change caused unexpected seasonal disasters such as heavy precipitation in specific regions otherwise droughts in other regions. Besides, increased sea levels, and more frequent and stronger hurricanes are threatening our life. To address this issue, greenhouse gas pollution must be reduced. Since carbon dioxide is the principal cause of pollution, CO_2 regulation is essential tasks all over the world. CO2 emissions in the United States grew starting in 1850 and peaked in 2005, as seen in Figure 1. The emission level is still high although the graph shows the decreasing trend. Fortunately, CO2 emissions appear to get severe with the growth of industries, but decreased amount of emissions are expected according to the technological innovations such as renewable energy and energy source substitution instead of gasoline [1]. The white house announced the new target with this expectation, the 2030 Greenhouse Gas Pollution Reduction Target [2]. The target aims to achieve a 50-52 percent reduction from 2005 levels in economy-wide net greenhouse gas pollution in 2030. Here, we built models to predict if collected CO2 emission data can represent the half-decreased in CO2 emissions by 2030.

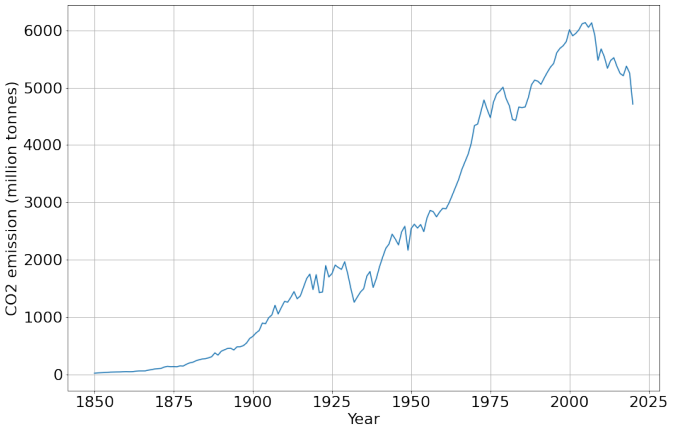


Fig. 1: United States total CO2 emission by year

1.1 CO2 reduction methods

1.1.1 Growing renewable energy business

The aim of US Energy Information Agency is to make renewables to be the primary source for new electricity generation out to 2050 [3]. Their short goal is to rapidly grow the renewable business by 2030 to reach around 50GW of net renewable generating capacity globally. Wind and solar are the major sources for it. The main idea for the wind business is to build fans offshore that are expected to power more than two million homes. Solar energy is the representative renewable source accounting for more than 40% of all new electricity generating capacity. Solar panels are efficient and easy to install anywhere.

1.1.2 Electric vehicles

Transportation is a primary factor of greenhouse gas pollution with the largest portion because most vehicles require gasoline [4]. Road transport from passenger vehicles such as cars and buses contributes almost half of CO2 emissions from transport. To lessen CO2 emissions from transportation, electric vehicles (EVs) which support the transition to renewable energy have developed and the supply is increasing. An increase in EVs usage can improve air quality and this improvement would be more effective if the electricity for charging is from renewable sources. This innovations will leads decline of emissions trends.

1.1.3 Carbon sequestration

Capturing and storing atmospheric carbon dioxide are crucial steps since carbon dioxide is already in excess in the air. Geologic and biologic ways are utilized for this approach [5] [6]. CO2 is required for enhanced oil recovery through storing CO2 in underground reservoirs. This process has some environmental risks such as leakage and impacts on drinking water, so this should be with sophisticated estimation and evaluation. CO2 is also utilized for photosynthesis in aquatic and land systems.

2 Related Work

There are many works on analyzing and predicting the CO2 emission around the world. For example, [7] analyzes the past CO2 emission data around the world and suggests that the global CO2 emissions have been flat for the recent decade. The authors explains the trend by breaking down

major CO2 emission sources and analyzing them through policy and social changes. They points out the major factors are revised land-use emissions, China and India rise in fossil CO2 emissions, coal and gas usage, and Changes to CO2 sources and sinks.

A software HOMER for technical analysis to reduce CO2 emissions provides the investigation that hybrid energy utilization composed of photovoltaic (PV), wind turbin, diesel generator, and battery has the best energy efficiency with a high return on investment [8]. The hybrid system is the most efficient way to reduce CO2 emission per household for now although electricity generation is a challenge. Active utilization of renewable energy and applying appropriate implementation of environmental friendly technologies have potential benefits; CO2 reduction, greater sustainable electricity generation and providing an economic justification for skateholders to invest in the renewable energy sector.

Vaclav Smil, a well-known scientist and policy analyst called for a "rapid decarbonization in order to combat global warming" [9]. According to the 26th United Nations Climate Change Conference held in Glaslow in 2021, it was recommended that the carbon dioxide emissions get reduced by 45 percent by 2030. However, this energy transition is seen as impossible as CO2 emissions are still exponentially rising with eight years left. Vaclav Smil considers carbon dioxide as one our deadliest problems and the article goes on to explain how complete decarbonization is not realistic. Though, carbon emissions can be regulated with a net carbon emission. This net carbon emission would be created with increased efforts of carbon sequestration.

Comparing to the related works, our work focuses on the CO2 emission only in the United States and trying to determine if the White House could reach their CO2 emission reduction goal by 2030, by using various numerical analysis techniques.

3 Linear Regression Model

In order to approximate future CO2 emission, it is informative to understand the trend among current and past CO2 emission data. Real life CO2 emission was calculated by kilotons of CO2. There are a large number of contributors to the United States' CO2 emission: electricity consumed and reduced, gallons of gasoline demand, gallons of diesel demand, gasoline-powered passenger vehicles per year, thermal units of natural gas, etc [10]. A linear regression model was created to see if there was relationship between CO2 emission and year.

3.1 Methodology

The emission factors and consumption of the largest variables were taken from the year 2020 to visually show any trends among the variables and CO2 emission. For example, the electricity consumption emission factor in 2020 was $4.33 * 10^{-4}$ metrics tons of CO2 per kilaWatt hour [11]. In the year of 2020, the United States had consumed 93 quadrillion British thermal units [12]. In terms of kiloWatt hours, it was 3.93e+12. Multiplying this value by the emission factor found in 2020, 1701690000 metric tons of CO2 was emitted in 2020. Another large contributor was the gallons of gasoline consumed in 2020. In 2020, Americans consumed about 123 billion gallons of motor gasoline [13], which had a emission factor of 8.887e-3 metric tons of CO2 per gallon of gasoline. This calculates to 1198234210 metric tons of CO2 emitted due to gasoline consumption in 2020. Another variable for CO2 emission is gallons of diesel consumed. The emission factor for diesel consumption is 10.180e-3 metric tons of CO2 per gallon of diesel [13].

3.2 Result

Graphing CO2 emission, gallons of gasoline consumed, diesel fuel consumed and electricity consumed, there are some notable trends to point out in Figure 2. There is a visible trend where as consumption of electricity, diesel fuel, and gasoline increases, so does the overall CO2 emission. However despite the increasing pattern in the consumption of these variables, there is still a overall decrease in CO2 emission. This is because of increased efforts in Carbon sequestration. The decrease in CO2 emission despite the soar in consumption of the largest CO2 contributors is due to alarming observations made by our carbon footprint. For example, for simply one acre of forest preserved from conversion to cropland, that is 148.26 metric tons of CO2 removed [14]. As long as our consumption of positive CO2 emission variables increase, if the United States were to increase their carbon sequestration as well, CO2 emission would be at a "net zero." For example, from the calculated CO2 emission contributions: electricity consumption and gallons of gasoline were the largest factors in CO2 emission. However with the long list of other variables emitting CO2, these variables are somewhat balanced by carbon sequestration efforts. That is why in Figure 2, we see trend lines of CO2 emitting variables increasing with positive slopes. Although despite this increasing trend, CO2 emission from 2009 to 2018 decreased with a negative value for the slope (or the rate of CO2 emission change per year). If carbon sequestration efforts were ever to surpass the emission of CO2, then the United States would continue to see CO2 emissions decreasing per year.

4 Autoregressive Model

All the approximation models we have learned so far have the output variable depends on a single independent variable (e.g., linear or polynomial approximation), or a set of variables (e.g., multi linear regression). These techniques often provides a smooth prediction line that is close enough to the real observed data. They are good in describing data (i.e., interpolation), but might fall short on prediction the future (i.e., extrapolation).

The main idea behind *Autoregressive Model* is to predict the next data point based on past observation. The model is useful for predicting real-world time series data because it can adapt to the unexpected changes in the observation data. Autoregressive model can predicts periodic or irregular data better than tradition linear or polynomial models.

Autoregressive models, however, have some weaknesses inherent from its self-referential nature. First, it is hard of even impossible to get a closed form formula, which could make the model harder to interpret. Still, simple linear

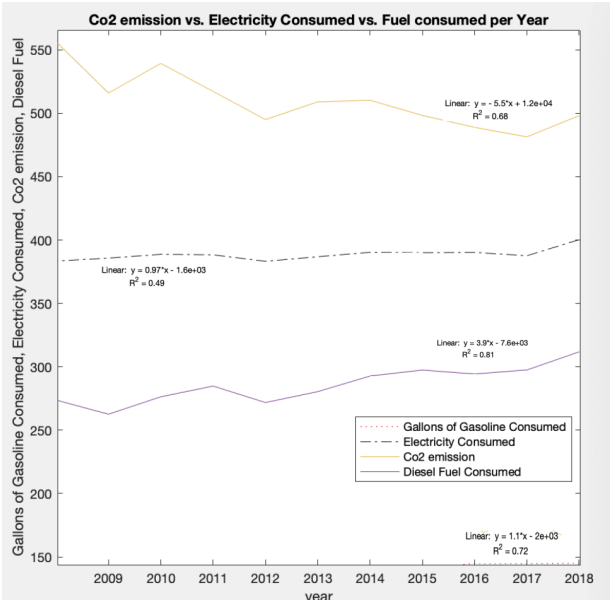


Fig. 2: Linear Regression Model

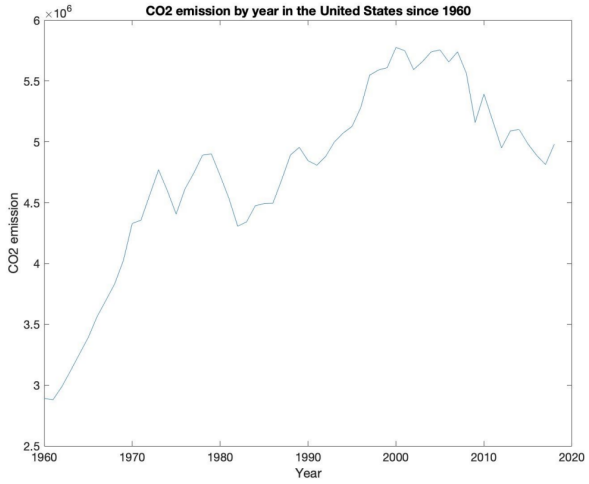


Fig. 3: United States total CO2 emission by year (starting at 1960)

models (i.e., the one shown in Equation 1) could be be solved [15] for a closed form formula and get some useful information out of it. Second, the model can only predicts a few steps into the future before we run out of data. Of course, we can pretend that the predicted values as the observed data and run the model in the self-feedback mode, but the accuracy will suffer. This is a common problem with all models, but especially problematic with autoregressive model because it only looks into n past observation and might miss the overall trend if the data is cut off at a bad spot. We demonstrate this issue more clearly in Section 4.3.

Despite weaknesses mentioned above, we found that autoregressive model is useful for our use case and it can predict the data with least amount of error. In this section, we will walk through our process and methods of choosing and fitting the autoregressive model to predict the CO2 emission of United States in the next 50 years.

4.1 Methodology

We use the open source data from [16] to construct and test the models. Specifically, we use the total CO2 emission per year by the United States from 1850 to 2020.. We test and compare the models (persistence model, linear autoregressive, and autoregressive integrated (ARI) models) via two methods:

1) Split the data into *training data* (year 1850 to 1950) and *test data* (from 1950 to 2020). We fit the model with only training data and then use the test data to validate the accuracy of the model.
2) We fit the model with all available data (year 1850 to 2020) and calculate the *root mean square error* (RMSE). Naturally, lower RMSE means the model can predicts the data better (i.e., close to the observed values).

We use Python open source libraries (NumPy, Pandas, Matplotlib, and SKLearn) to fit the model as well as visualize the data.

4.2 Persistence model

To establish a reference to compare the accuracy of later models, we use the *persistence model* as the baseline for comparison. The *persistence model* is the simplest autoregressive which would predict the next value equals the previous one, i.e., $f(t) = f(t-1)$.

We presents the persistence model result in Figure 4. The red line (predicted) is shifted one year to the right compared to the blue lien (observed). Persistence model is obviously not a good model because it does not predict the future well (notice the line is flat after 2020). However, it serves well as a baseline to compare with our other models.

4.3 Linear autoregressive model

The simplest one is the linear auto regressive model, i.e.,

$$f(t) = a_0 + a_1 f(t-1) + \cdots + a_n f(t-n) \tag{1}$$

To utilize an autoregressive model, first, we need to do some quick statistical check to verify the autocorrelation property of the raw data. We graph the data in Figure 1 in a *lag plot* as shown in Figure 6. Essentially, it is a scatter plot with the x-axis is $f(t-1)$ and the y-axis is $f(t)$. Visually, we can see a strong positive linear correlation between $f(t-1)$ and $f(t-1)$. That lead us to the idea of a simplest linear regressive model with $n = 1$ (i.e., we only look into only one past observation).

$$f(t) = a_0 + a_1 f(t-1) \tag{2}$$

We do a standard linear regression on the lag plot in Figure 6 and found the linear autoregression relation:

$$f(t) = 34.25402 + 0.99726 \times f(t-1) \tag{3}$$

We presents the result of this simple linear autoregressive model in Figure 5. Visually, it looks similar to the persistence model, except that it predicts the upward trend in emission after 2020. However, the linear autoregressive model performs marginally better than our baseline with

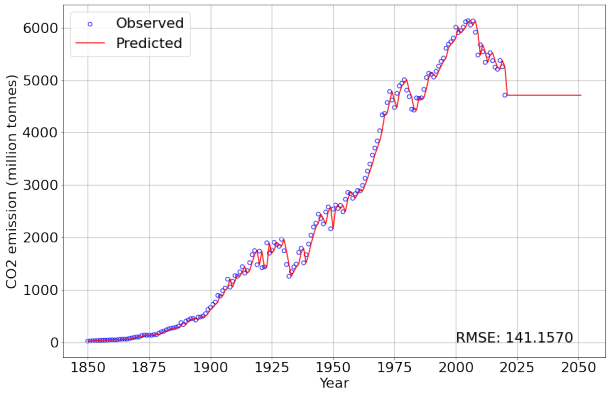


Fig. 4: Persistence model

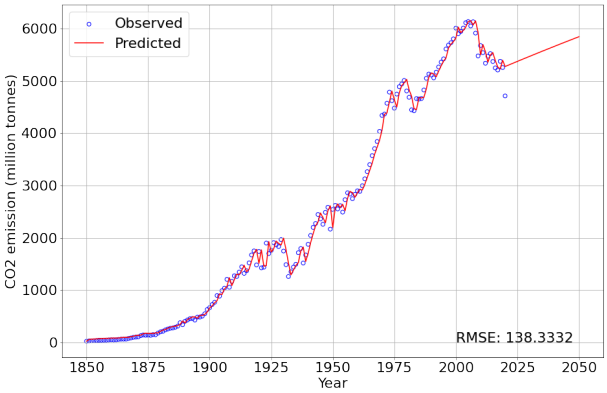


Fig. 5: Linear autoregressive model

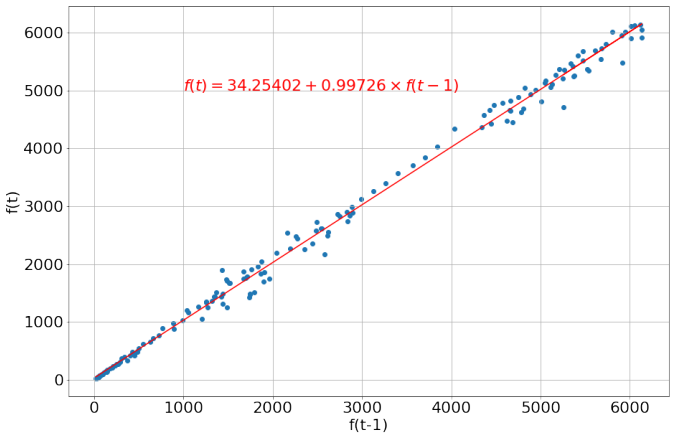


Fig. 6: Lag plot of Figure 1 and the linear regression line

RMSE of 138.3332 compared to RMSE of persistence model, 141.1570.

Residual is the deviations the observations from the predicted value. Another way to analyze the accuracy is using the kernel density estimation (KDE) graph of the residual. The persistence and linear regression models residual KDE is presented in Figure 9a and 9b. The residual density of both models follow approximately a normal distribution center around 0 and spread of around 500.

This simple linear autoregression model has a crtical weakness: the output only depends on the last provided data point. Hence, the model is very sensitive to the last provided data point and could go off course quickly if got bad data. Figure 7 shows the model prediction when the input data is cut off after 1950. It did predicts an upward trend, but way far off from the observed values. The RMSE in this experiment is 1058.0810.

4.4 Autoregressive integrated (ARI) model

Based the idea of the linear autoregressive model presented in Section 4.3, we arrive at a more advanced one, *Autoregressive integrated* (ARI) model. We will use more than one one past observation to eliminate the risk of noisy data points. The *integrated* part means that we works with the differencing of the raw observations. The rationale is that our model the output variable grows over time, hence we use differentiation to make our model stationary, hence easier to work with. We choose the parameters $(p, d) = (15, 2)$ to fit the ARI model. $p = 15$ means that we looks into past 15 observations to make the next year prediction. $d = 2$ means that we use the second-order derivative of the raw data. We choose those parameters because it works best for our data set and computing constraint. We presents the ARI model result in Figure 10, which is much better than the two previous model with RSME of only 73.8004.

The ARI model show can predict well both the upward and downward trend in the data. In Figure 10, when fitting with all available data, the model predicts the continuing of downward trend and the 50% reduction of 2005 emission level goal could be archived by 2036. Also look at the residual KDE graph in Figure 9c, the residual density curve centered around zero and spread of only 200, narrower than the baseline and linear autoregressive model residual density curve.

We also tried to fit ARI model with only data from 1850 to 1950 and let it predicts the following years. The result is presented in Figure 8. The predicted value matches closely with the observed data with RMSE of 622.7556. Of course, the model could not predict a little bump around 1975 or the sudden drop after 2005 due to government policy change, but the model still performs better than expected given insufficient input data.

4.5 Result

We summarizes the result of three models presented in this section via the Table 1. Based on the result, we conclude that the ARI model is the best among three, and it can predicts the future (i.e., do extrapolation) well. Using the ARI model fitting with all data from 1850 to 2020 (Figure 10), we predict that the United State could reach 50% emission level compared to 2005 by 2036 (from 6135 million tones in 2005 to 2984 million tones in 2036).

Model	Fit with data until 1950	Fit with all data
Persistence (4.2)	1733.333	141.157
Linear (4.3)	1058.081	138.333
ARI (4.4)	622.756	73.800

TABLE 1: RMSE of different models presented in Section 4.

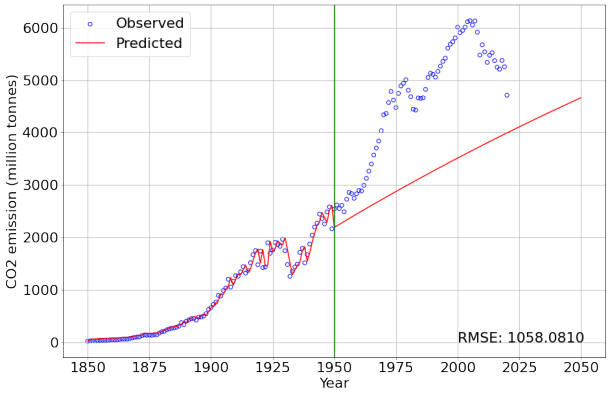


Fig. 7: Linear autoregressive model

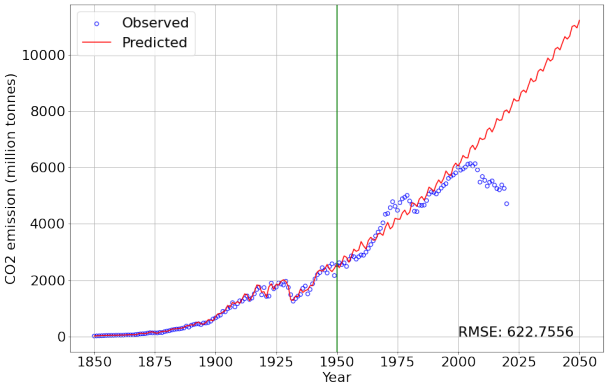
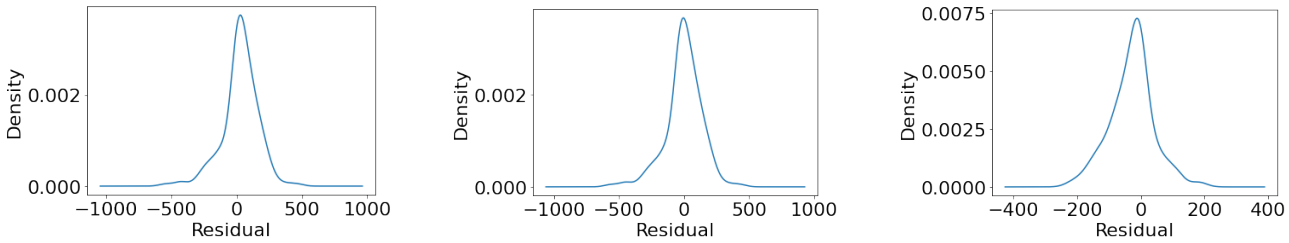


Fig. 8: Autoregressive integrated (ARI) model

Models fitted with data from year 1850 to 1950 and then make prediction for years 1950 to 2050.



(a) Persistence (baseline) (b) Linear autoregressive (c) Autoregressive integrated (ARI)

Fig. 9: Residual kernel density estimation (KDE) of different models fitting with all available data.

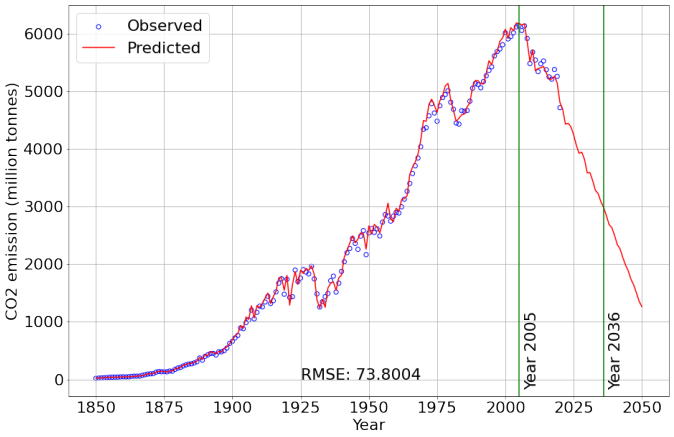


Fig. 10: ARI model results

5 Polynomial Approximation Model

Polynomial least squares approximations are widely used approaches with a higher degree above the linear approximation. The advantage of this approach is to provide approximated functions with very low errors. The general polynomial approximation function with a set of data, $(x_i, y_i)|i = 1, 2, ..., m$,

$$P_n(x) = a_n x^n + a_{n-1} x^{n-1} + ... + a_1 x + a_0, \quad \text{where } n < m - 1 \qquad (4)$$

is determined with optimal degree n minimizing the error. Built polynomial model can predict the outside of the dataset. The complicated processes getting coefficients can be easily performed through Matlab or Python functions.

5.1 Methodology

The simulation is performed with Matlab functions polyfit() and polyval(). Polyfit() returns the coefficients for a polynomial P(x) of degree n. The number of coefficients is n+1. Polynomial curve function $P_n(X)$ with three different degrees are shown in Figure 11.

Polyval() evaluates the polynomial function at each point in x. The approximated CO2 emission level is determined from this function. We set the interval of x from 1945 to 2020 for getting RMSEs, and from 1945 to 2040 for the year prediction of half of the CO2 emission level. The code for polynomial approximation models are available in "Polynomial approximations.mlx".

5.2 Result

Polynomial approximations to predict the year approaching over 50% CO2 reduction also show impressive results. Relatively low degrees such as two and three did not approximate as expected, but five, six, and seven degrees can make better regression models. As seen in Fig 12, the estimated 50% CO2-reduced years are 2037, 2029, and 2027 for degrees five, six, and seven, respectively. The polynomial approximation with degree seven shows the most ideal model with the lowest root means squared error. RMSE for the polynomial approximations is 237.0447, 204.3705, and 202.0975 for degrees five, six, and seven, respectively.

$$P_5(X) = -2.2432 * 10^{-7}x^5 + 0.00212x^4 - 8.03754x^3 + 15205.8224x^2 - 14378813.2413x + 5436923566.0437 \quad (5)$$

$$P_6(X) = -3.4373 * 10^{-9}x^6 + 3.9167 * 10^{-5}x^5 - 0.1859x^4 + 470.5422x^3 - 669733.500852505x^2 + 508277610.988749x - 160688761340.578 \quad (6)$$

$$P_7(x) = -1.5712 * 10^{-11}x^7 + 2.0663 * 10^{-7}x^6 - 0.00116x^5 + 3.6429x^4 - 6836.8787x^3 + 7695937.3254x^2 - 4811002569.511x + 1288470446801.27 \quad (7)$$

Fig. 11: Equations of polynomial approximation functions with five, six, and seven degrees. n indicates degrees

In particular, the difference between predicted years with degrees six and seven and the target year 2030 is a few years. With these results, we can predict 50% reduction in CO2 will be achieved between 2027 and 2037.

6 Conclusion

Persistent effort and interest in overcoming environmental issues are essential with a focus on reducing CO2 emissions being a crucial part of this. The performed analysis of the CO2 emission levels indicates the positive prediction that the emission levels will be decreased around the target year, 2030. Predictions and actual future CO2 emission levels may be able to differ because there are many factors affecting gas pollution. However, the given data is showing the efforts for decreasing CO2 levels in recent years and this supports our analysis. In conclusion, we believe that the goal can be achieved in the light of our findings if more developed businesses in renewable and ecologically technology are accompanied.

References

[1] "Global co2 emissions from transport," Oct 2020. [Online]. Available: https://ourworldindata.org/co2-emissions-from-transport

[2] "Fact sheet: President biden sets 2030 greenhouse gas pollution reduction target aimed at creating good-paying union jobs and securing u.s. leadership on clean energy technologies," Apr 2021. [Online]. Available: https://www.whitehouse.gov/briefing-room/statements-releases/2021/04/22/fact-sheet-president-biden-sets-2030-greenhouse-gas-pollution-reduction-target-aimed-at-creating-good-paying-union-jobs-and-securing-u-s-leadership-on-clean-energy-technologies/

[3] "International energy outlook 2021," Oct 2021. [Online]. Available: https://www.eia.gov/outlooks/ieo/

[4] "Turning the corner: Greenhouse gas emissions are up, but help is on the way," Jan 2021. [Online]. Available: https://ecology.wa.gov/Blog/Posts/January-2021/Turning-the-corner

[5] "Utilization of carbon and other energy gases - geologic research and assessments," Mar 2021. [Online]. Available: https://www.usgs.gov/faqs/what-carbon-sequestration

[6] "Utilization of carbon and other energy gases - geologic research and assessments," Oct 2018. [Online]. Available: https://www.usgs.gov/programs/land-change-science-program/science/landcarbon

[7] "Global co2 emissions have been flat for a decade, new data reveals," Nov 2021. [Online]. Available: https://www.carbonbrief.org/global-co2-emissions-have-been-flat-for-a-decade-new-data-reveals

[8] A. Razmjoo, L. G. Kaigutha, M. V. Rad, M. Marzband, A. Davarpanah, and M. Denai, "A technical analysis investigating energy sustainability utilizing reliable renewable energy sources to reduce co2 emissions in a high potential area," *Renewable Energy*, vol. 164, pp. 46–57, 2021.

[9] D. Marchese, "This eminent scientist says climate activists need to get real." [Online]. Available: https://www.nytimes.com/interactive/2022/04/25/magazine/vaclav-smil-interview.html

[10] U. S. E. P. Agency, "Greenhouse gases equivalencies calculator - calculations and references." [Online]. Available: https://www.epa.gov/energy/greenhouse-gases-equivalencies-calculator-calculations-and-references

[11] ——, "Emissions & generation resource integrated database." [Online]. Available: https://www.epa.gov/egrid

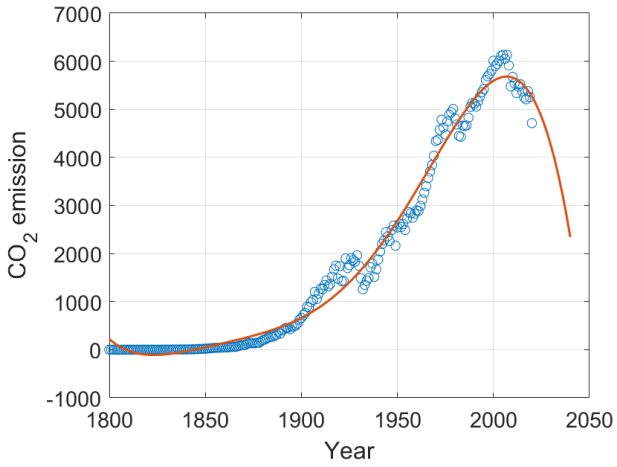
[12] U. E. I. Administration, "U.s. energy consumption fell by a record 7 percent in 2020." [Online]. Available: https://www.eia.gov/todayinenergy/detail.php?id=47397#:~:text=U.S.%20energy%20consumption%20fell%20by%20a%20record%207%25%20in%202020&text=In%202020%2C%20total%20U.S.%20energy,to%20EIA's%20Monthly%20Energy%20Review.

[13] U. S. E. I. Administration, "Gasoline explained." [Online]. Available: https://www.eia.gov/energyexplained/gasoline/use-of-gasoline.php#:~:text=Gasoline%20is%20the%20main%20U.S.,million%20gallons%20of%20aviation%20gasoline.

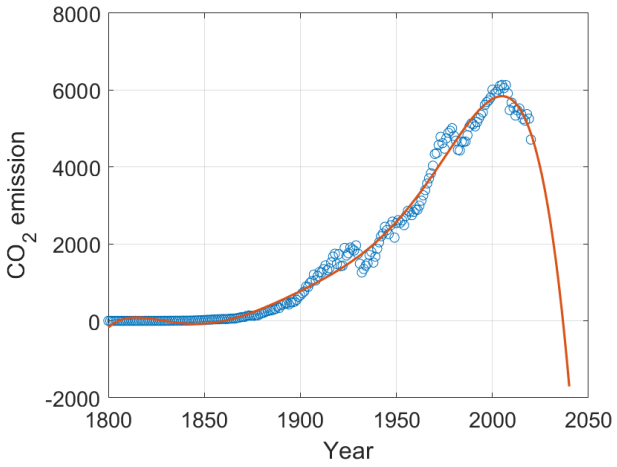
[14] "U.s. greenhouse gas emissions and sinks," 1999-2019. [Online]. Available: https://www.epa.gov/sites/default/files/2021-04/documents/us-ghg-inventory-2021-main-text.pdf

[15] O. Levin, "Discrete mathematicsan open introduction." [Online]. Available: http://discrete.openmathbooks.org/dmoi2/sec_recurrence.html

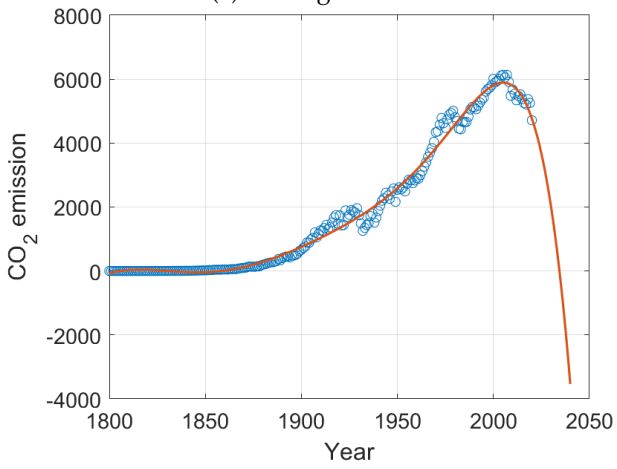
[16] Owid, "Owid/co2-data: Data on co2 and greenhouse gas emissions by our world in data." [Online]. Available: https://github.com/owid/co2-data



(a) the degree is 5



(b) the degree is 6



(c) the degree is 7

Fig. 12: Polynomial approximations with different degrees. Blue scattered circles indicate CO2 emission by year and orange lines are approximated polynomial regression models.