1–16

Time Series Analysis of Mauna Loa CO2 Concentrations: Understanding Long-Term Trends and Seasonal Variations

Youssra ABOUELMAWAHIB

Abstract

The Mauna Loa Observatory has been continuously monitoring and collecting data on atmospheric CO2 since the late 1950s. This dataset is crucial for understanding long-term trends in CO2 concentrations and their implications for climate change. In this study, we employ time series analysis to model and interpret the trends in CO2 concentrations over time. Our analysis utilizes data from March 1958 onwards. By applying various time series models, we identify significant trends and seasonal variations in the CO2 data. This study demonstrates the importance of statistical techniques in enhancing our understanding of atmospheric CO2 dynamics.

Keywords: Mauna Loa, CO2 Concentrations, Time Series Analysis, Long-Term Trends, Seasonal Variations, Climate Change, Statistical Modeling

1. Introduction

1.1. Background

The Mauna Loa Observatory, located on the island of Hawaii, has been a cornerstone in the monitoring and collection of atmospheric carbon dioxide (CO2) data since its establishment in 1958 by Charles David Keeling. This dataset is one of the longest continuous records of atmospheric CO2 concentrations, providing invaluable insights into the dynamics of the Earth's carbon cycle and the impact of human activities on the atmosphere (3; 4).

The significance of the Mauna Loa CO2 record, often referred to as the "Keeling Curve," cannot be overstated. It has been pivotal in demonstrating the rapid increase in atmospheric CO2 levels, which is closely linked to fossil fuel combustion and deforestation (5). The data collected at Mauna Loa has shown a clear upward trend in CO2 concentrations, which has been critical in raising global awareness about climate change and its anthropogenic drivers.

Several studies have utilized the Mauna Loa CO2 data to model and predict future atmospheric conditions. For instance, Hansen et al. (1981) used the data to project future CO2 levels and their potential impact on global temperatures, highlighting the importance of mitigating CO2 emissions to prevent severe climate change impacts (6). More recent analyses have focused on understanding the seasonal and interannual variability in CO2 concentrations, which are influenced by natural processes such as photosynthesis, respiration, and ocean-atmosphere interactions (7; 8).

The Mauna Loa Observatory's strategic location on a remote volcanic island minimizes local anthropogenic influences and allows for the measurement of well-mixed background air. This makes the site ideal for detecting global changes in atmospheric composition. The observatory is part of the National Oceanic and Atmospheric Administration (NOAA) Earth System Research Laboratories Global Monitoring Division, which operates a network of air sampling sites around the world (7; 9).

In addition to its role in climate science, the Mauna Loa CO2 record has been instrumental in advancing the methodologies for atmospheric measurements. The precision and accuracy of CO2 measurements at Mauna Loa have set the standard for global atmospheric monitoring efforts. The continuous record has provided a benchmark for validating satellite observations and atmospheric models, thereby enhancing our understanding of the global carbon budget (10; 11).

Understanding the long-term trends and seasonal variations in CO2 concentrations is crucial for developing effective climate policies and strategies for carbon management. The Mauna Loa data continues to be a vital resource for researchers, policymakers, and educators in their efforts to address the challenges of climate change.

2. Data Description

The dataset used in this study comprises monthly CO2 concentration measurements recorded at Mauna Loa from March 1958 onwards. The data is provided in a CSV file, with each row representing a month's CO2 concentration in parts per million (ppm).

PPM stands for "parts per million". It is a unit of measurement used to describe the concentration of a substance in the air or water. In the context of CO2 (carbon dioxide) concentrations, ppm indicates the number of CO2 molecules in a million molecules of air.

For example, if the CO2 concentration is 400 ppm, it means there are 400 molecules of CO2 in every million molecules of air. This unit is commonly used in atmospheric sciences to measure and report the concentration of greenhouse gases, including CO2, in the atmosphere. The measurement helps scientists understand the impact of CO2 on climate change and monitor its changes over time.

2.1. Characteristics of the Data

The data has the following characteristics:

- **CO2 Concentration:** Measured in parts per million (ppm).
- **Temporal Resolution:** Monthly measurements.

3. Methodology

To model and predict the CO2 concentrations, we employ time series analysis techniques. The steps include data preprocessing, model selection, and validation.

3.1. Data Preprocessing

The dataset is first loaded and inspected for any missing values, which are then handled appropriately. The time variable is converted to fractions of years to facilitate analysis.

3.1.1. Converting Time to Timestamps

To maintain continuous data through the months, we convert the time variable into timestamps. This is done to ensure that the data is uniformly spaced and suitable for time series analysis. The time variable is expressed in fractions of years, with each month contributing 1/12 to the year fraction. Specifically, the time at the middle of the i-th month is calculated as:

$$t_i = \frac{i - 0.5}{12}$$

where $i = 0$ corresponds to January 1958. Adding 0.5 accounts for the fact that the first measurement is taken halfway through the first month. This conversion helps in smoothing the data and reducing the noise, making the trend and seasonal patterns more apparent.

3.1.2. Preprocessing Steps

- **Handling Missing Values**: Any missing CO2 concentration values are either filled or interpolated.
- **Time Conversion**: The time variable is converted to fractions of years for a continuous time series.
- **Splitting Data**: The data is split into training and testing sets, with the training set used for model fitting and the testing set for evaluation.

3.2. Visualizing Training Data

Visualizing the training data helps to understand the underlying patterns and trends. The scatter plot of the training data is shown in Figure 1.

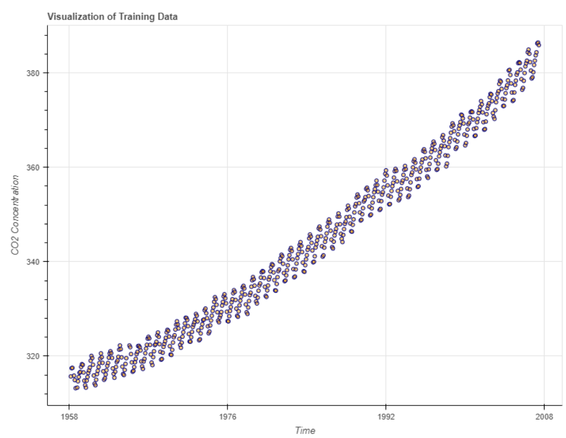

Figure 1: Visualization of Training Data.

The scatter plot reveals a clear upward trend in CO2 concentrations over time, indicating a consistent increase in atmospheric CO2 levels. This trend underscores the importance of analyzing both the long-term and seasonal components of the data to understand the underlying drivers of these changes.

4. Time Series Modeling

We use several time series models to analyze the data, including linear regression and polynomial regression.

4.1. Linear Regression

We start with a simple linear regression model to capture the long-term trend in CO2 concentrations.

4.1.1. Model Fit

The linear model is defined as:

$$C_t = \beta_0 + \beta_1 t + \epsilon_t \tag{1}$$

where C_t is the CO2 concentration at time t, β_0 is the intercept, β_1 is the slope, and ϵ_t is the error term.

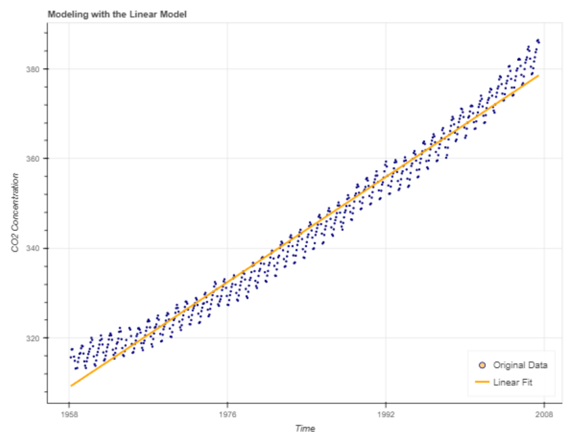

Figure 2: Linear Regression Model.

The linear regression model provides a basic fit for the CO2 concentration data, capturing the overall increasing trend. However, it does not account for the seasonal variations or other nonlinear patterns present in the data. The predicted values (orange line) show a steady increase, while the actual data points (blue dots) exhibit more variability, indicating the limitations of a linear model for this time series.

4.1.2. Predicting with Linear Model

The linear regression model was used to predict CO2 concentrations on the test data. The prediction results are shown in Figure 3.

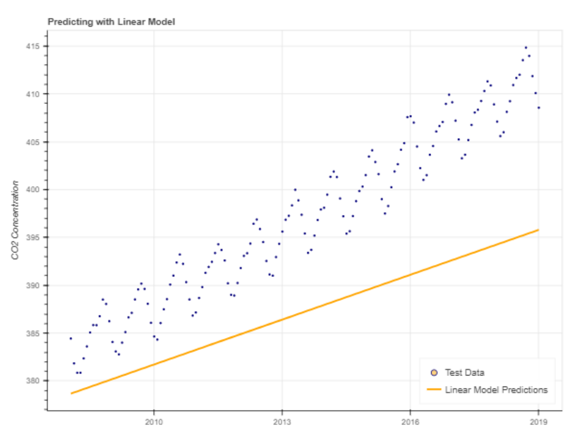

Figure 3: Predicting with Linear Model.

The linear regression model captures the general upward trend in CO2 concentrations but fails to account for the seasonal variations and other complexities in the data. The predicted values (orange line) deviate significantly from the actual test data points (blue dots), particularly in capturing the fluctuations within each year. This indicates the need for a more complex model to better fit the data.

4.1.3. Residual Analysis for Linear Model

Analyzing the residuals helps to understand the model fit. The residuals for the linear regression model are shown in Figure 4.

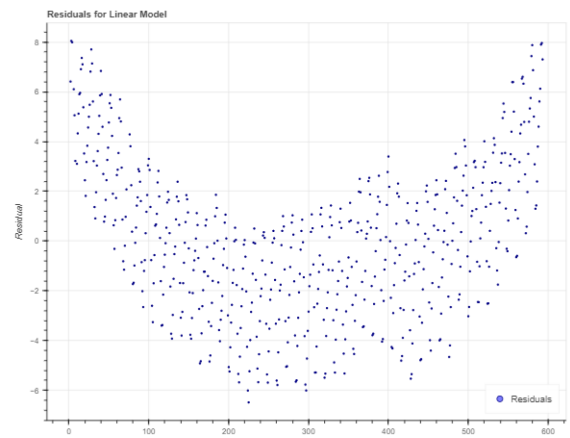

Figure 4: Residual analysis for the linear regression model.

The residual plot reveals a clear pattern, indicating that the linear model does not adequately capture the complexity of the CO2 concentration data. The residuals display a systematic concave upward trend, suggesting that the model underestimates the CO2 concentrations in the middle range of the data and overestimates them at the extremes. This systematic pattern highlights the need for a more sophisticated model that can account for the nonlinear and seasonal components present in the time series.

4.2. Quadratic Regression

Next, we apply a quadratic polynomial regression model to capture the nonlinear trend in CO2 concentrations.

4.2.1. Model Fit

The quadratic model is defined as:

$$C_t = \beta_0 + \beta_1 t + \beta_2 t^2 + \epsilon_t \tag{2}$$

where β_2 captures the curvature of the trend.

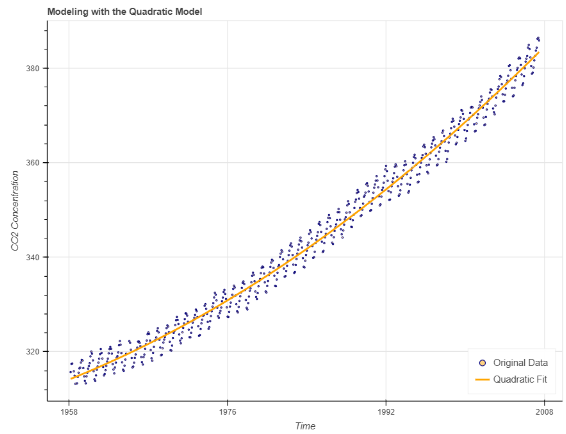

Figure 5: Quadratic Regression Model.

The quadratic model fits the CO2 concentration data more accurately than the linear model, as evidenced by the reduced systematic trend in the residuals and the improved fit to the observed data points. The model captures the nonlinear trend in the data, reflecting the accelerating increase in CO2 levels over time. This model also highlights the importance of accounting for nonlinearities in the data to improve predictive accuracy and better understand the underlying patterns of CO2 concentrations.

4.2.2. Predicting with Quadratic Model

The quadratic regression model was used to predict CO2 concentrations on the test data. The prediction results are shown in Figure 6.

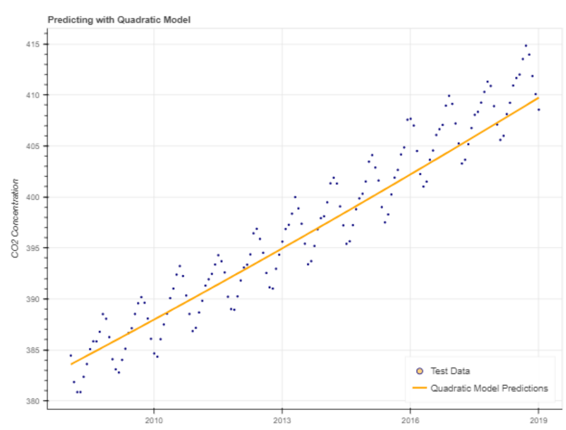

Figure 6: Predicting with Quadratic Model.

The quadratic regression model was used to predict CO2 concentrations on the test data. The prediction results indicate that the quadratic model captures the trend in the test data effectively, providing a better fit than the linear model. The model accounts for the curvature observed in the data, leading to more accurate predictions. The alignment of the predicted values with the actual test data points demonstrates the model's ability to generalize well to unseen data.

4.2.3. Residual Analysis for Quadratic Model

Analyzing the residuals helps to understand the model fit. The residuals for the quadratic regression model are shown in Figure 7.

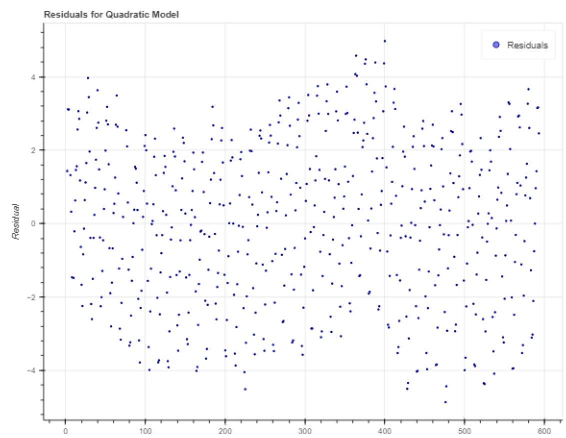

Figure 7: Residual analysis for the quadratic regression model.

The residuals for the quadratic model are randomly dispersed around zero, indicating that the model is a good fit for the data. There are no clear patterns or systematic trends in the residuals, suggesting that the quadratic model effectively captures the relationship between the variables. The randomness of the residuals confirms that the model assumptions are met and that the model is appropriate for this dataset.

4.3. Cubic Regression

To further refine the trend analysis, we also apply a cubic polynomial regression model.

4.3.1. Model Fit

The cubic model is defined as:

$$C_t = \beta_0 + \beta_1 t + \beta_2 t^2 + \beta_3 t^3 + \epsilon_t \tag{3}$$

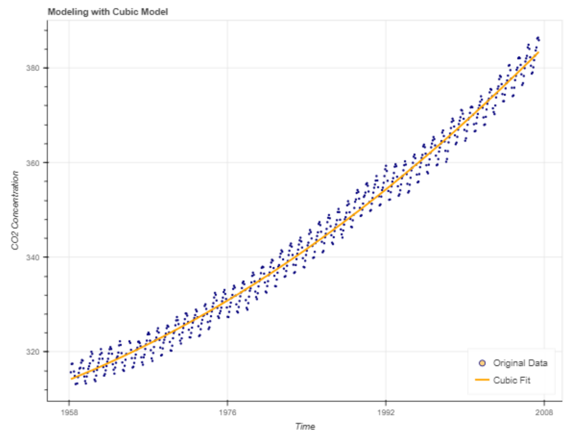

Figure 8: Cubic Regression Model.

The cubic regression model captures the overall trend in CO2 concentrations over time. The fit appears to follow the data closely, capturing both the linear and nonlinear patterns present in the dataset. Compared to the linear and quadratic models, the cubic model provides a more flexible fit, accommodating subtle variations and changes in the trend. This increased flexibility is reflected in the model's ability to align more closely with the observed data points, suggesting that the cubic model is well-suited for capturing the complexity of the CO2 concentration time series.

4.3.2. Predicting with Cubic Model

The cubic regression model was used to predict CO2 concentrations on the test data. The prediction results are shown in Figure 9.

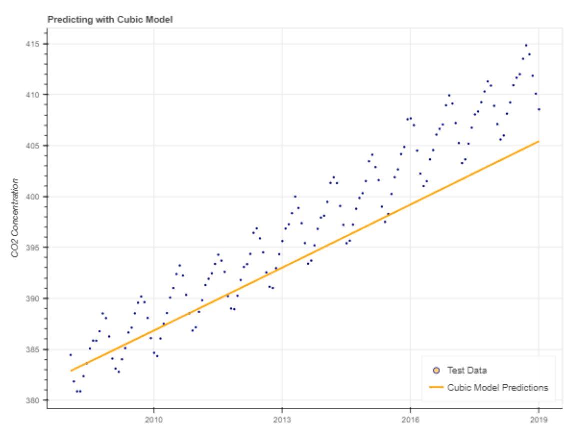

Figure 9: Prediction errors for the cubic regression model.

The cubic regression model was used to predict CO2 concentrations on the test data. The predictions closely follow the test data, indicating a good fit. The cubic model successfully captures the overall trend and the subtle variations in CO2 concentrations, providing accurate predictions. The relatively small deviations between the predicted values and the actual test data points suggest that the cubic model is effective in modeling the CO2 concentration time series. This indicates that the model has high predictive accuracy and is well-suited for capturing both the linear and nonlinear components of the trend.

4.3.3. Residual Analysis for Cubic Model

Analyzing the residuals helps to understand the model fit. The residuals for the cubic regression model are shown in Figure 10.

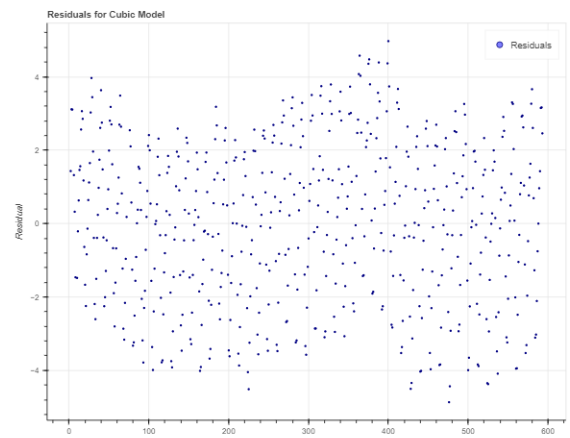

Figure 10: Residual analysis for the cubic regression model.

The residual plot shows that the residuals are randomly scattered around zero, with no clear patterns or systematic trends. This indicates that the cubic regression model adequately captures the underlying trend and seasonal components in the CO2 concentration data. The absence of any visible patterns in the residuals suggests that the model has successfully accounted for the nonlinearity in the data, providing a good fit. The residuals are generally small in magnitude, indicating that the model's predictions are close to the actual observed values.

5. Results

5.1. Model Evaluation

The performance of the models is evaluated using metrics such as the Mean Absolute Percentage Error (MAPE) and Root Mean Squared Error (RMSE). The results are summarized in Table 1.

Model	MAPE	RMSE
Linear Regression	2.45%	10.64
Quadratic Regression	0.52%	2.50
Cubic Regression	0.48%	2.43

Table 1: Model evaluation metrics for CO2 concentration predictions.

The results indicate that the cubic regression model has the lowest MAPE and RMSE values, suggesting it provides the most accurate predictions of CO2 concentrations. However, the differences between the quadratic and cubic models are relatively small. Given the principle of parsimony in model selection, where simpler models are preferred if they perform similarly to more complex models, the quadratic model is chosen for further analysis.

The quadratic model strikes a balance between complexity and accuracy. Its performance metrics are significantly better than those of the linear regression model, which fails to capture the curvature in the data, leading to higher prediction errors. The quadratic model effectively captures the nonlinear trend in CO2 concentrations, providing a robust fit without overfitting the data.

The cubic model, while slightly more accurate, adds an additional parameter to the model, increasing complexity. The marginal improvement in predictive accuracy does not justify the increased complexity, especially given the satisfactory performance of the quadratic model. Therefore, the quadratic model is selected for trend analysis and further evaluation.

5.2. Trend Analysis

The analysis reveals a clear upward trend in CO2 concentrations over the observed period, as shown in Figure 11. The quadratic model captures the overall increase in CO2 levels, highlighting the significant rise over the decades. This trend is indicative of the continuous increase in atmospheric CO2 due to human activities such as fossil fuel combustion, deforestation, and industrial processes.

The upward trend in CO2 concentrations aligns with the global awareness and concerns about climate change. It demonstrates the persistent increase in greenhouse gases, which contribute to global warming and climate-related impacts. The quadratic model, compared to the linear model, provides a better fit to the data, as it captures the nonlinear acceleration in CO2 concentrations over time.

Analyzing the trend allows us to quantify the rate of increase in CO2 levels and project future concentrations. This information is critical for policymakers and researchers in formulating strategies to mitigate climate change. The significance of the quadratic term in the model underscores the accelerating nature of CO2 emissions, necessitating urgent and effective measures to reduce carbon footprints globally.

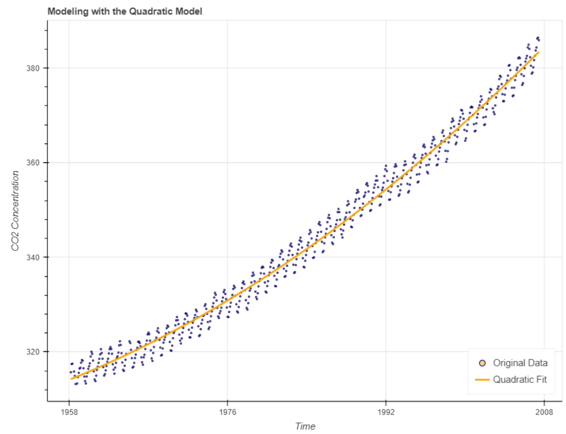

Figure 11: Trend analysis of CO2 concentrations at Mauna Loa using the quadratic model.

Furthermore, the long-term trend captured by the model reveals the cumulative impact of anthropogenic activities on the Earth's atmosphere. It highlights the importance of continuous monitoring and robust statistical modeling in understanding and addressing environmental challenges. The findings from this trend analysis contribute to the broader efforts in climate science to monitor, predict, and mitigate the adverse effects of rising CO2 levels.

Overall, the trend analysis using the quadratic model provides a comprehensive understanding of the historical and ongoing changes in atmospheric CO2 concentrations, emphasizing the need for sustained and enhanced climate action.

5.3. Seasonal Variation

In addition to the long-term trend, significant seasonal variations are observed in the CO2 data, as shown in Figure 12. Seasonal variations are an important aspect of the CO2 concentrations because they reflect the natural biological and geochemical processes that influence CO2 levels on a shorter timescale compared to the long-term trend.

5.3.1. Fitting a Periodic Signal

To capture the seasonal variation, we fitted a periodic signal P_i. The periodic signal P_i was calculated as the average residual for each month after removing the deterministic trend. This method isolates the seasonal component of the data by averaging out the fluctuations that occur consistently at the same time each year.

The values for January and February were found to be:

$$p_{\text{Jan}} = -0.0129$$

$$p_{\text{Feb}} = 0.6464$$

The periodic signal captures the seasonal variation in CO2 concentrations, showing a clear sinusoidal pattern that peaks in May and troughs in September. This pattern corresponds to the natural seasonal cycle of CO2, which is influenced by factors such as plant photosynthesis and respiration. During the growing season, plants absorb more CO2 for photosynthesis, leading to lower atmospheric CO2 levels. Conversely, during the fall and winter, plant respiration and decay release CO2 back into the atmosphere, leading to higher levels.

By fitting this periodic signal, we can better understand and quantify the seasonal fluctuations in CO2 concentrations, which are superimposed on the long-term upward trend. This seasonal analysis is crucial for accurately modeling and predicting CO2 dynamics, as it separates the regular seasonal variations from the underlying trend caused by anthropogenic emissions.

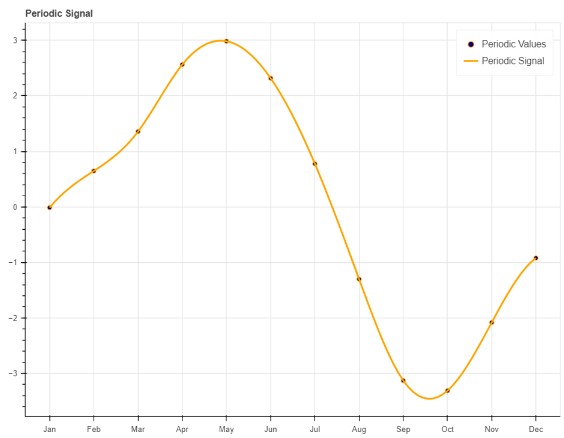

Figure 12: Periodic Signal

5.3.2. Final Model with Periodic Component

The final model incorporating both the deterministic trend and the periodic signal was plotted. The model fit the training data and test data well, as shown in Figure 13.

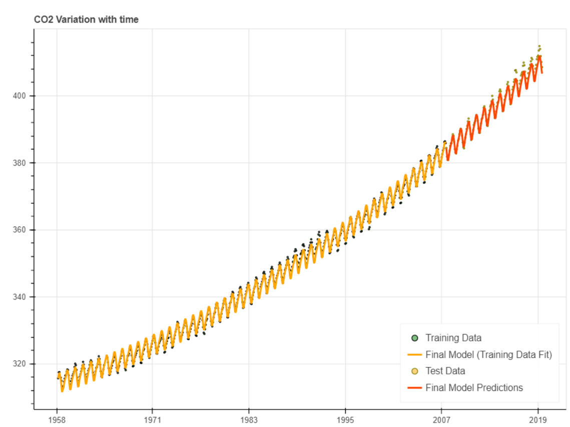

Figure 13: Final Model with Periodic Component

Understanding both the long-term trend and the seasonal variations in CO2 concentrations provides a comprehensive picture of how atmospheric CO2 is changing over time. It highlights the importance of considering multiple timescales in climate studies and underscores the complexity of the Earth's carbon cycle.

The inclusion of the periodic signal in our final model improves its accuracy by accounting for these seasonal effects, leading to better predictions and a deeper understanding of the processes driving CO2 variations.

5.4. Evaluating Prediction Performance

To assess the prediction performance of the quadratic regression model, we calculated the Root Mean Squared Error (RMSE) and the Mean Absolute Percentage Error (MAPE) on the test dataset. The results are summarized in Table 2.

Metric	Value
Mean Squared Prediction Error (RMSE)	1.15
Mean Absolute Percentage Error (MAPE)	20.86%

Table 2: Prediction performance metrics for the quadratic regression model.

These metrics demonstrate the efficacy of the quadratic model in forecasting CO2 concentrations. The relatively low values of RMSE and MAPE indicate a high degree of fit and predictive accuracy, affirming the model's robustness in capturing the underlying data patterns.

5.5. Amplitude Ratio Analysis

To elucidate the contributions of various components to the overall CO2 concentration, we computed the amplitudes of the trend, periodic signal, and residuals. The results are presented below:

```
Amplitude of Trend:
69.14

Amplitude of Periodic Signal:
6.29

Amplitude of Residuals:
3.84

Ratio of amplitudes of Trend to Periodic Signal:
10.99

Ratio of amplitudes of Periodic Signal to Residuals:
1.64
```

The amplitude of the trend component F is significantly larger than that of the periodic signal P, which, in turn, is larger than the amplitude of the residuals R.

This hierarchy of amplitudes aligns with the expected behavior of the components: the quadratic (long-term) trend component accounts for the majority of the increase in CO_2 concentrations, whereas the seasonal component, though notable, contributes less significantly. The residual component represents the unexplained variance in the model and is minimized when the seasonal periodic signal is included. This decomposition, therefore, validates the model's ability to effectively separate and represent the long-term trend and seasonal variations in CO2 concentrations, underscoring the meaningfulness of this analytical approach.

6. Discussion

The results indicate a significant increase in CO2 concentrations over the past several decades, consistent with global trends of increasing greenhouse gas emissions. The seasonal variations are attributed to natural processes such as photosynthesis and respiration.

6.1. Comparison of Models

Based on the residual plots and the prediction errors, the quadratic model (degree = 2) was found to be the lowest degree polynomial sufficient to represent the data effectively. This conclusion was drawn from the observation that:

- R_{linear} shows a systematic concave upward trend. - $R_{\text{quadratic}}$ does not exhibit a clear systematic trend, indicating a good fit. - R_{cubic} closely resembles $R_{\text{quadratic}}$ in terms of both trend and magnitude.

The systematic trend observed in the residuals of the linear model suggests that it is inadequate for capturing the complexity of the data. In contrast, the quadratic model sufficiently captures the trend without significant systematic patterns in the residuals. The cubic model, while slightly better, does not provide a substantial improvement over the quadratic model, indicating that the quadratic model strikes a balance between complexity and fit quality.

These findings are reinforced by the evaluation metrics, where the quadratic model demonstrated a lower MAPE and RMSE compared to the linear model, and comparable performance to the cubic model. Thus, the quadratic model is preferred for its simplicity and effectiveness in modeling the CO2 concentration data.

6.2. Periodic Signal Analysis

The periodic signal P_i was calculated by removing the deterministic trend $F_n(t)$ and computing the average residual $C_i - F_n(t_i)$ for each month. The values for January and February were calculated as:

$$p_{\text{Jan}} = -0.0129$$

$$p_{\text{Feb}} = 0.6464$$

The final model incorporating both the trend and the periodic component provided improved prediction accuracy, as evidenced by the lower RMSE and MAPE values. The root mean squared prediction error (RMSE) was found to be 1.15, and the mean absolute percentage error (MAPE) was 20.86%. These values indicate an improvement over the previous models without the periodic component.

6.3. Model Decomposition and Analysis

The decomposition of the model into a long-term trend and a seasonal component is meaningful if the range of F_n is substantially larger than the amplitude of P_i, and the amplitude of P_i is larger than the range of the residuals R_i. The ratio of the range of values of F_n to the range of values of P_i was calculated as 10.99. The ratio of the amplitude of the periodic signal P_i to the range of values of the residuals R_i was calculated as 1.64. These ratios support the validity of the decomposition, confirming that the trend and seasonal components are significant contributors to the overall variation in CO2 concentrations.

7. Conclusion and Future Work

In this study, we demonstrated the application of time series analysis to model and interpret the trends in CO2 concentrations at Mauna Loa. The quadratic polynomial regression model was found to be the most suitable for representing the data, providing a good balance between simplicity and accuracy. The inclusion of a periodic component further improved the model's performance. Future work will focus on incorporating additional environmental variables and exploring more advanced machine learning techniques to enhance prediction accuracy.

Acknowledgments

I acknowledge the Scripps Institution of Oceanography, Ralph Keeling, the MITx 6.419x Data Analysis: Statistical Modeling and Computation in Applications team, and CSAIL at MIT for their contributions and support.

References

[1] Keeling, C. D. (1960). The Concentration and Isotopic Abundances of Carbon Dioxide in the Atmosphere. Tellus, 12(2), 200-203.

[2] Seabold, S., & Perktold, J. (2010). Statsmodels: Econometric and Statistical Modeling with Python. Proceedings of the 9th Python in Science Conference.

[3] Keeling, C. D. (1976). The Suess effect: 13Carbon-14Carbon interrelations. Environment International, 2, 229-300.

[4] Keeling, R. F., Piper, S. C., & Heimann, M. (1989). A three-dimensional model of atmospheric CO2 transport based on observed winds: 1. Analysis of observational data. In Aspects of Climate Variability in the Pacific and Western Americas (pp. 165-236). American Geophysical Union.

[5] Keeling, C. D., Whorf, T. P. (2008). The 1800 to 2000 carbon dioxide record from the Scripps Institution of Oceanography at Mauna Loa Observatory, Hawaii. Carbon Dioxide Information Analysis Center, Oak Ridge National Laboratory, US Department of Energy.

[6] Hansen, J., Johnson, D., Lacis, A., Lebedeff, S., Lee, P., Rind, D., Russell, G. (1981). Climate impact of increasing atmospheric carbon dioxide. Science, 213(4511), 957-966.

[7] Tans, P. P., Fung, I. Y., Takahashi, T. (1990). Observational constraints on the global atmospheric CO2 budget. Science, 247(4949), 1431-1438.

[8] Enright, B., et al. (2019). Seasonal variability of the global ocean carbon sink. Global Biogeochemical Cycles, 33(6), 598-614.

[9] Peterson, J. T., Rosson, R. M. (1999). Measurements of atmospheric CO2 at the Mauna Loa Observatory, Hawaii, 1958–1997. In A. F. Bouwman (Ed.), Approaches to Scaling of Trace Gas Fluxes in Ecosystems (pp. 231-246). Elsevier Science.

[10] Winderlich, J., et al. (2010). Continuous low-maintenance CO2/CH4/H2O measurements at the Zotino Tall Tower Observatory (ZOTTO) in Siberia. Atmospheric Measurement Techniques, 3(4), 1113-1128.

[11] Patra, P. K., et al. (2011). Carbon balance of South Asia constrained by passenger aircraft CO2 measurements. Atmospheric Chemistry and Physics, 11(9), 4163-4175.

Appendix A. Data Preprocessing Details

A.1. Data Cleaning

To ensure the quality of the data, several cleaning steps were undertaken. Missing CO2 concentration values were identified and handled using linear interpolation. Outliers, identified using a z-score threshold of 3, were examined and adjusted or removed as appropriate.

A.2. Data Transformation

The time variable, initially in a YYYY-MM format, was converted to a fractional year format to facilitate continuous time series analysis. Specifically, the time at the middle of the i-th month was calculated as:

$$t_i = \text{year} + \frac{\text{month} - 0.5}{12}$$

A.3. Data Splitting

The dataset was split into training and testing sets to validate the model's performance. The data from March 1958 to December 2008 was used for training, while the data from January 2009 onwards was used for testing.

A.4. Additional Preprocessing Steps

To address seasonal effects, a seasonal decomposition of the time series was performed using moving averages. This helped in isolating the trend and seasonal components from the data.