# W271 Lab 2: CO2 Present

## Contents



### 0.1 (3 points) Task 0b: Introduction

In our 1997 report, we developed a linear models and an ARIMA model to forecast $CO_2$ through 2020 and beyond. Since then the $CO_2$ level increased from 360 ppmv in December 1997 to 420 ppmv in February 2023. The 420 ppmv level was at the higher end of our forecast range (385-424 ppmv) with 95% confidence. In this follow-up study, we will compare our 1997 forecast to the actuals, and further enhance the model performance.

The original data set was a monthly series and the new data set has both monthly and weekly frequency. Also, due to the eruption of the Mauna Loa Volcano, measurements from Mauna Loa Observatory were suspended as of 11/29/2022. Observations starting in December 2022 are from a site at the Maunakea Observatories, ~21 miles north of the original site. We believe the site change did not pose a significant impact on the data generation process.

### 0.2 (3 points) Task 1b: Create a modern data pipeline for Mona Loa CO2 data.

We sourced the weekly and monthly data set from the United States' National Oceanic and Atmospheric Administration data page [here]. We noted 18 observations with -999.99 value (missing values) in the weekly dataset. We replaced the 18 invalid weekly values with their corresponding month's value in the monthly data set. Then we generated two clean time series data sets from May 1974 to March 2023 (one weekly and one monthly time series).

Similar to our 1997 study, we explored the updated time series, see the charts in Figure 1. We continue to observe an increasing $CO_2$ trend and seasonal variability. However, the trend line seems to steepen after mid-2000s. The annual growth rates are mostly in the range of 0 to 1%, with a modest upward trend (also observed in 1997). We noticed that the long run average of the annual growth rates is 0.495%, higher than the prior long run average of annual growth rate of 0.37% in the 1997 data set. This further confirms that $CO_2$ levels increased faster in the recent years. The Decomposition graph continued to show an upward trend, strong seasonality, and the irregular effect.

We also noted that the peak $CO_2$ month is April (instead of May in the 1997 report) and the lowest month is September (instead of October). The seasonal difference between the peak and trough month is (6.67 ppmv), higher than the 5.46 ppmv seasonal difference observed in the original data.

### 0.3 (1 point) Task 2b: Compare linear model forecasts against realized CO2

In our 1997 report, we developed the two linear models: a linear model using time as the single variable, and a polynomial model using the third polynomial degree of trend and seasonality. We used these models to

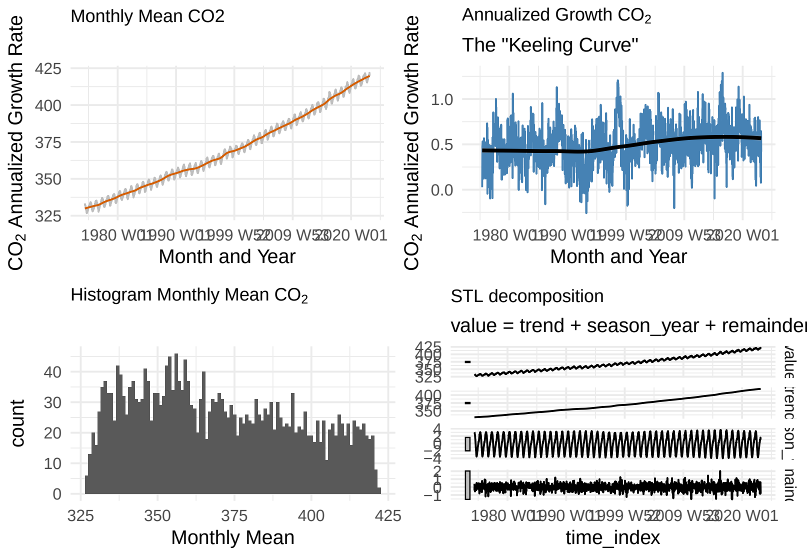


Figure 1: Atmospheric CO2 Level Time Series Overview

forecast the $CO_2$ level through 2020 and beyond. In comparison of the forecast results to the actuals through the present period (February 2023), we noted that both models underestimated the $CO_2$ levels. In a relative term, the linear model outperformed the polynomial model in the long run, which indicated the over-fitting issues of the third degree polynomial model. See the comparison in right char in Figure 2.

The actual level $CO_2$ was 420 ppmv in February 2023. The linear model projected a straight trend line (without any seasonal effect) and forecasted $CO_2$ to reach 395 ppmv in February 2023. The polynomial model captured the seasonal effect and projected increasing levels until the mid-2021 and then flattened and shifted downward. The polynomial model forecasted 384 ppmv in February 2023.

## 0.4 (1 point) Task 3b: Compare ARIMA models forecasts against realized CO2

In our 1997 report, we also fitted an ARIMA model (ARIMA(0,1,1)(1,1,2)[12]). This model captured the seasonal effect and projected well until 2005 but under-forecasted after 2005. The ARIMA model forecasted 405 ppmv vs. the actual level of 420 ppmv in February 2023. However, the ARIMA model performs much better than the two linear models. See the comparison in the left chart in Figure 2.

## 0.5 (3 points) Task 4b: Evaluate the performance of 1997 linear and ARIMA models

In our 1997 report, our ARIMA model projected the first time that CO2 would cross 420 ppm would be in May 2031. Sadly we already crossed this level in April 2021, 10 years earlier than our projection.

To quantify the model projection errors/biases, we calculated the RMSE of the models. Among the three models, the ARIMA model has the lowest RSME (8.2), in comparison to the linear model (14.1) and the polynomial model (17.6). See the comparison in the right chart in Figure 2.

## 0.6 (4 points) Task 5b: Train best models on present data

While the 1997 ARIMA model outperformed the linear models, it still under-predicted actual $CO_2$ levels after 2005. We decided to refit this model using the actuals to capture the accelerated $CO_2$ increase. In addition, we also seasonally adjusted the weekly series, and split both seasonally-adjusted (SA) and non-seasonally-adjusted (NSA) series into training and test sets, using the last two years of observations as the test sets.

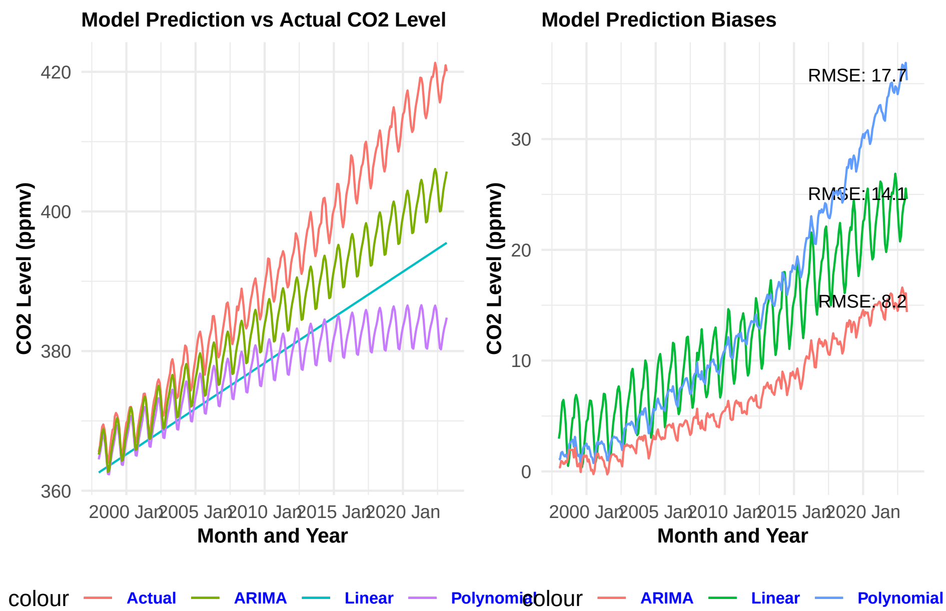


Figure 2: Model Prediction vs Actual (Left) and Model Prediction Biases (Right)

In our 1997 model fit process, we discussed the rationale of using BIC as the goodness-of-fit assessment criteria to choose the best model fit (Task 3a). We believe this process is still appropriate for the SA and NSA data series. We will choose the best ARIMA models with the lowest BIC score.

Using this model optimization process, we selected the best ARIMA model for the weekly NSA data series ("ARIMA Non-Seasonal Model") as ARIMA(0,1,1)(2,1,0)[52]. This model has seasonal parameters. The best ARIMA model for the weekly SA series ("ARIMA Seasonal Model") is ARIMA(1,1,1) with drift, which does not have seasonal parameters given the data was seasonally adjusted. All coefficients for the ARIMA models are significant, see the model results below. We also fitted a third degree polynomial model to the SA data series ("Polynomial Seasonal").

```
## Series: value
## Model: ARIMA(0,1,2)(2,1,0)[52]
##
## Coefficients:
##           ma1      ma2     sar1     sar2
##       -0.6269  -0.1664  -0.6639  -0.3297
## s.e.   0.0200   0.0202   0.0199   0.0201
##
## sigma^2 estimated as 0.2198:  log likelihood=-1538.46
## AIC=3086.92   AICc=3086.95   BIC=3115.69
```

```
## Series: seasonal_adj_value
## Model: ARIMA(1,1,1) w/ drift
##
## Coefficients:
##          ar1      ma1  constant
##       0.1990  -0.8309    0.0279
## s.e.  0.0255   0.0136    0.0013
##
## sigma^2 estimated as 0.1312:  log likelihood=-958.38
## AIC=1924.77   AICc=1924.78   BIC=1947.87
```

We performed diagnostic analysis of the model residuals to check for residuals stationary and any assumption violations. The two ARIMA models have stationary residuals and a normal distribution of the residuals, with some significant lags on the ACF plot. The Polynomial Seasonal model residuals are non-stationary and have significant and decaying autocorrelation lags, which is expected as we concluded above. The polynomial model is not a great fit to our data, and not suitable to capture the long term trend of $CO_2$.

Once fitted, We compared the model residuals for both the in-sample and out-sample periods for the three models, shown in Figure 3. Visually, the polynomial model has higher errors than the two ARIMA models for in-sample and out-sample periods.

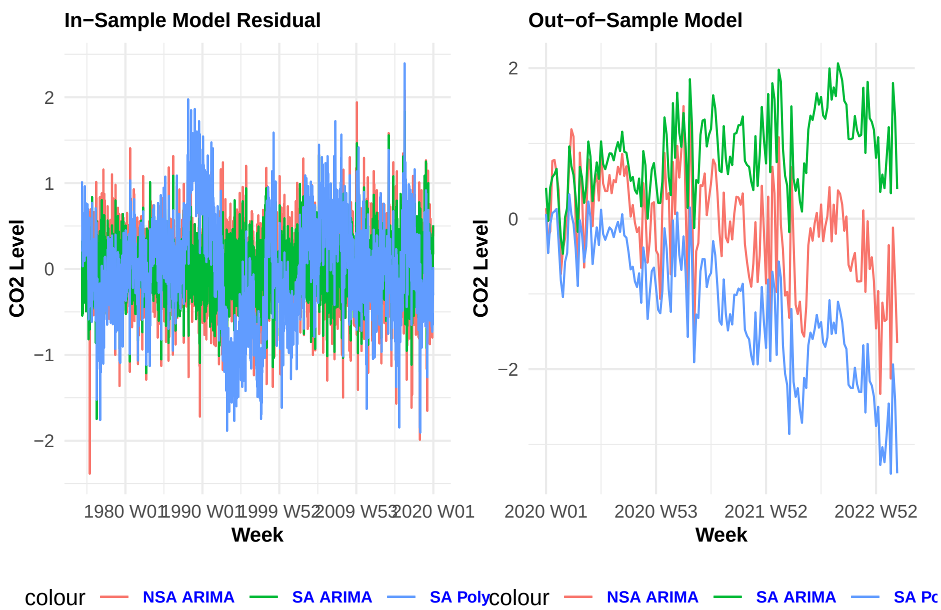


Figure 3: In-Sample Model Residuals (Left) Out-Sample Model Residuals (Right)

We calculated each model's RSME for in-sample and out-of-sample periods, see the table below. The ARIMA Seasonally Adjusted model has the lowest RSME for the in-sample periods, and the ARIMA Non-Seasonally Adjusted model has the lowest RSME for the out-of-sample periods. Out-of-sample period performance is more important, given the forecasting power. We concluded that the ARIMA Non-Seasonally Adjusted model is the best model for this data set.

| model | RMSE_In_Sample | RMSE_Out_Sample |
|---|---|---|
| ARIMA Non-Seasonal | 0.463 | 0.715 |
| ARIMA Seasonal | 0.362 | 1.033 |
| Polynomial Seasonal | 0.631 | 1.438 |

## 0.7 (3 points) Task Part 6b: How bad could it get?

We used the ARIMA Non-Seasonal model to predict the $CO_2$ levels through 2122. We projected $CO_2$ to reach 420 ppmv in Week 18 of 2021, which matches the actual data (420 ppmv in April 2021). Our model projects the $CO_2$ level to reach 500 ppmv in Week 32 of 2056, and reach 671 ppmv by 2122.

As we observed from the Figure 4, the estimated confidence interval widens as the forecast horizon increases. We are quite confident about the short term forecast with a tight confidence internal. For example, for week 18 of 2021, our estimated mean is 420 ppmv and variance is 1.6, with a tight range of 418 and 423 ppmv at 95% confidence. However, we are not confident about the long term forecast such as 2122, which has the estimated mean of 671 ppmv and a variance of 73013! This will yield a wide range between 141 and 1200 ppmv at 95% confidence, which means the $CO_2$ level could swing wildly and is not very useful information.

The increasing forecast uncertainty is expected for a time series forecast model. The more time goes by, the more uncertain the forecast becomes.

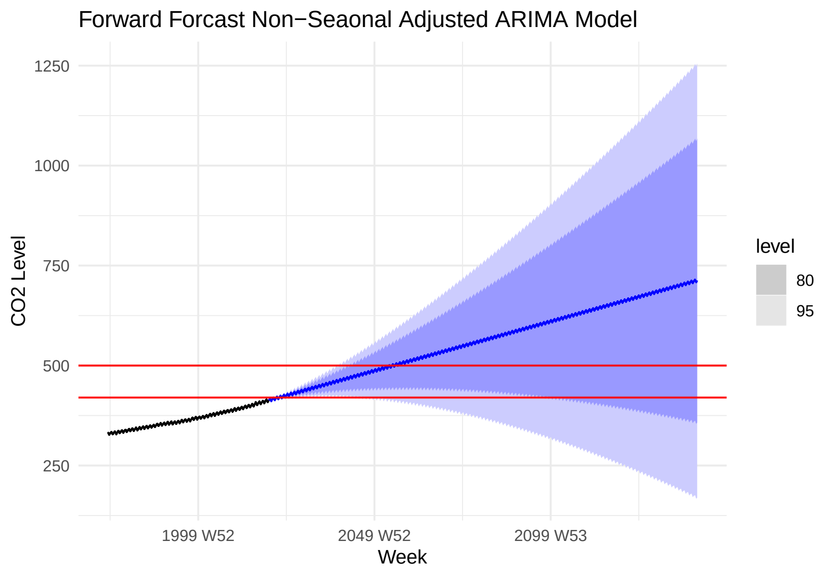


Figure 4: Non-Seasonally Adjusted ARIMA forward forecast model

| Time | Forecast CO2 level | Confidence Interval |
|---|---:|---|
| 2021 W18 | 420.0434 | [417.9934, 422.0934]95 |
| 2056 W34 | 499.7883 | [406.6748, 592.9019]95 |
| 2122 W01 | 668.4743 | [250.0466, 1086.9019]95 |