Global CO_2 Emissions in 1997

By Carolyn Dunlap
Ayda Nayeb Nazar
Qian Qiao
Hector Rincon*

This study analyzes the time series of atmospheric CO2 concentrations from monitoring stations around the world that managed by the National Oceanic and Atmospheric Administration (NOAA). Our goal is to develop a model that accurately captures the trend and seasonality of CO2 concentrations over time, as well as any other relevant patterns or anomalies. To accomplish this, various time series modeling techniques are employed, including linear regression models and seasonal autoregressive integrated moving average (ARIMA) models. The analysis finds that a seasonal ARIMA model with a trend and seasonal difference term provides the best fit to the CO2 time series data. This model reveals a clear upward trend in CO2 concentrations over time, with seasonal variations superimposed on top of this trend. *Keywords: Replication, Modern Science*

Climate change is one of the most pressing issues of our time. One of the key contributors to climate change is the increase in atmospheric carbon dioxide (CO2) levels. The Keeling Curve, created by Charles David Keeling, shows the steady increase in atmospheric CO_2 levels over time. What the trend and seasonality lies in the atmospheric CO_2 changing? How can we use the trend and seasonality to to model the future impacts of climate change? These are the core questions we want to address.This analysis will help us to have a better understanding on a changing climate and to develop strategies for mitigating its effects.

I. Background

A. Carbon Emissions

Carbon emissions are a major environmental concern because they contribute to the Earth's rising temperature, which can cause a range of negative impacts on the environment and human society. These impacts include sea-level rise, more frequent and severe weather events, changes in precipitation patterns, and the loss of biodiversity and ecosystem services.

* Dunlap: UC Berkeley, School of Information, cadunlap@ischool.berkeley.edu. Nayeb Nazar: UC Berkeley, School of Information, ayda@ischool.berkeley.edu. Qiao: UC Berkeley, School of Information, qianqiao@ischool.berkeley.edu. Rincon: UC Berkeley, School of Information, hector@ischool.berkeley.edu. The authors would like to thank their instructors from MIDS 271.

II. Measurement and Data

A. Measuring Atmospheric Carbon

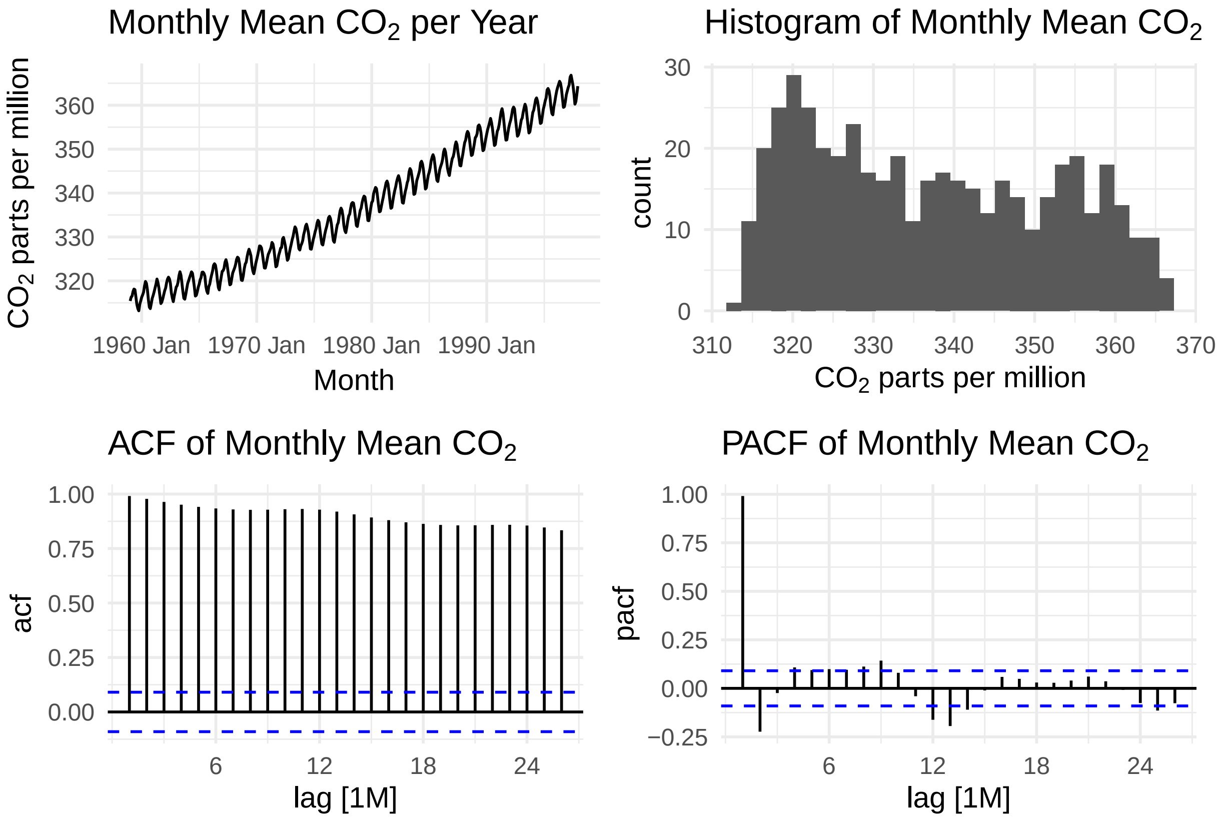
Our analysis will draw on atmospheric CO_2 data. The data on atmospheric carbon dioxide (CO2) levels is generated by the Global Monitoring Laboratory (GML) of the National Oceanic and Atmospheric Administration (NOAA), which operates a network of over 80 monitoring stations around the world. The data is collected through continuous measurements of atmospheric CO2 concentrations using highly precise instruments, such as gas chromatographs and infrared analyzers. One of the longest and most well-known records of this data is the Mauna Loa CO2 record, which has been collected since 1958 at the Mauna Loa Observatory in Hawaii.

These stations collect CO2 measurements on a continuous basis, with data typically collected every 10 to 15 minutes and measure the concentration of atmospheric CO2 using highly precise instruments, such as gas chromatographs and infrared analyzers.

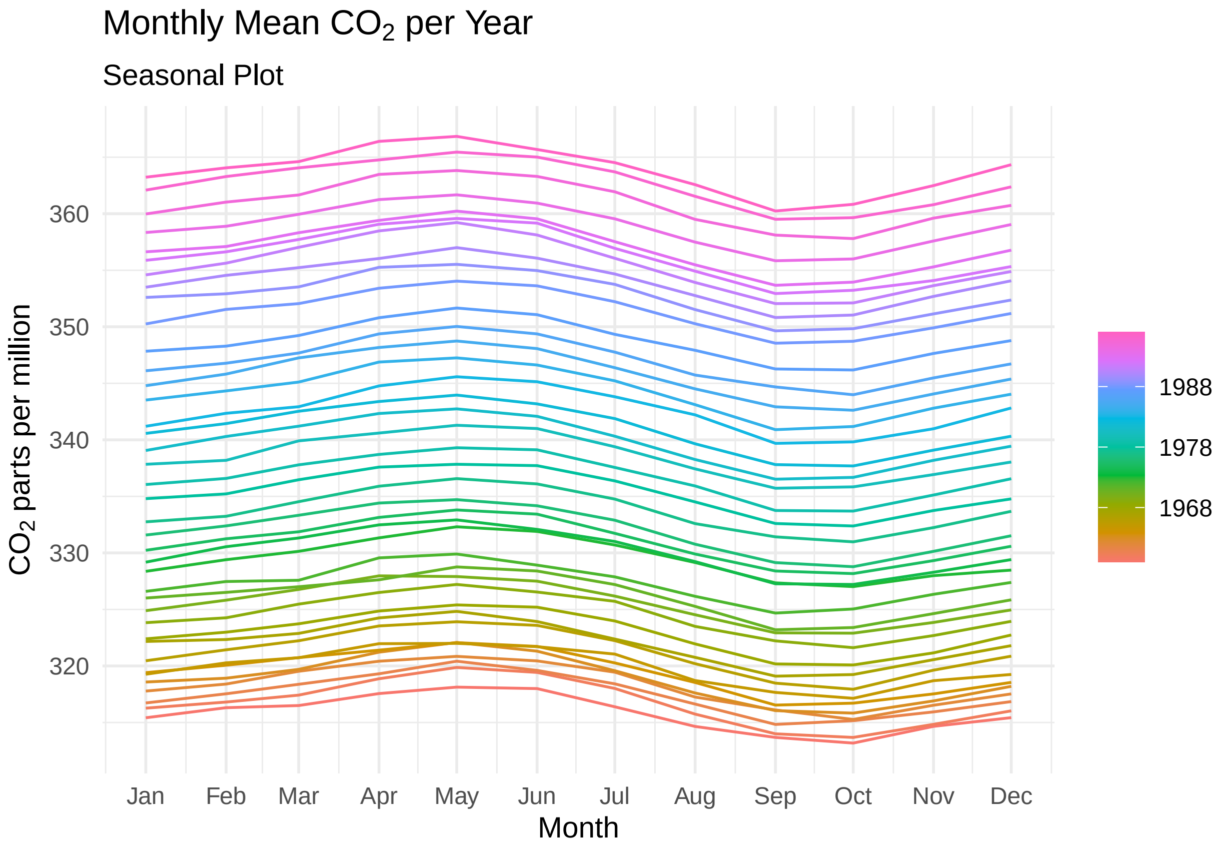
B. Historical Trends in Atmospheric Carbon

Monthly Atmospheric carbon from 1959 to 1997 is plotted in plot below. The time series plot shows an increasing trend from 1959 to 1997 and there is a seasonal fluctuations every year.

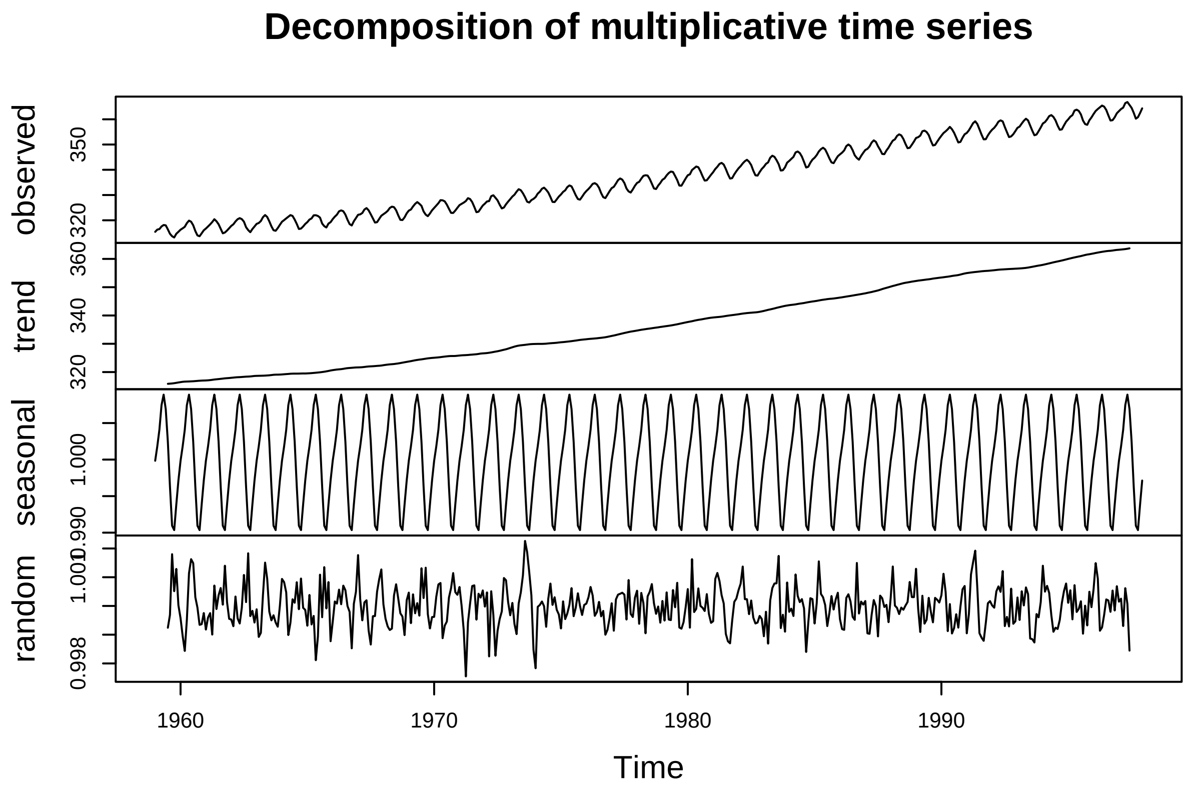
The ACF plot shows a slow decline and while the PACF plot drops sharply after the first lag and shows significance up to three lags, which confirmed our observation from the time series plot that the CO_2 data has a trend pattern.



The seasonal plot displays a clear seasonality for each year. The atmospheric CO_2 level went to peak in April and May, and had the yearly lowest CO_2 pmm in September and October.



Through the multiplicative composition plot, we observed a clear upward trend in overall CO_2 levels over time as well as seasonal patterns.



III. Models and Forecasts

To investigate the trend, seasonal, and irregular elements of the CO_2 data, we used linear regression and ARIMA method to model the long-term increase in CO_2 levels over time.

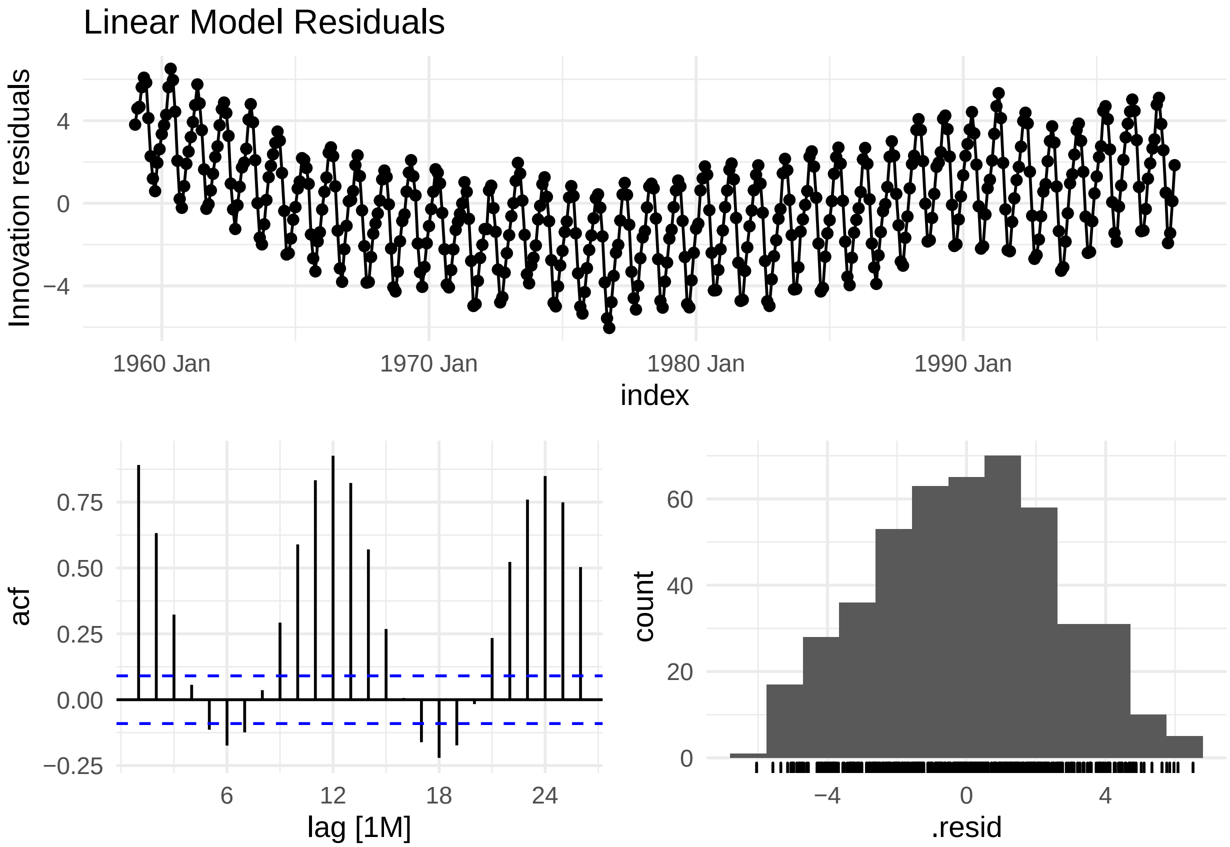
A. Linear Models

We begin by fitting the following simple linear model motivated by the linear trend observed in the EDA:

$$\text{CO}_2 = \beta_0 + \beta t \tag{1}$$

The model parameters are then estimated in the following way,

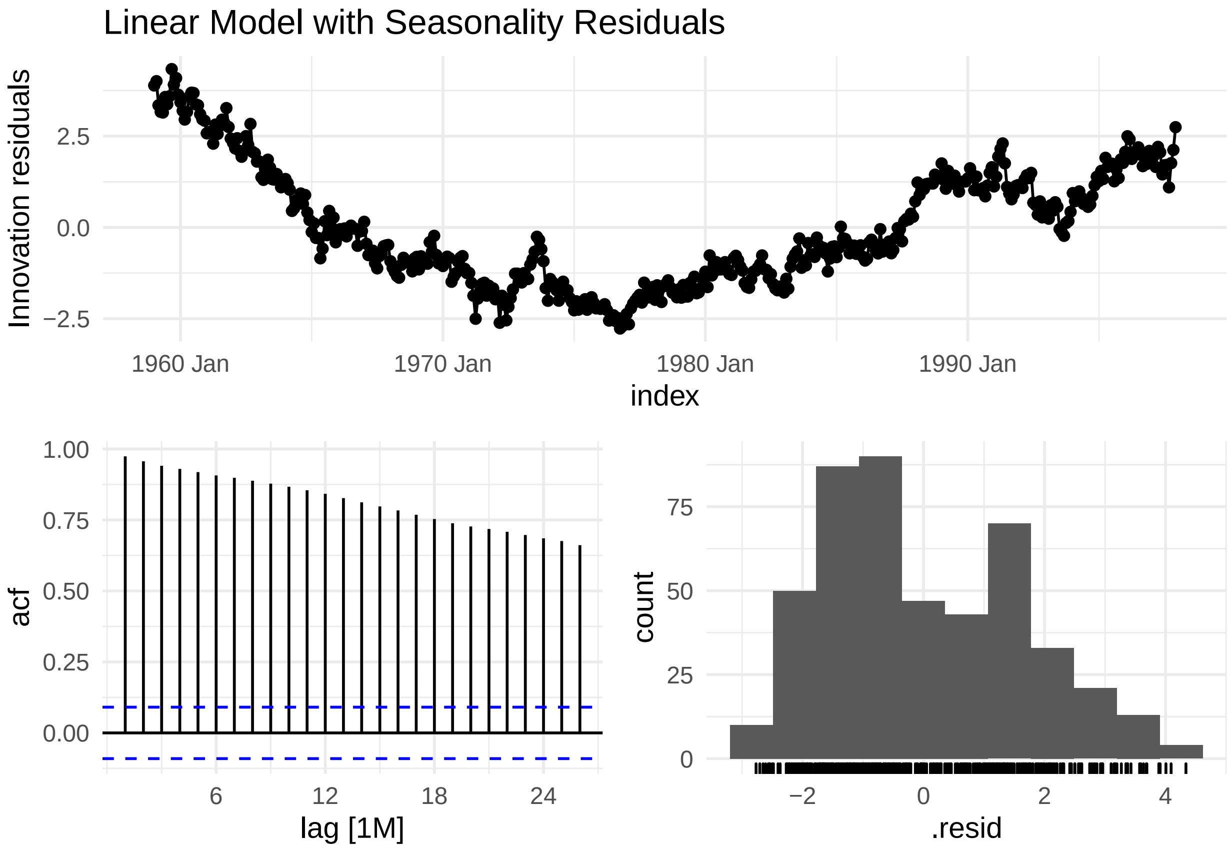
```
co2_reg <- ts_co2 %>%
    model(TSLM(value ~ trend())) %>%
  report()
```



Despite the strong linear trend observed in the time series plot, the residuals of the simple linear model to not appear to be white noise as showcases by the numerous significant lags and strong seasonal pattern in the ACF plot. We can attempt to rectify this by incorporating seasonal dummy variables into our model,

```
co2_reg_season <- ts_co2 %>%
    model(TSLM(value ~ trend() + season())) %>%
  report()
```

```
## Series: value
## Model: TSLM
##
## Residuals:
##    Min     1Q Median     3Q    Max
##  -2.77  -1.28  -0.41   1.26   4.34
##
## Coefficients:
##                 Estimate Std. Error t value Pr(>|t|)
## (Intercept)    311.42208    0.29171 1067.57  < 2e-16 ***
## trend()           0.10921    0.00056  195.00  < 2e-16 ***
## season()year2     0.66336    0.37054    1.79  0.07408 .
## season()year3     1.40543    0.37054    3.79  0.00017 ***
## season()year4     2.53597    0.37054    6.84  2.5e-11 ***
## season()year5     3.01445    0.37054    8.14  4.0e-15 ***
## season()year6     2.35140    0.37055    6.35  5.4e-10 ***
## season()year7     0.83039    0.37055    2.24  0.02551 *
## season()year8    -1.23728    0.37056   -3.34  0.00091 ***
## season()year9    -3.06162    0.37056   -8.26  1.6e-15 ***
## season()year10   -3.24441    0.37057   -8.76  < 2e-16 ***
## season()year11   -2.05490    0.37058   -5.55  5.0e-08 ***
## season()year12   -0.93744    0.37059   -2.53  0.01176 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.64 on 455 degrees of freedom
## Multiple R-squared: 0.988,   Adjusted R-squared: 0.988
## F-statistic: 3.22e+03 on 12 and 455 DF, p-value: <2e-16
```

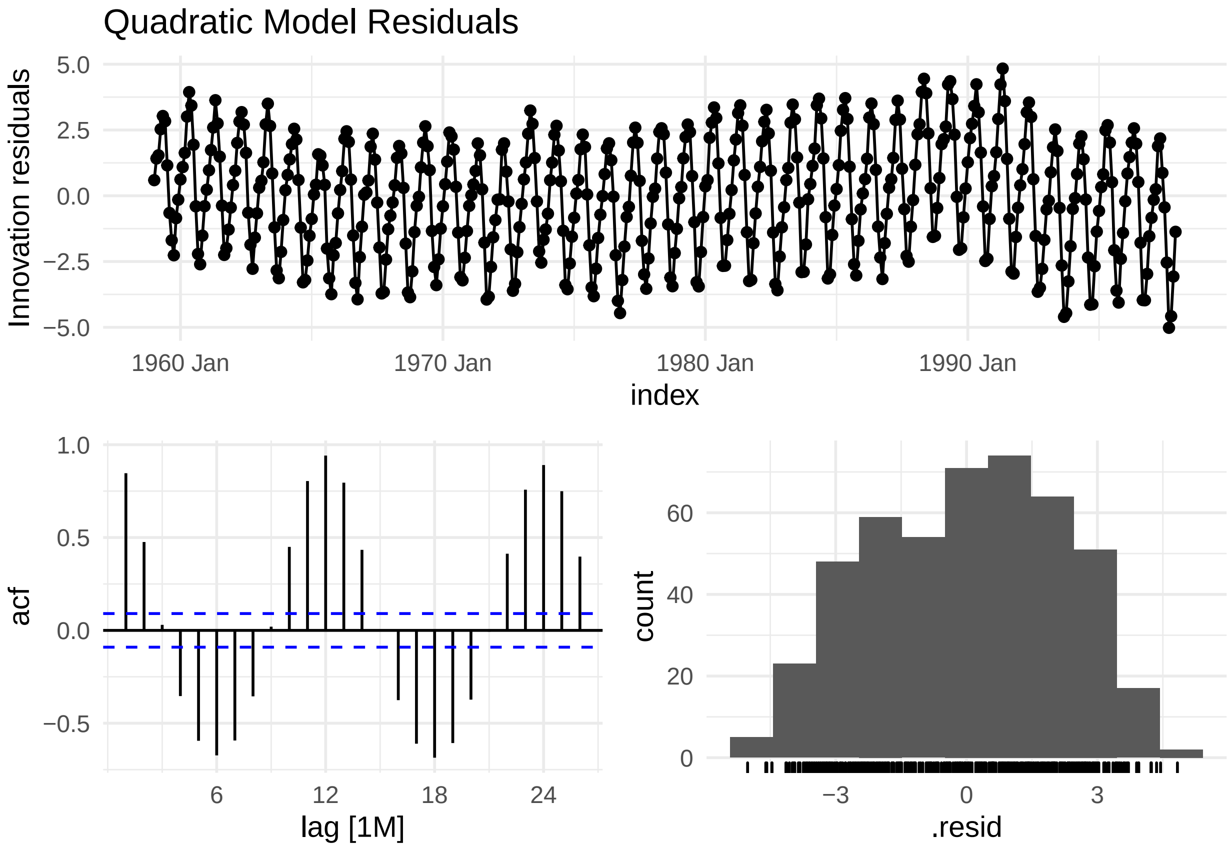


The residuals for this model, however, result in lags that are all autocorrelated and a residual plot that forms a parabolic pattern, which is evidence against the hypothesis that a linear model can appropriately fit the data. A polynomial model may therefore be a more sensible option in order to capture the non-linearities in the data, so we then fit the following quadratic model:

$$\mathrm{CO}_2 = \beta_0 + \beta_1 t + \beta_2 t^2 \tag{2}$$

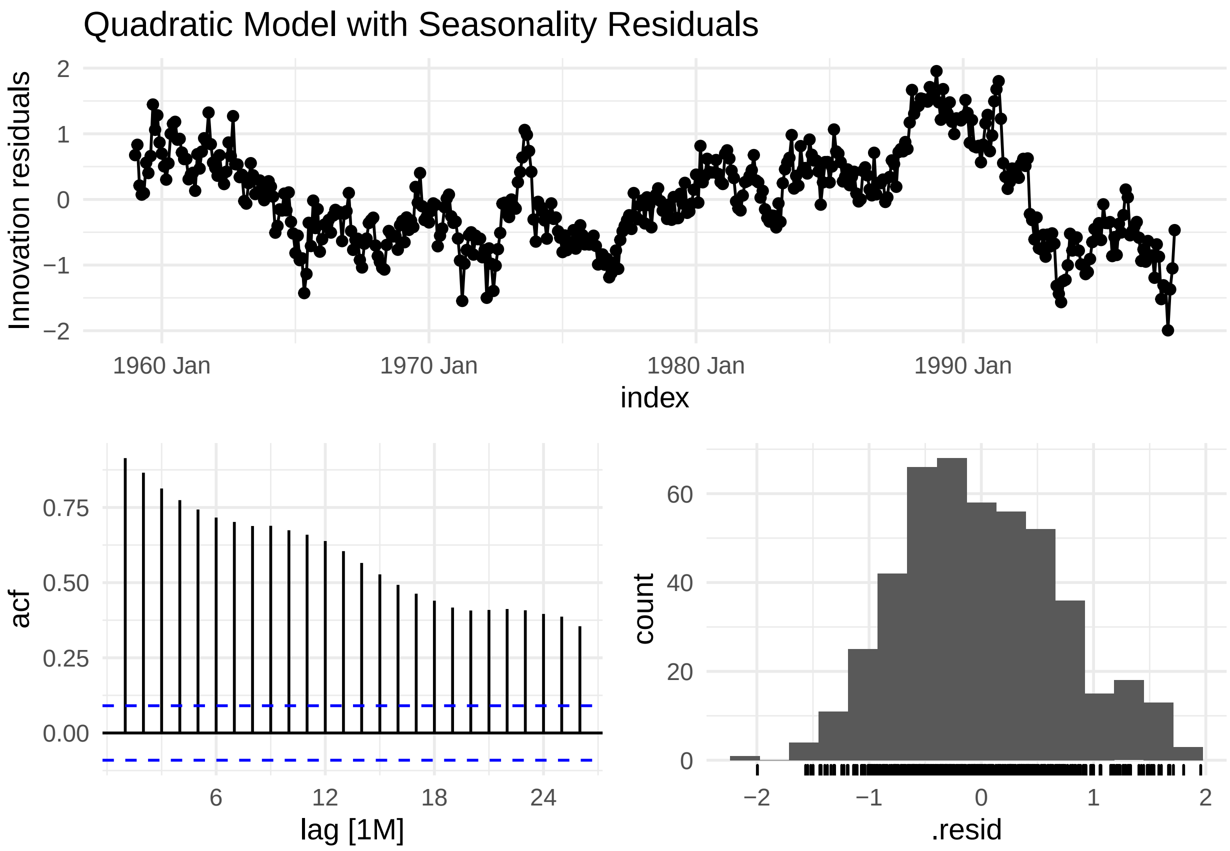
Estimating the parameters as follows,

```
co2_quadratic <- ts_co2 %>%
    model(TSLM(value ~ trend() + I(trend()^2))) %>%
    report()
```

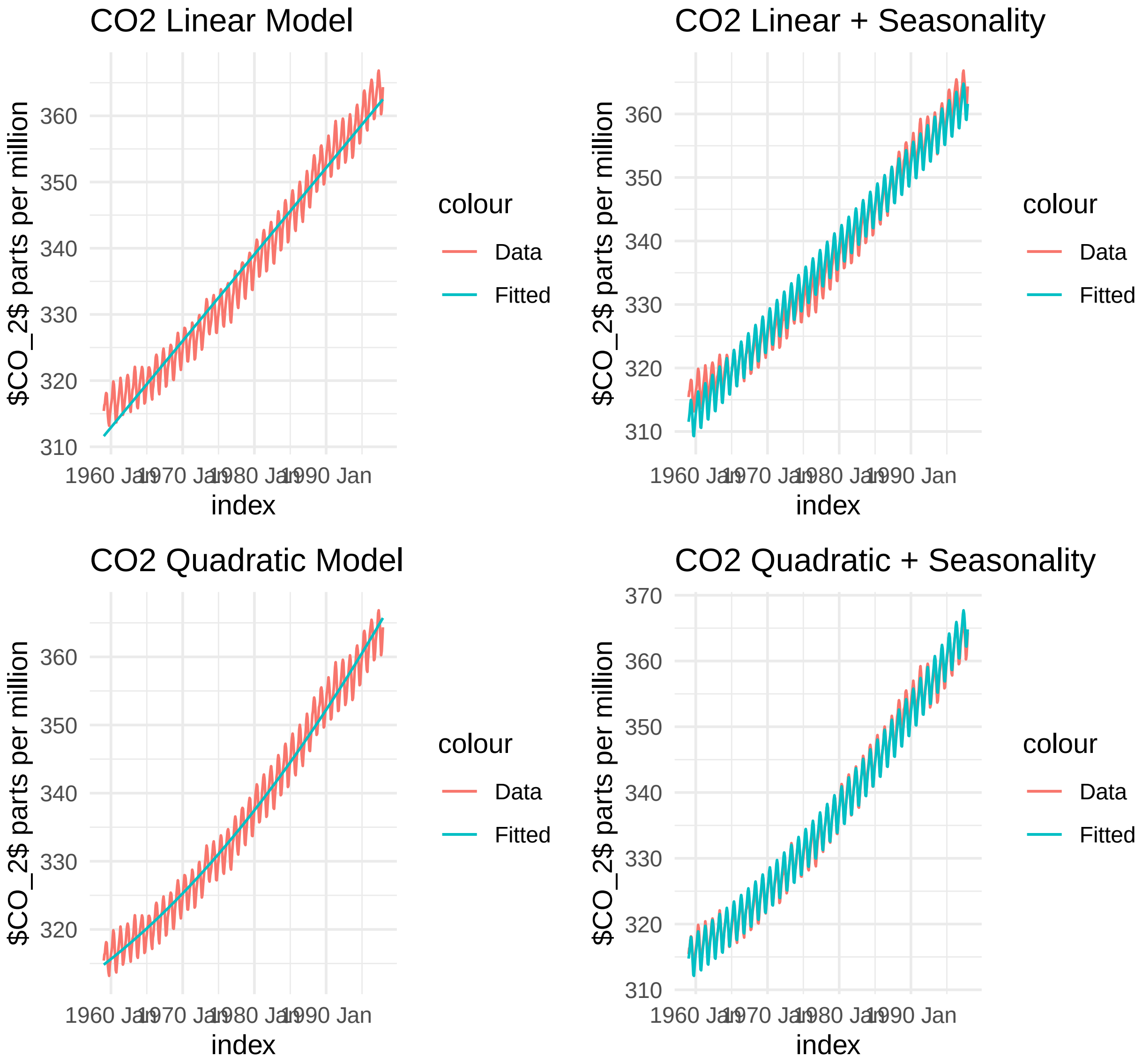


The result is a more white-noise-like residual plot, but with strong seasonality still being showcased in the ACF plot. We can now attempt to rectify for the seasonality once more with the addition of a seasonal dummy variable,

```
co2_quadratic_season <- ts_co2 %>%
    model(TSLM(value ~ trend() + I(trend()^2) + season())) %>%
    report()
```



but despite the addition of the seasonal term, a subtle seasonal pattern can still be observed in the ACF, along with all significant lags and strong autocorrelation in the ACF.



From the fitted value plots, we can see that the simple linear and quadratic models capture the trend quite well, but fail to capture the seasonal fluctuations present in the data, although it's important to note that the quadratic model was slightly more successful at this than the linear model. The linear model corrected for seasonality attempts to better capture the seasonal effect and does much better job at doing so, but it underestimates the observed data. In the end, our final quadratic model with the additional seasonal dummy variable seems to do a good job capturing both trend and seasonal movement in the data,

adj_r_squared	CV	AIC	AICc	BIC	name
0.998	0.541	-286	-285	-224	Quadratic + Seasonality
0.988	2.759	476	477	534	Linear + Seasonality
0.979	4.794	735	735	752	Quadratic
0.969	6.889	905	905	917	Linear

and a quick calculation of the AIC, AICc, and BIC of all tested models confirms this theory as it has the lowest value for all aforementioned information criterion.

.model	lb_stat	lb_pvalue
TSLM(value ~ trend())	2974	0
TSLM(value ~ trend() + season())	8011	0
TSLM(value ~ trend() + I(trend()^2))	3838	0
TSLM(value ~ trend() + I(trend()^2) + season())	4441	0

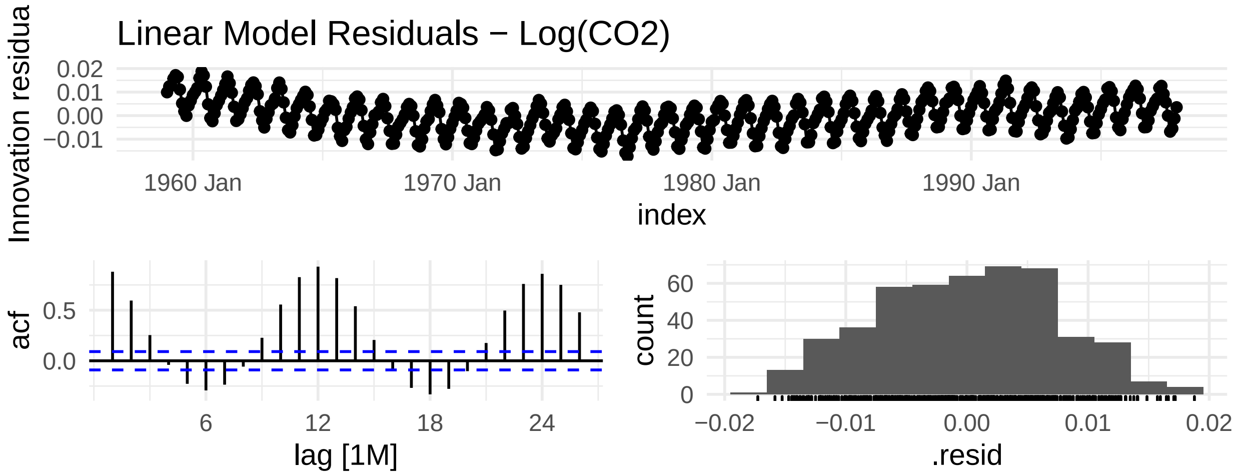
From the Ljung Box test results, the statistic for all models are high and the p-value are small, we reject the null hypothesis of no autocorrelation. We can conclude that the residuals are auto-correlated.

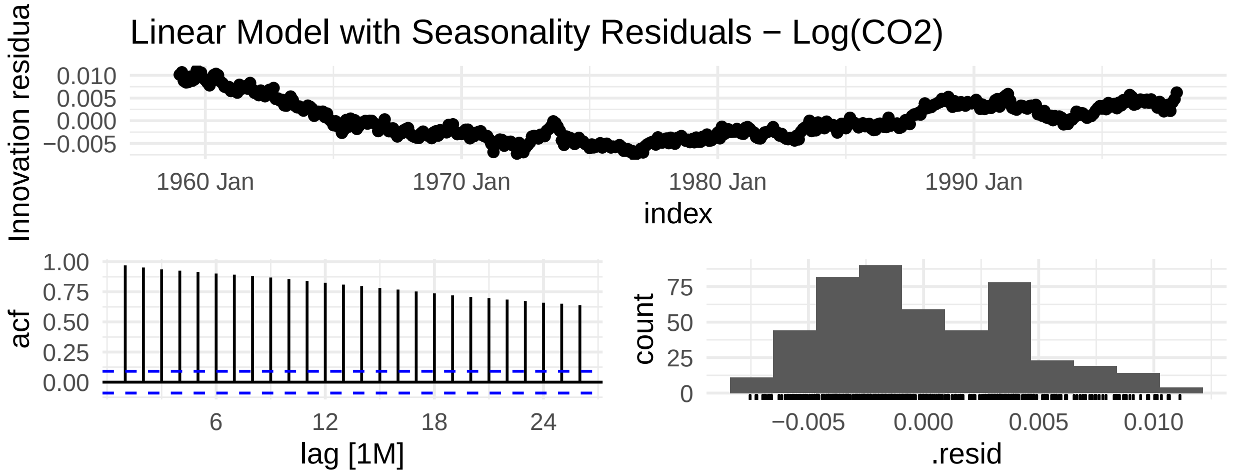
We then evaluate a logarithmic transformation of the data in order to gauge whether it would be a worthwhile transformation to consider in trying to improve the fit of our models.

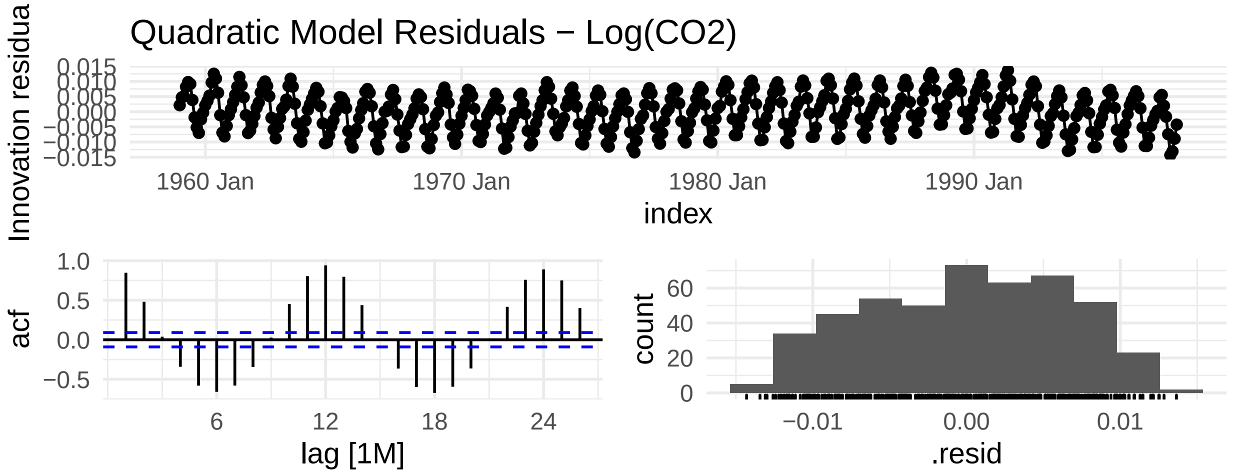
```r
ts_log_co2 <- ts_co2 %>%
      mutate(log_co2 = log(value))
head(ts_log_co2) %>% knitr::kable()
```

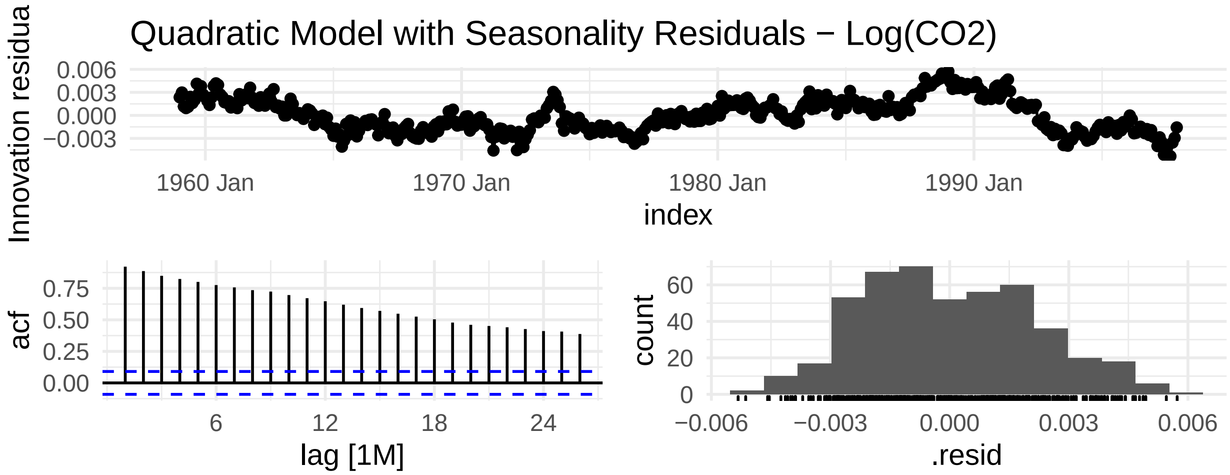
index	value	log_co2
1959 Jan	315	5.75
1959 Feb	316	5.76
1959 Mar	316	5.76
1959 Apr	318	5.76
1959 May	318	5.76
1959 Jun	318	5.76

Then new log-transformed data is now used to change the response variable of CO_2 for the same four previous models,


Linear Model Residuals − Log(CO2)
Innovation residua
0.02
0.01
0.00
−0.01
1960 Jan
1970 Jan
1980 Jan
1990 Jan
index
acf
0.5
0.0
6
12
18
24
lag [1M]
count
60
40
20
0
−0.02
−0.01
0.00
0.01
0.02
.resid


Linear Model with Seasonality Residuals − Log(CO2)
Innovation residua
0.010
0.005
0.000
−0.005
1960 Jan
1970 Jan
1980 Jan
1990 Jan
index
acf
1.00
0.75
0.50
0.25
0.00
6
12
18
24
lag [1M]
count
75
50
25
0
−0.005
0.000
0.005
0.010
.resid

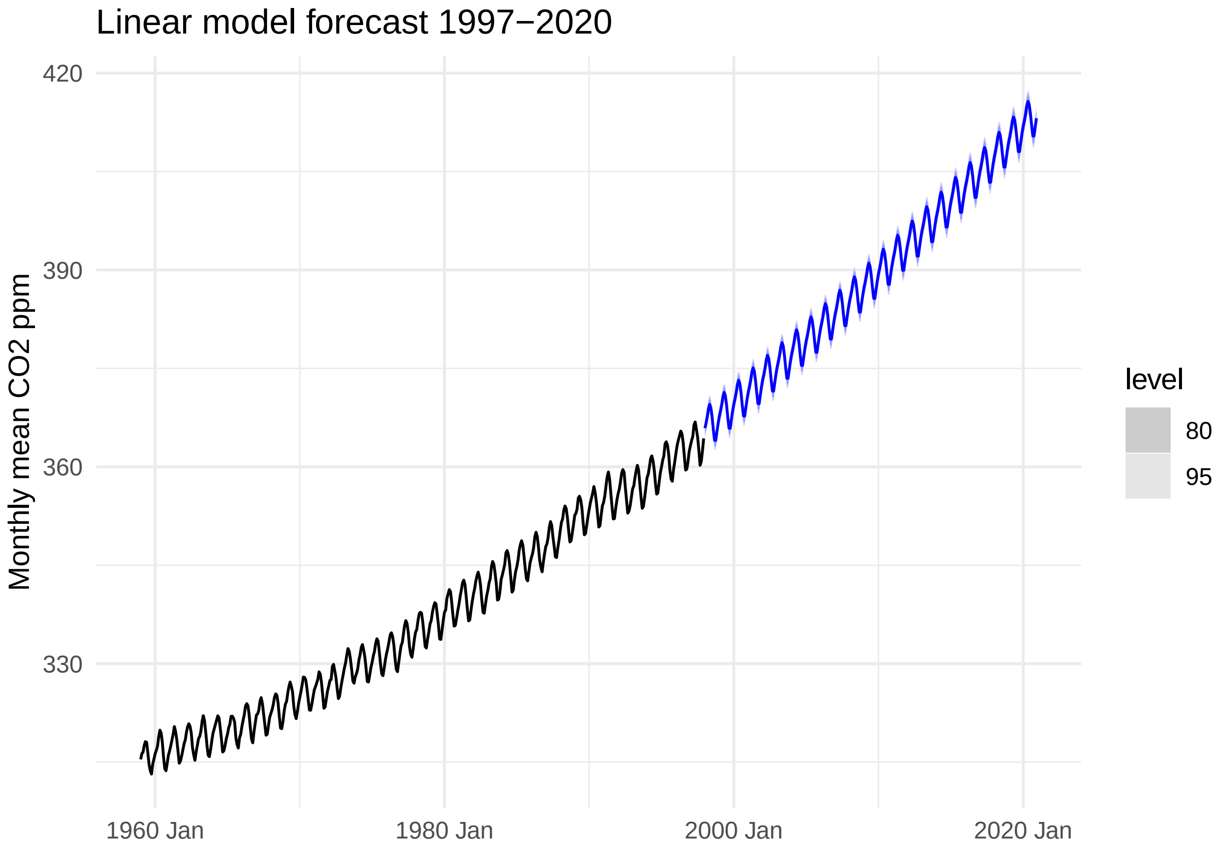

Quadratic Model Residuals − Log(CO2)
Innovation residua
0.015
0.010
0.005
0.000
−0.005
−0.010
−0.015
1960 Jan
1970 Jan
1980 Jan
1990 Jan
index
acf
1.0
0.5
0.0
−0.5
6
12
18
24
lag [1M]
count
60
40
20
0
−0.01
0.00
0.01
.resid



and no significant change can be observed in the residuals of the models. This is to be expected, however, as the exploratory data analysis above did not show a clear exponential trend in the CO_2 data that would need to be smoothed and correct with a logarithmic transformation, and the variance did not appear to increase or decrease over time. Therefore, a logarithmic transformation is not exactly necessary for modeling the CO_2 data.

We further put our best quadratic with seasonality model to the test by generating forecasts to the year 2020, and graphing the results:

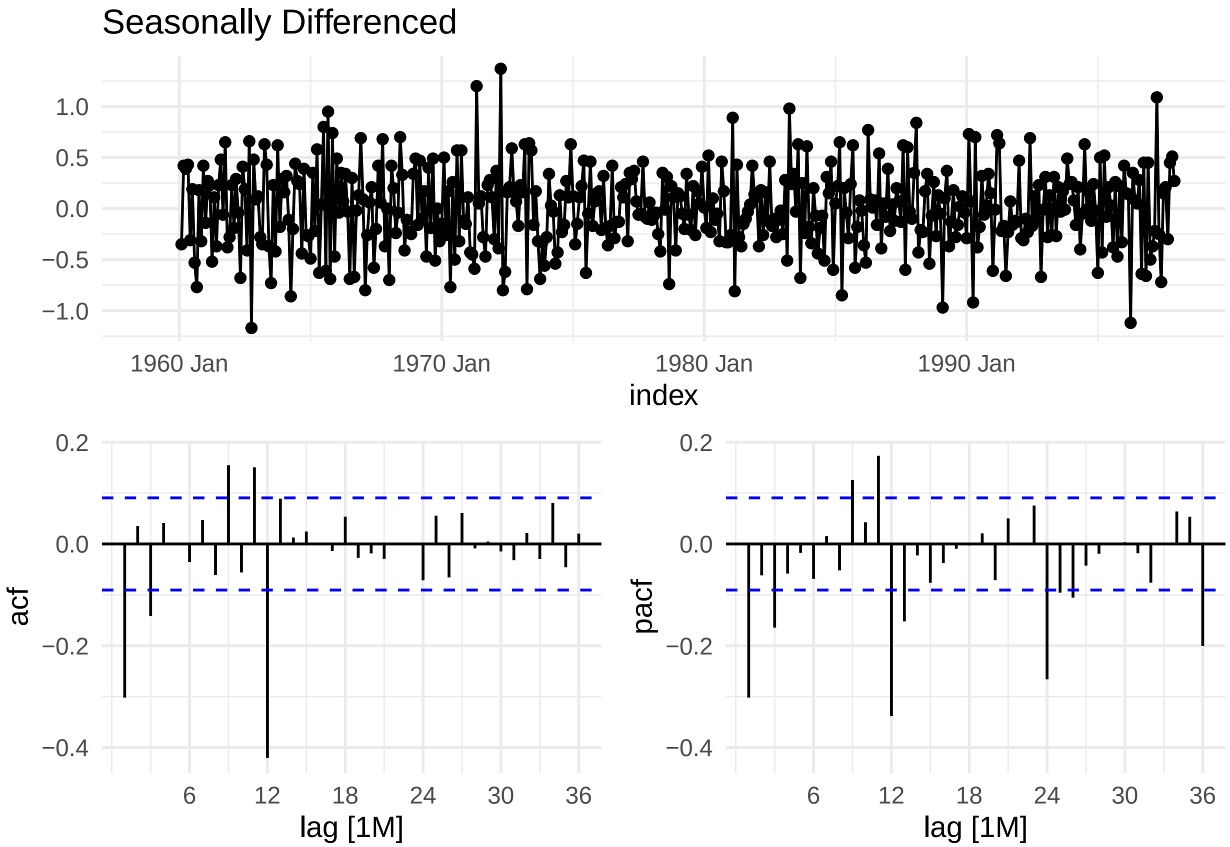
```
# Create a forecast object based on the model and forecast to the year 2020
save(co2_quadratic_season, file='1997_quad_seasonal_modelfit.RData')
quad.forecast <- fabletools::forecast(co2_quadratic_season, h = "23 years")
# Forecast + previous data graph
forecast.graph <- quad.forecast %>% autoplot(ts_co2) + labs(title="Linear model forecas
forecast.graph
```



Although the forecast seems to follow along the same patterns as past CO_2 data, there's strong evidence from the high autocorrelation observed in the ACF of the model residuals and the lack of observed white noise residuals that this model does not perfectly fit our model. This therefore motivates exploring an ARIMA model and checking the first difference.

B. ARIMA Models

We begin by first differencing the CO_2 data:



From the EDA, we found the CO_2 time series had obvious trend and seasonality. Based on the time series plot of the first difference, we now observe that the data's first difference may be stationary, and from the ACF plot, we see seasonal fluctuations every 12 lags (months). A Phillips-Perron and Augmented Dickey-Fuller test can then be run to test the alternate hypothesis that the time series is stationary, as opposed to the null hypothesis that it is explosive, or non stationary.

```
PP.test(co2_diff)
```

```
##
##  Phillips-Perron Unit Root Test
##
## data:  co2_diff
## Dickey-Fuller = -9, Truncation lag parameter = 5, p-value = 0.01
```

```
adf.test(co2_diff, alternative = "stationary")
```

```
##
##  Augmented Dickey-Fuller Test
```

```
##
## data:  co2_diff
## Dickey-Fuller = -30, Lag order = 7, p-value = 0.01
## alternative hypothesis: stationary
```

Both the Phillips Perron and Augmented Dickey Fuller (ADF) tests show a p-value that is less than 0.05, therefore providing strong evidence for us to reject the null hypothesis that this time series is non stationary. Thus, the time series with the first difference is stationary and we can start to build the model with 1 difference.

```
fit %>%
  report() %>% arrange(AIC) %>%
  select(-ar_roots, -ma_roots) %>% knitr::kable()
```

```
## Warning in report.mdl_df(.): Model reporting is only supported for individual
## models, so a glance will be shown. To see the report for a specific model, use
## ‘select()‘ and ‘filter()‘ to identify a single model.
```

.model	sigma2	log_lik	AIC	AICc	BIC
auto	0.085	-83.4	177	177	197
arima012011	0.086	-85.5	179	179	196
arima111112	0.086	-84.4	181	181	205
arima210011	0.087	-87.7	183	184	200
arima214000	0.290	-373.3	763	763	796
arima111000	0.631	-554.5	1115	1115	1128
arima121000	0.785	-603.7	1213	1213	1226

```
fit$auto[[1]]$fit$spec %>% knitr::kable()
```

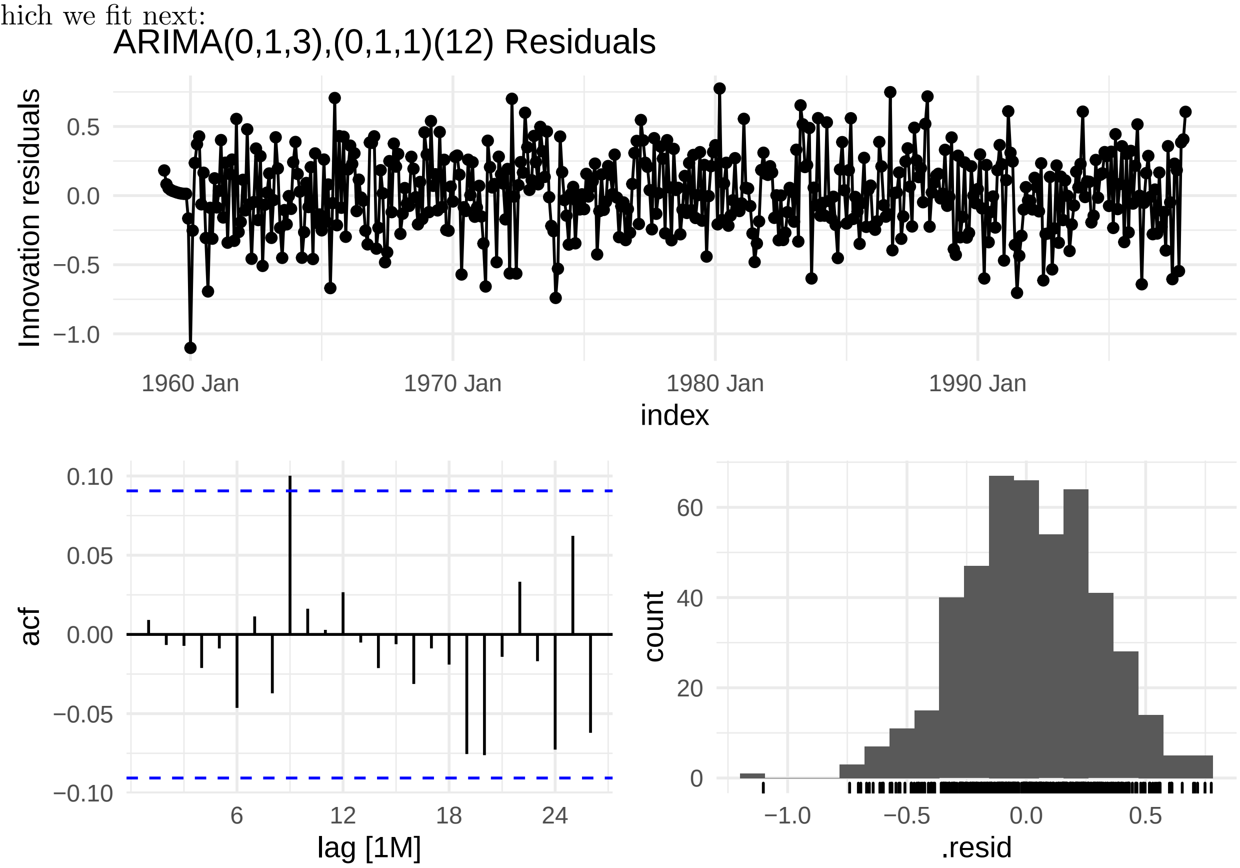
p	d	q	P	D	Q	constant	period
0	1	3	0	1	1	FALSE	12

```
fit$auto[[1]]$fit$model
```

```
##
## Call:
## .f(x = ..1, order = ..2, seasonal = ..3, xreg = ..4, include.mean = FALSE, fixed = .
##     method = ..5)
##
```

```
## Coefficients:
##          ma1     ma2     ma3    sma1
##       -0.339  -0.018  -0.097  -0.854
## s.e.   0.048   0.050   0.047   0.026
##
## sigma^2 estimated as 0.0852:  log likelihood = -83.4,  aic = 177
```

After fitting numerous ARIMA models with varying specifications and running the auto model as well, we find that based on the AIC, AICc scores displayed in the table above, the auto model found to be ARIMA(0,1,3),(0,1,1)(12) has the lowest scores and is therefore the best fit model. When looking at the BIC scores, however, the lowest value was found to be for the ARIMA(0,1,2),(0,1,1)(12) model. Since we are aiming to complete more of a prediction task with our forecasting efforts rather than an explanation task and that AIC is most optimal in minimizing the mean squared error of predictions, we choose to use AIC/AICc as our main information criteria. The best fit model is therefore ARIMA(0,1,3),(0,1,1)(12), which we fit next:

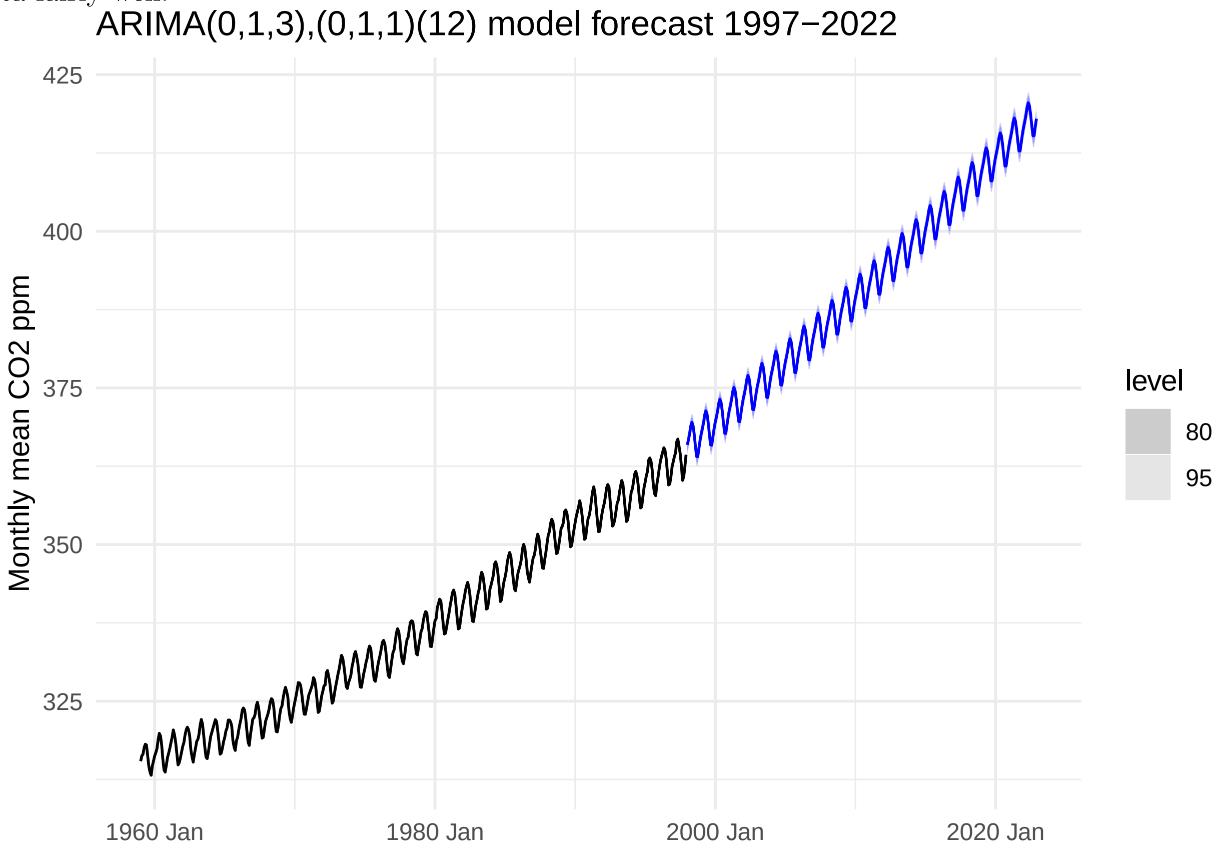


The flat mean of the residuals at 0, the roughly normally distributed residual counts, and the minimally autocorrelated/significant lags in the ACF are all strong evidence that we have white noise residuals and that the ARIMA(0,1,3),(0,1,1)(12) model fits the CO_2 model well.

```
augment(arima013011) %>%
    features(.resid, ljung_box, lag = 10, dof = 0)
```

```
## # A tibble: 1 x 3
##   .model      lb_stat lb_pvalue
##   <chr>         <dbl>     <dbl>
## 1 arima013011    7.01     0.724
```

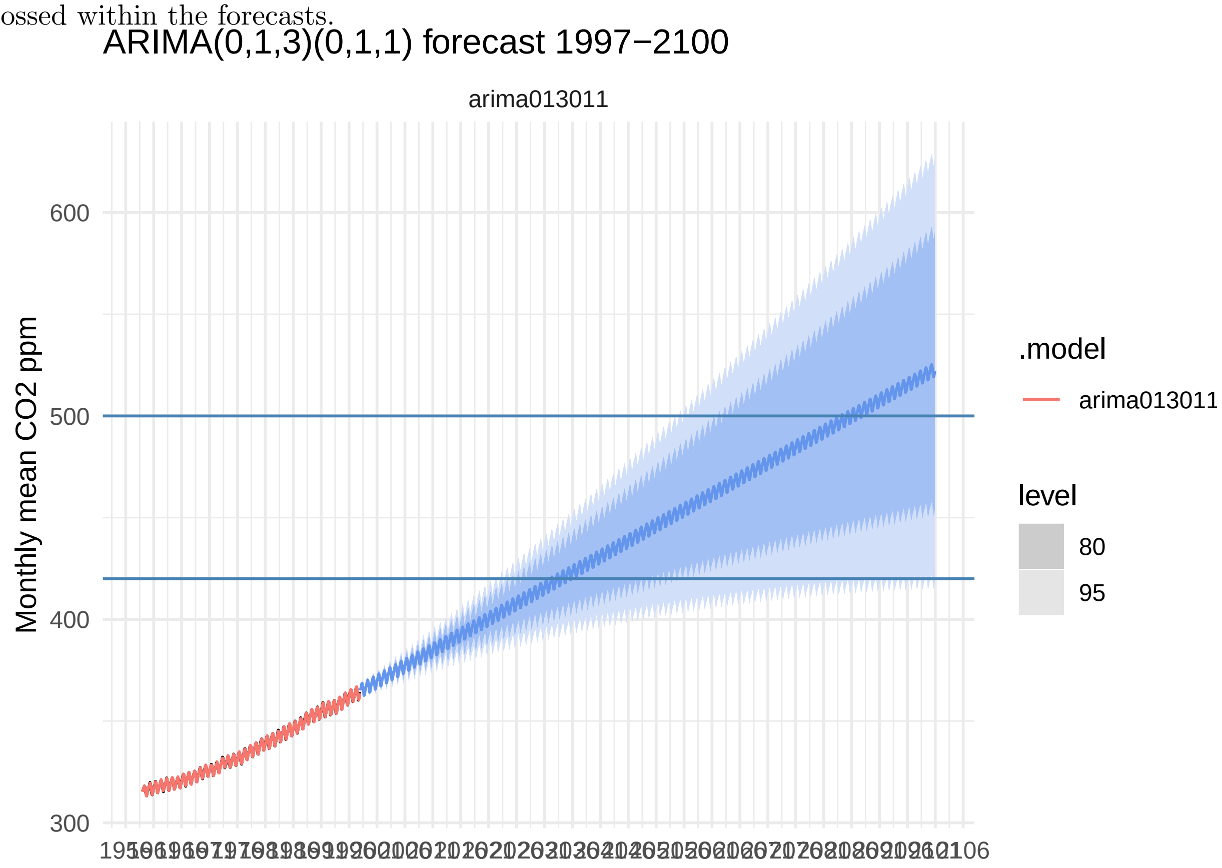
From the Ljung Box test, we got a large p-value that failed to reject the null hypothesis, further supporting that our best model's residuals are randomly distributed, thus making it a good model fit. Generating forecast to the year 2022 with the new ARIMA model shows that the model follows the pattern of past data fairly well.



C. Forecasts

Further forecasts can be generated to the year 2100 to test our model and make predictions about the future. We see the performance within 80% and 95%

confidence intervals below, and the times the 420 and 500 ppm thersholds are crossed within the forecasts.



target	index	.mean	ci.80.lower	ci.80.upper	ci.95.lower	ci.95.upper
420	2032 Apr	420	405	436	397	444
420	2036 Sep	421	403	439	393	449
500	2084 Apr	500	447	553	419	582
500	2088 Oct	501	444	558	413	588

The forecasts for the year 2100 in particular can be seen below:

index	.mean	ci.80.lower	ci.80.upper	ci.95.lower	ci.95.upper
2100 Jan	522	454	589	419	625
2100 Feb	522	455	590	419	626

index	.mean	ci.80.lower	ci.80.upper	ci.95.lower	ci.95.upper
2100 Mar	523	456	591	420	627
2100 Apr	525	457	592	421	628
2100 May	525	458	593	422	629
2100 Jun	525	457	592	421	628
2100 Jul	523	455	591	419	627
2100 Aug	521	453	589	417	625
2100 Sep	519	451	587	415	623
2100 Oct	519	451	587	415	623
2100 Nov	521	453	589	416	625
2100 Dec	522	454	590	418	627

Although it follows the pattern of past data well, it's difficult to say if these are accurate predictions for the future, but if the current trends persist through to the year 2100, then these predictions may not be far off from the true values we will observe in CO_2 then.

IV. Conclusions

From the modeling and analysis on atmospheric CO_2 data from 1959 to 1997, we achieved our goals to answer the initial questions. Firstly, we observed and demonstrated that CO2 concentrations in the Earth's atmosphere have been increasing steadily from 1959 to 1997 and will keep increasing based on both the quadratic model and ARIMA forecast results. Secondly, both linear model and ARIMA with seasonality have better performance than models that didn't considering the seasonality. Finally, the quadratic model with seasonality and ARIMA(0,1,3),(0,1,1)(12) model fit CO_2 data well, but there's strong evidence that quadratic model has the high autocorrelation observed in the ACF of the model residuals. Therefore, the most plausible model that we estimate is ARIMA(0,1,3),(0,1,1)(12) model and we can use this model to forecast the atmospheric CO_2 changing.

Appendix: Model Robustness