Submitted to Ms. Dara Baltin for MPEC Data Analysis position

Monthly Carbon Dioxide Level Analysis Report

Minjiao Yang

Table of Contents

Introduction

Carbon dioxide is a greenhouse gas, which absorbs heat. Warmed by sunlight, Earth's land and ocean surface continuously radiate thermal infrared energy (heat). Unlike oxygen or nitrogen (which make up most of our atmosphere), greenhouse gases absorb that heat and release it gradually over time. Increases in greenhouse gases have tipped the Earth's energy budget out of balance, trapping additional heat and raising Earth's average temperature. According to the State of the Climate in 2018 report from NOAA and the American Meteorological Society, global atmospheric carbon dioxide was 408.53 $\pm$0.1 ppm in 2019. Carbon-dioxide levels today are higher than at any point in at least the past 800,000 years.

Levels of carbon dioxide (CO2) are monitored at several sites around the world to investigate atmospheric changes. One of the sites is at Alert, Northwest Territories, Canada, near the arctic circle. A record from the carbon dioxide level monitoring site shown the Monthly CO2 levels from January 1994 to December 2004. We want to identify trends and seasonality of this data set and fit a time series model to forecast the future carbon dioxide level.

Stationary and Transformation of Data

The data set contains 132 records of monthly carbon dioxide levels from January 1994 to December 2004.

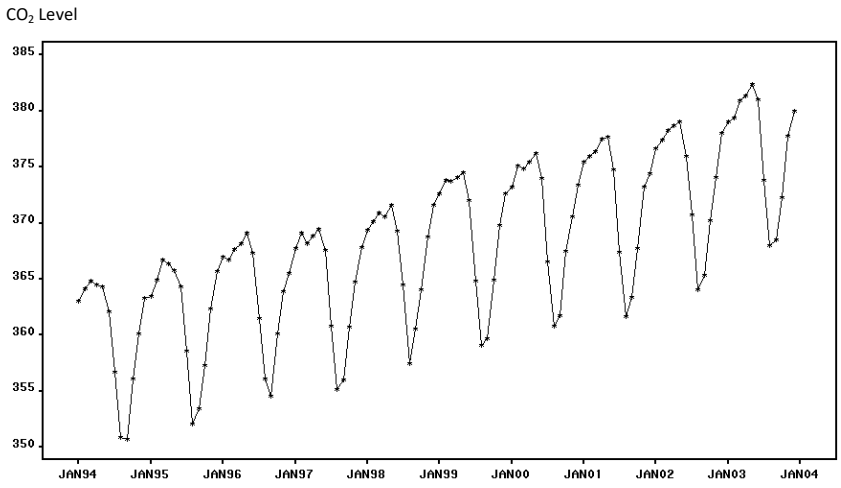

Figure 1. Times series graph of monthly CO2 level from Jan.1994 to Dec.2004

In figure 1, the level of CO_2 increases over the year. The overall maximum Co_2 level is 383.58ppm in April 2004. And the CO_2 level reaches a peak in May every year. So, the time series appears to have an overall trend and seasonal effects.

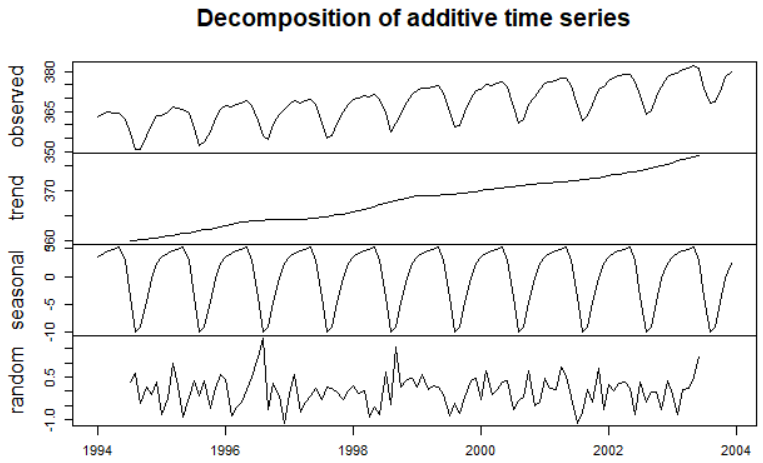

Figure 2. Trend and Seasonality decomposition of Time Series

In figure 2, the time series is strongly seasonal with an overall increasing trend and non-stationary. We know that stationary is important because, in its absence, a model describing the data will vary inaccuracy at different time points. To achieve stationary, we performed a box-cox transformation for the data.

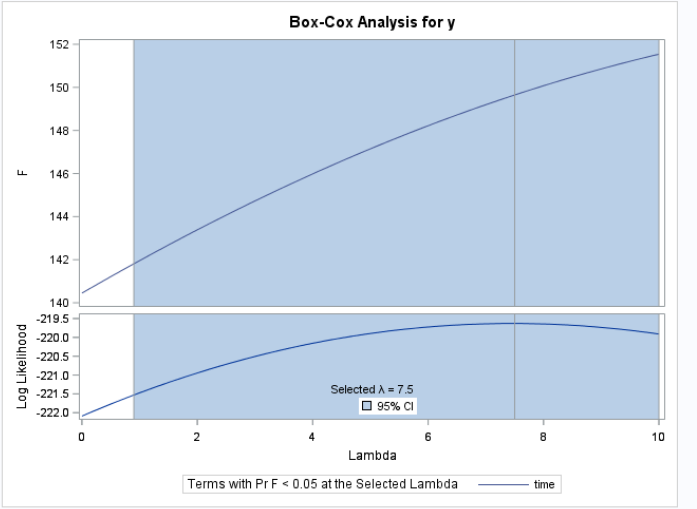

Figure 3. Box-Cox Transformation for y

However, BOX-COX transformation suggested the best lambda is 7.5. lambda=1 is included in a 95% Confidence Interval of log-likelihood indicating transformation on data is unnecessary. Then, we did simple differencing on data to remove overall trend and seasonal differencing to remove seasonality, the resulting time series finally achieved stationary in both mean and variance (as shown in Figure 5).

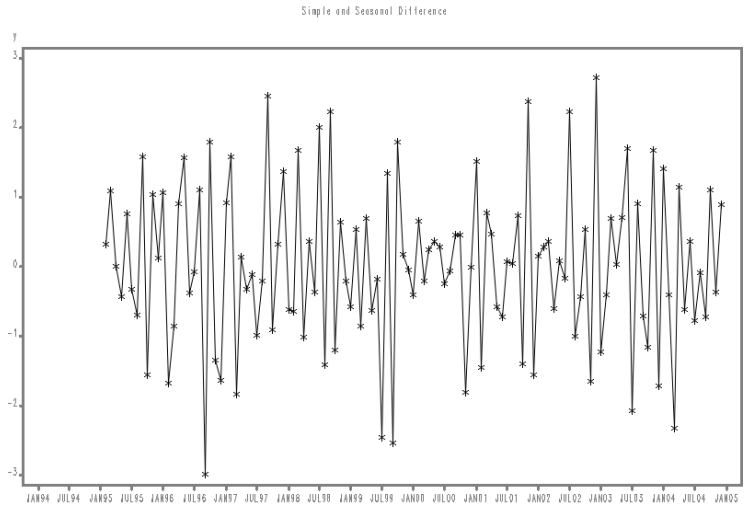

Figure 5. time series graph after adjustment

Model Selection By AIC/BIC/LBP Test

1) Candidate 1: ARIMA(0,1,1)x$(0,1,1)_{12}$

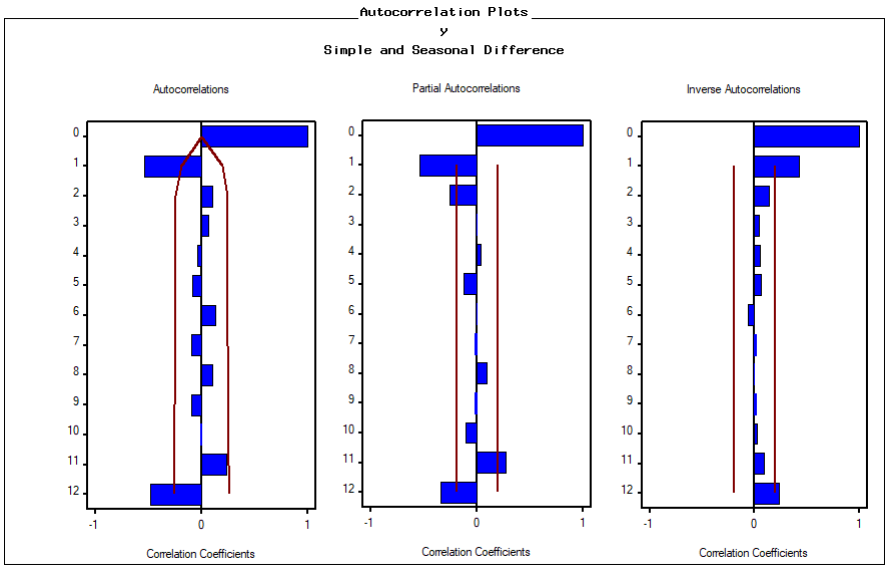

Figure 6. Autocorrelation plots of simple and seasonally differenced data (lag<12)

In the autocorrelation plots of simple and seasonally differenced data, In the non-seasonal lags (lag<12), there are 2 significant spikes in PACF at lag1& 2, while ACF decays exponentially. This suggests a possible AR(2) term. On the other hand, there is 1 significant spike in ACF at lag 1, while PACF decays exponentially. This suggests a possible MA(1) term.

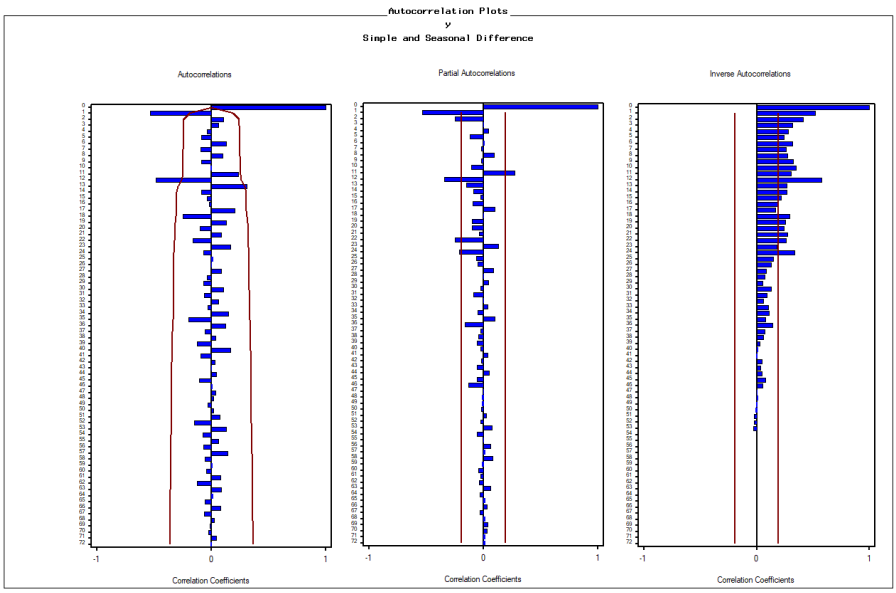

Figure 7. Autocorrelation plots of simple and seasonally differenced data (more lags)

Besides, there are spikes in the PACF at lag 12 and 24, while ACF decays exponentially at lag 12, 24, etc. This may be suggestive of a seasonal term AR(2) term. On the other hand, there is a spike in the ACF at lag 12, while PACF decays exponentially at lag 12, 24. This may be suggestive of a seasonal term MA(1).

Consequently, this initial analysis suggests that possible models for these data are ARIMA(2,1,0)x(2,1,0)$_{12}$, ARIMA(2,1,0)x(2,1,0)$_{12}$, ARIMA(0,1,1)x(2,1,0)$_{12}$, and ARIMA(0,1,1)x(0,1,1)$_{12}$. We fit these models, along with some variations on them (in total 25 models), compute the Akaike information criterion (AIC=2k-2ln(loglikelihood)), Bayesian information criterion (BIC=2nk-2ln(loglikelihood) and look at the Ljung–Box test (LBP-test) (details shown in Appendix). The models ARIMA(0,1,1)x(0,1,2)$_{12}$, ARIMA(0,1,1)x(1,1,1)$_{12}$, and ARIMA(0,1,1)x(0,1,1)$_{12}$ are selected as the best three models based on AIC/BIC criteria (shown in table.1).

MODEL	AIC	BIC	LBP TEST$_{(LAG>0.05)}$	SIGNIFICANT VBLES
ARIMA(0,1,1)X(0,1,2)$_{12}$	254.182	262.201	Lag 6, 12, 54, 60 pass	MA at lag 24 is not
ARIMA(0,1,1)X(1,1,1)$_{12}$	254.184	262.202	Lag 6, 12, 54, 60 pass	AR at lag 12 is not
ARIMA(0,1,1)X(0,1,1)12	252.248	257.593	Lag 6, 12, 48,54, 60 pass	All variables are significant

Table 1. best 3 Models among 25 Comparison (all models without intercept)

Among three models, ARIMA(0,1,1)x(0,1,1)$_{12}$ is selected as the first candidate model because:

- A preferred model should have the smallest AIC and BIC. ARIMA(0,1,1)x(0,1,1)$_{12}$ has the least AIC/BIC among 3 models.
- Residuals of a model should behave like a white noise series. LBP-test is to check if residuals are correlated with each other with null hypothesis “data are independently distributed”. If the p-value of lags is all greater than 0.05 then it passes the LBP-test (accept null hypothesis), the residuals are not correlated with each other.

None of the 3 models passed LBP Test at all lags. While, ARIMA(0,1,1)x(0,1,1)$_{12}$ passed LBP test at more lags.

- It's the only one with all significant variables. (p-value < 0.05)
- If we take out the insignificant variables from the other two models, both of them will end up with ARIMA(0,1,1)x(0,1,1)$_{12}$ model.

So, the first candidate model is:

$$(1 - B^{12})(1 - B)y_t = (1 - \Theta_1 B)(1 - \theta_1 B)w_t$$

And, the fitted model is:

$$(1 - B^{12})(1 - B)y_t = (1 - 0.91106B)(1 - 0.56180B)w_t$$

A detailed summary of the model can be found in the Appendix.

Then we look at the residual diagnose of the ARIMA (0,1,1)x(0,1,1)$_{12}$.

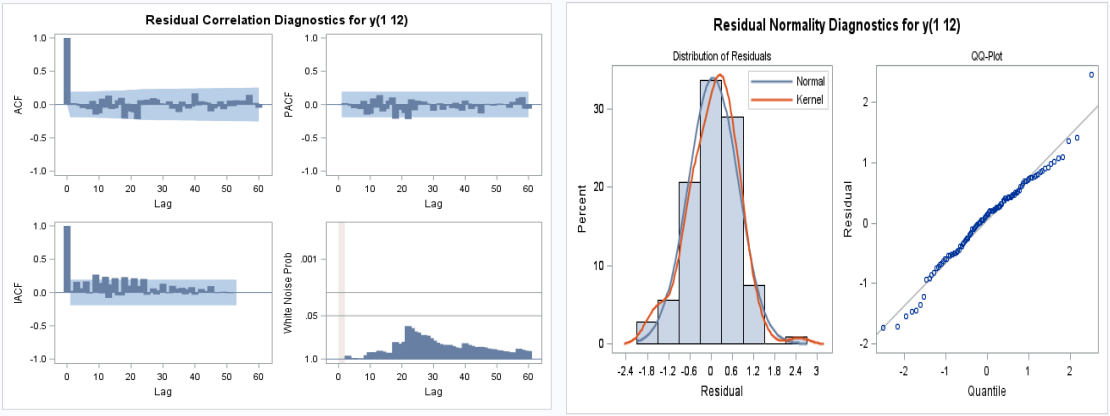

Figure 8. Residual Diagnose for ARIMA(0,1,1)x(0,1,1)_12

Residuals of ARIMA(0,1,1)x(0,1,1)$_{12}$ seems normally distributed in histogram and QQ-plot. But if we take a closer look at residual autocorrelation plots (ACF&PACF), not all spikes stay within the significance limits, so the residuals do not appear to be white noise. The Ljung-Box test also shows that the residuals have remaining autocorrelations.

2) Candidate 2&3: ARIMA(0,1,(1,9,18))x(0,1,1)$_{12}$ & ARIMA(0,1,(1,18))x(0,1,1)$_{12}$

Both the ACF and PACF show significant spikes at lag 18, indicating that some additional non-seasonal terms need to be included in the models. So model ARIMA(0,1,18)x(0,1,1)$_{12}$ was fitted and backward model reduction was perform (take out insignificant variables), the resulting models ARIMA(0,1,(1,9,18)x(0,1,1)$_{12}$ and ARIMA(0,1,(1,18)x(0,1,1)$_{12}$ are shown in table 2.

MODEL	AIC	BIC	LBP TEST$_{(LAG>0.05)}$	SIGNIFICANT VBLES
ARIMA(0,1,(1,9,18))X(0,1,1)$_{12}$	247.314	258.000	All pass	All significant
ARIMA(0,1,(1,18))X(1,1,1)$_{12}$	248.792	256.810	Lag 24 not pass	All significant

Table 2. Developed models from ARIMA(0,1,1)x(0,1,1)_12 (all models without intercept)

So, the second candidate model is:

$$(1 - B^{12})(1 - B)y_t = (1 - \Theta_1 B)(1 - \theta_1 B + \theta_2 B^9 + \theta_3 B^{18})w_t$$

And, the fitted model is:

$$(1 - B^{12})(1 - B)y_t = (1 - 0.92412B)(1 - 0.64312B + 0.18646B^9 + -0.15927B^{18})w_t$$

A detailed summary of the model can be found in the Appendix.
Then we look at the residual diagnose of the ARIMA (0,1,(1,9,18))x(0,1,1)$_{12}$.

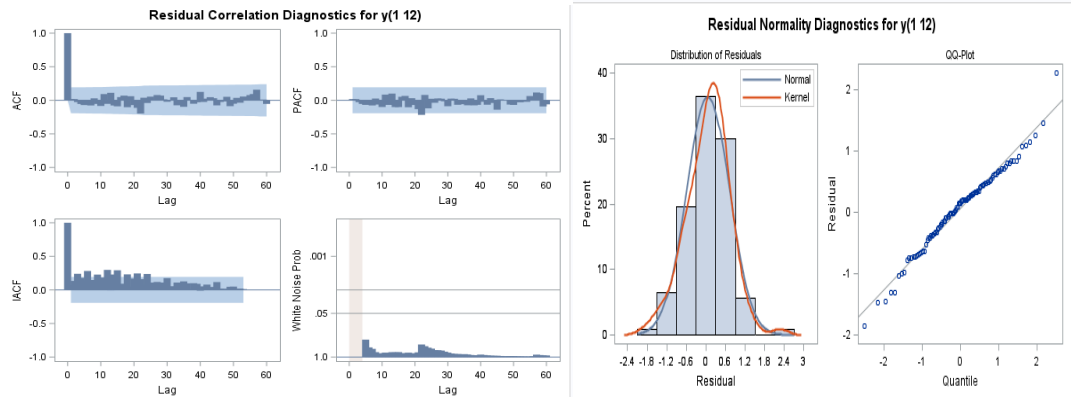

Figure 9. Residual Diagnose for ARIMA(0,1,(1,9,18))x(0,1,1)_12

Residuals of ARIMA(0,1,(1,9,18))x(0,1,1)$_{12}$ seems normally distributed in histogram and QQ-plot (figure 9). All spikes are within the significance limits in residual autocorrelation plots, so the residuals appear to be white noise. The Ljung-Box test also shows that the residuals have no remaining autocorrelations.

The third candidate model is:

$$(1 - B^{12})(1 - B)y_t = (1 - \Theta_1 B)(1 - \theta_1 B + \theta_2 B^{18})w_t$$

And, the fitted model is:

$$(1 - B^{12})(1 - B)y_t = (1 - 0.87968B)(1 - 0.60903B + 0.18523B^{18})w_t$$

A detailed summary of the model can be found in the Appendix.
Then we look at the residual diagnose of the ARIMA (0,1,(1,18))x(0,1,1).

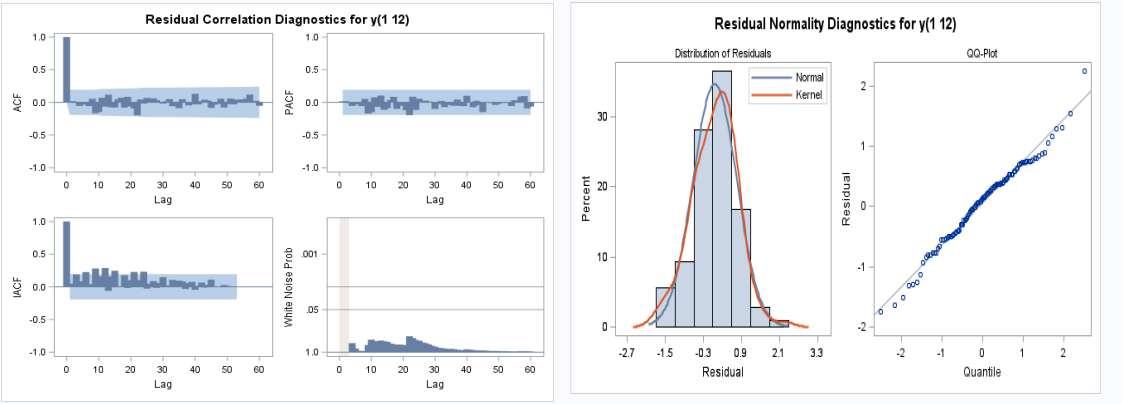

Figure 10. Residual Diagnose for ARIMA(0,1,(1,18))x(0,1,1)_12

Residuals of ARIMA(0,1,(1,18))x(0,1,1)$_{12}$ seems normally distributed in histogram and QQ-plot. Most spikes are within the significance limits in residual autocorrelation plots. It failed the LBP-test at lag 24, but the model can still be the candidate forecasting model, just the prediction intervals may not be accurate due to the correlated residuals.

3) Select the best model among three candidates

MODEL	AIC	BIC	RMSE	LBP TEST$_{(LAG>0.05)}$	SIGNIFICANT VBLES $_{(P<0.05)}$
ARIMA(0,1,1)X(0,1,1)$_{12}$	252.248	257.593	0.71	Lag 6, 12, 48,54, 60 pass	MA at lag 24 is not
SARIMA(0,1,(1,9,18))X(0,1,1)$_{12}$	247.314	258.000	0.67	All pass	AR at lag 12 is not
SARIMA(0,1,(1,18))X(0,1,1)$_{12}$	248.792	256.810	0.70	Lag 24 not pass	All variables are significant

Table 3. Candidates models comparison (all models without intercept)

Among three candidate models, ARIMA(0,1,(1,9,18))x(0,1,1)$_{12}$ is selected as the best model because:

- It has the least AIC (248.79) among the 3 models. Although its BIC (258.00) is slightly larger than the other two models.
- It has a lower standard error (0.56).
- It's the only one who passed the LBP Test at all lags.
- All variables are significant.
- The residuals are white noise. No ACF/PACF is significantly different from zero.
- QQ-plot and histogram of residual distribution indicate the residuals are asymptotically normally distributed.

Model Selection By MAPE

The mean absolute *percentage* error (MAPE) is also often useful for comparing the model, which expresses the prediction accuracy of the forecasting model as a percentage of the error. By computing the MAPE value for the two best models from the previous step using the data from January 2004 to December 2004, ARIMA(0,1,(1,9,18)x(0,1,1)$_{12}$ is selected as the best model with the least MAPE value. Since the percentage error is smaller, the prediction accuracy of the ARIMA(0,1,(1,9,18)x(0,1,1)$_{12}$ model is higher than the other one.

MODEL	MAPE
ARIMA(0,1,(1,9,18))X(1,1,1)$_{12}$	0.01695
ARIMA(0,1,(1,18))X(0,1,1)12	0.01952

Table 4. MAPE values of competing models

Forecast

Forecasts from the best model ARIMA(0,1,(1,9,18)x(1,1,1)$_{12}$ for next year each month (year 2004) are shown in Figure.11. Forecasts have shown in red dashed lines, and blues lines represent upper and lower 95% confidence interval for forecast values.

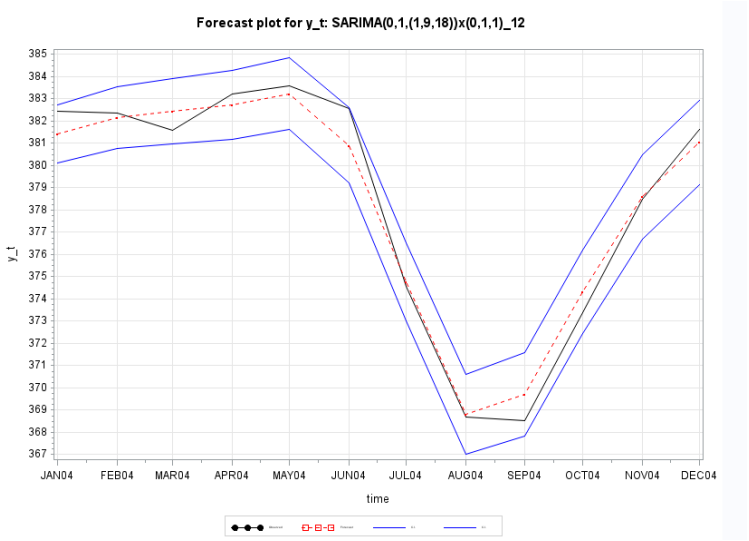

Figure 11. MAPE values of competing models

The forecasts follow the recent trend in the data, reach a peak in May 2004 at carbon dioxide level 383 ppm, and drop rapidly till 368ppm in August then raise again.

Conclusion

The time-series data of carbon dioxide is simple differenced and seasonal differenced for stationary. Under model selection criteria, ARIMA(0,1,(1,9,18)x(1,1,1)$_{12}$ is selected as the best model with least AIC, RMSE, MAPE and the only one passed LBP-test at all lags. According to the selected best model, the next peak of carbon dioxide level next year (in 2004) will be in May again with 384ppm.

Reference

Eleanor Imster and Deborah Byrd. (2019. June 17). Atmospheric CO2 hits a record high in May 2019. Retrieved from https://earthsky.org/earth/atmospheric-co2-record-high-may-2019.

Rebecca Lindsey. (2019. Sep 19). Climate Change: Atmospheric Carbon Dioxide. Retrieved from https://earthsky.org/earth/atmospheric-co2-record-high-may-2019.

Appendix

1) Summary of 25 possible models

	Model	AIC	BIC	LBP Test (lag>0.05)	Insignificant vbles
(2,1,0)x	(2,1,0)12	263.61	274.3	lag 18,24 not pass	significant
(2,1,0)x	(0,1,2)12		264< <268		
(2,1,0)x	(1,1,1)12	257.99	268.68	lag 54,60 pass	AR_12 at lag 12
(2,1,0)x	(0,1,1)12	255.99	264.0169	lag 18,24,30,36,42 not pass	significant
(2,1,0)x	(1,1,0)12	272.97	280.98	lag6,12,48,54 pass	significant
(0,1,2)x	(2,1,0)12		266.87< <280.65		
(0,1,2)x	(0,1,2)12	256.1802	266.8716	lag6,12,54pass	MA at lag2 , MA_12 at lag 24
(0,1,2)x	(1,1,1)12	256.1819	266.8732	lag6,12,54pass	MA at lag2 , AR_12 at lag 12
(0,1,2)x	(0,1,1)12	254.247	262.266	lag6,12,54,60pass	MA at lag2
(0,1,2)x	(1,1,0)12	272.63	280.65	lag6,12,48,54 pass	MA at lag2
(1,1,1)x	(2,1,0)12	262.93	273.62	lag6,12,48,54,60pass	AR at lag1
(1,1,1)x	(0,1,2)12	256.1799	266.8712	lag6,12,54pass	AR at lag1, MA_12 at lag 24
(1,1,1)x	(1,1,1)12	256.1816	266.8729	lag6,12,54pass	AR at lag1, AR_12 at lag 12
(1,1,1)x	(0,1,1)12	254.2476	262.266	lag6,12,54,60pass	AR at lag1
(1,1,1)x	(1,1,0)12	272.4656	280.4841	lag6,12,48,54,60pass	AR at lag1
(1,1,0)x	(2,1,0)12	266.3108	274.3293	lag42,48,54,60pass	significant
(1,1,0)x	(0,1,2)12		269< <274		
(1,1,0)x	(1,1,1)12	261.2034	269.2219	lag60pass	AR_12 at lag 12
(1,1,0)x	(0,1,1)12	259.2603	264.6059	lag60pass	significant
(1,1,0)x	(1,1,0)12	275.7962	281.1418	nopass	significant
(0,1,1)x	(2,1,0)12		262.2223<,	pass	significant
(0,1,1)x	(0,1,2)12	254.1824	262.2009	lag6,12,54,60pass	MA_12 at lag24
(0,1,1)x	(1,1,1)12	254.1838	262.2023	lag6,12,54,60pass	AR_12 at lag 12
(0,1,1)x	(0,1,1)12	252.2477	257.5933	lag6,12,48,54,60pass	significant
(0,1,1)x	(1,1,0)12	272.0084	277.354	lag6pass	significant

2) Summary of ARIMA (0,1,1)x$(0,1,1)_{12}$

Maximum Likelihood Estimation					
Parameter	Estimate	Standard Error	t Value	Approx Pr > |t|	Lag
MA1,1	0.56180	0.07415	7.58	<.0001	1
MA2,1	0.91106	0.20690	4.40	<.0001	12

Variance Estimate	0.50562
Std Error Estimate	0.71107
AIC	252.2477
SBC	257.5933
Number of Residuals	107

Autocorrelation Check of Residuals									
To Lag	Chi-Square	DF	Pr > ChiSq	Autocorrelations					
6	0.41	4	0.9820	-0.001	0.026	0.018	-0.014	-0.048	0.010
12	7.12	10	0.7137	-0.066	0.057	-0.147	-0.129	0.090	-0.047
18	17.83	16	0.3340	0.134	-0.065	-0.053	0.064	0.099	-0.211
24	28.32	22	0.1653	0.012	-0.104	-0.117	-0.219	0.065	-0.005
30	31.48	28	0.2961	0.074	0.016	0.098	0.077	0.021	0.000
36	34.75	34	0.4320	0.020	0.052	-0.059	-0.003	-0.108	0.047
42	41.07	40	0.4234	-0.025	-0.056	-0.028	0.170	-0.010	0.053
48	44.67	46	0.5281	0.042	0.029	-0.101	-0.055	-0.030	0.047
54	47.59	52	0.6479	-0.070	-0.027	0.059	-0.024	0.064	0.014
60	55.46	58	0.5705	0.065	0.066	0.150	0.007	0.029	-0.043

3) Summary of ARIMA $(0,1,(1,9,18))x(0,1,1)_{12}$

Maximum Likelihood Estimation					
Parameter	Estimate	Standard Error	t Value	Approx Pr > |t|	Lag
MA1,1	0.64312	0.08338	7.71	<.0001	1
MA1,2	0.18646	0.08235	2.26	0.0236	9
MA1,3	0.15927	0.08046	1.98	0.0478	18
MA2,1	0.92412	0.22839	4.05	<.0001	12

Variance Estimate	0.450885
Std Error Estimate	0.67148
AIC	247.3136
SBC	258.0049
Number of Residuals	107

Autocorrelation Check of Residuals									
To Lag	Chi-Square	DF	Pr > ChiSq	Autocorrelations					
6	1.04	2	0.5943	0.028	0.026	-0.017	-0.045	-0.062	-0.039
12	4.20	8	0.8391	-0.069	0.034	-0.049	-0.080	0.096	-0.049
18	9.65	14	0.7870	0.113	-0.072	-0.086	0.049	0.070	-0.102
24	17.91	20	0.5931	0.061	-0.068	-0.109	-0.193	0.043	-0.035
30	20.24	26	0.7800	0.059	0.000	0.082	0.072	0.018	-0.014
36	23.12	32	0.8744	0.010	0.052	-0.067	-0.001	-0.089	0.053
42	26.69	38	0.9156	0.008	-0.044	-0.023	0.129	-0.018	0.032
48	29.11	44	0.9591	0.054	0.035	-0.071	-0.022	-0.015	0.054
54	32.35	50	0.9750	-0.085	-0.049	0.037	-0.034	0.058	0.018
60	43.13	56	0.8961	0.084	0.100	0.164	0.014	0.018	-0.046

MAPE value for SARIMA (0,1,(1,9,18))x(0,1,1)_12 for logY without Intercept

Obs	_TYPE_	_FREQ_	mape3
1	0	12	.001695085

4) Summary of ARIMA $(0,1,(1,18))x(0,1,1)_{12}$

Maximum Likelihood Estimation					
Parameter	Estimate	Standard Error	t Value	Approx Pr > |t|	Lag
MA1,1	0.60903	0.07452	8.17	<.0001	1
MA1,2	0.18523	0.08305	2.23	0.0257	18
MA2,1	0.87068	0.15249	5.71	<.0001	12

Variance Estimate	0.491467
Std Error Estimate	0.701047
AIC	248.7916
SBC	256.8101
Number of Residuals	107

Autocorrelation Check of Residuals									
To Lag	Chi-Square	DF	Pr > ChiSq	Autocorrelations					
6	0.44	3	0.9314	0.023	0.023	-0.001	-0.046	-0.026	-0.006
12	7.88	9	0.5460	-0.052	0.056	-0.162	-0.145	0.072	-0.064
18	12.75	15	0.6216	0.118	-0.073	-0.074	0.047	0.071	-0.080
24	20.83	21	0.4692	0.084	-0.060	-0.092	-0.195	0.047	-0.015
30	22.39	27	0.7171	0.072	0.005	0.064	0.036	0.005	-0.013
36	24.97	33	0.8408	0.023	0.055	-0.054	-0.005	-0.084	0.051
42	28.70	39	0.8872	0.000	-0.040	-0.026	0.134	-0.026	0.024
48	31.45	45	0.9371	0.047	0.028	-0.082	-0.034	-0.023	0.057
54	34.18	51	0.9661	-0.078	-0.027	0.059	-0.030	0.045	0.002
60	40.21	57	0.9551	0.051	0.063	0.127	0.019	0.029	-0.045

MAPE value for SARIMA (0,1,(1,18))x(0,1,1)_12 for logY without Intercept

Obs	_TYPE_	_FREQ_	mape2
1	0	12	.001951742