CO2 Emmisions Forecasting with Time Series Analysis*

Liam Wall 1007674991

1 Introduction

1.1 Background and Global Context

The monitoring and prediction of carbon dioxide (CO2) and greenhouse gas emissions has played a great role in global policy in the twenty-first century. In 2007, the Intergovernmental Panel on Climate Change received the Nobel Peace Prize for its efforts to increase global awareness of climate change and its scientific foundations. The IPCC releases comprehensive assessments on climate change every five to seven years and continues to highlight the increasing amount of CO2 and greenhouse gases in the atmosphere, the most recent being in March of 2023. The IPCC also denotes a carbon budget, or the threshold amount of greenhouse gases in Earth's atmosphere at which the global temperature will increase by 1.5 degrees Celsius relative to the pre-industrial period (about 1750). The IPCC predicts that we will surpass the carbon budget by around 2050, and maybe even sooner if countries' climate change policy pledges are not met, which I will discuss later. (IPCC 2023).

The greatest producers of greenhouse gas emissions is the burning of coal, oil, and gas for energy, which is driven by China, the United States, and India. The United States and Europe and developing countries rely heavily on natural gas and oil for energy, whereas China, India, and most of Asia rely on coal for energy production. The alternatives besides coal, natural gas and oil are renewable energy sources like solar, wind, and hydro, as well as nuclear power production. Currently, coal is the most polluting energy source, followed by oil and then natural gas. (Ritchie, Rosado, and Roser 2020).

The link between CO2 concentrations and the increasing global temperature has many effects. The most obvious being hotter temperatures, droughts, rising ocean levels, ice cap and glacial melting, and more severe weather systems. These physical effects can have devastating consequences on food production, cause home displacement and these weather systems and storms

*Code and data are available at: https://github.com/Lwall02/co2_forecast.

have the potential to cause destruction to structures. It is widely accepted that we should do our best in order to mitigate the future harmful effects of climate change. And so this leads to one of the core assumptions of this paper: the largest factor and effect of climate change is global policy.

1.2 The Importance of Short-Term Forecasting

That short introduction into the concerns and drivers of climate change research was to highlight the important role of global policy. International agreements like the Kyoto Protocol, later replaced by the Paris Agreement, as well as the EU Green Deal and other smaller or individual country agreements all center around a core idea. In order to meet the carbon budget, that is in order to not raise the global temperature by 1.5 degrees Celsius, we cannot continue to emit greenhouse gas and CO2 into the atmosphere at our current rate. In particular, in order to accomplish this, countries must report their emissions and aim to achieve net zero emissions (typically by 2025). Net zero emissions means that the same amount of CO2 and greenhouse gases produced equals the amount removed from the atmosphere. These agreements also focus on clean energy investment, sector-specific emission goals and pledges, and penalties and carbon taxes. (UNFCCC (2018), Office (2020))

The intervention of international agreements, carbon capture and greenhouse gas removal technologies, and the overall awareness of the effects of climate change means that the environment in which emissions are produced and emitted is constantly changing. What China, the United States, or India chooses to do regarding policy of energy production and emissions (or even global pandemics) can have drastic effects on the annual concentration of greenhouse gases and the rate at which they will continue to increase. For this reason, instead of analysis of the driving factors of CO2 and greenhouse gas emissions, and resulting long term forecasts, which is the work of the hundreds of IPCC researchers and scientists devoted to this work, I will instead employ a time series analysis and analyze the short term trends and forecast. This paper' will develop time series models of greenhouse gas emissions and attempt to make short term predictions. It will combine the benefits of time series and stochastic modelling while not requiring a future knowledge of global policy.

1.3 Research Objectives

This paper's goal is two fold:

1. To learn from a valid model how the trend of CO2 emissions has changed in recent time periods and perhaps why.
2. To determine the amount of confidence we can place in these model predictions and the reasons why.

2 Literature Review

The forecasting of CO2 and greenhouse gas emissions is a well established area of statistical research because of its global effects highlighted by the reasons previously discussed. After looking at many examples of modeling and forecasting in the area of greenhouse emissions, previous research mainly dives into the analysis of the driving factors of greenhouse gas emissions and the resulting forecasts. For example, the work done by Samara, Jeong, and Beaupre (2025) to forecast the United States CO2 emissions employs a multivariate regression model using the polluting factors or different types of energies. Kumar and Jain (2010) and Zhong et al. (2024) use ARIMA forecasting methods for global CO2 and other air pollutant concentrations. We also see a combination of the above methods in Althobaiti (2025), who employs a spectrum analysis and ARIMA hybrid model in order to capture and forecast the local pollutant levels in Bahrain, which is experiencing recent air quality issues.

The differences in the underlying methods of these mentioned research papers is in regard to the focus of the paper. Specifically, Samara, Jeong, and Beaupre (2025) and their regression model aims to be a tool for policy makers with regard to energy use policy in the US. The ARIMA forecasting approach of Kumar and Jain (2010) aims to accurately deliver short term air quality warnings, similar to Althobaiti (2025). And Zhong et al. (2024) works with daily CO2 concentration data to build and refine an ARIMA model with validated prediction results. The conclusion of these papers, and the common result of the research in greenhouse gas emissions forecasting, is that emissions are continuing to increase and it is becoming ever more important to enact and enforce climate change policy.

The focus of this paper differs from some of the above papers in that it is not interested in the analysis of the variables that drive CO2 emissions, like the effects of money or wealth or the changes among the energy sectors. In order to employ a model with those well defined variables, and especially in order to be able to forecast CO2 levels using these predictors, much more research is required. For example, forecasting CO2 levels based on the United State's annual real GDP or based on China's annual coal production would require future predictions, which have no guarantee of accuracy. It is especially difficult when accounting for future policy, natural disasters, war, and so on.

3 Data and Variables

This paper will use a linear regression model with autocorrelated errors. Specifically, I will regress the annual global population against annual CO2 concentrations from 1950 to 2023. Both the population data and the CO2 data are sourced from the Our World in Data: CO2 and Greenhouse Gas Emissions project (Ritchie, Rosado, and Roser (2023), Friedlingstein et al. (2025)). The population estimates are compiled from various sources with processing and integration conducted by the Our World in Data team. The CO2 concentrations are the annual total emissions of carbon dioxide, excluding land-use change, measured in metric tonnes. The entire dataset of the CO2 levels and population counts from 1950-2023 is shown in Figure 1.

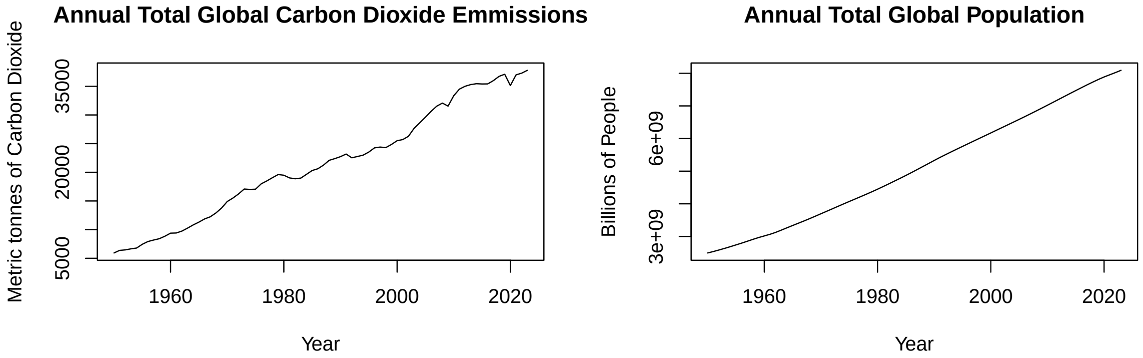

Figure 1: Time series plots of the annual global population and total carbon dioxide levels from 1950 to 2023.

The research into population as a significant factor of rising CO2 levels does not get as much attention as economic and other factors. Twenty years ago there was less focus on population being a significant factor, although publications like Martínez-Zarzoso and Bengochea (2007), and Shi (2001) both find that population has been a driving factor for CO2 levels for individual countries across the globe and the European Union. Shi (2001) goes so far as to say that half of the increase in emissions from 1996 to 2025 will be due population growth. The more recent population-CO2 analyses have mixed results. Zhou and Liu (2016) finds that, in China, income per capita as opposed to population is a better predictor. Similarly, Sulaiman and Abdul-Rahim (2018), finds that economic growth in Nigeria, a developing and rapidly growing country, is also a better predictor. However, on the contrary for developed countries we also find recent studies indicating population as a significant driver of CO2 levels (Dong et al. 2018; Dodson et al. 2020).

4 Modeling Framework

This paper's analysis uses a regression with AutoRegressive Integrated Moving Average (ARIMA) errors model. The regression with ARIMA errors model combines the strengths of standard regression with the time series modeling capabilities of ARIMA. A standard linear regression model assumes that the error term is independent and identically distributed white noise. However, the errors are often correlated. The regression with ARIMA errors model addresses this issue by constructing a two-part model:

- A Regression Component: This part models the relationship between the dependent variable, here this will be the log transformation of the annual CO2 level, against the

annual global population and the indicator function for the COVID-19 pandemic.
- An ARIMA Error Component: The error term from the regression, x_t, is not assumed to be white noise. Instead, it is modeled as an ARIMA(p,d,q) process, which captures its underlying autocorrelation structure.

$$\begin{aligned}
y_t &= \beta_0 + \beta_1 \times \text{population}_t + \beta_2 \times \mathbb{1}_{t\in[2020,2021]}(t) + x_t \\
x_t &\sim \text{ARIMA}(p, d, q) \\
\nabla^d x_t &= \phi_1 x_{t-1} \dots \phi_p x_{t-p} + w_t + \theta_1 w_{t-1} + \theta_q w_{t-q} \\
t &= 1950, \dots, 2023 \quad w_t \sim \text{WN}(0, \sigma_w^2)
\end{aligned}$$

4.1 COVID-19 Indicator

The inclusion of the indicator variable for the years 2020 and 2021 is a deliberate choice for the model trained on the 1950-2021 data. This is justified for two primary reasons: addressing a structural break and enhancing forecast reliability.

First, the COVID-19 pandemic represents a classic exogenous shock to the global economic system and therefore also to CO2 emissions. The resulting lockdowns and economic slowdown caused a sharp, sudden drop in emissions that was not a product of the underlying long-term trend, or any energy or policy related variables. Second, using an indicator variable is essential for improving prediction accuracy and generating sensible forecasts when the prediction interval is so close to the COVID-19 period. It allows the regression component of the model to isolate the pandemic's unique effect. This prevents the emissions drop from being misinterpreted by the model as an element of the trend as opposed to an autocreelated error.

4.2 Model Justification

The reason the regression with ARIMA errors was chosen is because the goal is to build accurate and interpretable short term CO2 predictions. The predictions simply use the trend between population and CO2 levels in combination with the correlated errors to offer forecasts that do not account for potential large future changes. This model helps analyze the current trends and what could happen if no drastic changes are made. A complete discussion of the forecasts will take place later.

5 Data Analysis and Model Fitting

In this section we will fit three models. One trained on population and CO2 data from 1950 to 2007, a second on data from 1950 to 2014, and a third on data from 1950 to 2021. I will call them respecitively, the first, second and third models. We will discuss the purpose and results of the three models in the Discussion section. Firstly, note that we take the log of the annual metric tonnes of CO2 data in order to reduce exponential increases and decreases, reduce variance, and improve the model fit. You can find a comparison of the information criterias of the log transformed and not log transformed models in the appendix. Next, we fit a linear regression model on these training data sets. We can see the fitted line in Figure 2 as well as their residuals plots. As expected, the fitted line is not perfect and we see that the residuals display a clear pattern and are not white noise.

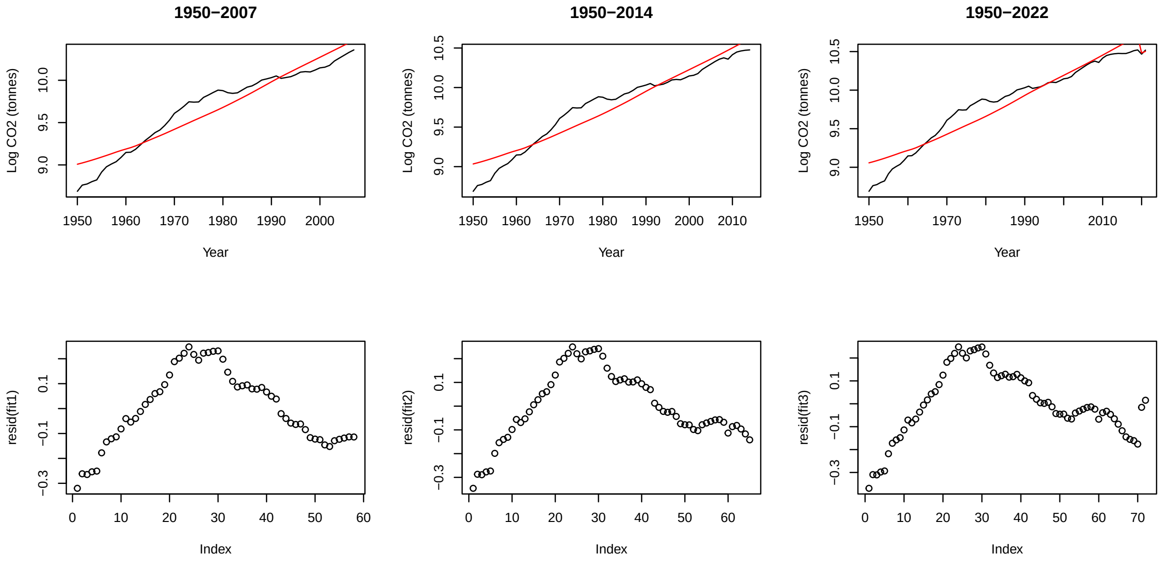

Figure 2: This plot shows the log transformed CO2 data and the fitted linear models on the three training data sets. Below each is their respective residuals plots.

To choose the differencing, autoregressive, and moving average, parameters d, p and q we consider the ACF and PACF plots of the fitted residuals. Refer to the Appendix for a visual of all three model's ACF and PACF plots. After examination we can see the clear choice of model is $d = 2$, where the ACF cuts off at 1 and the PACF tails off. This corresponds to model three having ARIMA$(0, 2, 1)$ errors. Below is the ACF and PACF plot of the third model. Figure 3 shows the resulting twice differenced residuals, as well as the ACF and PACF plots.

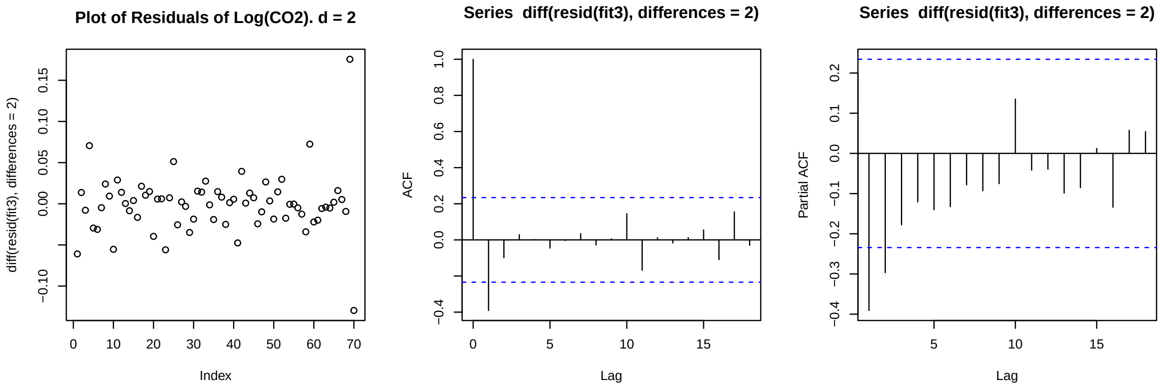

Figure 3: From left to right: residuals of the twice differenced third model, the corresponding ACF plot, and the corresponding PACF plot.

Now that we have chosen the regression with $\text{ARIMA}(0, 2, 1)$ errors model, we need to make sure the model assumptions are satisfied before forecasting. Refer to the Appendix for complete statistics on the residuals of the regression with $\text{ARIMA}(0, 2, 1)$ errors. The residual diagnostics display residual with no significat lags. We also see a QQ plot where the values lie all very close to the diagonal indicating white noise, as well as all values above the threshold on the Ljung Box Test.

6 Forecasting and Results

Now that we have three fitted models over three different training data sets, and we have completed the residual diagnostic tests. We may examine the forecasted results of the fitted model. What we find from the performance of these three models is that both model one and model two have very similar forecasts. We can see, in Figure 4 and Figure 5, that the blue forecasted line is almost the exact same in both models. Most noticeably, we see that is clearly overestimates that observed CO2 levels in the forecasted years. More importantly however, model one and two both significantly over estimate the forecasted value for 2022 and 2023. On the contrary, model 3 very accurately forecasts the post COVID-19 years.

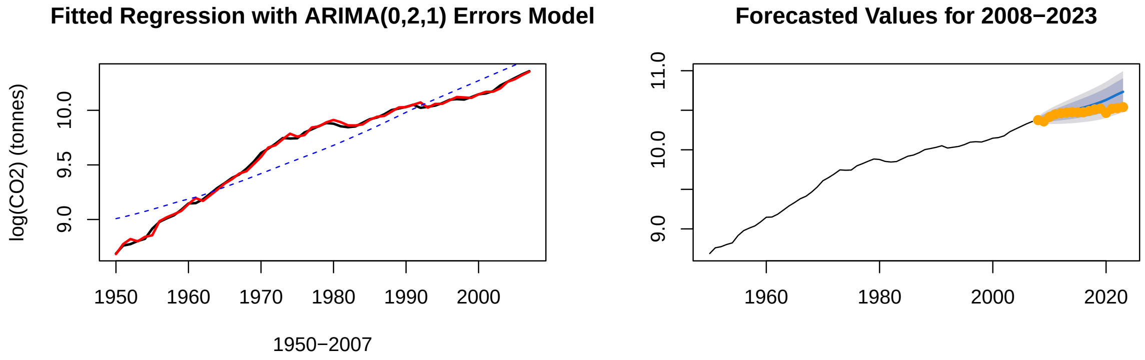

Figure 4: The plot on the left shows the fitted values (red) against the real observed log(CO2) values for model one (black). We can see the linear regression in dotted blue. The plot on the left shows the forecasted values for 2008 to 2023 as well as the observed CO2 levels in orange.

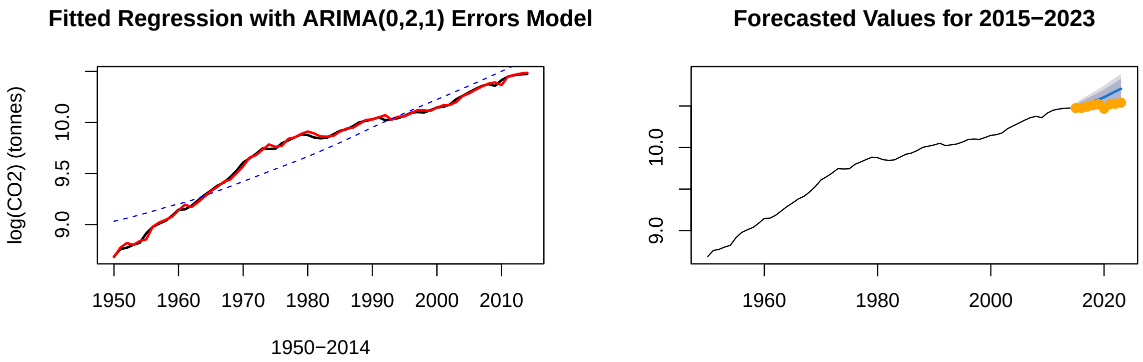

Figure 5: The plot on the left shows the fitted values against the real observed log(CO2) values for model two. The plot on the left shows the forecasted values for 2015 to 2023 as well as the observed CO2 levels in orange.

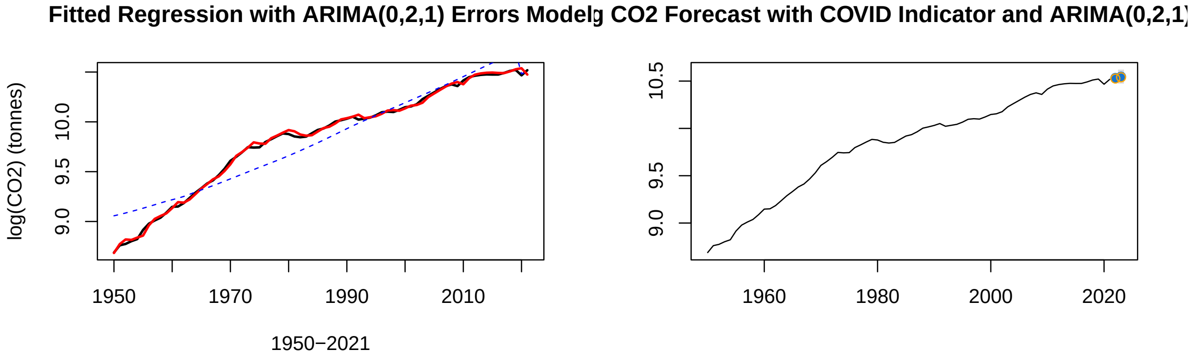

Figure 6: The plot on the left shows the fitted values against the real observed log(CO2) values for model two. The plot on the left shows the forecasted values for 2022 and 2023 as well as the observed CO2 levels in orange.

Table 1: A table comparing the information criterias of fitted models of population against CO2 and log(CO2).

	Fit from 2007	Fit from 2014
(Intercept)	8.226	−7206.507
	(0.062)	(518.878)
pop_train2	0.000	0.000
	(0.000)	(0.000)
Num.Obs.	65	65
R2	0.914	0.979
R2 Adj.	0.913	0.978
AIC	1209.7	1115.2
BIC	1216.2	1121.7
Log.Lik.	32.489	−554.586
F	670.431	2878.311
RMSE	0.15	1228.02

7 Appendix

Log versus not log models of CO2 regressed against population were assesed in order to choose and baseline for the regression with ARIMA errors. We can see in Table 1 that the log models performed much better with respect to information criterias. Note that because our time variable is valued at 1950 to 2023, the β_1 coefficient of population is near 0.

When choosing the parameters we looked at the ACF and PACF plots of the models when $d = 0$, $d = 1$, and $d = 2$. Figure 3, 4 and 5 shows the ACF and PACF plots of the third model with those three differences. The ACF and PACF plots of the first two models behave the same way.

Below are the residual diagnostic plots for the fitted ARIMA(p,d,q) models.

```
initial  value -3.533908
iter   2 value -3.614650
iter   3 value -3.686327
iter   4 value -3.707695
iter   5 value -3.708172
iter   6 value -3.708489
iter   7 value -3.708558
iter   8 value -3.708645
```

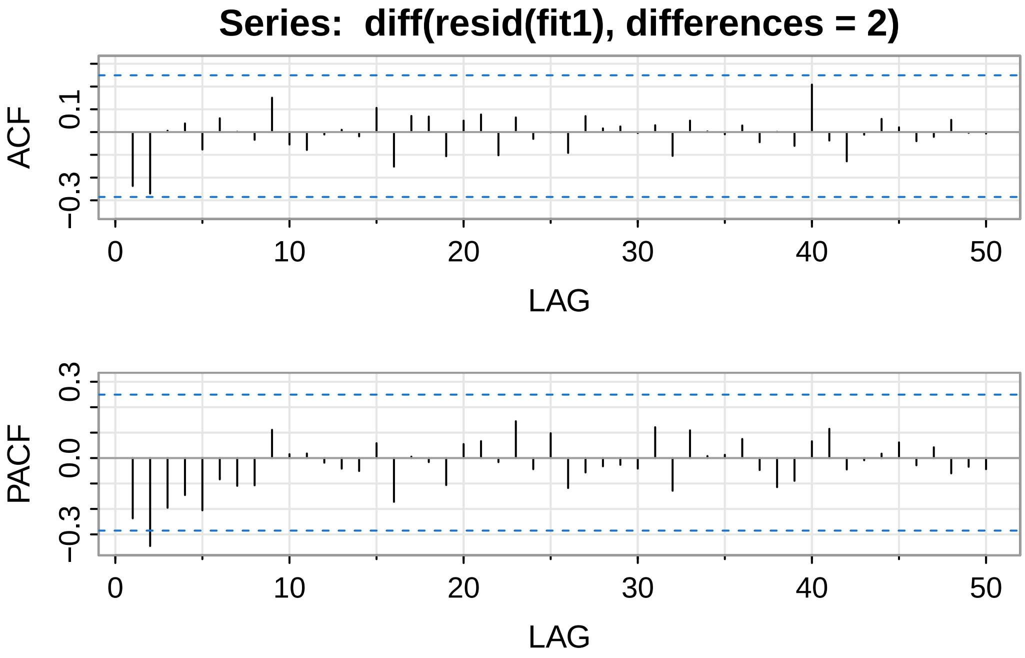

Figure 7: This plot shows the ACF and PACF plots of all three fitted models.

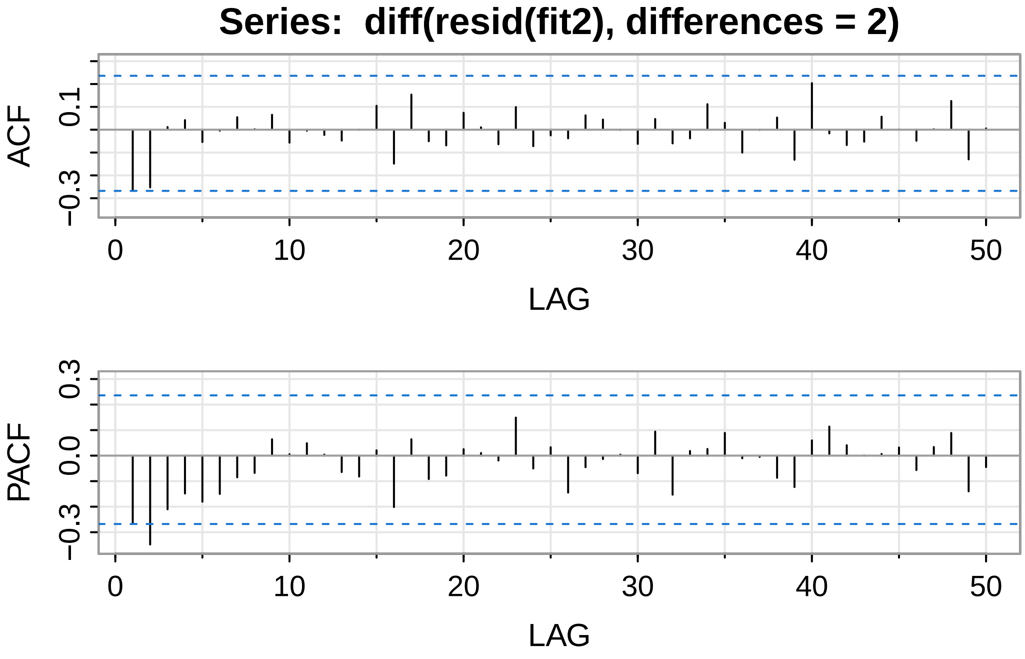

Figure 8: This plot shows the ACF and PACF plots of all three fitted models.

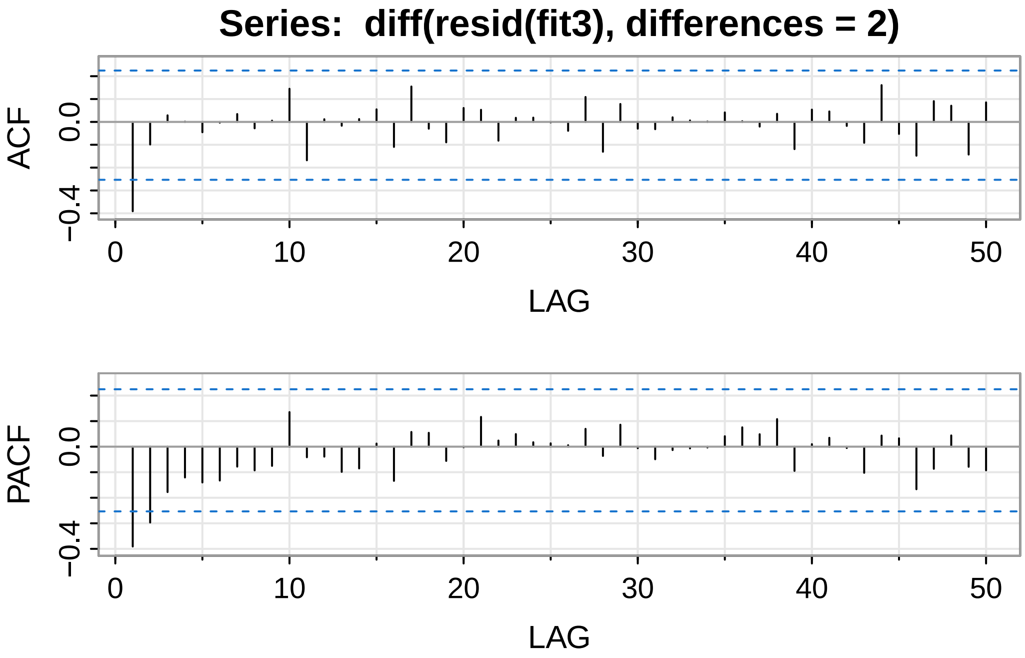

Figure 9: This plot shows the ACF and PACF plots of all three fitted models.

```
iter   9 value -3.708693
iter  10 value -3.708693
iter  10 value -3.708693
iter  10 value -3.708693
final  value -3.708693
converged
initial  value -3.726160
iter   2 value -3.727969
iter   3 value -3.728202
iter   4 value -3.729548
iter   5 value -3.729672
iter   6 value -3.729682
iter   6 value -3.729682
final  value -3.729682
converged

Warning in sqrt(diag(fitit$var.coef)): NaNs produced
Warning in sqrt(diag(fitit$var.coef)): NaNs produced

<><><><><><><><><><><><><><>
```

```
Coefficients:
     Estimate  SE t.value p.value
ma1   -0.8572 NaN     NaN     NaN
xreg   0.0000 NaN     NaN     NaN

sigma^2 estimated as 0.0005652067 on 68 degrees of freedom

AIC = -4.535772  AICc = -4.533213  BIC = -4.439408
```

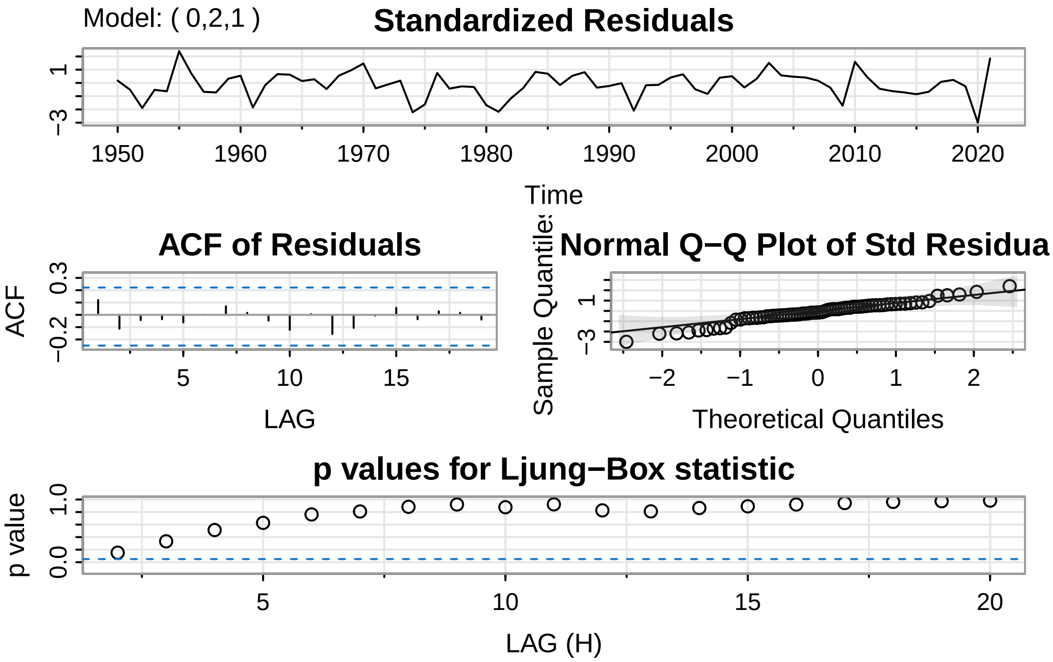

Figure 10

References

Althobaiti, Zahrah Fayez. 2025. "Forecasting Carbon Dioxide Emissions of Bahrain Using Singular Spectrum Analysis and ARIMA Hybrid Model." *bioRxiv*. https://doi.org/10.1101/2025.03.26.645400.

Dodson, Jenna C., Patrícia Dérer, Philip Cafaro, and Frank Götmark. 2020. "Population Growth and Climate Change: Addressing the Overlooked Threat Multiplier." *Science of The Total Environment* 748: 141346. https://doi.org/https://doi.org/10.1016/j.scitotenv.2020.141346.

Dong, Kangyin, Gal Hochman, Yaqing Zhang, Renjin Sun, Hui Li, and Hua Liao. 2018. "CO2 Emissions, Economic and Population Growth, and Renewable Energy: Empirical Evidence Across Regions." *Energy Economics* 75: 180–92. https://doi.org/https://doi.org/10.1016/j.eneco.2018.08.017.

Friedlingstein, P., M. O'Sullivan, M. W. Jones, R. M. Andrew, J. Hauck, P. Landschützer, C. Le Quéré, et al. 2025. "Global Carbon Budget 2024." *Earth System Science Data* 17 (3): 965–1039. https://doi.org/10.5194/essd-17-965-2025.

IPCC. 2023. "Climate Change 2023: Synthesis Report. Contribution of Working Groups i, II and III to the Sixth Assessment Report." *Intergovernmental Panel on Climate Change*.

Kumar, Ujjwal, and V. Jain. 2010. "ARIMA Forecasting of Ambient Air Pollutants (O3, NO, NO2 and CO)." *Stochastic Environmental Research and Risk Assessment* 24 (July): 751–60. https://doi.org/10.1007/s00477-009-0361-8.

Martínez-Zarzoso, Inmaculada, and Aurelia Bengochea. 2007. "The Impact of Population on CO2 Emissions: Evidence from European Countries." *Environmental and Resource Economics* 38 (February): 497–512. https://doi.org/10.1007/s10640-007-9096-5.

Office, Constanze Fetting ESDN. 2020. "The European Green Deal." *EDSN Report*.

Ritchie, Hannah, Pablo Rosado, and Max Roser. 2020. "CO Emissions by Fuel." *Our World in Data*.

———. 2023. "CO and Greenhouse Gas Emissions." *Our World in Data*.

Samara, Kamil, Yunhwan Jeong, and Thomas Beaupre. 2025. "Forecasting CO2 Emission in the US Using Regression Models." *Journal of Data Science and Intelligent Systems*, April. https://doi.org/10.47852/bonviewJDSIS52024482.

Shi, Anqing. 2001. "Population Growth and Global Carbon Dioxide Emission," January.

Sulaiman, Chindo, and A. S. Abdul-Rahim. 2018. "Population Growth and CO2 Emission in Nigeria: A Recursive ARDL Approach." *SAGE Open* 8 (2): 2158244018765916. https://doi.org/10.1177/2158244018765916.

UNFCCC. 2018. "Paris Agreement." *United Nations Climate Change*.

Zhong, Weiyi, Dengshuai Zhai, Wenran Xu, Wenwen Gong, Chao Yan, Yang Zhang, and Lianyong Qi. 2024. "Accurate and Efficient Daily Carbon Emission Forecasting Based on Improved ARIMA." *Applied Energy* 376: 124232. https://doi.org/https://doi.org/10.1016/j.apenergy.2024.124232.

Zhou, Yang, and Yansui Liu. 2016. "Does Population Have a Larger Impact on Carbon Dioxide Emissions Than Income? Evidence from a Cross-Regional Panel Analysis in China."

Applied Energy 180: 800–809. https://doi.org/https://doi.org/10.1016/j.apenergy.2016.08.035.