2024 SICASH
SLIIT INTERNATIONAL CONFERENCE ON ADVANCEMENTS IN SCIENCES & HUMANITIES

Forecasting Global Annual Average CO_2 Concentrations

Rasanjali R.P.B*[1], Tharupathi M.D.G[2], Dharmarathne S.R.J.M[3], Weerakoon M.M[4], Peris T.S.G[5]

[1 1,2,3,4,5] *Sri Lanka Institute of Information Technology*

Email address of the corresponding author - *rasanjaleebuddhika7gmail.com

Abstract

This study aims to enhance the accuracy of CO_2 level forecasts, compare the efficacy of different predictive models, and provide insights for policy development. Employing time series and regression analysis techniques, the study uses historical data from global monitoring stations (1979- 2022) to model the annual mean concentration of atmospheric CO_2. The results reveal that the ARIMA (1,1,1) model outperforms the simple linear regression model in predictive accuracy. Nevertheless, the regression model came across a technical problem as residuals are significantly autocorrelated. The Augmented Dickey-Fuller test was applied to ensure stationarity of the first difference of the original series. The model was trained using data from 1979 to 2022 and validated for 2023. The errors of the ARIMA(1,1,1) was found to be white noise. The ARIMA model projected CO_2 concentrations of 419.5, 421.8 and 424.2 for the years 2023, 2024, and 2025 respectively, with a percentage error of just 0.048% for the 2023. In contrast, the corresponding percentage of error for the simple linear regression model was -1.236%. These findings underscore the ARIMA model's superior performance in forecasting future CO_2 levels and its suitability for environmental monitoring and climate change mitigation strategies. This research provides a valuable methodological framework for future atmospheric science studies and informs policy decisions aimed at addressing rising CO_2 concentrations.

Keywords: ARIMA; CO_2; Forecasting; Regression; Time series

Introduction

Each year, human activities release more carbon dioxide (CO_2) into the atmosphere than natural processes can remove, causing a continuous increase in atmospheric CO_2 levels (Schwartz, 2018). In 2023, the global average CO_2 concentration reached a record high of 419.3 parts per million (ppm), marking a 50% increase since the pre-Industrial Revolution era (Lindsey,R. 2024). This dramatic rise is largely due to the burning of fossil fuels like coal and oil, which release carbon that plants sequestered over millions of years (Estes, 2023). It further highlighted that over the past 60 years, the annual rate of increase in atmospheric CO_2 has been about 100 times faster than natural increases observed at the end of the last ice age, 11,000-17,000 years ago (NOAA Global Monitoring Laboratory, n.d.).

The ocean absorbs a significant portion of this CO_2, leading to a drop in pH by 0.1 units, a 30% increase in acidity. Despite the natural "sinks" on land and in the ocean that absorb about half of the CO_2 emissions, they cannot keep up with the volume of emissions, causing the total atmospheric CO_2 to rise annually. CO_2 is Earth's most crucial greenhouse gas, absorbing and radiating heat. Without it, Earth's natural greenhouse effect would be too weak to maintain a global average surface temperature above freezing. The additional CO_2 is amplifying this effect, leading to global temperature increases. In 2021, CO_2 was responsible for about two-thirds of the total heating effect from all human-produced greenhouse gases.

https://doi.org/10.54389/NCIX3883

Furthermore, CO_2 dissolves into the ocean, forming carbonic acid and lowering the ocean's pH, a process known as ocean acidification.

In the recent past, annual CO_2 emissions at the international level were examined from various perspectives by many authors (IPCC, 2023; Vollmer & Eberhardt, 2024). Those models are either complex, or accuracy was very low. In this paper, a simple model is developed to predict annual CO_2 concentrations with high accuracy.

Materials and Methodology

Secondary data

The data utilized in this research originate from the US Government's Earth System Research Laboratory, Global Monitoring Division, and these datasets consist of annual CO_2 concentrations in parts per million (ppm) from 1979 to 2023 (NOAA Global Monitoring Laboratory, n.d.). Data was analysed using Minitab and EViews software. The study employed ARIMA models in EViews software and a simple linear regression model in Minitab software.

Methodology

A regression model determines the relationship between a dependent variable and independent variables. In this research, linear regression uses the equation $Y = \beta_0 + \beta_1 X + \varepsilon$ where Y is the dependent variable, X is the independent variable, β_0 is the intercept, β_1 is the slope and ε is the error term (James, Hastie, & Tibshirani,2013).

The ARIMA model is used for time series forecasting. It combines Autoregression (AR), Differencing (I) to achieve stationarity, and a Moving Average (MA). ARIMA models are denoted as ARIMA (p,d,q), where p is the number of lags, d is the degree of differencing, and q is the order of the moving average (Box, Jenkins, Reinsel, & Ljung, 2015).

Results and Discussion

Descriptive Analysis

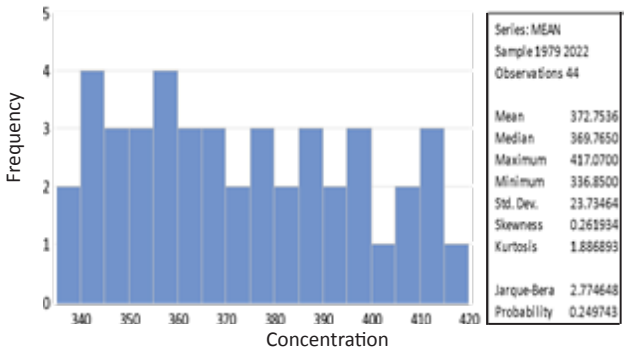

Figure 01. Descriptive Statistics of Annual Average CO_2 Concentration

The annual average CO_2 concentration data ranges from a minimum of 336.85 ppm to a maximum of 417.07 ppm, with a median value of 369.765 ppm, indicating that half of the years have CO_2 concentrations below this level. The mean annual average CO_2 concentration is 372.7536 ppm, with a standard deviation of 23.7346 ppm, showing moderate variability around the mean. The dataset exhibits slight positive skewness, with a skewness value of 0.2619, and the non-significance of Jarque-Bera test (p = 0.2497) suggests that the data does not significantly deviate from a normality. These statistics highlight an overall upward trend in CO_2 levels over the years, with occasional higher concentration outliers.

Fitting a linear regression model

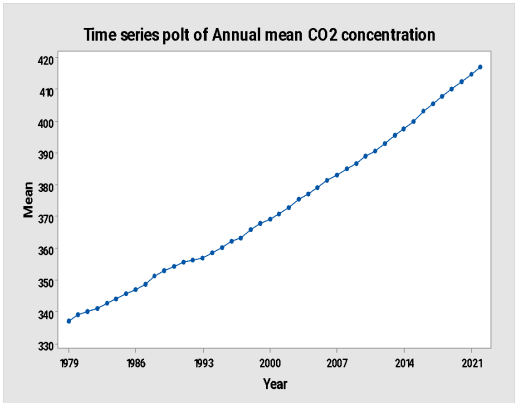

Figure 02 . Annual Average CO_2 Concentration 1979-2022

Figure 02 clearly shows a simple linear relationship between average CO_2 concentration (y) and time(t). This further justified by the highly significant correlation between time and CO_2 ($r = .995$, $p < 0.05$). Based on the regression analysis, the fitted model is $y = -3305.121 + 1.838*t$ ($R^2 = 99\%$). Thus, it can be concluded with 95% confidence that the fitted model explains 99% of the observed variability of average CO_2. The percentage errors for the training set (1979 to 2022) vary between -6.27% and 2.56. The percentage error for 2023 is -1.24% (Table 01).

Table 01. Forecasted Values Using the Fitted Regression Model

Year	Predicted value	Actual value	Percentage error
2023	414.119	419.3	-1.236
2024	415.958	-	-
2025	417.796	-	-

However, Durbin-Watson statistic of 0.656 suggests that errors are not random confirming the fitted regression model is statistical not valid and slight positive autocorrelation in the residuals. Thus, it is necessary to find an alternative approach and we developed the ARIMA model as described below.

Fitting a Time Series Model

Model Selection

The Augmented Dickey-Fuller (ADF) test was applied to the data and value was 4.9760 yielding a p-value of 1, which is greater than the 0.05 significance level. This result indicates that we cannot reject the null hypothesis that the series has a unit root, confirming that the CO_2 concentration data is non-stationary. This implies that the time series has a unit root, and its mean and variance are not constant over time, necessitating an appropriate differencing method for accurate time series modelling and analysis.

Due to the non-stationarity of the original series, we consider the first difference series. The ADF test showed a value of -3.945 and shows that the first

Table 02.

Sample (adjusted): 1980 2022
Included observations: 43 after adjustments

Autocorrelation	Partial Correlation		AC	PAC	Q-Stat	Prob
		1	0.424	0.424	8.2712	0.004
		2	0.255	0.092	11.345	0.003
		3	0.399	0.322	19.062	0.000
		4	0.265	-0.007	22.558	0.000
		5	0.204	0.057	24.668	0.000
		6	0.060	-0.202	24.859	0.000
		7	0.210	0.233	27.227	0.000
		8	0.272	0.092	31.324	0.000
		9	0.125	0.039	32.212	0.000
		10	0.243	0.119	35.679	0.000
		11	0.141	-0.153	36.876	0.000
		12	0.046	-0.091	37.007	0.000

According to the correlogram analysis of the first difference series of the annual average CO_2 concentration data from 1979 to 2022, the Autocorrelation Function (ACF) indicates that the 1st and 3rd lags are significant, while the others are not significant. Similarly, the Partial Autocorrelation Function (PACF) also shows significance at the 1st and 3rd lags, with other lags not showing significant correlations. These findings suggest that there are significant autocorrelations at these specific lags in the first difference series, which play a key role in identifying MA and AR components in time series modelling. Thus, the three parsimonious models (Table 02) were considered.

Model Identification

Table 02. A Comparison of Different Statistics Among the Identified Three Models

Model	Parameter		AIC	SBIC	HQIC	Adj R squared	Log-likehood
	AR 1	MA1					
ARIMA (1,1,0)	Sig.	-	1.6350	1.7579	1.6803	0.1426	-32.1536
ARIMA (0,1,1)	-	Sig	1.6494	1.7723	1.6947	0.1310	-32.4628
ARIMA (1,1,1)	Sig.	Sig	1.6085	1.7722	1.6688	0.1895	-30.5817

Among those possible models (Table 04), There are 3 models with all significant parameters, those being ARIMA (1,1,0), ARIMA (0,1,1), ARIMA (1,1,1). The lowest values of AIC, SBIC, and HQIC, and the maximum log likelihood can be identified from the model ARIMA (1,1,1). Thus, the ARIMA (1,1,1) is the best fitted model. The equation of the best fitted model can be written as,

$(1- B)Y_t = \varepsilon_t(1+ 0.34B+ 0.24 B^2)$. _____ (3)

Date: 06/24/24 Time: 23:09
Sample (adjusted): 1980 2022
Q-statistic probabilities adjusted for 2 ARMA terms

Autocorrelation	Partial Correlation		AC	PAC	Q-Stat	Prob
		1	0.075	0.075	0.2609	
		2	-0.152	-0.158	1.3454	
		3	0.144	0.174	2.3432	0.126
		4	-0.016	-0.076	2.3553	0.308
		5	-0.059	0.003	2.5346	0.469
		6	-0.270	-0.324	6.3420	0.175
		7	0.058	0.161	6.5208	0.259
		8	0.184	0.063	8.3840	0.211
		9	-0.042	0.079	8.4859	0.292
		10	0.223	0.227	11.401	0.180
		11	0.090	-0.025	11.889	0.220
		12	-0.028	-0.018	11.937	0.289

Figure 03. ACF and PACF of the Residuals of the Best Fitted Model

Based on Figure 04, the Q statistic for the residuals' probabilities was not statistically significant ($p > 0.05$). This means there is 95% confidence that the errors are random and uniformly distributed. Additionally, the scatter plot between the predicted values and the residuals showed no systematic pattern, indicating that the residuals have a constant variance.

Dependent Variable: D(MEAN)
Method: ARMA Maximum Likelihood (OPG - BHHH)
Date: 06/24/24 Time: 23:08
Sample: 1980 2022
Included observations: 43
Convergence achieved after 37 iterations
Coefficient covariance computed using outer product of gradients

Variable	Coefficient	Std. Error	t-Statistic	Prob.
C	1.893611	0.347431	5.450324	0.0000
AR(1)	0.934621	0.130865	7.141872	0.0000
MA(1)	-0.720619	0.252565	-2.853200	0.0069
SIGMASQ	0.239662	0.050506	4.745256	0.0000

R-squared	0.247461	Mean dependent var	1.865581
Adjusted R-squared	0.189573	S.D. dependent var	0.571011
S.E. of regression	0.514045	Akaike info criterion	1.608455
Sum squared resid	10.30547	Schwarz criterion	1.772288
Log likelihood	-30.58178	Hannan-Quinn criter.	1.668871
F-statistic	4.274848	Durbin-Watson stat	1.840243
Prob(F-statistic)	0.010569		

Inverted AR Roots	.93
Inverted MA Roots	.72

Figure 04. Residual Plot of the Best Possible Model

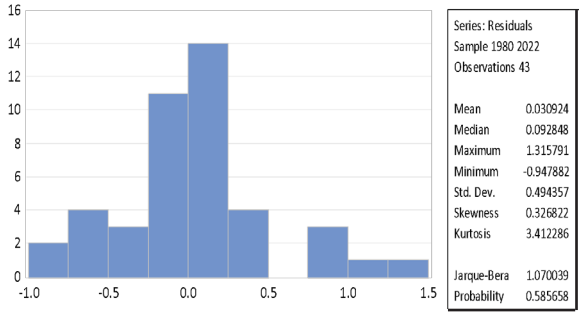

Figure 05. Residual Plot of the Best Possible

Since p-value (p = 0.586) is greater than 0.05, we can accept H0 at 5% level of significance. Therefore, it can be concluded that errors are not significantly deviated from normality.

Heteroskedasticity Test: ARCH

F-statistic	0.789944	Prob. F(12,17)	0.6555
Obs*R-squared	10.73969	Prob. Chi-Square(12)	0.5513

Since the p value (0.2094) of heteroskedasticity test is greater than 0.05, it can be concluded with 95% confidence that there is no ARCH effect. Therefore, it can be confirmed that the variance of the errors is Homogeneous. Hence, it can be concluded that the residuals of the model are white noise.

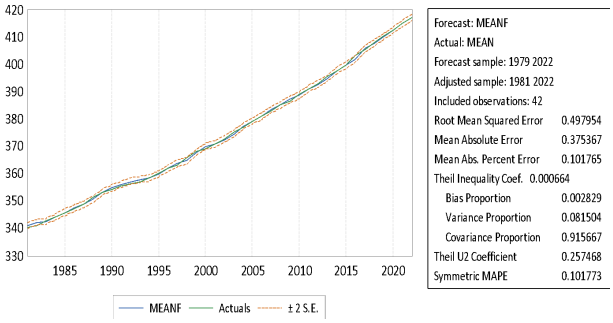

Figure 06. Actual and Forecasted Average CO_2 Concentration

The time series forecast demonstrates a strong alignment between the predicted and actual observed values of annual average CO_2 concentration from 1979 to 2022. The narrow confidence intervals indicate a high level of certainty in the predictions. Performance metrics support the model's accuracy, with a low Root Mean Squared Error (0.4980) and Mean Absolute Error (0.3754) reflecting minor

discrepancies. The Mean Absolute Percent Error (0.1018) suggests moderate accuracy. The high covariance proportion (0.9157) and low bias (0.00283) and variance (0.0815) proportions suggest that most errors are unsystematic. Overall, the model is reliable and provides accurate forecasts of CO_2 concentrations (Nagendrakumar et al., 2021).

Table 03. Forecast of 2023, 2024 and 2025

Year	Forecast value	Actual value	Percentage error (%)
2023	419.5	419.3	0.048
2024	421.8	-	-
2025	424.2	-	-

The last step is to predict the future values using the ARIMA (1,1,1) model. Using the AR and MA components, the CO_2 concentration for 2023, 2024 and 2025 was forecasted as in the table. The small difference and percentage error (0.048%) for 2023 suggest that the ARIMA model used for forecasting is accurate for this time series data. The percentage errors for the training set (1979 to 2022) vary between -2.415% and -0.224%. Overall, the model showed an increasing trend in future values with respect to 2022.

Conclusion

In conclusion, our study aimed to model and predict CO_2 concentrations over time using two statistical approaches: a linear regression model and an ARIMA (1,1,1) time series model. The linear regression model showed a strong linear relationship with an R-squared value of 99%, indicating a good fit with historical data. However, its assumption of a strictly linear relationship limits its accuracy for future predictions, as it does not account for potential changes in trends over time and the percentage error is higher.

To address these limitations, we used an ARIMA (1,1,1) model, which better captures the underlying patterns and fluctuations in CO_2 concentrations. This model includes autoregressive and moving average components, and differencing to ensure stationarity, making it more robust for future predictions by considering temporal dependencies and trends.

In summary, while the linear regression model is effective for explaining past data, the ARIMA (1,1,1) model provides a more reliable method for forecasting future CO_2 concentrations. This highlights the importance of selecting appropriate modelling techniques based on the analysis purpose, with time series models like ARIMA being better suited for future predictions.

Acknowledgement

Gratitude is extended to Sri Lanka Institute of Information Technology, Malabe for providing the opportunity to undertake this research. Appreciation is conveyed to all lecturers in the Department of Mathematics and Statistics, including the Head of the Department, for their invaluable guidance, support, and encouragement throughout this study. The successful completion of this work was made possible by their expertise and insights.

References

Biancalani, Francesco & Gnecco, Giorgio & Metulini, Rodolfo & Riccaboni, Massimo. (2023). Prediction of annual CO2 emissions at the country and sector levels, based on a matrix completion optimization problem. *Optimization Letters. 18.* 2203-2219. 10.1007/s11590-023-02052-2. NOAA Global Monitoring Laboratory (n.d.) *Data Visualization*. https://gml.noaa.gov/dv/data.html

Box, G. E. P., Jenkins, G. M., Reinsel, G. C., & Ljung, G. M. (2015). *Time Series Analysis: Forecasting and Control*. John Wiley & Sons.

Emerson Electric Co. (n.d.*) Using a CO2 Flow Meter for Carbon Dioxide Capture – CO2 Metering*. https://www.emerson.com/en-us/automation/measurement-instrumentation/common-applications/carbon-dioxide-co2-metering

Estes, R. J. (2023). Global change and indicators of social development. *Handbook of Community Practice*: *Global Change and Indicators of Social Development*, *2. http*://dx.doi.org/10.4135/9781412976640.n28

IPCC (2023). *Carbon Dioxide: Projected emission and concentration*.

James, G., Witten, D., Hastie, T., & Tibshirani, R. (2013). *An Introduction to Statistical Learning: with Applications in R*. Springer Science & Business Media.

Lindsey, R. (2024). *Climate Change: Atmospheric Carbon Dioxide.* https://www.globalchange.gov/indicators/atmospheric-carbon-dioxide

Nagendrakumar, N., Lokeshwara, A. A., Gunawardana, S. A. D. C. K., Kodikara, U. P., Rajapaksha, R. W. N. H., & Rathnayake, K. R. M. C. S. (2021). Modelling and Forecasting Tourist Arrivals in Sri Lanka. *SLIIT Business Review, 1*(2), 95-120. https://doi.org/10.54389/GKED9337

Roether, W. (1980). The effect of the ocean on the global carbon cycle. *Experientia, 36*, 1017–25.

Schwartz, S. E. (2018). Resource letter GECC-2: the greenhouse effect and climate change: the intensified greenhouse effect. *Am. J. Phys., 86* (4), 645–56.

Vollmer, M. & Eberhardt, W. (2024). A simple model for the prediction of CO_2 concentrations in the atmosphere, depending on global CO_2 emissions*. Eur. J. Phys*. 4, 1-16.

World Economic Forum (2021, March 22*). Met Office: Atmospheric CO2 now hitting 50% higher than pre-industrial levels*. https://www.weforum.org/agenda/2021/03/met-office-atmospheric-co2-industrial-levels-environment-climate-change/