Forecasting Carbon Emission Using ETS Exponential Smoothing, ARIMA and Regression with ARIMA errors Techniques

Jiali Feng[1,*] and Baoli Huang[2]

[1]University of Malaya, Kuala Lumpur, Malaysia
[2]Guangdong University of Finance & Economics, Guangzhou, China
Email: gillian.feng123@gmail.com (J.F.)
[*]Corresponding author
Manuscript received March 13, 2024; revised April 17, 2024; accepted May 21, 2024; July 11, 2024

***Abstract*—Forecasting Carbon Emissions Using Time Series Analysis Global warming is one of the most difficult and complex problems facing the world today, and forecasting carbon emissions has become a worldwide challenge. In this study, we try to use three models Exponential Smoothing (ETS) model, seasonal ARIMA and error regression Autoregressive Integrated Moving Average (ARIMA) model to train the data of carbon dioxide emissions in a region of the United States from 1990 to 2015, to simulate and forecast the carbon emissions in the United States, and to find out the optimal forecasting model.**

***Keywords*—carbon emissions, Exponential Smoothing (ETS), Autoregressive Integrated Moving Average (ARIMA), Seasonal Autoregressive Integrated Moving Average (SARIMA), ARIMA with error**

I. Introduction

In recent years, climate change has become one of the important issues on the international political agenda. Global warming is one of the toughest and most complex issues facing the world today. Accumulating scientific evidence indicates that increasing concentrations of Greenhouse Gases (GHGs) in the atmosphere since the industrial age have led to rising global temperatures and changes in climate patterns, mainly carbon emissions in the atmosphere [1, 2]. Carbon dioxide (CO_2) is a Greenhouse Gas (GHG), which is mainly derived from human activities and exists in large quantities in the atmosphere, causing global ecological problems and weather change [3].

Since 1750, it is estimated that about two-thirds of anthropogenic carbon emissions (the most important anthropogenic greenhouse gases) come from fossil fuel combustion, and these emissions have continued to increase in recent years. While power generation (including power generation and heat supply) is a fossil fuel One of the main sources of carbon emission produced by combustion. In 2018, carbon emissions from the US power generation sector ranked second in total emissions, reaching 27% [4].

The prediction of carbon emission has become a worldwide problem, because greenhouse gases have the greatest impact on the earth's environmental problems. Predicting carbon emission is also an important key to raising public awareness on how to solve environmental problems [5]. Understanding the past carbon emission path of the United States and making reliable projections of its future emissions is critical. This study attempts to use three models to model and forecast carbon emission in the United States and find out the most effective forecasting model.

In this project, our goal is to apply some time series models to predict the carbon emissions of power generation in the US context, and find out the most accurate forecasting methods, these methods include exponential smoothing method, ARIMA method. and regression with Autoregressive Integrated Moving Average (ARIMA) errors method. The project objectives are:

1) Find the optimal model for applying the Exponential Smoothing (ETS) exponential smoothing method to US carbon emissions.

2) Find the optimal model for applying the ARIMA method to US carbon emissions.

3) Find the optimal model for US carbon emissions using regression with seasonal ARIMA errors method.

4) Find out the most accurate model based on these three methods.

II. Math

A. Sample Data

Data used in this project comes from the U.S. Energy Information Administration (EIA), an independent statistical and analytical agency within the U.S. Department of Energy. It collects, analyzes and publishes information related to energy production, consumption and prices, as well as forecast data related to energy market and policy trends from 1990 to 2015.

B. Definition of Variables

Carbon emissions from electricity generation in the United States refer to the release of carbon dioxide (CO_2) and other greenhouse gases into the atmosphere as a result of producing electricity from various sources. Electricity generation is a significant contributor to greenhouse gas emissions, including carbon emissions, which are a major driver of climate change. In the United States, the primary sources of carbon emissions in electricity generation include fossil fuels such as coal, natural gas, and petroleum. The data is divided into two sections, one for initialization set (1990 Q1 to 2013 Q4), the other for test set (2014 Q1 to 2015 Q4).

The data is available from first quarter of 1990 to fourth quarter of 2015. However, we will only employ data period from first quarter of 1990 to fourth quarter of 2013 as the Initialization dataset to develop the model since we hold 8 test data points (first quarter of 2014 to fourth quarter of 2015) as the out-sample data for the model's forecast performance assessment purpose.

DOI: 10.7763/IJET.2024.V16.1267

C. Empirical Model

This project will use three popular methods to develop time series models for the collected data sets, they are exponential smoothing method, ARIMA method and regression with ARIMA errors method technique. The developed models are then compared and evaluated which has better predictive performance. Finally, the methods used in this project can be assimilated to other datasets by using them, and the rationale for these methods will be illustrated next.

The ETS exponential smoothing method is a more comprehensive and flexible exponential smoothing method for time series forecasting [6]. The core idea is to decompose time series data into error term, trend term and seasonal term, and predict their future values by exponential smoothing. In the ETS method, the smoothness of each component is controlled by a smoothing parameter (smoothing coefficient), which can be adjusted according to the characteristics of the data. In addition, ETS methods can estimate confidence intervals for predictions to provide a measure of forecast accuracy. It has wide applications in time series forecasting, especially for data with obvious trends and seasonality. It can flexibly adapt to different data patterns and provide reliable prediction results. However, choosing an appropriate ETS model and parameters still needs to be evaluated and adjusted according to the specific situation.

The ARIMA model combines a combination of Autoregressive (AR), differencing (I), and Moving Average (MA) to model and forecast various time series patterns [7]. The core idea is to use historical observations and forecast errors to predict future values. It is based on the autocorrelation (autoregressive term) and moving average term of the time series, and removes non-stationarity through difference operations.

The seasonal ARIMA method is an extension of the ARIMA model for dealing with time series data with pronounced seasonal patterns. It incorporates a combination of Autoregressive (AR), differencing (I) and Moving Average (MA) of the ARIMA model, and introduces seasonal differencing and seasonal components to better capture seasonal variations. Seasonally differencing time series data to remove seasonal patterns and transform them into a stationary time series. Then, apply an ARIMA model to the differenced data to build a SARIMA model. Values at future time points can be predicted and restored to the original seasonal timescale. The express equation for SARIMA as followed [8]:

$$\begin{aligned}&(1-\varphi_1 L-\ldots-\varphi_P L^P)(1-\Phi_1 L^S-..\Phi_P L^{P_S})(1-L)^d(1-L^S)^D y_t \\ &=(1+\theta_1 L+\ldots+\theta_q L^q)(1+\Theta_1 L^S+..+\Theta_Q L^{Q_S})\varepsilon_t\end{aligned} \quad (1)$$

Linear and non-linear usually are used to model trend pattern, and non-linear include quadratic trend models and exponential trend. And regression on seasonal dummies is used to model seasonal pattern. Combined with both methods, we can construct the regression model for the data with trend and seasonal components. The express equation for the regression as followed [9]:

Linear trend and Seasonality:

$$y_t = \beta_1 \times t + \sum_{i=1}^{s} y_i D_{it} + v_t \quad (2)$$

Forecast:

$$y_{T+h} = \beta_1 \times (T+h) + \sum_{i=1}^{s} y_i D_{it} + v_{T+h} \quad (3)$$

Quadratic trend & Seasonality:

$$y_t = \beta_1 \times t + \beta_2 \times t^2 + \sum_{i=1}^{s} \beta_i D_{it} + v_t \quad (4)$$

Forecast:

$$y_{T+h} = \beta_1 \times (T+h) + \beta_2 \times t^2 + \sum_{i=1}^{s} \beta_i D_{i,t+h} + v_{T+h} \quad (5)$$

In above equations, denotes a full set of a seasonal dummies. Including the intercept and a full set of a seasonal dummies will lead to perfect mulitcollinearity problem, hence we drop the intercept terms in the above regression modes.

But sometimes error will be ignored, that if error exist autocorrelation problem, it will be affect the forecast accuracy of model and fitness of model. Then ARIMA errors will be adopted to fit the regression model. Regression with ARIMA errors method combines the techniques of regression analysis and ARIMA modeling to deal with regression models with Autoregressive Integral Moving Average (ARIMA) errors. In this method, a regression model is first established to describe the linear relationship between the dependent variable and the independent variable. Then, analyze the regression model's residuals, which are the differences between the observed values and the values predicted by the regression model. Due to the possible autocorrelation and non-stationarity of the residuals, an ARIMA model was used next to capture the temporal dependence of these errors. Integrating the forecast error from the ARIMA model back into the regression model yields a revised forecast or estimate. By combining the regression model with the ARIMA model, the regression and ARIMA error method can more accurately deal with the temporal correlation of residuals and improve the ability to predict future observations.

IV. Empirical Result

A. Data Descriptive

The Fig. 1 shows the monthly data trend of carbon emissions used for electricity generation in the United States from 1990 Q1 to 2015 Q4. It can be seen that the carbon emissions show a repeated pattern or peaks and troughs in the same time interval, and show a trend of rising first and then falling. Repeating patterns, or peaks and troughs, indicate that there is a seasonality to carbon emissions, meaning that emissions may be higher or lower during a particular season or time period. For example, warmer seasons may be accompanied by higher energy demands, such as the use of air conditioning and cooling systems, resulting in higher carbon emissions. On the other hand, the trend of rising first and then falling indicates that carbon emissions show an overall rising and then falling change within a period of time. This likely reflects the influence of a range of factors, such as economic development, energy policy, technological progress, and changes in environmental awareness. At the same time has a systematic trend.

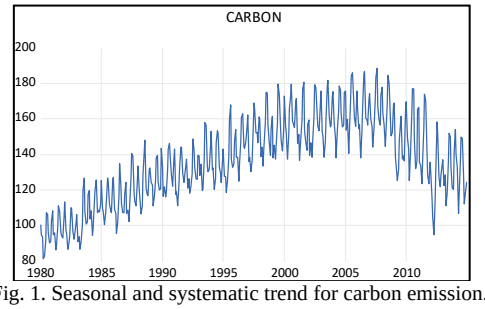

Fig. 1. Seasonal and systematic trend for carbon emission.

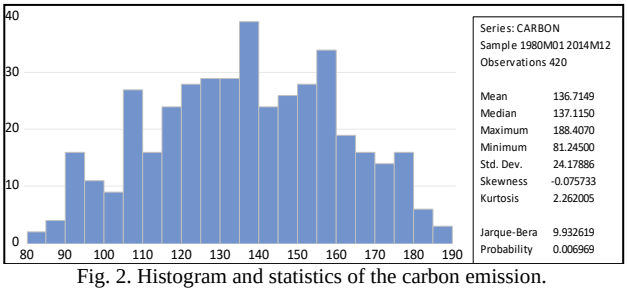

Fig. 2. Histogram and statistics of the carbon emission.

Based on the observation of the Fig. 1, it is confirmed that the multiplicative form is appropriate. And there is obvious systematic trend and seasonal patterns. According to Fig. 2, the data is normal distribution, p-value for Jarque-Bera Test (The null hypothesis: The series is normal distribution) is 0.0069, which less than the 5% significant level.

B. ETS Smoothing Model

We can confirm that the error component is multiplicative (M). As a result, there are 15 different ETS combinations to consider. From the below table, by comparing the AIC values, the best result is determined to be ETS (M,N,A).

ETS (M,N,A) modeling was conducted on carbon data from the first quarter of 1990 to the fourth quarter of 2013. Based on this model, carbon data from the first quarter of 2014 to the fourth quarter of 2015 was predicted, and the forecast accuracy is presented in Table 1. The MAE, MSE, and MAPE values for the prediction results are 35.0917, 2113.6385, and 10.6017, respectively.

Table 1. Result of ETS smoothing

Items	**Seasonal Components**		
Trend	N (None)	A (Additive)	M (Multiplicative)
N (None)	(N,N) [1113.96]	(N,A)* [958.111]	(N,M) [959.038]
A (Additive)	(A,N) [1109.54]	(A,A) [986.508]	(A,M) [962.615]
Ad (Additive damped)	(Ad,N) [1107.74]	(Ad,A) [962.508]	(Ad,M) [964.031]
M (Multiplicative)	(M,N) [1109.56]	(M,A) [963.053]	(M,M) [963.972]
Md (Multiplicative damped)	(Md,N) [1107.75]	(Md,A) [962.615]	(Md,M) [964.115]
Model	MAE	MSE	MAPE
ETS (M,N,A)	35.0917	2113.6385	10.6072
The values in [ ] denote AIC			

C. SARIMA Model

Though the test of correlogram, getting the result as shown in Figs. 3 and 4 ,there exist a trend and a very pronounced seasonal pattern. But the auto-correlations illustrate clearly that the series is non-stationary (the values of auto-correlations stay large). The series is seasonal (the values of auto-correlations at lags 4, 8, and 12 are all larger than their adjacent auto-correlations). Therefore, we apply the seasonal differencing on the series.

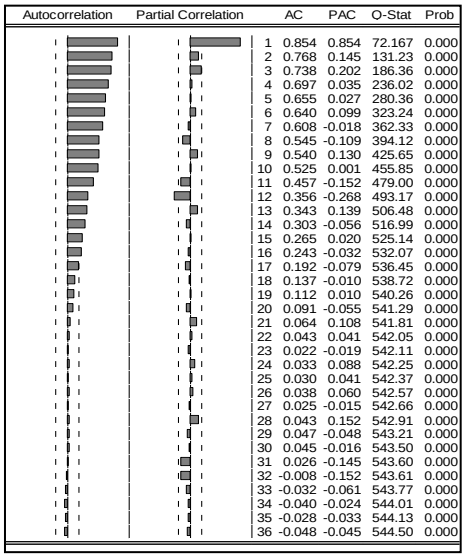

Fig. 3. The Correlogram of SD_CARBON.

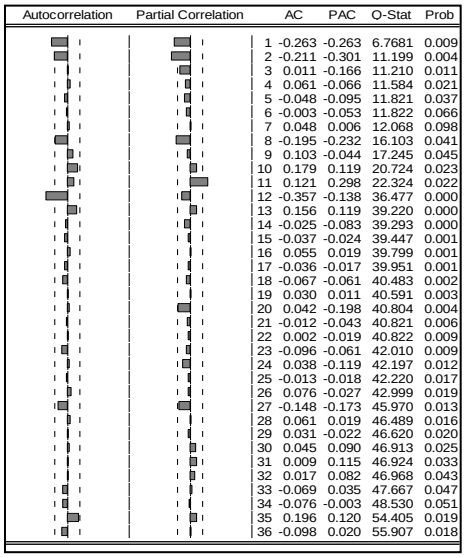

Fig. 4. The Correlogram of D_SD_CARBON.

The seasonal differenced carbon data represents the change in carbon between quarters of consecutive years, the seasonal differenced series is now much closer to being stationary than before. The seasonality is also much less obvious, although still present as shown by spikes at lags 4, 8, and 12 in the PACF. To achieve stationarity, non-seasonal differencing can be applied.

The seasonal difference carbon data after the first difference (D_SD_CARBON) is stationary.

We performed seasonal difference and first difference on the data for one time respectively, the letter “d” denotes non-seasonal differencing order, and “D” denotes seasonal differencing order, hence we confirm that d = D = 1.Then, we have identified the model to be an where values for p, q, P, and Q are yet to be determined.In order to select the optimal model, we constructed a grid table listing all parameter combinations and compared the values of AIC and SIC, choosing the minimum values. Therefore, we set the

maximum values as q = 2 and Q = 3 for a comprehensive examination. Consequently, the following combinations were selected for model testing.

The SARIMA models with the values of "p" in the range of 0 to 2 and the "Q" in the range of 0 to three, and the relevant results are shown in Table 2. Based on AIC and SIC, $ARIMA(0,1,2)(0,1,3)_4$ has the smallest values, which is the best fit model. The model passes the model diagnosis, containing invertible test and stationary test. According to Fig. 5, no root lies outside the unit circle and this report has 96 sample data, focusing on the neighborhood of $\sqrt{T} = \sqrt{96} \approx 10$ is often reasonable. The model is invertible, hence the model is invertible and adequate.

Table 2. Comparison of SMARIMA models

Model	Variable	Coefficient	*p*-value	AIC	SIC
$SARIMA(0,1,1)(0,1,1)_4$	MA(1)	−0.5057	0.0000	8.4949	8.5751
	SMA(4)	0.0582	0.6292		
$SARIMA(0,1,2)(0,1,1)_4$	MA(1)	0.3883	0.0002	8.4716	8.5785
	MA(2)	0.2166	0.0397		
	SMA(4)	0.0982	0.4340		
$SARIMA(0,1,1)(0,1,2)_4$	MA(1)	0.4505	0.0000	8.5043	8.6111
	SMA(4)	0.0448	0.7180		
	SMA(8)	0.1615	0.1659		
$SARIMA(0,1,1)(0,1,3)_4$	MA(1)	−0.2744	0.0082	8.3784	8.512
	SMA(4)	0.0607	0.5142		
	SMA(8)	−0.0989	0.4639		
	SMA(12)	−0.5150	0.0000		
$SARIMA(0,1,2)(0,1,2)_4$	MA(1)	-0.3671	0.0003	8.4885	8.6221
	MA(2)	−0.1972	0.0551		
	SMA(4)	0.0213	0.8681		
	SMA(8)	−0.1027	0.3870		
$SARIMA(0,1,2)(0,1,3)_4$	MA(1)	−0.2383	0.0210	8.3701*	8.5304*
	MA(2)	−0.1873	0.0722		
	SMA(4)	0.0746	0.5510		
	SMA(8)	−0.0774	0.5682		
	SMA(12)	−0.4887	0.0000		

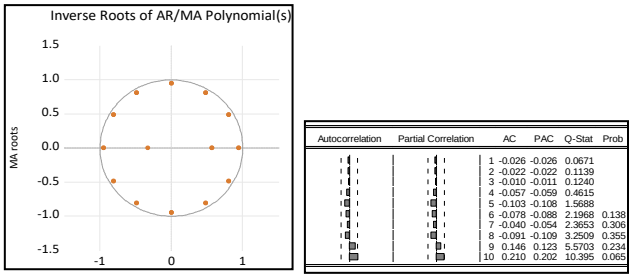

Fig. 5. Model diagnosis of the $SARIMA(0,1,2)(0,1,3)_4$.

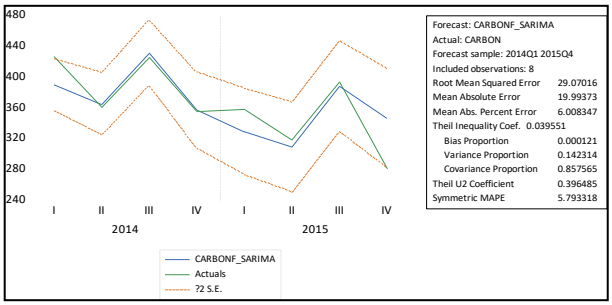

Fig. 6. The forecast of the $SARIMA(0,1,2)(0,1,3)_4$ model.

SMARIMA modeling was conducted on carbon data from the first quarter of 1990 to the fourth quarter of 2013. Fig. 6 represents the predicted values and actual values obtained from the model. Based on this model, the predictions for carbon data from Q1 2014 to Q4 2015 are within the confidence space. The MAE, MSE, and MAPE values are 19.9937, 845.0742, and 6.008347, respectively.

D. Regression with Error ARIMA

We fit the linear trend and the non-linear model to these data and compared with AIC and SIC, we obtain that the non-linear model (quadratic trend) is better. Furthermore, the quadratic trend and seasonality regression suffer positive autocorrelation problem with the low DW statistic (0.514) in Table 3, which suggests the presence of potential significant errors. Therefore, it is necessary to perform error correction on this model. Similarly, we continue to focus on the neighborhood of the root of samples, $\sqrt{T} = \sqrt{96} \approx 10$ from Fig. 7, and *p*-value at the lag 10 is equal to 0.494 which is not smaller than 0.05, in addition the residuals in this model are a white noise process, thus it is an adequate model.

Table 3. Comparison of regression, regression with error ARIMA

Regression Model	Variable	Coefficient	*p*-value	AIC	SIC	DW
Quadratic Trend &Seasonality	@TREND+1	11.5190	0.0000	8.8535	9.0138	0.5144
	$(@TREND+1)^2$	−0.0474	0.0000			
	@QUARTER=1	−223.2068	0.0000			
	@QUARTER=2	−244.2068	0.0000			
	@QUARTER=3	−177.8434	0.0000			
	@QUARTER=4	−224.3191	0.0000			
Quadratic Trend & Seasonality with Error ARIMA	@TREND+1	11.0781	0.0000	8.1064*	8.3202*	1.8664*
	$(@TREND+1)^2$	−0.0458	0.0000			
	@QUARTER=1	−195.9392	0.0337			
	@QUARTER=2	−216.3411	0.0194			
	@QUARTER=3	−149.6815	0.1018			
	@QUARTER=4	−195.8383	0.033			
	AR(1)	0.7412	0.0000			

Regression with ARIMA errors was conducted on carbon data from the first quarter of 1990 to the fourth quarter of 2013. Fig. 8 represents the predicted values and actual values obtained from the model. Based on this model, carbon data from the first quarter of 2014 to the fourth quarter of 2015 was predicted. The MAE, MSE, and MAPE values for the results are 23.2457, 1072.4230, and 6.9823, respectively.

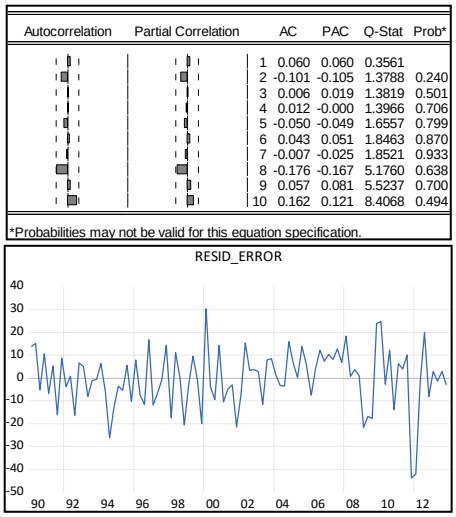

Fig. 7. Model diagnosis of regression with error ARIMA.

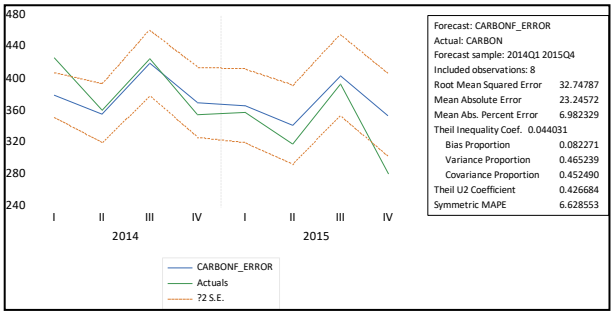

Fig. 8. The forecast of the regression with error ARIMA.

V. Conclusion

The project intends to use 96 time series data observations as the initialization set (from the first quarter of 1990 to the fourth quarter of 2013) to build a forecasting model to predict the monthly emissions of carbon dioxide in the United States. However, 104 time series data observations were actually collected (from Q1 1990 to Q4 2015),The quarterly observations of the last 2 years (the last 8 quarters) are reserved as a test set, which is used to evaluate the out-of-sample forecast performance of the established model, thereby calculating the out-of-sample forecast accuracy of the model.

Next, the project employed different regression techniques such as (1) ETS Exponential Smoothing, (2) ARIMA method and (3) Regression with ARIMA errors techniques to develop different types of forecasting models. Then, the most suitable model will be selected from each regression technique to perform the prediction. Finally, the out-of-sample forecast performance of each model will be compared against each other, so we can know which technique provides the best forecast accuracy for quarterly carbon emissions. You have ensured that all selected models are appropriate models by performing some diagnostic checks. The following shows the appropriate model to choose for each regression technique:

1) Model 1: ETS method: ETS(M,N,A).
2) Model 2: ARIMA method: ARIMA $(0,1,2)(0,1,3)_4$.
3) Model 3: Regression with ARIMA errors.

Table 4. Comparison of the three techniques

	MSE	MAE	MAPE
ETS(M,N，A)	2113.6385	35.0917	10.6072
SARIMA$(0,1,2)(0,1,0)_4$	845.0742	19.9937	6.0083
Regression with Error ARIMA	1072.423	25.2457	6.9823

In addition, according to the comparison of the best model prediction results in the various methods shown in Table 4, it is observed that ARIMA$(0,1,2)(0,1,3)_4$ using seasonality is the best model for predicting quarterly emissions of carbon dioxide. This is because this model has the lowest out-of-sample predicted MAE, MSE, and MAPE values compared to using the other 2 models. It can be seen that the SARIMA model used in this project is the best model for predicting the quarterly emission of carbon dioxide. At the same time, the method can be assimilated to other data sets for prediction.

Conflict of Interest

The authors declare no conflict of interest

Author Contributions

Feng Jiali conducted the research, analyzed the data and wrote the paper. Huang Baoli has contributed to editing and improving this paper. Both authors had approved the final version.

References

[1] W. G. Bonga and F. Chirowa, “Level of cooperativeness of individuals to issues of energy conservation,” *Social Science Research Network*, 2014.
[2] A. Hossain, M. A. Islam, M. Kamruzzaman, M. A. Khalek, and M. A. Ali, “Forecasting carbon dioxide emissions in Bangladesh using Box-Jenkins ARIMA models,” in *Proc. International Conference on Bioinformatics and Biostatistics for Agriculture, Health and Environment*, Department of Statistics, University of Rajshahi, 2017.
[3] IPCC, “Contribution of Working Groups I, II and III to the 5th assessment report of the IPCC,” Synthesis Report, Geneva, 2014.
[4] Environmental Protection Agency, *Inventory of U.S. Greenhouse Gas Emissions and Sinks Fast Facts*, 2020.
[5] L. Abdullah and H. M. Pauzi, “Methods in forecasting carbon dioxide emissions: A decade review,” *Sciences & Engineering*, vol. 75, no. 1, pp. 67–82, 2015.
[6] S. Slawek, “A hybrid method of exponential smoothing and recurrent neural networks for time series forecasting,” *International Journal of Forecasting*, vol. 36, issue 1, pp. 75–85, 2020.
[7] Y. Wang, *Applied Time Series Analysis*, Beijing: Chinese University Press, 2016.
[8] R. Hyndman and G. Athanasopoulos, *Forecasting: Principles and Practice*, OTexts, 2018.
[9] A. M. Bianco, B. M. García, E. J. Martínez, and V. J. Yohai, “Outlier detection in regression models with ARIMA errors using robust estimates,” *J. Forecast.*, vol. 20, pp. 565–579, 2001.