Proceedings of ICEMGD 2025 Symposium: Innovating in Management and Economic Development
DOI: 10.54254/2754-1169/2025.LH26584

Carbon Emission Forecasting in Electric Power Sector in Shanxi Province Based on ARDL Model and Machine Learning Algorithm

Yitong Liu

Taiyuan University of Technology, Taiyuan, China
luochenliu702@gmail.com

Abstract. With the nation's introduction of the "double carbon" objective, given that the power industry's total carbon emissions constitute a major part of these emissions, precise forecasting of these carbon emissions becomes crucial. Consequently, this document develops a combined model utilizing the autoregressive distributed lag model (ARDL) and machine learning algorithm (Transformer) for forecasting Shanxi's power sector's carbon emissions. Shanxi, a significant coal resource province, boasts a precise carbon emission measurement technique, making it an ideal location for gathering the dataset. Initially, data pertaining to Shanxi's electric power sector over the last decade undergoes preprocessing, followed by normalization of yearly electricity production figures. Subsequently, these figures are fed into the time series ARDL model in a suitable manner to examine the long-term and short-term correlations between electricity production and carbon emissions. Ultimately, the residuals are processed by entering the data into the ARDL and Transformer models, leading to the creation of the fusion model. model of fusion. The study demonstrates the hybrid model's extensive benefits in merging predictive precision, capturing long-term trends, and theoretical clarity, making it ideal for policy simulations and intricate energy structure analyses. Consequently, this paper's methodology introduces an innovative approach for forecasting carbon emissions in other coal-dependent provinces, serving as a benchmark for future prediction techniques.

Keywords: carbon emission forecasting, ARDL modeling, machine learning, electricity data, Shanxi Province

1. Introduction

In recent years, the nation has been steadily advancing towards eco-friendly and reduced-carbon growth, with the rise of the “carbon peak, carbon neutral” strategy elevating the precision of carbon emission forecasts to a major national concern [1]. The power sector, a significant contributor to carbon emissions, represents over 40% of the nation's total carbon output. It's both the central and challenging aspect of reaching the "double carbon" target, and is crucial for the advancement of double carbon initiatives. While traditional statistical models such as time series and regression analysis are effective for linear relationships and can yield precise forecasts [2], they face challenges

DOI: 10.54254/2754-1169/2025.LH26584

in handling complex nonlinear relationships [3]; in contrast, machine learning approaches excel at capturing nonlinear dependencies and extracting features, compensating for the limitations of statistical methods. Given Shanxi's status as a significant coal resource province in China, its power industry's carbon emission traits are indicative. With Shanxi's coal reserves making up roughly a quarter of its coal reserves and its thermal power capacity leading the nation, the carbon emission patterns and characteristics during energy use hold substantial importance for other coal resource provinces. Therefore, by precisely forecasting Shanxi's power industry's carbon emissions, it serves as a practical model for other coal resource-dependent provinces to achieve low-carbon development [4]. This document introduces a novel hybrid model, integrating the ARDL and machine learning models, derived from Shanxi province's electric power industry's carbon emission data. Four distinct models are employed for comparative analysis, leading to the selection of the most effective model for forecasting Shanxi province's electric power industry's carbon emissions. Concurrently, Shanxi Province's electricity usage data underwent preliminary processing, including data cleansing, standardization, and time series alteration. Subsequently, the ARDL model was employed to examine the enduring link between electricity usage and carbon emissions, and the Transformer model was utilized to refine the residual segment, enhancing the model's predictive precision accordingly.

2. Literature review

Forecasting carbon emissions is of great importance in the environmental and energy economics and has attracted considerable scholarly interest. Traditional methods of carbon emission forecasting use statistical modeling, including linear regression, time series analysis, etc. Traditional methods of regression analysis usually model the explanatory variables on the right side of the equation. The ARDL model is a time series model, which can observe the fitted relationship between variables. The ARDL model has been used by a large number of scholars as a primary target for carbon emission forecasting. In addition, Pata applied the Bootstrap ARDL method to examine urbanization and renewable energy impacts on emissions [5,6]. However, as the amount of carbon emission data from the electricity sector grows, the complexity of the data grows. With the decrease of the accuracy of traditional statistical models, the technical machine learning method represented by transformers and convolutional neural networks has become one of the hot spots in the field of carbon emission prediction in recent years [7,8]. The transformer model based on the self-attention mechanism can be called the long- and short-term dependent relation model of time series data, which has wide application prospects, such as natural language processing and time series data analysis. There are currently few studies using machine learning methods combined with ARDL models to predict carbon emissions. This paper tries to combine the deep learning model with the traditional econometric model to improve the prediction accuracy and the model's interpretability. In this paper, the residual ARDL model is used to improve the short-term fluctuation emission prediction. Most of the existing literature is on macro carbon emission prediction, but few provincial carbon emission projections are available. Shanxi is a large coal province in China, and the carbon emissions from the power industry in the province are highly representative. Therefore, the combination of ARDL and machine learning methods, especially provincial carbon emissions prediction, needs further study.

DOI: 10.54254/2754-1169/2025.LH26584

3. Methodology

3.1. Research design

This study aims to develop a carbon emissions prediction model that both high interpretability and high accuracy. Shanxi Province, as a coal-based energy base with representative carbon emissions patterns, was selected as the research subject. While traditional statistical models can effectively capture long-term relationships, they struggle to handle nonlinear features; machine learning models excel in this regard but lack economic interpretability. To fully leverage the complementary advantages of both approaches, this study proposes a hybrid ARDL-Transformer framework. By converting annual thermal power generation data from 2015 to 2024 into carbon emissions (using standard emission factors), the study conducted data preprocessing, model construction, and model performance evaluation based on MAPE, RMSE, and MAE.

3.2. Data collection and processing

3.2.1. Interpreting and cleansing data

The initial data table in the article, organized by columnar data, is spread out horizontally annually, while the remaining data follows a descending sequence, including annual and thermal power generation. Missing 2022 data was replaced with minimal values.

3.2.2. Projection of carbon emissions and data unit

For the efficient transformation of power generation data into carbon emissions, employing carbon emission coefficients that align with the National Bureau of Statistics (NBS)'s standard power generation unit is essential. This research involves choosing and transforming Shanxi Province's total thermal power output based on the mean carbon emission factor produced per unit of power generation (kWh). The fundamental equation for conversion reads: Carbon Emission = Electricity Generation × Carbon Emission Factor.

In China's Electric Power Industry's Annual Development Report, the carbon emission factor is based on the mean thermal power generation emission factor, measured in kg/kWh. This transformation allows for the consistent transformation of raw power generation data into analogous carbon emission figures, serving as the focal variable for further analytical modeling.

This study utilizes power generation figures (in billion kWh) to normalize carbon emission calculations, aligning them with IPCC's suggested carbon emission factor (0.991 kg CO_2/kWh), and initially calculates the annual total carbon emissions. Given that Shanxi Province generates approximately 90% of thermal power, it's necessary to modify the overall emission level, leading to the following final formula for calculation:

$$\left[\text{Carbon Emissions } (10^4 \text{ tons } CO_2) = \frac{\text{Power Generation } (10^8 \text{ kWh}) \times 10^8 \times 0.9 \times 0.991}{10^6}\right] \tag{1}$$

3.2.3. Service for normalizing data

Annually, the scale of power production varies significantly. To mitigate the negative impact of diverse quantitative frameworks of power generation and carbon emissions on the model, this study

DOI: 10.54254/2754-1169/2025.LH26584

implements a normalization process. Both variables were normalized to [0,1] to improve training consistency and generalization.

3.2.4. Building of sequence samples

Constructing the sliding window method sequentially involves using data from the past 5 years (2015-2019) as the input, forecasting the carbon emission figures for the 6th year, and continuing in this manner, thereby creating various training examples for the model to understand the carbon emission pattern and timing. The technique enhances the utilization of time series data and promotes the gathering and variety of time-based training samples in the model.

3.3. Hybrid model development

3.3.1. Selection of models

The subsequent phase involves choosing the appropriate model, wherein the author employed both the traditional ARDL model and the recently developed Transformer model in tandem. Each of these models possesses unique advantages and disadvantages, and merging them can leverage their respective strengths to circumvent their shortcomings.

ARDL model: This model primarily aims to investigate the enduring correlation between the use of electricity and carbon emissions. The purpose of the ARDL model is to investigate the enduring link between the use of electricity and carbon emissions, and to demonstrate the impact of electricity usage on the variations in carbon emissions. The ARDL model excels in depicting the enduring balance among variables, owing to the delayed impact of electricity usage and carbon emissions [9].

The Transformer model excels in managing nonlinear connections within time series data. This method effectively identifies both immediate and prolonged data dependencies via the self-attention process, making it ideal for forecasting short-term variations in electricity usage.

Wang and colleagues suggested a hybrid approach using ARDL + Transformer, merging ARDL and Transformer for carbon emission analysis, and confirmed the practicality of this model in managing both long-term patterns and short-term variations. Initially, the study utilizes the ARDL model to explore the long-term dynamics; subsequently, the transformer model merges both long-term and short-term data, optimizing short-term residuals for an optimal amalgamation, and integrating long- and short-term data yields more precise predictive outcomes [10].

3.3.2. Development of the model

This study aims to create a concurrent forecasting model that accounts for both the long-term patterns and short-term variations in carbon emissions by developing a hybrid prediction model that integrates ARDL-Transformer. During the modeling phase, the ARDL model is initially set up to establish a prolonged cointegration link between electricity usage and carbon emissions; subsequently, the subsequent ARDL model is developed with carbon emissions as the dependent factor and electricity generation as the independent factor:

$$[y_t=\alpha+\sum_{i=1}^{p}\beta_i y_{t\text{-}i}+\sum_{j=0}^{q}\gamma_j x_{t\text{-}j}+\varepsilon_t] \tag{2}$$

To depict the temporary variations in the model's residual, the ARDL model's forecast value is derived post-fitting, and this residual is calculated by deducting the forecasted value from the real value:

DOI: 10.54254/2754-1169/2025.LH26584

$$[\varepsilon_t = y_t - \hat{y}_t^{ARDL}] \tag{3}$$

Subsequently, to enhance the precision of short-term predictions, a Transformer neural network was employed for fitting the residual series. To gather sample data, a sliding window is employed, utilizing the residuals from five successive periods as the input sequence for determining the prediction period's residuals. The Transformer model, made up of two layers of Encoder stacking, was also used by Zhang et al. [11] to forecast short-term carbon emissions through the Transformer model refined by the attention mechanism, showcasing its superiority in nonlinear residual modeling. Within this framework, every stratum is made up of a multi-head self-attention and a feed-forward neural network, with inputs being transformed into a stable dimensional realm via linear embedding. The interdependencies among time series are extracted using the multi-head self-attention method, culminating in the derivation of predicted residual values.

During the forecasting phase, the ARDL model's output is overlaid with the remaining correction term forecasted by Transformer, culminating in the hybrid model's ultimate prediction outcome:

$$[\hat{y}_t^{Hybrid} = \hat{y}_t^{ARDL} + \hat{\varepsilon}_t^{Transformer}] \tag{4}$$

The hybrid approach merges the ARDL model's capacity for elucidating enduring structural connections with the Transformer model's proficiency in simulating nonlinear short-term variations, achieving cohesive modeling of carbon emission data across various time frames, thereby greatly enhancing the model's predictive accuracy and functional worth.

3.4. Assessment of the model

To delve deeper into evaluating the predictive accuracy of each model, these typical indicators were primarily mentioned.

The Mean Absolute Percentage Error (MAPE) serves to assess the discrepancy between forecast and actual figures. A lower value indicates improved predictive accuracy of the model.

Root Mean Square Error (RMSE): RMSE quantifies the magnitude of a prediction error, and a higher RMSE indicates a greater error in prediction; yet, more precise model prediction results in a reduced RMSE.

Mean Absolute Error (MAE): This term denotes the mean discrepancy between the forecast and actual values, indicative of the model's total impact. Utilizing these metrics, one can discern the pros and cons of each model, leading to the selection of the most appropriate model for predicting carbon emissions in this project post-evaluation.

4. Model implementation and experimental results

After data preprocessing, the ARDL model, Transformer model, CNN model and hybrid model are realized respectively; from several common error evaluation indexes (MAPE, RMSE, MAE) which are commonly used to compare the performance of the models, experiments are conducted to obtain the results of the model error analysis as shown in Table 1 in order to demonstrate the performance of each model on carbon emission prediction.

Table 1. Model error analysis table

Model	MAPE (%)	RMSE	MAE
ARDL Model	3.07	135.21	133.16
Transformer Model	0.24	12.25	10.17
CNNModel	1.13	55.43	4949 49.49
hybrid model（ARDL + Transformer）	0.31	14.26	13.50

Table 1 shows that the ARDL model captures the long-term relationship between electricity consumption and carbon emissions but struggles to model short-term fluctuations effectively.

Secondly, in contrast, the Transformer model shows a very obvious superiority, in which MAPE = 0.24%, RMSE = 12.25, and MAE = 10.17, thus being able to reflect its better ability to grasp the fluctuations within a short period of time.At the same time,the CNN model performs moderately, capturing local features but lacking in overall predictive accuracy. Among the models listed in this study, the hybrid model performs outstandingly. The MAPE, RMSE, and MAE of the hybrid model are all at a low level, indicating that the hybrid model has higher forecasting accuracy and high interpretability after combining the forecasting ability of the ARDL model for long-period fluctuations and the ability of the Transformer model to correct for short-period fluctuations. This proves that the hybrid model is more advantageous than a single model, and has better structural flexibility and applicability when modeling both long and short cycles within the same model.

5. Conclusion

Utilizing the ARDL model and Transformer neural network, a combined model for forecasting carbon emissions in Shanxi Province's power sector was developed. This model's practicality was evaluated through empirical methods. Benchmarked against standalone ARDL, CNN, and Transformer models, the hybrid system achieves superior accuracy on every metric. By fusing classical econometrics with advanced machine-learning techniques, it overcomes the linear limitations of ARDL and captures the nonlinear dynamics essential to robust forecasting and conceptual innovation. Simultaneously, it seeks to compensate for the limitations of the hybrid model in precisely forecasting sewage volume to bolster current policies and aid Shanxi in reaching the carbon peak carbon neutrality goal. However, this document also presents some limitations. Initially, this study focuses on Shanxi Province's power industry carbon emission data, which is indicative of the broader context, yet its applicability to other provinces and countries requires additional validation. Secondly, despite the high predictive precision of the hybrid model presented in this paper, its practical application demands further refinement for verification. Future work should expand the model geographically—adding more provinces and countries—while advancing along three fronts: (1) embedding external drivers such as climate-change variables; (2) fusing diverse machine-learning techniques to sharpen short-term forecasts; and (3) coupling carbon-emission data with granular low-carbon policy information to strengthen evidence-based policymaking.

References

[1] Wang Q., Su M. (2020) A preliminary assessment of the impact of COVID-19 on environment–A case study of China. Science of The Total Environment, 728: 138915.

DOI: 10.54254/2754-1169/2025.LH26584

[2] Zhang C., Liu Y. (2020) Predicting CO_2 emissions using ARIMA: empirical evidence from China. Energies, 13(16): 4142.
[3] Li Y., Zhao H., Chen M. (2023) Limitations of linear models in nonlinear CO_2 forecasting: a comparative study. Heliyon, 9(5): e14987.
[4] Huang G., Lin J., Zhou J. (2023) A regional carbon emission prediction approach based on electricity–carbon coupling. Energy Reports, 9: 1781–1794.
[5] Shahbaz M., Lean H.H., Shabbir M.S. (2012) Environmental Kuznets curve hypothesis in Pakistan: cointegration and Granger causality. Renewable and Sustainable Energy Reviews, 16(5): 2947-2953.
[6] Rahman M.M. (2017) Modeling CO_2 emissions in South Asia using ARDL approach. Journal of Cleaner Production, 161: 1235-1245.
[7] Li X., Chen W., Li Y. (2022) A hybrid meta-learning and differential LSTM method for industrial carbon emissions prediction. PLOS ONE, 17(9): e0274052.
[8] Zhang Y., Xu J., Huang H. (2023) Transformer-based long-term prediction of carbon emissions under multiple uncertainties. Frontiers in Environmental Science, 11: 1270248.
[9] Rahman M.M. (2018) Long-run dynamics of CO_2 emissions in Bangladesh: An ARDL bound testing approach. Journal of Environmental Management, 210: 379–387.
[10] Liu S., Peng L., Qian Y. (2022) Hybrid ARDL–LSTM model for carbon emission forecasting with macroeconomic indicators. Scientific Reports, 12: 20388.
[11] Fan J., Yu Y., Zhang T. (2023) Integrating carbon market policies into machine learning prediction of CO_2 emissions. Journal of Cleaner Production, 414: 137742.