Pol. J. Environ. Stud. Vol. XX, No. X (XXXX), 1-16
DOI: *10.15244/pjoes/199608* ONLINE PUBLICATION DATE:

Original Research

Incorporating Hybrid Prediction with Feature Selection to Estimate Carbon Emissions with Limited Data

Weiru Wang*, Xueting Cheng, Xiao Chang, Rui Li, Tan Wang

State Grid Shanxi Electric Power Research Institute, Taiyuan, Shanxi 030000, China

Received: 12 July 2024
Accepted: 29 December 2024

Abstract

Carbon dioxide (CO_2) emission forecasting is crucial for efficient carbon reduction management. The majority of carbon emission prediction models are developed based on limited data, which are often collected annually and spatially sparse, and hence face the problems of overfitting and low robustness. Aiming at reliable estimation of CO_2 emissions with a small-scale of data, we propose a CO_2 prediction framework that incorporates a hybrid predictor with feature selection. The hybrid predictor, formed by a fractional-order Grey multivariate model (FGM) and an ensemble learning model, XGBoost, can capture both linear and nonlinear variations of CO_2 emissions, demonstrating strong predictive ability. ReliefF is used for feature selection due to its ability to balance features' importance and diversity, which helps reduce model overfitting. The forecasting effect of the proposed framework is validated on the county-level CO_2 emissions in Shanxi Province, China, from 2012-2022. The results show that the proposed model is superior to other linear and machine learning prediction models and achieves a good forecasting effect, with RMSE, MAE, and R^2 values of 1.73, 1.03, and 0.93, respectively. The Likelihood Ratio (LR) test, the soundness test, and the heterogeneity test have confirmed the generalizability and stability of our proposed hybrid model for CO_2 emissions predictions, as well as the effectiveness of feature selection. Consequently, the prediction results of Shanxi's CO_2 emissions provide a reliable basis for spatial correlation analysis using Moran's Index.

Keywords: CO_2 emissions, feature selection, grey multivariate model, ensemble learning model, spatial autocorrelation.

Introduction

Since September 2020, when China announced its goal of reaching a carbon peak by 2030 and carbon neutrality by 2060, plenty of studies have been conducted to predict national [1,2], provincial [3,4], or municipal carbon dioxide (CO_2) emissions [5,6], and provide strategies for carbon reduction. These studies mainly focused on exploring various influencing

*e-mail: hdwangweiru@163.com
Tel.: +86-351-4263035
Fax: +86-351-4263035

factors and/or developing prediction algorithms for CO_2 emissions, utilizing restricted statistical data.

In terms of forecasting methods, three main categories can be grouped: statistical econometric models, machine learning models, and hybrid models with multiple stages [7-9]. The widely used econometric models include the Grey Model (GM) [9] and its variations [10-11], autoregressive integrated moving average (ARIMA) [12], vector autoregressive (VAR) [13], multivariate linear regression (MLR), and the Bayesian model [14]. Econometric models are often based on economic theory, making it possible to explain the causal relationship behind the prediction. Moreover, some models, such as MLR, can be statistically extrapolated. However, they operate under the premise that the correlation between variables is linear and stable, which cannot always be ensured for carbon emissions fluctuating with complex factors. Machine learning models have the ability to handle complex nonlinear data. The models used for CO_2 emission forecasting include back propagation (BP) neural networks [15-16], support vector regression (SVR) [17], random forest (RF) [8], extreme learning machine (ELM) [18] and long short-term memory (LSTM) [19]. These models can learn from historical data and achieve high-accuracy predictions of CO_2 emissions. However, they require large amounts of data for training. When the data is limited, they easily overfit, resulting in low generalization. Multi-stage hybrid models combine different methods to achieve better prediction performance, which is often done through multiple stages. Lin et al. [20] combined the multivariable Grey Model (GM) with genetic programming (GP) in the second stage to lower the forecasting error. Wen and Yuan [15] established a BP neural network prediction model with the first step of index quantization selection using random forest and then performance optimization using particle swarm optimization (PSO). Li et al. [21] proposed a complicated hybrid model, DNCHAE, which includes variational mode decomposition (VMD), neural networks, ensemble empirical model decomposition, error correction, and least squares support vector and runs diverse optimization for each step in order to reach excellent performance in terms of evaluation metrics. However, such a complex model will be excessively sensitive to parameters.

To develop prediction models of carbon emissions, most studies rely on the annual data of CO_2 emissions and their influencing factors, such as energy consumption, production of fossil fuels, GDP, and population [17, 20, 22]. For example, Qiao et al. [17] forecasted the carbon emissions of 12 countries with 53 data points for each country, of which 48 are used as a training set, and 5 are used as a test set. Some projects, such as Carbon Monitor [23], provide daily carbon emission data, but only at the country level. A significant problem is raised: how can a data-driven model achieve reliable predictions with the limited carbon estimations data?

Feature selection is a crucial step in machine learning and statistics that involves choosing a subset of relevant features to improve model performance, reduce overfitting, and enhance interpretability. Studies have shown that the prediction factors selected based on the random forest can improve the prediction accuracy of carbon emissions [8]. Kong et al. [18] employed a two-stage feature selection composed of partial autocorrelation function (PACF) and ReliefF to select appropriate inputs for the CO_2 prediction model. The feature selection approaches, broadly classified into filter methods, wrapper methods, and embedded methods, are based on distinct mathematical theories and possess unique properties. Therefore, the appropriateness of feature selection methods for a particular task depends on factors such as dataset characteristics, the nature of the problem, and the available computational resources.

By reviewing existing studies, it can be seen that the majority of researchers prioritize the accuracy of the prediction model while neglecting the stability of the forecast [17]. Moreover, prediction models are often established using annually collected and spatially sparse data. Thus, they face the problem of overfitting with a small-scale dataset.

To solve this problem, we propose a CO_2 emissions prediction framework by incorporating a hybrid prediction method with feature selection. Taking multi-source data of NPP/VIIRS nighttime light, socioeconomic statistics, and meteorological data as the input, we first evaluate the potentials of various feature selection methods in determining the most important or relevant features for CO_2 prediction. Then, we establish a hybrid predictor by incorporating a multivariate Grey model with the XGBoost model, which uses the selected features to predict CO_2 emissions. Lastly, Moran's Index is applied to analyze the multivariate spatial pattern of CO_2 emissions. To verify the effectiveness of the proposed framework, we take the carbon emission prediction of Shanxi Province, China, as an example. Shanxi Province is a major coal production and energy chemical base in China [24], ranking fourth in terms of per capita carbon emissions and carbon intensity per unit of GDP in China in 2020 [25]. In this paper, we analyze and predict the state of county-level carbon emissions in Shanxi Province. The predictive model can be used for more granular carbon emissions prediction, providing a scientific basis for developing differentiated carbon reduction policies at the city-county level.

The main contributions of this study are threefold:

(1) We propose a research framework for CO_2 emissions prediction, which can reach accurate predictions with limited historical data. This framework effectively explains the comprehensive associations of CO_2 emissions by determining independent variables from multi-source indicators, which are derived from nighttime light images, socioeconomic statistics, and meteorological records. It realizes accurate prediction by incorporating the abilities of multivariate Grey prediction FGM and ensemble learning XGBoost,

considering both linear and nonlinear relationships between multi-variables and CO_2 emissions.

(2) We comprehensively study the significance of feature selection for CO_2 prediction with small data. The results raise awareness about the diversity of feature selection, and key feature selections importantly determine the performance of feature selection techniques. We point out that Relief is an adequate feature selection method in this study.

(3) We thoroughly analyze the prediction model performance through statistical tests, including the LR test, the soundness test, and the heterogeneity test. The LR test confirms the effectiveness of feature selection. The soundness test and heterogeneity test confirm that our model has good generalization performance and prediction stability.

Material and Methods

Data Sources

The research data used in this study are sourced from multi-modal data of Shanxi Province, China, from 2012 to 2021. Shanxi Province is situated between latitudes 34°36′N and 40°44′N and longitudes 110°15′E and 114°32′E. Covering a total area of 156,700 square kilometers, the province had a resident population of 34.8 million people as of the end of 2022. Administratively, Shanxi is divided into 11 prefecture-level cities and 117 county-level units. Shanxi experiences four distinct seasons, notable climate differences between the north and south, and significant diurnal temperature variations, ranging from 4.2°C to 14.2°C. Shanxi is abundant in mineral resources and holds an important position in the national mining economy as a resource-rich province. The province's economic and industrial structure has long relied on coal, resulting in significant total carbon emissions and high emissions intensity. Moreover, the province faces prominent issues related to energy resource structure [26].

The data sources include nighttime lights imagery, administrative boundary vector maps of Shanxi Province, socioeconomic statistical data, meteorological data, and county-level carbon emissions data. Table 1 summarizes the sources and specific uses of these data.

Data Preprocessing

We preprocessed these data sources and extracted features to construct our dataset. As shown in Table 2, a total of 19 factors were extracted as potential independent variables for CO_2 emissions prediction. The details of data preprocessing are described below.

VIIRS Nighttime Lights Data

We first extract nighttime lights (NL) indexes from remote sensing NL imagery. The NL index, which effectively indicates the intensity of human social activities, has demonstrated a strong correlation with carbon emissions [16,27,28]; therefore, it can reflect regional carbon emission levels.

The annual NL imagery used in this study is from version 2 of the Visible Infrared Imaging Radiometer Suite (VIIRS VNL V2) dataset. The NL images are 500 meters in spatial resolution and produced by the Earth Observation Group (EOG) [29].

Since the study area is Shanxi Province, global NL images need to be cropped only for Shanxi Province. To do so, the global NL images and the administrative boundary vector maps of Shanxi Province are first projected to the GCS_WGS_1984_UTM coordinate system, and then the NL images, which retain only the data of Shanxi Province, are cropped based on the administrative boundary using ArcGIS software. Next, to mitigate the impact of extreme outliers and noise in the NL data, pixel values exceeding a radiance threshold of 472.86 are smoothed using an 8-neighborhood denoising technique, and pixel values below 0.5 are considered background noise and set to zero [30]. Finally, seven NL indexes are computed for each county

Table 1. Multi-source data and their usage.

Data name	Source	Data description	Data usage
Administrative boundary of Shanxi	Department of Natural Resources of Shanxi Province	Shapefile, a digital vector format for storing geometric location and attributes	Defining the administrative units for analysis
Nighttime lights imagery	Earth Observation Group (https://eogdata.mines.edu/products/vnl/#annual_v2)	Annual nighttime lights imagery, spatial resolution ~500m, 2012-2021	Mapping nighttime light indicators to carbon emission for each city and county
County-level socioeconomic data	China Statistical Yearbook (County-level)	Annual statistical data from 2010-2021	Influencing factors of carbon emissions
Meteorological data	Shanxi Meteorological Bureau (http://shanxi. weather. com.cn)	Daily weather data of all counties in Shanxi from 2012-2021	Providing climate factors for each city and county
County-level carbon emissions data	Emissions Database for Global Atmospheric Research (https://edgar.jrc.ec.europa.eu/)	Annual CO2 emissions, 0.1°x0.1°, from 1997-2021	Providing empirical data for carbon emissions prediction

Table 2. Features for carbon emissions prediction.

Source	Indicator	Description
Nighttime light (NL) data	NLsum	Sum of NL digital number (DN) values
	NLavg	Average of DN values
	NLstd	Standard deviation (Std) of DN values
	NLmax	Maximum of DN values
	NLmin	Minimum of DN values
	NLmed	Median of DN values
	NLcount	Number of DN values > 0
Socioeconomic statistical data	Area	Area of a region /km^2
	Pep	Population of a region/ 10^{4}person
	GDP	GDP of a region /10^{4}Yuan
	GDPc	per capita GDP of a region /Yuan
Meterological data	Tmean	Average of air temperature /°C
	Tstd	Std of air temperature
	Hmean	Average of relative humidity /%
	Hstd	Std of relative humidity
	Wmean	Average of wind speed /m/s
	Wstd	Std of wind speed
	Pmean	Average of air pressure /hpa
	Pstd	Std of air pressure

of Shanxi from the denoised NL images, as shown in Table 3. These indexes are simple to compute; here, we only provide the equations for the sum of NL values NLsum, the average of NL, and the total number of pixels for which digital values are greater than zero in (1), (2), and (3), respectively.

$$\mathrm{NLsum} = \sum_{i=1}^{m} (D_i) \tag{1}$$

$$\mathrm{NLavg} = \frac{1}{m} \sum_{i=1}^{m} (D_i) \tag{2}$$

$$\mathrm{NLcount} = \sum_{i=1}^{m} \mathrm{II}(D_i > 0) \tag{3}$$

where, D_i represents the brightness value of the ith pixel in a city or county, m represents the total number of pixels in the corresponding area, and II(·) denotes if the condition in the parentheses is met, it equals 1; otherwise, it equals 0.

Socioeconomic Statistical Data

Socioeconomic statistical data are often annually reported, containing a wide range of information related to a region's economic and social conditions, which can help make informed decisions to promote balanced and sustainable development. We obtained the socioeconomic data for Shanxi Province's regions from the "China Statistical Yearbook (County-level)" and the "China City Statistical Yearbook," published by China Statistics Press [31,32]. Although the data contain tens of data fields, many are not directly correlated with carbon emissions, and significant data gaps exist for various counties and districts. After eliminating some data fields and filling in the missing data by bilinear interpolation, we retain four key statistical indicators: Area, Gross Domestic Product (GDP), population, and per capita GDP (i.e., the ratio of GDP to the total population of the region).

Meteorological Data

The meteorological data of Shanxi province are reported daily and mainly include temperature, air pressure, relative humidity, and wind speed. To maintain the same dimensions as other data, we processed the daily meteorological factors into annual average values

Table 3. The description of feature selection methods.

Method	Criterion	Type
PearsonC	Pearson correlation	Univariate Filter
NMI	Mutual information	Univariate Filter
CFS	Symmetrical uncertainty correlation	Multivariate Filter
mRMR	Maximize relevance and minimize redundancy	Multivariate Filter
ReliefF	Ability to discriminate between instances of different target values	Multivariate Filter (weighting)
RFE	Impact on a prediction model performance	Wrapper
Lasso	L1 Norm regularization	Embedded
RForest	Average decrease in the impurity of the forest	Embedded

for each city and county. We computed the standard deviations of the four indicators, considering that the change of weather may affect or be affected by carbon emissions.

Dataset Construction

We aim to use all the indicators extracted from NL images, statistical data, and meteorological data (as listed in Table 2) to predict CO_2 emissions. Therefore, the data must be on the same scale.

The county-level CO_2 emissions data are sourced from the IEA Emissions Database for Global Atmospheric Research (IEA-EDGAR) v8.0, which can be downloaded from https://edgar.jrc.ec.europa.eu/emissions_data_and_maps. The IEA-EDGAR CO_2 emission data are provided in 0.1deg x 0.1deg annual gridmaps. We processed the gridmaps according to the administrative boundary vector maps of Shanxi Province to generate the CO_2 emissions (in a million tons /Mt) for each county/district of Shanxi. Since the available NL data started in 2012, we also used the CO_2 emissions data from 2012, covering a 10-year period from 2012 to 2021.

Since 2012, there have been some changes in Shanxi's administrative divisions. The number of districts and counties changed from 119 (23 districts, 11 cities, and 85 counties) to 117 (26 districts, 11 cities, and 80 counties). To be consistent, we took the administrative divisions of 2020 as the baseline and integrated the data from previous years according to this baseline. The 11 prefecture-level cities include Taiyuan (TY), Datong (DT), Yangquan (YQ), Changzhi (CZ), Jincheng (JC), Shuozhou (SZ), Jinzhong (JZ), Yuncheng (YC), Xinzhou (XZ), Linfen (LF), and Lvliang (LV). In the end, a total of 117 county-level cities and counties were retained for analysis.

Overall, the dataset is constructed with 1170 data samples, each with 19 features. Therefore, the data from 2012 to 2019 will be used as the training set, while the data from 2020 to 2021 will serve as the test set to evaluate the performance of the carbon emissions prediction model. This study faces the challenge of small-scale available data.

Method

The overall framework of CO_2 prediction is demonstrated in Fig. 1. The process includes four steps: (1) data preparation; (2) feature selection; (3) prediction modeling; and (4) spatial pattern analysis of CO2. The data preparation, as explained in Section "Data Preprocessing", produces a set of indicators to be used as features for predicting CO_2 emissions. Let (x_i,y_i) be for $i = 1,...,N$ be for N independent identically distributed samples, $x_i \in \mathbb{R}^k$ be the k-dimensional vector, and y_i be the true value of CO_2 emissions. The input matrix to the feature selection method is denoted by $X = (x_1, x_2, \ldots, x_N) \in \mathbb{R}^{N\times k}$, and $Y = (y_1, y_2, \ldots, y_N) \in \mathbb{R}^N$. Feature selection filters out a subset of features based on feature importance, which is measured differently using various feature selection methods. We will use $X_p \in \mathbb{R}^{N\times p}, p < k$ to refer to the feature matrix after the feature selection step. Taking X_p as the input, the prediction model will learn from the historical time-series data and forecast future CO_2 emissions. The prediction model's performance is then verified by statistical tests for its effectiveness, generalizability, and stability. Finally, Moran's Index is adopted to analyze the spatial patterns of CO_2 emissions in Shanxi Province.

Feature Selection Algorithms

Feature selection methods are broadly categorized into filter, wrapper, and embedded methods based on their interaction with the prediction algorithm. Filter methods are model-blind and rely on data properties; wrapper methods involve training models and selecting features based on model performance, and embedded methods integrate feature selection into the model training process using regularization or inherent model properties. Feature selection methods can also be considered based on the metric used to determine the feature importance, e.g., distance, correlation,

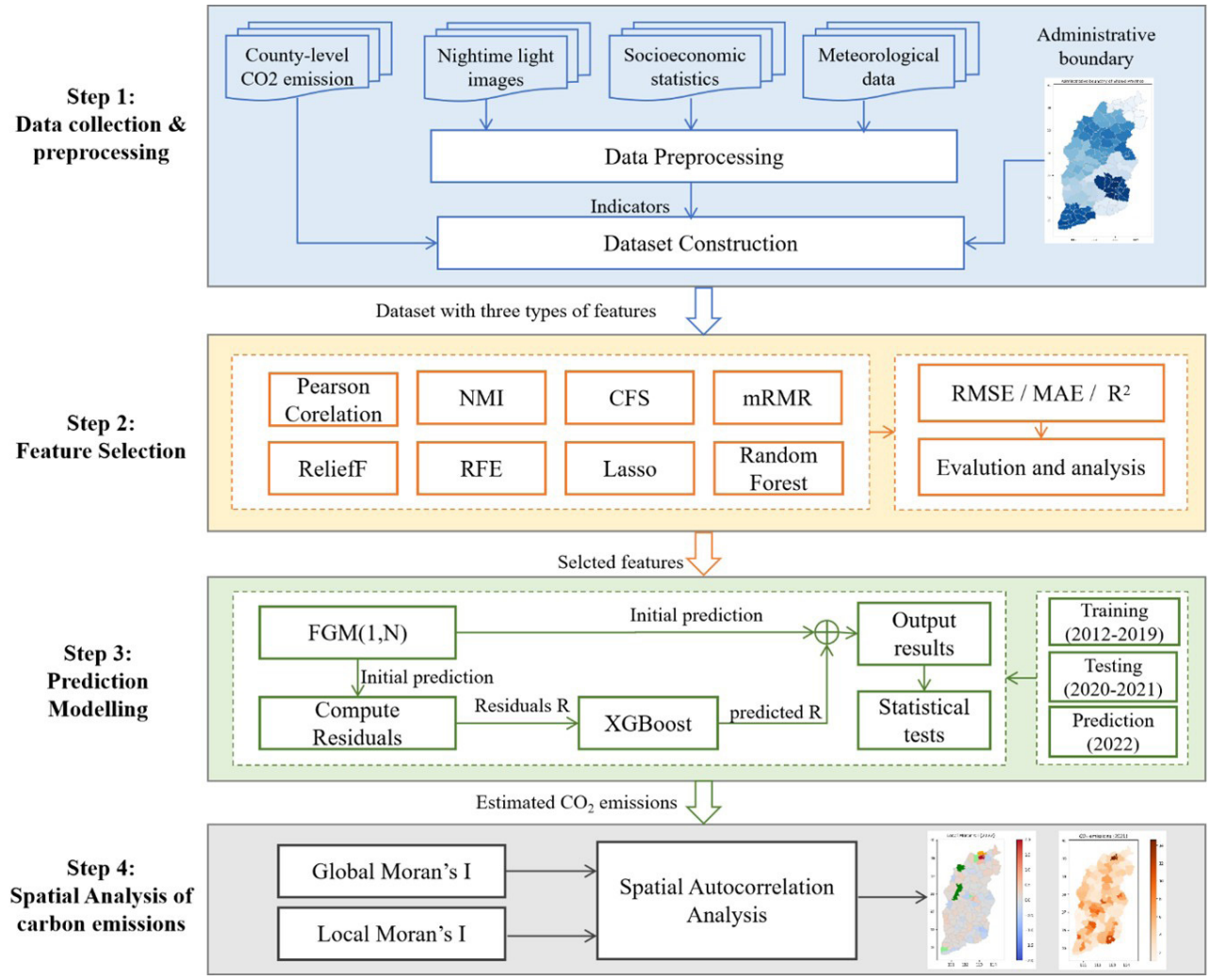


Fig. 1. Flowchart of the methodology.

similarity, information entropy, and weights in linear models. In this study, we investigated eight different feature selection methods, as listed in Table 3.

Pearson coefficient (PearsonC). This is a filter method based on the Pearson coefficient ρ, which is computed between every feature X_i, and the target of CO_2 emission Y:

$$\rho(X_i, Y) = \frac{Cov(X_i, Y)}{\sigma(X_i)\sigma(Y)} \tag{4}$$

where, Cov is the covariance and σ is the standard deviation. The range of ρ is [-1, 1]. A value of ρ close to 0 indicates that the two variables are not correlated.

Normalized Mutual Information (NMI). The NMI method is non-parametric (it does not assume a specific form or distribution) based on entropy estimation. It goes beyond the Pearson correlation, which is a linear model sensitive to outliers. The NMI is defined as:

$$NMI(X_i, Y) = \mathrm{MI}(X_i, Y) + \frac{\mathrm{H}(Y)}{2} \tag{5}$$

where MI is the mutual information, measuring the amount of information obtained about one random variable through another random variable. H denotes the information entropy, and H(Y)=MI(Y, Y). The NMI is between 0 and 1. A value close to 1 indicates a high correlation between the two variables.

Correlation-based Feature Selection (CFS). This is also a multivariate filter method, which selects a feature subset instead of individual features. CFS assumes that a good subset of features should contain highly relevant features to the target and not be related to each other. It uses a symmetrical uncertainty (SU) correlation coefficient to measure the correlation between variables, given by:

$$SU(X, Y) = 2 \times \frac{H(X) + H(Y) - H(X|Y)}{H(X) + H(Y)} \tag{6}$$

Then, a merit metric is computed to rank each subset S containing k features:

$$Merit_s = \frac{k\overline{SU_{yf}}}{\sqrt{k + k(k-1)\overline{SU_{ff}}}} \tag{7}$$

where, $\overline{SU_{yf}}$ means the average SU correlation between the feature ($f \in X$) and the target, and $\overline{SU_{ff}}$ is the average SU feature-feature intercorrelation. CFS aims to find an optimal subset of features that maximize $Merit_s$.

Max-Relevance Min-Redundancy (mRMR). It uses a greedy algorithm to select features that maximize relevance and minimize redundancy. This involves iteratively adding the feature with the highest relevance to the target while ensuring it is minimally redundant with the already selected features. The mRMR score for feature X_i can be expressed as:

$$mRMR(X_i) = MI(X_i, Y) - \frac{1}{|S|}\sum\nolimits_{X_s \in S} MI(X_i, X_S) \tag{8}$$

where $\frac{1}{|S|}\sum_{X_s \in S} MI(X_i, X_S)$ represents the average MI between feature X_i and all the features already in the subset S.

ReliefF. This is a feature weighting algorithm, where the feature weight is defined as the ability to discriminate close observations. Features with weights below a certain threshold will be removed. As ReliefF was originally designed for a multi-class classification problem, it is necessary to discretize the dependent variable carbon emission when using it to select carbon emission factors.

In each step, the Relief method randomly takes data point R from the training set, then finds k, the nearest neighbor of R with the same target class (i.e., NearHit), $H_j (j = 1,2,...,k)$, and k, the nearest neighbor of R with different target classes (i.e., NearMiss), $M_j (j = 1,2,...,k)$. The weights for each feature are then updated as follows:

$$W(X_i) = W(X_i) + \frac{1}{k}\sum_{j=1}^{k}\left[dif\left(X_i, R, H_j\right) - dif\left(X_i, R, M_j\right)\right] \tag{9}$$

where $W(X_i)$ denotes the weights of X_i, dif(X_i,R_1,R_2) refers to the difference between R_1 and R_2 in the feature X_i.

Suppose the distance between R and NearHit on a feature is less than between R and NearMiss. In that case, it indicates that the feature is beneficial in distinguishing the nearest neighbors of the same class from those of different classes, and the weight of the feature is increased. Otherwise, the weight of the feature is reduced. The above process is repeated m times, and finally, the average weight of each feature is obtained.

Recursive Feature Elimination (RFE). RFE, which belongs to the wrapper methods, iteratively removes the least important features based on model performance. It requires an external predictor to assign weights to features (e.g., a linear model or a decision tree). In this study, we used the decision tree regressor. RFE starts with model fitting all features, and at every step, it deletes the least important feature from the subset. That procedure repeats recursively until reaching the desired number of features to select.

Least Absolute Shrinkage and Selection Operator (Lasso). As an embedded method, the Lasso feature selection method is essentially a linear regression model with L1 norm regularization. Because the L1 norm penalization in linear models tends to shrink the feature coefficients of some features to zero, it results in sparse solutions. The features with non-zero weights are the remaining important features.

Random Forest (RForest). RF can be used as a tree-based embedded feature selection method. The RF algorithm creates multiple decision trees using different samples obtained by random sampling with replacement (bootstrap sampling). At each split in the tree, a random subset of features is selected, and the best feature from this subset is chosen to make the split. This process introduces more diversity among the trees. Each time a feature is used to split a node in a decision tree, the algorithm measures the improvement in the splitting criterion (e.g., Gini impurity). The importance of a feature is calculated as the average decrease in impurity over all the trees in the forest.

Prediction Model

In this paper, we consider multi-variable, multi-step prediction for achieving more accurate prediction of CO_2 emissions with limited data. With the selected features as the independent variables, our prediction method combines a fractional-order Grey multivariate model, FGM(1,N) [33], and an ensemble model, eXtreme Gradient Boosting (XGBoost), in order to capture both the linear and nonlinear variants of CO_2 emissions and obtain stable prediction results.

Different from the typical first-order Grey multivariate prediction model GM(1,N), FGM(1,N) is a fractional-order accumulation Grey model that adjusts the time-series data accumulation weights, which reflect the different importance of long- and short-term information in time-series data. Given the time-series CO_2 emission data $\left\{x_1^{(0)}(1), x_1^{(0)}(2), \ldots, x_1^{(0)}(t)\right\}$, and the sequences of the associated features are:

$$\begin{cases} X_2^{(0)}(i) = \left\{x_2^{(0)}(1), x_2^{(0)}(2), \ldots, x_2^{(0)}(t)\right\} \\ \cdots \\ X_{p+1}^{(0)}(i) = \left\{x_{p+1}^{(0)}(1), x_{p+1}^{(0)}(2), \ldots, x_{p+1}^{(0)}(t)\right\} \end{cases} \tag{10}$$

Then, the r fractional-order accumulation operator is:

$$x_j^{(r)}(k) = \sum_{i=1}^{k} W_{k-i+r-1}^{k-i} x_j^{(0)}(i), k = 1,2,\dots,t \tag{11}$$

The weights are calculated as:

$$W_{r-1}^{0} = 1, W_k^{k+1} = 0, W_{k-i+r-1}^{k-i} = \frac{(k-i+r-1)(k-i+r-2)\dots(4+1)r}{(k-i)!} \tag{12}$$

The whitening equation is:

$$\frac{dx_1^{(r)}(k)}{dk} + b_1 x_1^{(r)}(k) = \sum_{j=2}^{p} b_j x_j^{(r)}(k) + u \tag{13}$$

The parameters b_1, b_2, ..., b_{p+1} and u can be solved using the least square method.

$$\begin{cases} x_i^{(r)}(1) = x_i^{(0)}(i) \\ x_i^{(r)}(i+1) = r x_i^{(0)}(i) + x_i^{(0)}(i+1) \end{cases} \tag{14}$$

Although the Grey prediction model has some advantages in the case of small-scale data and has good interpretability, it is based on linear assumptions and expects the relationship between the variables to be linear and describable. This does not always hold true in practical applications. Studies have shown the nonlinear associations between economic factors (e.g., GDP and population) and CO_2 emissions [34]. Moreover, the multivariate Grey model is sensitive to noise and outliers of input data. These noises and outliers may significantly affect the estimation of model parameters, which in turn affects the accuracy of the prediction results. Some advanced GM models have been proposed recently, such as the adjacent accumulation Grey multivariate convolution model (AGMC) [35] and adaptive weighted least squares model (AWLS) [36]. They improve the fitting performance of historical data and enhance the generalization performance of future trends. However, they are still essentially looking for better weights for feature combinations.

To utilize the advantages of the Grey model in capturing time trend information and making up for its weaknesses in nonlinear modeling, we incorporate it with the ensemble model XGBoost. XGBoost is a decision tree-based ensemble learning algorithm that improves the model performance by progressively building multiple weak learners (usually decision trees) and optimizing the errors of the previous step at each step, meanwhile using regularization to reduce model complexity and prevent overfitting.

The mathematical operation of XGBoost can be expressed as:

$$\hat{y}_i = \sum_{l=1}^{L} \alpha_l f_l(x_i) \tag{15}$$

where f_l is the l-th weak learner and α_l is its weight, L is the total number of learners.

The XGBoost aims to minimize the loss function, expressed as the following:

$$\mathcal{L} = \sum_{i=1}^{n} \|y_i, \hat{y}_i\|_2 + \sum_{l=1}^{L} \Omega(f_l) \tag{16}$$

where $\Omega(f_l)$ is a regularization term to control the model's complexity.

Historical data with selected features are used to train the hybrid model. We first solve the FGM using multi-variable time series with a time step of 5 and then use the FGM model to make initial predictions $\tilde{Y}$ on CO_2 emissions. Then, the residuals between the predictions and true values are computed, $R = Y - \tilde{Y}$. Next, taking the preliminary predicted residual as a new feature, XGBoost is trained to model the trend of residuals. With the trained hybrid model, the predicted CO_2 values are obtained by adding the initial predictions by the FGM model to the residual predictions by the XGBoost model.

Spatial Autocorrelation Model

We employ the widely-used Moran's index (Moran's I) to investigate the county-level CO_2 emission patterns from the spatial perspective [37]. Moran's I measures a variable's spatial autocorrelation and indicates whether similar values are clustered, dispersed, or randomly distributed. This is defined as the ratio of the covariance of a single variable between neighboring observations to the variance of that variable within the dataset, given by:

$$Moran's\ I = \frac{n \sum_i \sum_j \delta_{ij} (y_i - \bar{y})(y_j - \bar{y})}{(\sum_i \sum_j \delta_{ij}) \sum_i (y_i - \bar{y})^2} \tag{17}$$

where n is the total number of observations, $\bar{y}$ is the mean value of the variable, y_i and y_j are the values at observations i and j, and δ_{ij} is the spatial weight between i and j. Given a significant level (e.g., 0.05), the larger Moran's I is, the higher the spatial correlation degree.

Both global Moran's I and local Moran's I are used in this study. The former aims to compute the degree of clustering or disparity of CO_2 emissions across Shanxi Province, whereas the latter aims to identify

Table 4. Variable importance measures using different feature selection methods. (Up to six important features are highlighted in bold for each method).

Features	PearsonC	NMI	CFS	mRMR (×1e-1)	ReliefF	RFE	Lasso	RForest (×1e-2)	Times
NLmin	0.08*	0.01	0	-0.46	13	**1**	0	1.98	1
NLmax	0.13**	0.06	0	-0.25	11	**1**	0	2.86	1
NLavg	0.20**	**0.26**	0	-1.08	8	0	0	1.81	1
NLcount	**0.24****	**1.00**	0	0.02	**2**	0	**2.39**	**7.06**	5
NLsum	**0.36****	**0.20**	0	**0.30**	9	**1**	0	2.93	4
NLstd	**0.23****	0.07	0	-1.74	7	0	0	3.01	1
NLmed	0.11**	**0.27**	**1**	-1.91	12	0	0	2.74	2
Area	-0.16**	0.01	0	**24.23**	**1**	**1**	**-0.19**	**13.97**	5
PEP	**0.40****	**1.02**	0	**1.53**	**5**	**1**	**6.59**	**21.53**	7
GDP	**0.44****	**1.22**	0	**0.31**	10	**1**	**6.10**	**13.63**	6
GDPc	**0.30****	0.19	0	-0.87	**6**	0	**1.49**	**9.03**	4
Tmean	0.06	0.03	**1**	**3.87**	**3**	0	0	3.75	3
Tstd	0.02	0	0	-0.11	**4**	0	0	**5.18**	2
Wmean	-0.05	0.03	0	**0.47**	16	0	0	1.95	1
Wstd	-0.06	0	0	-5.72	15	0	0	1.42	0
Hmean	-0.03	0.01	**1**	-6.07	18	0	0	2.17	1
Hstd	-0.04	0	**1**	-2.82	19	0	0	1.26	0
Pmean	-0.05	0.02	**1**	-1.99	17	0	0	2.06	1
Pstd	-0.06	0	**1**	-8.12	14	0	0	1.66	1

Note: *p < 0.05, **p < 0.01

clusters of similar or dissimilar CO_2 emissions at the county level of Shanxi Province. By mapping the values of local Moran's I onto the geographical map, specific areas where spatial autocorrelation is strong can be highlighted, which might be overlooked by global measures. This detailed insight helps governors to target interventions for carbon emissions more effectively.

Results and Discussion

In this section, we first analyze the different behaviors of feature selection methods and the impacts of their selection results on CO_2 emission prediction. Then, we show the performance of our hybrid model with the optimal feature selection. We also carried out a Likelihood Ratio (LR) test, a soundness test, and a heterogeneity test to statistically verify the performance of the proposed hybrid model. Finally, we demonstrate the prediction results of the proposed model and provide a spatial analysis of Shanxi's CO_2 emissions in the last five years.

Evaluation Metrics

The performance of carbon emission prediction is evaluated using the root mean squared error (RMSE), mean absolute error (MAE), and coefficient of determination (R^2) as evaluation metrics. The formulas are as follows:

$$RMSE = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(\hat{y}_i - y_i)^2} \tag{18}$$

$$MAE = \frac{\sum_{i=1}^{N}|\hat{y}_i - y_i|}{n} \tag{19}$$

$$R^2 = 1 - \frac{\sum_{i=1}^{N}(\hat{y}_i - y_i)^2}{\sum_{i=1}^{N}(\bar{y}_i - y_i)^2} \tag{20}$$

where y_i represents the actual carbon emission values at the i-th area, $\hat{y}$ represents the predicted carbon emission values, $\bar{y}$ represents the mean of the actual

carbon emission values, and N represents the total number of areas.

RMSE is useful for evaluating the model's prediction error magnitude and prefers penalizing larger errors more heavily. MAE is less sensitive to outliers compared to RMSE. The smaller the RMSE and MAE, the more accurate the predicted results are. R^2 measures the proportion of the variance explained by the model. The value range of R^2 is [0,1], where a value closer to 1 indicates a better fit of the model to the data.

Analysis of Feature Selection

Table 4 shows the feature importance computed by eight diverse feature selection methods, including univariate filter methods Pearson correlation and NMI, the multivariate filter methods CFS, mRMR, and ReliefF, the wrapper methods RFE, and the embedded methods Lasso and RF. The bold values denote the top features selected by these methods. It should be noted that different methods provide different results. For Pearson correlation, NMI, mRMR, and RF, the feature importance is shown in scores; the higher the score, the more important the feature. For CFS and RFE, "1" and "0" indicate selected and not selected, respectively. For Lasso, the values represent the weights of a linear regression model; weights equal to zero mean the corresponding features are eliminated. ReliefF ranks features in descending order of importance.

Table 4 shows that none of these eight approaches yield identical outcomes, but some common choices are demonstrated. The last column lists the times the corresponding feature was selected. The most selected features are PEP, GDP, Area, NLcount, GDPc, and NLsum.

For the NL indicators, three methods, PearsonC, mRmR, and RForest, take NLsum as the most important feature, while NMI, ReliefF, and Lasso pick NLcount as the top feature. Only NLsum gains a Pearson correlation coefficient $\rho \geq 0.3$ ($p < 0.001$), indicating a statistically moderate positive correlation with CO_2 emissions in Shanxi Province.

All four socioeconomic statistical indicators are highly selected. In particular, PEP, GDP, and GDPc are moderately positively correlated with CO_2 emissions ($\rho \geq 0.3$, $p < 0.001$). However, there may be collinearity between these features. For example, GDPc, as the per capita GDP, is calculated by dividing the GDP by the population PEP. In the Lasso method, this phenomenon is avoided. The population PEP is chosen by seven methods, showing its significance to carbon emissions. Both PearsonC and Lasso show a linear negative influence of Area on CO_2 emissions. This reveals that counties with large areas in Shanxi Province are not necessarily large carbon emitters.

According to the Pearson coefficients, the meteorological indicators have no linear correlations with CO_2 emissions. However, Tmean and Tstd are selected three and two times, respectively. This suggests that climate factors should be considered when predicting CO_2 emissions.

Impact of Feature Selection on Prediction Performance

Dealing with the small-sample-size prediction problem in machine learning, we conducted comparative experiments to assess the performance of different prediction algorithms and their performance with and without feature selection. The compared algorithms include linear regression, support vector regression model (SVR), random forest (RF), XGBoost, and FGM. In this paper, SVR uses the Radial Basis Function (RBF) kernel, realizing nonlinear mapping. RF and XGBoost are both tree-based ensemble learning methods, but RF is based on the bagging strategy and builds trees in parallel, and XGBoost uses the boosting strategy and builds trees sequentially. All the models were trained with the training set from 2012 to 2019 and evaluated on the test set from 2020 to 2021.

Table 5 shows the comparison results of different predictors using the features (up to six) selected by various feature selection methods. The following phenomena can be observed: 1) Overall, the linear regression performs inferior to the others. The value of R^2 is less than 0.22, which is significantly lower than those of other models. This demonstrates that these features are not linearly related to CO_2 emissions. Moreover, the linear regression model performs best using all the features. This is more likely to overfit with more variables, resulting in poor generalization of unseen data. 2) CFS and Pearson correlation are the worst feature selection methods. Unlike Pearson correlation, CFS avoids redundancy between features. As these two methods rely on linear correlations between features and the target to make a selection, they potentially miss important nonlinear interactions that could be valuable for a machine learning model. Exceptionally, the FGM model with CFS-selected features achieves good results (R^2=0.74). This can be explained by FGM's ability to learn possible chronological dependencies between features and target variables. This indicates FGM has great potential for accurate CO_2 prediction. 3) Prediction models perform similarly with the features selected by Lasso and RForest because Lasso and RForest choose the same five features (as shown in Table 4). 4) ReliefF performs excellently with different predictors, except the linear regression model. In particular, it performs best with the ensemble algorithms RF and XGBoost. Analyzing the features selected by ReliefF, we can find that all three features (nighttime light, statistical, and meteorologic indicators) are included. RForest also includes three types but cannot handle collinearity between the selected PEP, GDP, and GDPc.

From the above analysis, we can conclude that ReliefF is a useful feature selection method; FGM can perform well by considering chronological information, and ensemble models such as XGBoost can handle

Table 5. Comparison of prediction performance of different models with and without feature selections. ("All" denotes using all features, "XXX selected" denotes using the features selected by the method XXX; The values in bold refer to the best results).

Features	Metrics	Lasso	SVR	RF	XGBoost	FGM
All	RMSE	**5.84**	3.38	3.92	4.16	4.62
	MAE	**3.96**	2.64	2.73	3.01	2.19
	R^2	**0.22**	**0.74**	0.65	0.6	0.55
PearsonC selected	RMSE	5.91	4.94	4.62	4.33	4.03
	MAE	3.99	3.11	2.86	2.9	1.51
	R^2	0.2	0.44	0.51	0.57	0.47
NMI selected	RMSE	5.91	4.99	4.66	4.55	3.14
	MAE	3.99	3.01	2.99	3.01	**1.37**
	R^2	0.2	0.43	0.5	0.52	**0.81**
CFS selected	RMSE	6.56	6.01	5.82	5.82	4.31
	MAE	4.47	3.54	3.65	3.84	2.11
	R^2	0.01	0.17	0.22	0.22	0.74
mRMR selected	RMSE	5.96	5.00	2.37	3.14	**3.07**
	MAE	3.84	2.67	1.55	2.25	1.53
	R^2	0.18	0.43	0.85	0.77	0.78
ReliefF selected	RMSE	5.93	**3.34**	**2.22**	**2.69**	4.19
	MAE	3.82	**1.92**	**1.08**	**1.91**	2.06
	R^2	0.19	**0.74**	**0.85**	**0.83**	0.72
RFE selected	RMSE	5.93	4.07	3.06	3.56	3.61
	MAE	3.96	2.34	1.83	2.19	1.52
	R^2	0.19	0.62	0.77	0.71	0.76
Lasso selected	RMSE	5.91	4.08	2.8	3.02	4.41
	MAE	3.96	2.32	1.66	1.98	1.83
	R^2	0.2	0.62	0.82	0.79	0.71
RForest selected	RMSE	5.91	4.07	2.93	2.96	4.08
	MAE	3.95	2.34	1.65	1.92	1.75
	R^2	0.20	0.62	0.8	0.8	0.75

nonlinear relationships and form a strong predictor by integrating multiple weak learners.

Performance of the Proposed Model

Our proposed hybrid model takes advantage of FGM and XGBoost and performs CO_2 emissions with the features selected by ReliefF. In the hybrid model, we use XGBoost instead of RF to be combined with FGM. The reason is that XGBoost is much faster than RF in prediction and focuses on reducing both bias and variance. On the test set of 2020 and 2021, the evaluation metrics of RMSE, MAE, and R^2 reach 1.73, 1.03, and 0.93, respectively. Fig. 2 shows the scatterplots and the coefficients of determination (R^2) of CO_2 emissions predicted by various models. These comparison prediction models are all at their best performance with the corresponding feature selection methods. Our hybrid model fits well with the actual values and outperforms other models.

To further verify the proposed hybrid model's performance, we conducted statistical tests, including a Likelihood Ratio (LR) test, a soundness test, and a heterogeneity test.

(1) LR Test

The LR test compares the goodness of fit between the model with selected features (simple) and the model without feature selection (complex). It assesses whether the additional features in the complex model significantly improve the fit. For the simple model, we

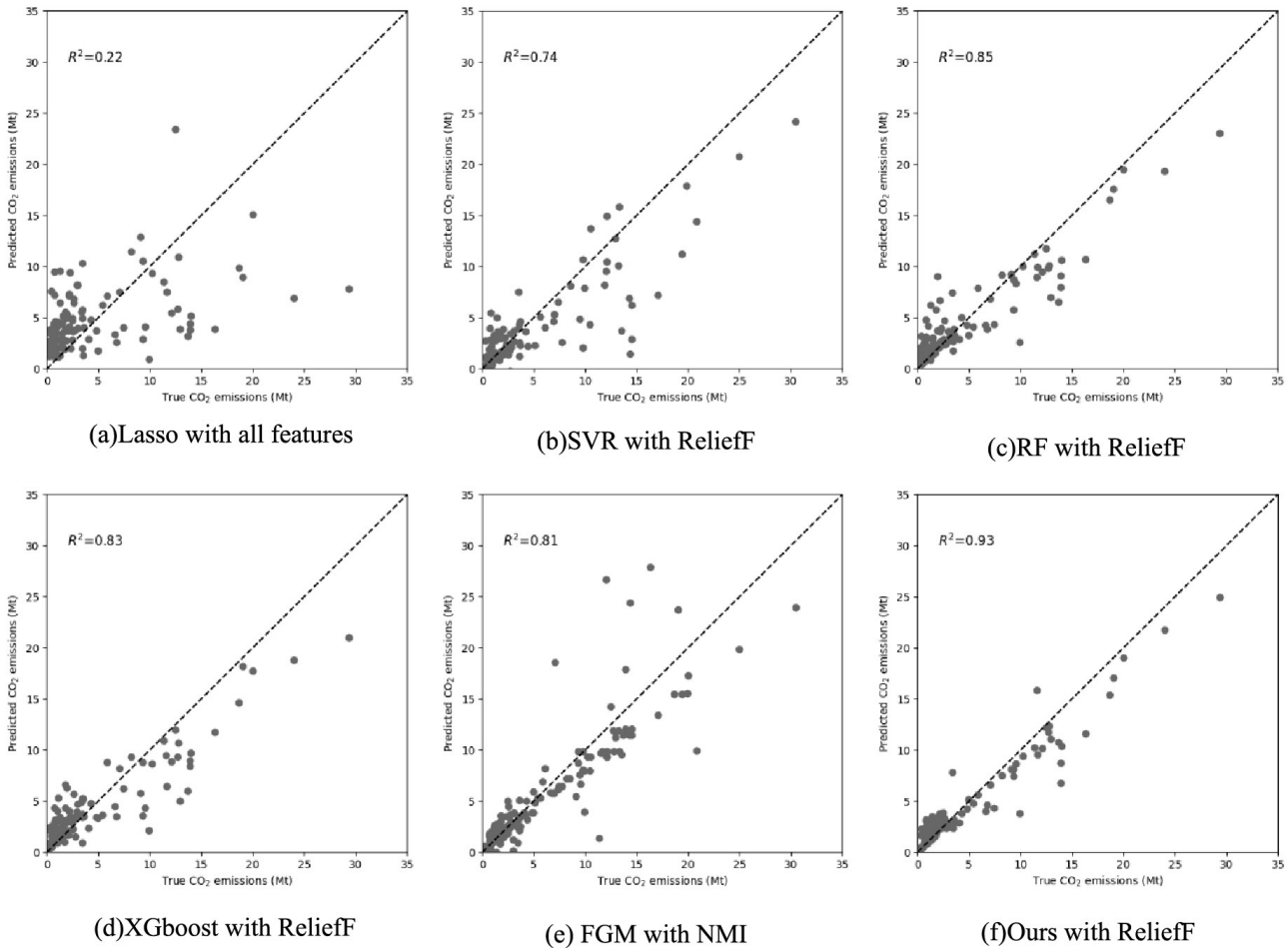


Fig. 2. Scatterplots of the true CO_2 emissions vs. the predicted CO_2 emissions by different prediction models with feature selection.

refer to the hybrid model with the top six ranked features by ReliefF. The complex model has more features added based on the feature ranking. The hypothesis is defined as: H0: The simpler model is good.

After computing the log-likelihoods of the two models and the LR statistic, the LR statistic is compared to a chi-squared distribution with degrees of freedom equal to the difference in the number of parameters between the models. If the LR statistic is smaller than the critical value, we cannot reject the null hypothesis, suggesting that the simple model provides a significantly better fit. The results are shown in Table 6. Taking the complex model with full parameters (19 features) as the example, LR statistic =-132.51<22.36 and a high p-value =1.0 suggest that the simpler model is sufficient, and the additional parameters in the full model do not provide a significantly better fit.

(2) Soundness test

The soundness test is used to evaluate the reliability and stability of a model under different assumptions or conditions. For a machine learning model, a soundness test often refers to the model's generalizability on unseen data. Cross-validation is commonly used to test whether a model can robustly generalize to unseen data through multiple data partitions. We performed 5-fold cross-validation and recorded the model performance of each fold with MAE, MSE, and R^2. Then, we ran a one-way ANOVA test on the metric values of five folds and received an F-statistic of 0.0965 and a p-value of 0.9813. This indicates that we cannot reject the null hypothesis: there is no significant difference in performance among the compared models.

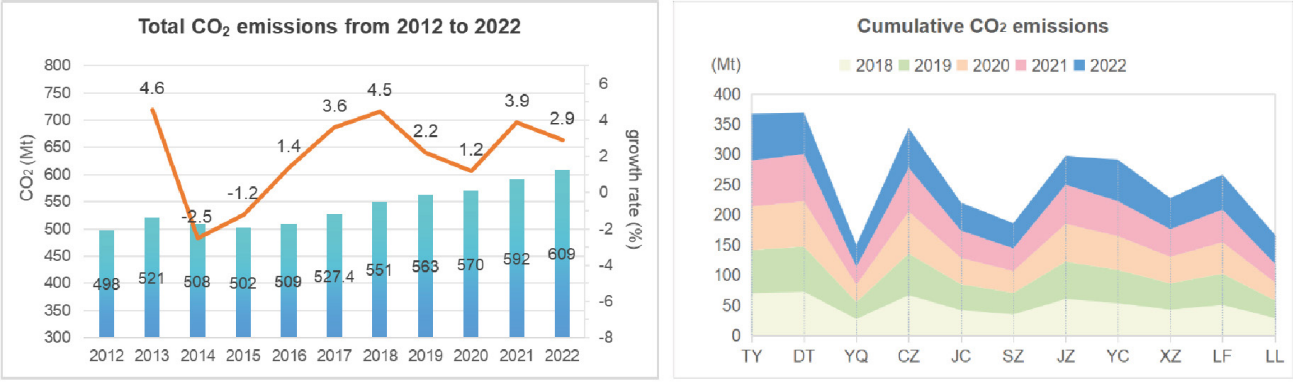
This can be attributed to our hybrid model integrating FGM and XGBoost. The FGM model has strengths in robustness in small-sample scenarios. XGBoost tends to generalize well to unseen data due to its ensemble nature and the ability to prevent overfitting with L2 regularization. This helps to reduce variance and improve the accuracy of unseen data.

(3) Heterogeneity test

A heterogeneity test is used to assess whether the model's performance exhibits significant variation across different groups of data. In this study, we split the data according to the "year" of CO_2 emission and then train subgroup models with these sub-datasets. Then, we run an independent t-test for each pair of groups to determine whether there are statistically significant differences in the performance metrics across the subgroups. The p-value for all the tests is greater than 0.25 (> 0.05), so we cannot reject the null hypothesis:

Table 6. LR test results.

Model pairs	Degree of freedom	Critical value for Chi-squared	LR statistic	p-value
Simple: 6 features Complex: 7 features	1	3.841	-125.20	1.0
Simple: 6 features Complex: 10 features	4	9.488	-6.91	1.0
Simple: 6 features Complex: 12 features	6	12.592	-21.62	1.0
Simple: 6 features Complex: 19 features	13	22.362	-132.51	1.0



(a) (b)

Fig. 3. (a) Total CO_2 emissions of Shanxi Province from 2012 to 2022. (b) Stacked diagram for the cumulative CO_2 emissions of 11 cities in Shanxi Province from 2018 to 2022.

the two models perform similarly. Besides, we split the data according to the 11 cities of Shanxi and executed the heterogeneity test. The results are also not significant (p-value >0.3), and our model performs similarly across subgroups. Therefore, we can conclude that there is no evidence of heterogeneity in our model.

Prediction of CO_2 Emissions

The prediction model was used to estimate the CO_2 emissions of Shanxi Province in 2022. The results show that the CO_2 emissions in 2022 reach 609Mt, with an increase rate of 2.9%. Fig. 3a shows the changes in total CO_2 emissions from 2012 to 2022. Generally, the CO_2 emissions of Shanxi Province have consistently increased since 2016 and reached a peak in 2022, with a significant growth of 22.29%. The 10-year average growth rate is 2.06%. By aggregating county-level carbon emissions into the cities they belong to, we obtain the CO_2 emissions of 11 cities.

The accumulated CO_2 emissions for 2018-2022 are illustrated in Fig. 3b. DT, TY, and CZ are ranked the top three in terms of accumulated CO_2 emissions, followed by JZ, YC, LF, JC, XZ, SZ, LL, and YQ. Historically, DT is known as the "Coal Capital" of China. Tens of thousands of tons of coal are produced daily in DT and are transported to every place that desperately needs energy. Therefore, DT is under great pressure to reduce carbon emissions. It is good to see that DT's CO_2 emissions in 2022 dropped significantly compared to 2021, reducing by 10Mt. This is inseparable from DT's efforts to develop new and renewable energy in recent years. As the capital city of Shanxi Province, TY has a high carbon emission intensity closely related to socioeconomic development. In contrast, YQ is the city with the smallest area, the smallest population, and the lowest GDP, and thus produces the lowest carbon emissions.

Spatial Analysis of CO_2 Emissions

We computed the global Moran's I to analyze the spatial autocorrelation of some key statistical factors in Shanxi Province. The results are shown in Table 7 and Fig. 4. Moran's I>0 indicates positive spatial correlation; the larger the value, the more obvious the spatial correlation; Moran's I<0 indicates negative spatial correlation; the smaller the value, the greater the spatial difference; otherwise, Moran's I=0 suggests a random spatial pattern.

As shown in Fig. 4, the Moran's I index for population and GDP are all significantly positive (Moran's I value > 0.18, p<0.001). This indicates that population and GDP are all higher in places with large spatial aggregation,

Table 7. Global Moran's I of Shanxi Province in 2012-2022

Year	Moran's I		
	CO_2	Population	GDP
2012	0.1613**	0.1844**	0.2353**
2013	0.1606**	0.1942***	0.2472**
2014	0.1602**	0.1873**	0.2749***
2015	0.1607**	0.1908**	0.3052***
2016	0.1615**	0.1939**	0.3274***
2017	0.1629**	0.1961**	0.3314***
2018	0.0499	0.3043***	0.3791***
2019	0.0489	0.3033***	0.4006***
2020	0.0491	0.2931***	0.3963***
2021	0.0491	0.3020***	0.3784***
2022	0.0445	0.3019***	0.3063***

Note: *p < 0.05, **p < 0.01, ***p<0.001

which is reasonable. We also note that Moran's I for GDP decreased from 2020. This may suggest that the overall economic level is declining. The Mordan's I of CO_2 emissions was significantly positive before 2018 but decreased thereafter. The Mordan's I is 0.0445 in 2022, and the p-value is 0.14. This suggests a slight positive

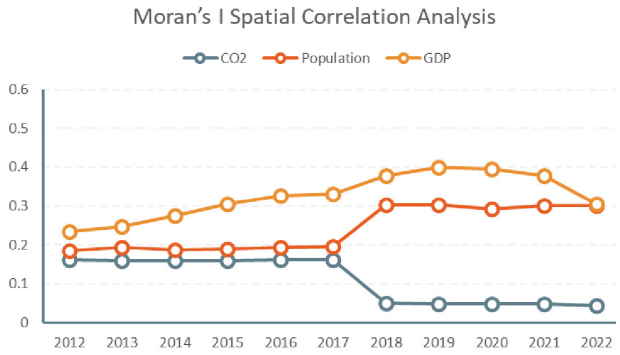


Fig. 4. Spatial analysis from year 2012 to 2022 using Moran's Indexes.

spatial autocorrelation of CO_2 emissions, but this spatial aggregation effect has been weakened. The first row of the figure illustrates the distribution of CO_2 emissions in Shanxi from 2018 to 2022. It can be observed that the high-emission regions are scattered along the central axis from north to south. The low carbon emission area on the west side is mainly the Luliang Mountains.

Local Moran's I provides detailed insights into local patterns and helps pinpoint the areas of interest [38]. Therefore, we computed local Moran's I for each county (or district) in 2018-2022 and mapped the index values in Fig. 5 (the second row). In the local Moran's I map, the counties (or districts) with high CO_2 emissions (> 10Mt) are highlighted in orange for local Moran's I>0 and in

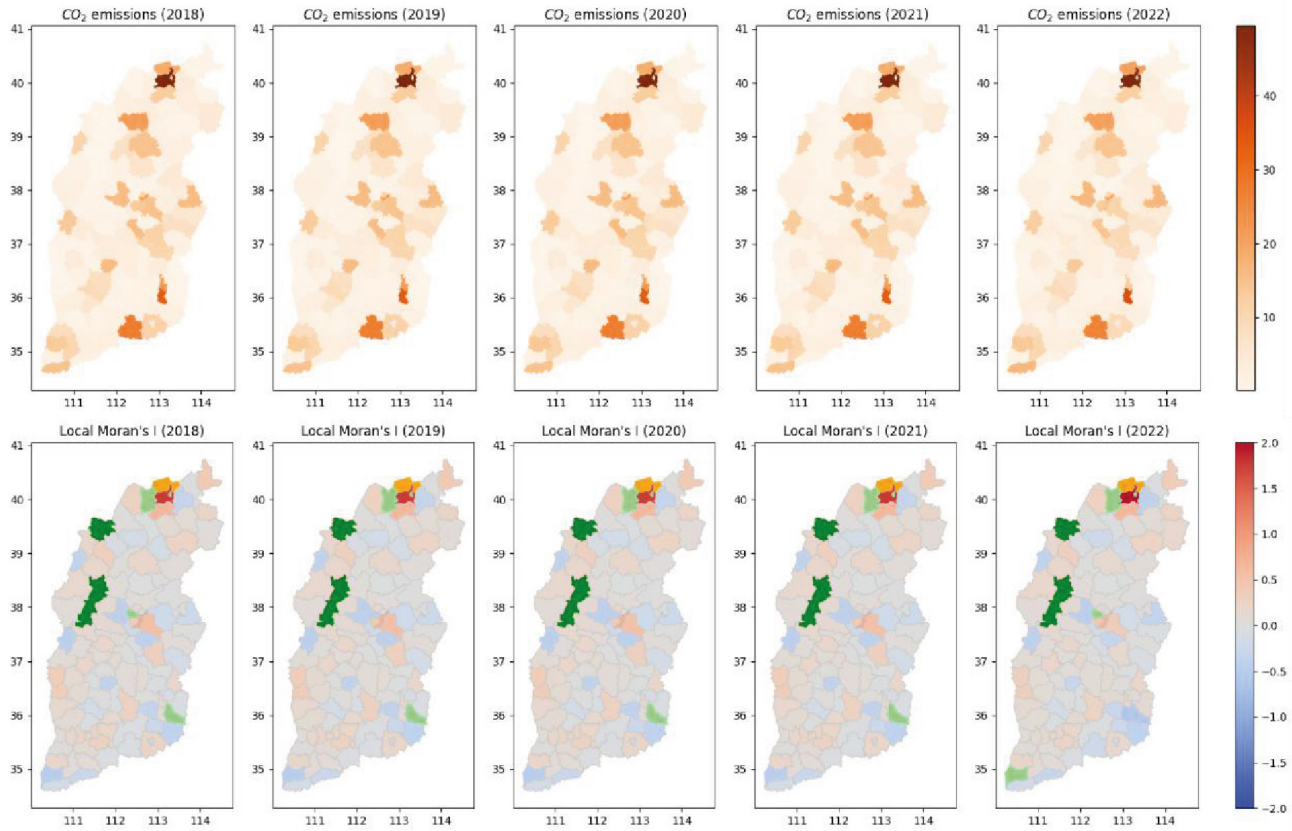


Fig.5. The distribution maps of CO_2 emissions (top) and Local Moran's I (bottom) from 2018 to 2022.

yellow for local Moran's I<0 with a significance level of <0.05; the counties (or districts) with low CO_2 emissions (<2Mt) are highlighted in green for local Moran's I>0 and in light green for local Moran's I<0 with a significance level of <0.05. We can observe that high neighboring clusters with high carbon emissions (orange blocks) are only gathered in the Xinrong district in DT. This is due to the multiple mining areas surrounding the area. There are no regions that have high carbon emissions while being surrounded by low carbon emissions neighbors. The low carbon emission regions are mainly clustered in Fangshan and Lan in LL, Pianguan, and Wuzhai in XZ (green blocks). Isolated low carbon emission regions are in Pingcheng and Zuoyun in DT, Wanbailin district in TY, Huguan in CZ, and Yongji in YC (light green blocks). These spatial patterns can help in developing policies for carbon emission reduction.

Conclusions

The accurate prediction of CO_2 emissions is of great significance to achieving the carbon reduction goal. This study proposes a framework for predicting CO_2 emissions with limited historical data. This framework integrates multi-source data of NPP/VIIRS nighttime light, socioeconomic statistics, and meteorological data as the input, selects the suitable features by comparing the performance of various feature selection approaches, predicts CO_2 emissions with a hybrid prediction model based on Grey theory and ensemble learning, and enables spatial correlation analysis of CO_2 emissions with Moran's I. Using Shanxi Province as a case study, we performed model training and testing with the data from 2012 to 2021, predicted county-level CO_2 emissions of 2022, and analyzed spatial clustering patterns of CO_2 emissions in counties. The comprehensive conclusions include:

(1) The selection of key features and the diversity of feature categories have greatly influenced the result of feature selection. In this study, ReliefF can select diverse types of features and demonstrates high prediction accuracy with various prediction models.

(2) The hybrid prediction method combines the fractional-order Grey model (FGM) and XGBoost to fully utilize the advantages of the two types of models. Based on a linear assumption, the FGM had a good fitting ability for the historical data. The XGBoost ensemble learning model is robust to nonlinear relationships. The hybrid model achieves better forecasting performance by integrating CO_2 emissions' trend prediction with nonlinear residual prediction. Statistical tests have verified the model's robustness and stability.

(3) According to the predicted CO_2 emissions in 2018-2022, high-emission counties (or districts) are distributed from north to south on the central line along the terrain of Shanxi, while low-emission counties are mainly clustered in the west and east mountains. Moran's I index confirms regional heterogeneity of CO_2 emissions. Therefore, to reduce carbon emissions, the government should focus on strengthening regional governmental coordination.

It is worth noting that this study has limitations. On the one hand, the data used in this study have a relatively large time scale with an annual resolution, so the analysis granularity is not detailed enough. Future research could incorporate monthly statistical data to monitor carbon emissions. On the other hand, carbon emissions are a complex process associated with socioeconomic activities, and further research can consider different industry characteristics to achieve a more accurate prediction of carbon emissions. In addition, we conducted a simple spatial autocorrelation analysis based on a binary weight matrix; it does not adequately account for the frictional coefficient of economy, commerce, labor, and capital movement that impact CO_2 emissions.

Acknowledgments

The authors acknowledge financial support from the Science and Technology Project of State Grid Shanxi Electric Power Company (Grant number: 520530220024).

Conflict of Interest

The authors declare no conflict of interest. The funders played no role in the design of the study, data collection, analysis or interpretation, manuscript writing, or decision to publish the results.

References

1. DUAN C., ZHU W., WANG S., CHEN B. Drivers of global carbon emissions 1990 – 2014. Journal of Cleaner Production, **371**, 133371, **2022**.
2. ALCÁNTARA V., PADILLA E., DEL RÍO P. The driving factors of CO_2 emissions from electricity generation in Spain: A decomposition analysis. Energy Sources, Part B: Economics, Planning, and Policy, **17** (1), 2014604, **2022**.
3. ZHU B.Z., ZHANG Y.L., ZHANG M.F., HE K.J., WANG P. Exploring the driving forces and scenario analysis for China's provincial peaks of CO_2 emissions. Journal of Cleaner Production, **378**, 134464, **2022**.
4. ZHAO Q., GAO W., SU Y., WANG T. Carbon emissions trajectory and driving force from the construction industry with a city-scale: A case study of Hangzhou, China. Sustainable Cities and Society, **88**, 104283, **2023**.
5. DAI S.Q., ZUO S.D., REN Y. A spatial database of CO_2 emissions, urban form fragmentation and city-scale effect related impact factors for the low carbon urban system in Jinjiang city, China. Data in Brief, **29**, 105274, **2020**.
6. YANG J., DENG Z., GUO S., CHEN Y. Development of bottom-up model to estimate dynamic carbon emission for city-scale buildings. Applied Energy, **331**, 120410, **2023**.
7. ZENG L.J., LU H.Y, LIU Y.P., ZHOU Y., HU H.Y.

Analysis of regional differences and influencing factors on China's carbon emission efficiency in 2005-2015. Energies, **12**, 3081, **2019**.
8. FANG Y., LU X.Q., LI H.Y. A random forest-based model for the prediction of construction-stage carbon emissions at the early design stage. Journal of Cleaner Production, **328**, 129657, **2021**.
9. PAO H.T., FU H.C., TSENG C.L. Forecasting of CO_2 emissions, energy consumption and economic growth in China using an improved Grey model. Energy, **40** (1), 400, **2012**.
10. WANG M., WANG W., WU L.F. Application of a new Grey multivariate forecasting model in the forecasting of energy consumption in 7 regions of China. Energy, **243**, 123024, **2022**.
11. LI X.M., ZHANG B.J., ZHAO Y., ZHANG Y.F., ZHOU S.W. A novel dynamic Grey multivariate prediction model for multiple cumulative time-delay shock effects and its application in energy emission forecasting. Expert Systems with Applications, **251**, 124081, **2024**.
12. KOUR M. Modelling and forecasting of carbon-dioxide emissions in South Africa by using ARIMA model. International Journal of Environmental of Science and Technology. **20**, 11267, **2023**.
13. XING Q.F., HE Z.W., WEI L. Evaluation of linkage relationships between carbon emissions and economic development based on the decoupling model and the VAR model: a case study of Shanxi Province (China). Environmental Science and Pollution Research, **30** (25), 66651, **2023**.
14. SANTOS T.M.O, JÚNIOR J.N.O, BESSANI M., MACIEL C.D. CO_2 emissions forecasting in multi-source power generation systems using dynamic Bayesian network. In proceedings of 2021 IEEE International Systems Conference (SysCon), IEEE: Vancouver, BC, Canada, **2021**.
15. WEN L., YUAN X. Forecasting CO2 emissions in China's commercial department, through BP neural network based on random forest and PSO. Science of The Total Environment, **718**, 137194, **2020**.
16. ZHOU L., SONG J., CHI Y.J., YU Q.Z. Differential spatiotemporal patterns of CO_2 emissions in eastern China's urban agglomerations from NPP/VIIRS nighttime light data based on a neural network algorithm. Remote Sensing, **15** (2), 404, **2023**.
17. QIAO W.B., LU H.F., ZHOU G.F., AZIMI M., YANG Q., TIAN W.C. A hybrid algorithm for carbon dioxide emissions forecasting based on improved lion swarm optimizer. Journal of Cleaner Production, **244**, 118612, **2020**.
18. KONG F., SONG J.B., YANG Z.Z. A daily carbon emission prediction model combining two-stage feature selection and optimized extreme learning machine. Environmental Science and Pollution Research, **29** (58), 87983, **2022**.
19. HOU L., CHEN H. The prediction of medium- and long-term trends in urban carbon emissions based on an ARIMA-BPNN combination model. Energies, **17**, 1856, **2024**.
20. LIN C.C., HE R.X., LIU W.Y. Considering multiple factors to forecast CO_2 emissions: A hybrid multivariable Grey forecasting and genetic programming approach. Energies, **11** (12), 3432, **2018**.
21. LI G.H., WU H., YANG H. A hybrid forecasting model of carbon emissions with optimized VMD and error correction. Alexandria Engineering Journal, **81**, 210, **2023**.
22. KALAYCI S, ARTEKIN AÖ. The linkage between truck transport, trade openness, economic growth, and CO_2 emissions within the scope of green deal action plan: An empirical investigation from Türkiye. Polish Journal of Environmental Studies, **33** (3), 3231, **2024**.
23. Global carbon emission data. Carbon Monitor. Available online: https://carbonmonitor.org.cn/ (accessed on 13 March 2024).
24. ZOU X., WANG R.F., HU G.H., RONG Z., LI J.X. CO_2 emissions forecast and emissions peak analysis in Shanxi Province, China: An application of the LEAP model. Sustainability, **14** (2), 637, **2022**.
25. SHANXI STATISTICS BUREAU. Shanxi Statistical Yearbook 2020. China Statistics Press: Beijing, China, **2020** [In Chinese].
26. ZANG X.L., ZHAO T., WANG J., GUO F. The effects of urbanization and household-related factors on residential direct CO_2 emissions in Shanxi, China from 1995 to 2014: A decomposition analysis. Atmospheric Pollution Research, **8** (2), 297, **2017**.
27. ZHANG X.Y., XIE Y.W., JIAO J.Z., ZHU W.Y., GUO Z.C., CAO X.Y., LIU J.M., XI G.L, WEI W. How to accurately assess the spatial distribution of energy CO_2 emissions? Based on POI and NPP-VIIRS comparison. Journal of Cleaner Production, **402**, 136656, **2023**.
28. ZUO C., GONG W., GAO Z.Y., KONG D.Y., WEI R., MA X. Correlation analysis of CO_2 concentration based on DMSP-OLS and NPP-VIIRS integrated data. Remote Sensing, **14** (17), 4181, **2022**.
29. ELVIDGE C.D., ZHIZHIN M., GHOSH T., HSU F.C., TANEJA J. Annual time series of global VIIRS nighttime lights derived from monthly averages: 2012 to 2019. Remote Sensing, **13** (5), 922, **2021**.
30. WEI W., ZHANG X.Y., ZHOU L., XIE B.B., ZHOU J.J., LI C.H. How does spatiotemporal variations and impact factors in CO_2 emissions differ across cities in China? Investigation on grid scale and geographic detection method. Journal of Cleaner Production, **321**, 128933, **2021**.
31. NATIONAL STATISTICS BUREAU OF CHINA. China City Statistical Yearbook. China Statistics Press: Beijing, China, **2022.** [In Chinese].
32. NATIONAL STATISTICS BUREAU OF CHINA. China Statistical Yearbook (County-level). China Statistics Press: Beijing, China, **2022** [In Chinese].
33. WU L.F., LIU S.F., YAO L.G., YAN S.L., LIU D.L. Grey system model with the fractional order accumulation. Communications in Nonlinear Science and Numerical Simulation, **18** (7), 1775, **2013**.
34. CONG X.P., WANG X.Y., XIE Q.C. Nonlinearity, heterogeneity and indirect effects in the CO_2 emissions-financial development relation from partial linear additive panel model. Polish Journal of Environmental Studies. **2024**.
35. HU Z.M., JIANG T. Innovative Grey multivariate prediction model for forecasting Chinese natural gas consumption. Alexandria Engineering Journal, **103**, 384, **2024**.
36. GU H.L., CHEN Y., WU L.F. A new Grey adaptive integrated model for forecasting renewable electricity production. Expert Systems with Applications, **251**, 123978, **2024**.
37. BOLEA L., ESPINOSA-GRACIA A., Jimenez S. So close, no matter how far: A spatial analysis of CO2 emissions considering geographic and economic distances. The World Economy, **47** (2), 544, **2024**.
38. DE ABREU DOS SANTOS D., LOPES T.R., DAMACENO F.M., DUARTE S.N. Evaluation of deforestation, climate change and CO_2 emissions in the Amazon biome using the Moran Index. Journal of South American Earth Sciences, **143**, 105010, **2024**.