Journal of
Water & Climate Change

IWA PUBLISHING

Journal of Water and Climate Change Vol 14 No 2, 477 doi: 10.2166/wcc.2023.331

Forecasting carbon dioxide emissions: application of a novel two-stage procedure based on machine learning models

Chunzi Wang [a,*], Moye Li[a] and Junpeng Yan[b]

[a] Shanghai Normal University Tianhua College, Shanghai City 201815, China
[b] Xianda College of Economics and Humanities, Shanghai International Studies University, Shanghai City 202162, China
*Corresponding author. E-mail: wcz2766@sthu.edu.cn

CW, 0000-0001-9617-3464

ABSTRACT

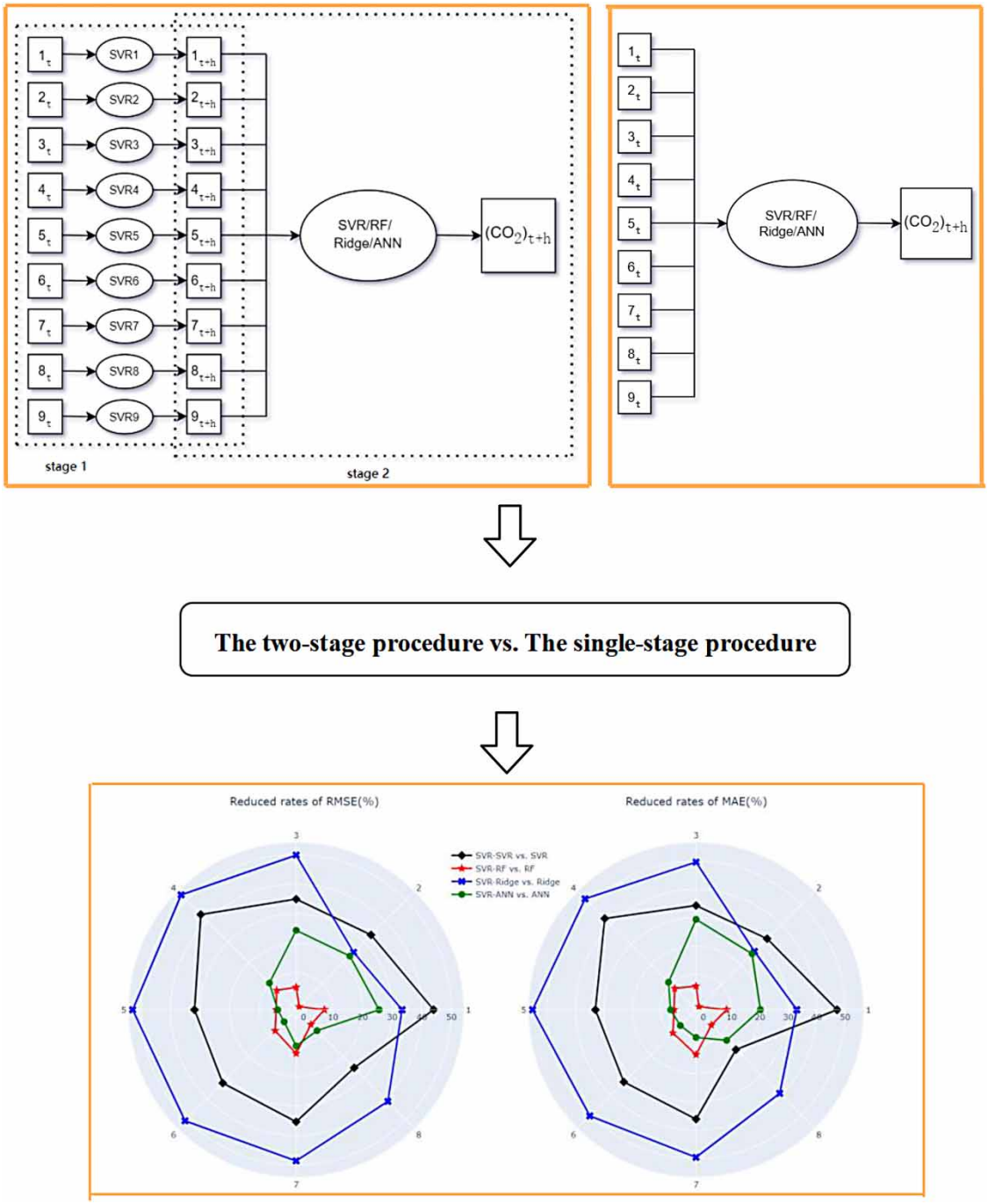
Accurate forecast of carbon dioxide (CO_2) emissions plays a significant role in China's carbon peaking and carbon neutrality policies. A novel two-stage forecast procedure based on support vector regression (SVR), random forest (RF), ridge regression (Ridge), and artificial neural network (ANN) is proposed and evaluated by comparing it with the single-stage forecast procedure. Nine independent variables' data (study period: 1985–2020) are used to forecast the CO_2 emissions in China. Our results reveal that, when the time gap, *h* increases from 1 to 8, the average root mean squared error (RMSE) and mean absolute error (MAE) of SVR–SVR, SVR–RF, SVR–Ridge, and SVR–ANN are almost uniformly lower than errors arising from their single-stage version, respectively. Among these two-stage models, SVR–ANN exhibits the lowest forecast errors, whereas SVR–RF admits the highest. The mean percentage decrease in forecast errors of SVR–SVR vs. SVR, SVR–RF vs. RF, SVR–Ridge vs. Ridge, and SVR–ANN vs. ANN are 36.06, 5.98, 43.05, and 14.81 for RMSE, and 36.06, 6.91, 43.27, and 15.35 for MAE. Our two-stage procedure is also suitable to forecast other variables, such as fossil fuel and renewable energy consumption.

Key words: artificial neural network, CO_2 emission forecast, random forest, ridge regression, support vector regression, two-stage forecast procedure

HIGHLIGHTS

- A novel two-stage forecast procedure is proposed and evaluated.
- Four hybrids of machine learning models based on SVR, RF, Ridge, and ANN are constructed to provide an accurate forecast of CO_2 emissions in China.
- SVR–ANN gives the lowest forecast errors in terms of RMSE and MAE.
- SVR–Ridge shows the highest performance improvement than Ridge.

GRAPHICAL ABSTRACT



INDEX OF NOTATIONS AND ABBREVIATIONS

ANFIS	adaptive neuro-fuzzy inference system
ANN	artificial neural network
ARIMA	autoregressive integrated moving average
BLSTM	bi-directional long short-term model
CART	Classification And Regression Tree
CO_2	carbon dioxide

CV	coefficient of variation
G20	Group of Twenty
GDP	gross domestic product
GM	grey model
HHO	Harris hawk optimization
KF	kalman filter
KLS	acronym of KF, LSTM, and SVM
KKT	Karush–Kuhn–Tucker
LSTM	long short-term model
MAE	mean absolute error
MAPE	mean absolute percentage error
ML	machine learning
RBF	radial basis function
Relu	rectified linear unit
RF	random forest
Ridge	ridge regression
RMSE	root mean squared error
SD	standard deviation
SOM	self-organizing map
SVD	singular value decomposition
SVM	support vector machine
SVR	support vector regression

1. INTRODUCTION

According to the Global Carbon Budget 2021 (Friedlingstein *et al.* 2022), carbon dioxide (CO_2) emitted to the atmosphere attributed to human activities keeps growing at an unprecedented rate. Of all CO_2 emissions during the last 70 years, 70% were emitted after the 1960s, and 33% were emitted in this millennium. The astronomical amount of CO_2 emitted is exacerbating global warming. Consequently, considerable impacts on the ecological environment are introduced including crop production, rainfall, water scarcity, ocean acidification, etc., which, conversely, causes severe harm to human beings (Valipour 2017; Bhatt & Hossain 2019; Rehman *et al.* 2021).

In the last 10 years (2012–2021), around 89% of total anthropogenic CO_2 emissions had originated from fossil fuel combustion (Friedlingstein *et al.* 2022). Consumption of fossil fuels, such as coal, oil, and natural gas, is a substantial factor that affects CO_2 emissions. Therefore, it is imperative to promote the transformation of energy consumption structure from fossil fuels to renewable energy such as hydropower (Kuriqi *et al.* 2019, 2020, 2021; Malka *et al.* 2022), solar energy (Rabaia *et al.* 2021; Djaafari *et al.* 2022), and wind energy (Sadorsky 2021).

China takes its share of responsibility in tackling the climate change issue. At the 75th United Nations General Assembly in 2020, China stated its dual carbon commitment to achieve carbon peaking before 2030 and carbon neutrality before 2060. To make informative decisions in achieving these ambitious goals, it is necessary to make an accurate forecast of CO_2 emissions coming from fossil fuel consumption.

1.1. Literature review of forecasting models

There is rich literature addressing the issue of CO_2 emission forecast, with various forecasting methods proposed. Classical methods such as linear regression (Al-Mulali *et al.* 2016; Wang *et al.* 2019), grey models (Lu *et al.* 2009; Lin *et al.* 2011; Lotfalipour *et al.* 2013), and the time-series models, e.g., autoregressive integrated moving average (ARIMA) (Nyoni & Bonga 2019; Yang & O'Connell 2020; Ning *et al.* 2021), are frequently used. It has been widely concerned that the performance of linear regression can be compromised by the violation of the independence or normality assumptions (Gallo *et al.* 2014). Grey models focus on dealing with small data and poor information. The forecasting performance of grey models may be decreased when there are many independent variables or data. Moreover, the often-used grey model GM(1,1), together with ARIMA, is, by definition, unworkable for the particular forecasting task when multiple independent variables are included.

In contrast, machine learning (ML) methods are more flexible and show great potential in modeling nonlinear and complex patterns. For example, Stamenković *et al.* (2015) use a back-propagation neural network and a general regression neural network to forecast emissions of methane. It is observed that both neural network models are more effective than multiple linear

regression. Saleh *et al.* (2016) utilize support vector regression (SVR) to predict the expenditure of CO_2 emissions based on energy consumption in Yogyakarta and acquire a low prediction error. Acheampong & Boateng (2019) establish an artificial neural network (ANN) model to forecast the carbon emission intensity for Australia, Brazil, China, India, and the USA. Their study also identifies the most sensitive variable to each country's carbon emission intensity.

Many studies suggest that a combination of two or more models exhibits better-forecasting performance than a single model in various applications. For instance, Mardani *et al.* (2020) propose a multi-stage method to forecast CO_2 emissions in G20 (Group of Twenty) countries. They use singular value decomposition (SVD) to forecast missing data, self-organizing map (SOM) to cluster data, and ANN and adaptive neuro-fuzzy inference system (ANFIS) to forecast CO_2 emissions based on the gross domestic product (GDP) and energy consumption. Compared against methods using only ANN, ANFIS, or multiple linear regression, the multi-stage approach yields the lowest forecast error. Wang & Zhu (2021) use the Johansen cointegration test and the neural network autoregression model to forecast China's CO_2 emissions based on assumptions of three levels of GDP growth rate, which results in a different amount increase in natural gas consumption and production. Morshed-Bozorgdel *et al.* (2022) find that a two-level ensemble of ML algorithms captures the high variation in wind speed. Farzin *et al.* (2022) use a bi-directional long short-term model (BLSTM) and the Harris hawk optimization (HHO) algorithm to optimize the model's hyperparameters to forecast the underground water table in Iran. They find that their BLSTM–HHO method is superior to benchmark methods regarding assessment criteria such as mean absolute error (MAE), root mean square error (RMSE), and forecast variance.

Li (2020) develops the KLS (acronym of KF, LSTM, and SVM) method, which is a fusion of Kalman filter (KF), long short-term memory (LSTM), and support vector machine (SVM), to forecast China's carbon emissions. In the meantime, ridge regression (Ridge) is used to select independent variables. The KLS method integrates time-series forecast and variable selection and has been confirmed to be more accurate than the four existing methods. Geevaretnam *et al.* (2022) apply three ML models, namely random forest (RF), SVM, and ANN to forecast global CO_2 emissions and forecast performance was compared in terms of MAE, RMSE, and mean absolute percentage error (MAPE). It reveals that SVM produces the lowest forecast errors.

Accordingly, we will consider a hybrid of ML models to improve the forecast accuracy of CO_2 emissions in China.

1.2. Contributions

We observe that, in the literature of CO_2 emission forecast, few studies are explicitly discussing the time domain of their dependent and independent variables (Saleh *et al.* 2016; Acheampong & Boateng 2019; Mardani *et al.* 2020). So it is possible that some of them train models on data sets in the form of $\{X_t,\ y_t\},\ t = 1, 2, \ldots, T$, where T is the observation length of the time series. Then, for the testing set, the value of independent variables at time $t+h$, namely X_{t+h}, is substituted into the trained model to forecast the value of the dependent variable at time $t+h$ to get $\hat{y}_{t+h}$. Such a procedure is suitable when investigating the dynamic links between independent variables and dependent variables. But it is inappropriate and unnecessary to forecast CO_2 emissions by giving independent variables at the same time domain. Because the released data of production-based CO_2 emissions are calculated by multiplying activity data (i.e. fuel consumption data in the industrial sector) by corresponding emission factors by the type of fuel (Liu *et al.* 2020; Friedlingstein *et al.* 2022). That means, when independent variables, i.e. activity data at time $t+h$ are released, the CO_2 emission data at time $t+h$ are also released.

In practical applications, policymakers are more interested in forecasting CO_2 emissions at time $t+h$ through independent variables at time t. That is, models should be trained on data sets in the form of $\{X_t,\ y_{t+h}\},\ t = 1, 2, \ldots, T$, where the forecasted value, $\hat{y}_{t+h}$, is obtained by given X_t. There is a time gap h in the dependent variable, so data sets in the form of $\{X_t,\ y_{t+h}\}$ will be called lagged data sets thereafter. For instance, Hou *et al.* (2022) aim to forecast China's short-term CO_2 emissions at time $t+w$ via independent variables at time t. It converts multivariate time-series data into labeled sample data using the sliding window method and different shallow learning approaches are tested to find the best one. Faruque *et al.* (2022) also train several deep-learning models on lagged data sets in their second step to forecast Bangladesh's CO_2 emissions.

We will explore the functional relationship between X_t and y_{t+h}. However, when forecasting CO_2 emissions at time $t+h$ using independent variables at time t, the forecast errors may increase in contrast with using independent variables at time $t+h$. To our knowledge, there is no work focused on lowering the forecast errors caused by the time gap, h, in the CO_2 emission forecast field. We propose a novel two-stage forecast procedure, adapting the procedure in Patel *et al.* (2015) by making

necessary modifications to reduce the forecast errors. The first stage forecasts independent variables at time $t+h$ with their historical values up to time t, i.e. $\hat{X}_{t+h}$ is calculated via $\{X_i\}$, $i = 1, 2, \ldots, t$. The second stage forecasts CO_2 emissions at time $t+h$ with forecasted independent variables at time $t+h$, i.e. $\hat{X}_{t+h}$. We make an effort to contribute to the existing knowledge in the following ways:

- Clarify the necessity to train models on lagged data sets when forecasting CO_2 emissions.
- Propose a novel two-stage forecast procedure based on ML models to reduce the forecast errors caused by the time gap in the dependent variable of lagged data sets. It is shown that four hybrids of ML models in the two-stage procedure almost all have lower forecast errors than four individual ML models in the single-stage procedure, respectively.
- Develop a forecasting tool for CO_2 emissions from a real-case data set in China. A model with high accuracy could provide a guiding tool for the formulation of China's carbon emission reduction policy.

There are several novelties of the proposed two-stage forecast procedure compared to that of Patel *et al.* (2015):

- The historical values of the independent variables are used to forecast their future counterparts in the first stage. Thus, this method introduces no exogenous variables to update independent variables and greatly simplifies the forecasting procedure of Patel *et al.* (2015).
- Patel *et al.* (2015) forecast stock prices in the financial field, while we attempt to apply a two-stage forecast procedure based on ML models to data sets dealing with macroeconomics and climatology. We expand the application domain of the two-stage forecast procedure.
- Four ML models, i.e. SVR, RF, Ridge, and ANN, are applied in this article, while Patel *et al.* (2015) only tried SVR, RF, and ANN.

2. DATA AND EVALUATION MEASURES

Inspired by Acheampong & Boateng (2019), nine variables that may have impacts on China's CO_2 emissions are chosen as independent variables, as shown in Table 1. Raw data of these 10 time series, i.e. nine independent variables as well as CO_2

Table 1 | Explanations and sources of nine independent variables as well as CO_2 emissions

Variable names	Variable explanations and units	Data sources
CO_2 emissions	Total production-based emissions of carbon dioxide, including fossil fuels and cement production, excluding land-use change. (million tons)	Global Carbon Project[a]
Foreign direct investment	The net inflows of investment to acquire a lasting management interest in an enterprise operating in an economy other than that of the investor. This series shows net inflows (new investment inflows less disinvestment) from foreign investors and is divided by GDP (%).	The World Bank[b]
Industry, value added	An increase in industrial activity. It comprises value added in mining, manufacturing, construction, electricity, water, and gas, and is divided by GDP (%).	The World Bank
Trade	The sum of exports and imports of goods and services measured as a share of GDP (%).	The World Bank
Urban population	Proportion of urban population to total population (%).	The World Bank
GDP per capita	Gross domestic product is divided by total population (current US$).	The World Bank
Energy consumption	Total energy consumption, including coal, oil, nature gas, and primary electricity and other energy sources. (10,000 tons of standard coal).	China Statistical Yearbook[c]
Coal	Proportion of coal consumption to total energy consumption (%).	China Statistical Yearbook
Oil	Proportion of oil consumption to total energy consumption (%).	China Statistical Yearbook
Gas	Proportion of nature gas consumption to total energy consumption (%).	China Statistical Yearbook

The unit of each variable is marked in parentheses.
[a]https://www.globalcarbonproject.org/.
[b]https://data.worldbank.org/country/china?view=chart.
[c]http://www.stats.gov.cn/tjsj/ndsj/.

emissions, are in annual form ranging from 1985 to 2020. To develop accurate models, the Chow–Lin method (Chow & Lin 1971) is used to convert the annual data into their corresponding quarterly versions, obtaining 144 quarterly samples for each variable. The sum of every four quarterly samples equals the raw annual observation. Hence, the quarterly data ranging from 1985Q1 to 2020Q4 are applied for this study. Each variable X is min–max normalized as below:

$$\text{Normalized}(X) = \frac{X - \min(X)}{\max(X) - \min(X)} \tag{1}$$

Figures 1 and 2 reveal the trend of the nine independent variables and CO_2 emissions. Urban population, GDP per capita, energy consumption, and gas all have an approximately monotonically increasing trend similar to CO_2 emissions. This phenomenon may affect the forecast accuracy of the RF method, and we will discuss this problem in the Results and discussion section later. Table 2 also summarizes some descriptive statistics of each variable.

The metrics measuring forecast performance are RMSE and MAE, i.e.,

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

$$\text{MAE} = \frac{1}{n}\sum_{i=1}^{n}|y_i - \hat{y}_i| \tag{3}$$

where y_i, $\hat{y}_i$, $i = 1, 2, \ldots, n$ represent the observed and forecasted values, respectively. RMSE and MAE both indicate the discrepancy between the observed and forecasted values. The lower the RMSE and MAE values, the better the model performance.

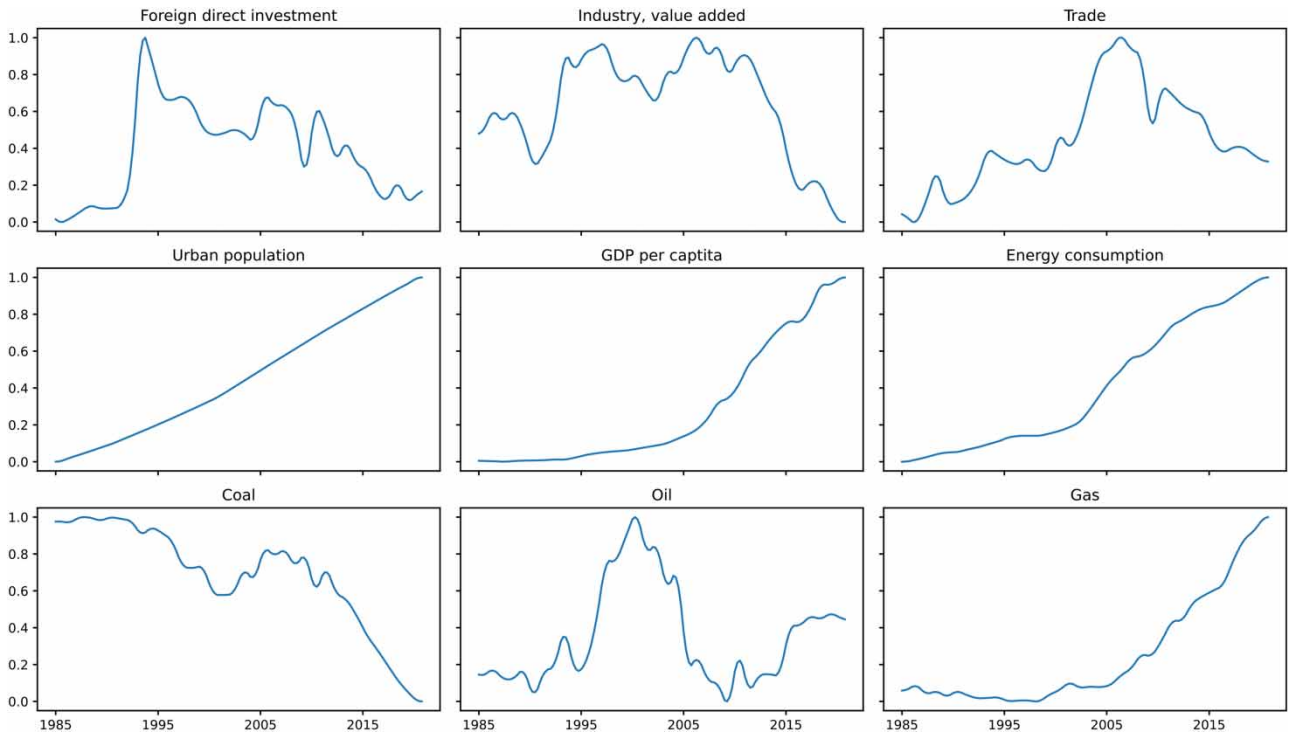


Figure 1 | Trends of nine independent variables. The shared *x*-axis represents the year ranging from 1985 to 2020, and the shared *y*-axis denotes the normalized values of nine independent variables.

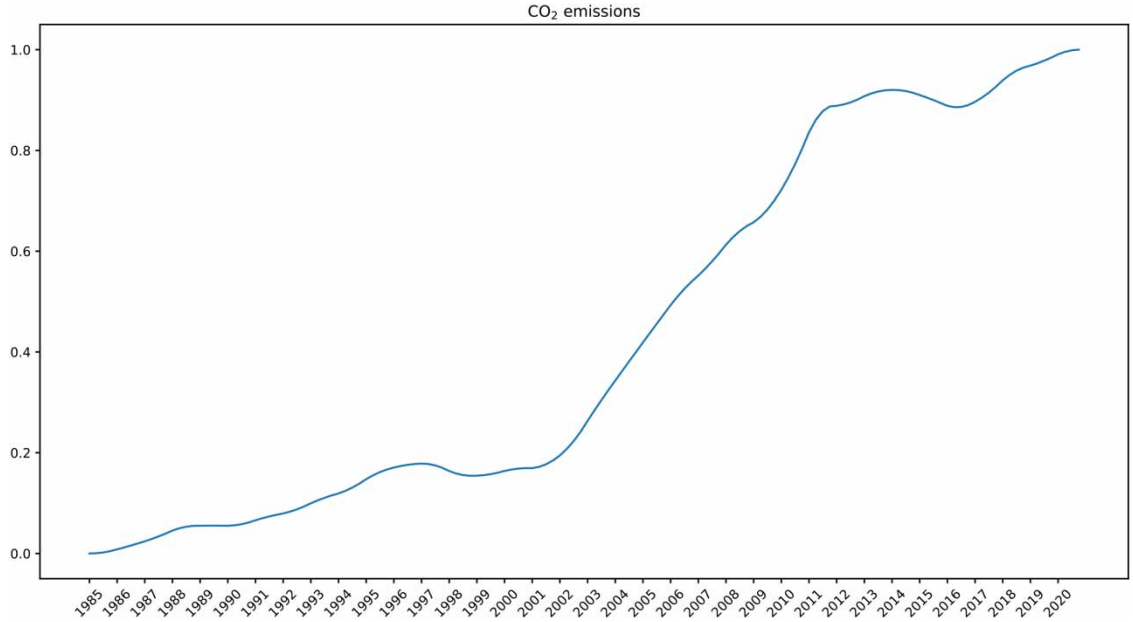


Figure 2 | Trend of CO_2 emissions. The *y*-axis denotes the normalized values of CO_2 emissions.

Table 2 | Descriptive statistics of non-normalized data

Variable	Count	Min	Max	Mean	SD	CV
CO_2 emissions (million tons)	144	495.59	2,675.07	1,432.60	782.79	0.545
Foreign direct investment (%)	144	0.13	1.69	0.74	0.39	0.635
Industry, value added (%)	144	9.44	11.92	11.03	0.68	0.416
Trade (%)	144	4.85	16.18	9.92	2.97	0.584
Urban population (%)	144	5.69	15.41	10.11	3.01	0.679
GDP per capita (current US$)	144	62.06	2,624.95	787.73	849.54	1.172
Energy consumption (10,000 tons of standard coal)	144	19,022.43	124,913.29	61,547.91	35,913.87	0.844
Coal (%)	144	14.17	19.07	17.54	1.37	0.393
Oil (%)	144	4.06	5.53	4.58	0.39	0.764
Gas (%)	144	0.44	2.12	0.86	0.49	1.194

SD, standard deviation; CV, coefficient of variation.

3. METHODOLOGY

3.1. Model building blocks

Four supervised ML models, namely SVR, RF, Ridge, and ANN, are introduced as the building blocks of our forecast procedure. Denote a training set as $\{(\boldsymbol{x}_1,\ y_1), (\boldsymbol{x}_2,\ y_2),\ \ldots, (\boldsymbol{x}_n,\ y_n)\} \subset R^D \times R$, where D is the number of features and n is the sample size.

3.1.1. Support vector regression

SVR maps data to a high-dimensional feature space through the function $\phi(\cdot)$ and then seeks a hyperplane $f(\boldsymbol{x}) = \boldsymbol{w} \cdot \phi(\boldsymbol{x}) + b$. Its goal is to make the hyperplane $f(\boldsymbol{x})$ as flat as possible while minimizing the ε-insensitive loss function (Vapnik 1999) described by

$$L_\varepsilon(y, f(\boldsymbol{x})) = \max(|y - f(\boldsymbol{x})| - \varepsilon,\ 0). \tag{4}$$

The value of ε-insensitive loss function is zero when the deviation between the observed value and the forecasted value is no greater than ε. Equation (4) makes SVR robust to outliers and can prevent overfitting to some extent. It is also verified that SVR has excellent performances in regression and time-series forecast applications (Smola & Schölkopf 2004).

Next, SVR is expressed in the following optimization problem

$$\min \quad \frac{1}{2}||\boldsymbol{w}||^2 + C\sum_{i=1}^{n}(\xi_i + \xi_i^*) \tag{5}$$

$$\text{s.t.} \quad \begin{cases} y_i - \boldsymbol{w}\cdot\boldsymbol{x}_i - b \le \varepsilon + \xi_i \\ \boldsymbol{w}\cdot\boldsymbol{x}_i + b - y_i \le \varepsilon + \xi_i^* \\ \xi_i, \xi_i^* \ge 0, \ i = 1, 2, \ldots, n \end{cases} \tag{6}$$

where $\boldsymbol{w}$ and b represent weights and bias of the hyperplane $f(\boldsymbol{x})$. The regularization parameter $C > 0$ determines the trade-off between the flatness of $f(\boldsymbol{x})$ and the value of ε-insensitive loss function. $\xi_i, \xi_i^*, \ i = 1, 2 \ldots, n$ are slack variables. The optimization problem expressed in (5) and (6) is then transformed into a dual optimization problem by the Lagrange multiplier method. Parameters are calculated by the sequential minimal optimization algorithm (Platt 1998) and KKT (Karush–Kuhn–Tucker) conditions (Bishop & Nasrabadi 2006). Finally, the hyperplane is represented as follows:

$$f(\boldsymbol{x}) = \sum_{i=1}^{n}(\alpha_i - \alpha_i^*)\kappa(\boldsymbol{x}_i, \ \boldsymbol{x}) + b \tag{7}$$

where $\kappa(\boldsymbol{x}_i, \ \boldsymbol{x}) = \phi(\boldsymbol{x}_i)\cdot\phi(\boldsymbol{x})$ is the kernel function, and $\alpha_i, \alpha_i^* \ge 0$ are Lagrange multipliers. Polynomial kernel and radial basis function (RBF) are used as alternative kernel functions for training, where

$$\text{Polynomial kernel: } \kappa(\boldsymbol{x}_i, \ \boldsymbol{x}_j) = (\boldsymbol{x}_i \cdot \boldsymbol{x}_j + 1)^d, \tag{8}$$

$$\text{RBF: } \kappa(\boldsymbol{x}_i, \ \boldsymbol{x}_j) = \exp\left(-\frac{1}{D}||\boldsymbol{x}_i - \boldsymbol{x}_j||^2\right). \tag{9}$$

Setting $\varepsilon = 0.1$, a grid search method (Mselmi *et al.* 2017; Papouskova & Hajek 2019) is conducted to determine the optimal parameters that give the smallest test error by choosing from all possible value combinations in Table 3.

3.1.2. Random forest

Ensemble learning takes advantage of the power of multiple ML models combined together to train on the same set of observations. Therefore, ensemble models can significantly improve forecast accuracy compared to individual ML models. RF is one of the most popular decision tree-based ensemble models (Ramasubramanian & Singh 2017) due to its high forecasting accuracy (Boulesteix *et al.* 2012; Belgiu & Drăguţ 2016; Ouedraogo *et al.* 2019). RF randomly fits multiple decision trees and then takes the average of each tree's forecast as forest's forecast. RF used in our research is implemented as follows:

1. Create N_1 bootstrap samples by sampling n elements from a given training set with replacement.
2. Use the Classification And Regression Tree (CART) algorithm (Loh 2011) to train a decision tree with randomly selected m ($m \le D$) features on each bootstrap sample.
3. Calculate the forecasted value of each trained decision tree on the test set and average them as the ultimate forecasted value.

Table 3 | Tested values of SVR parameters

Parameters	Tested values
Kernel function	Polynomial kernel, RBF
Polynomial kernel, d	{1,2,3,4,5}
Regularization parameter, C	100 numbers equally spaced between [0.01,10]

Table 4 | Tested values of RF parameters

Parameters	Tested values
Number of features, m	{3,4,5,6,7,8,9}
Number of decision trees, N_1	{10, 20, 30, …, 200}

The number of features (m) and the number of decision trees (N_1) are hyperparameters that need to be determined through experiments. A comprehensive number of experiments are carried out by varying the parameter values as shown Table 4.

3.1.3. Ridge regression

Ridge adopts a regularized loss function to compress the linear regression coefficients resulting in reduced variance but increased bias of coefficient estimators. Compared to linear regression, this method can avoid overfitting and the ridge estimator is preferably effective at enhancing the least-squares estimate when there is multicollinearity (Arashi *et al.* 2019).

Let $x_{ij},\ i = 1, 2, \ldots, n;\ j = 1, 2, \ldots, D$ represent D independent variables with a sample size n. Let $\{y_i\}_{i=1}^{n}$ denote the dependent variable, the regression equation is

$$\hat{y}_i = \beta_0 + \beta_1 x_{i1} + \ldots + \beta_D x_{iD}. \tag{10}$$

Minimizing the regularized loss function

$$\text{Loss} = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 + \lambda \sum_{j=0}^{D} \beta_j^2 \tag{11}$$

gives the ridge estimator

$$\hat{\boldsymbol{\beta}}_{\text{Ridge}} = (\boldsymbol{X}^T \boldsymbol{X} + \boldsymbol{\lambda I})^{-1} \boldsymbol{X}^T \boldsymbol{Y} \tag{12}$$

where $\hat{\boldsymbol{\beta}}_{\text{Ridge}} = (\beta_0, \beta_1, \ldots, \beta_D)^T$, $\boldsymbol{X}$ is a $n \times (D+1)$ matrix whose ith row is $(1, x_{i1}, x_{i2}, \ldots, x_{iD})$, $i = 1, 2, \ldots, n$, $\boldsymbol{Y} = (y_1, y_2, \ldots, y_n)^T$, and $\boldsymbol{I}$ is a $(D+1) \times (D+1)$ identity matrix. The regularization parameter is $\lambda > 0$ and it needs to be determined through experiments. One hundred numbers equally spaced between [0.001,10] are tested to find the optimal value of λ.

3.1.4. Artificial neural network

ANN mimics the central nervous system of the human brain. The schematic diagram of a fully connected feed-forward neural network designed in this study is shown in Figure 3. It contains an input layer, two hidden layers, and an output layer. The input layer contains nine neurons, which are the observations of nine independent variables. The output layer has only one neuron, which is the forecasted value of CO_2 emissions. The two hidden layers are set to share the same activation function and the output layer has no activation function. It has been theoretically verified that a three-layer neural network can reflect any continuous relationship with desired precision (Han *et al.* 2019).

The number of neurons for each hidden layer and the optimal activation function are hyperparameters determined by choosing the combination that gives the smallest test error. To that end, set {1, 2, 3, 4, 5, 6, 7, 8, 9, 10} for the number of neurons and three activation functions, i.e. sigmoid, hyperbolic tangent (tanh), and rectified linear unit (Relu), are tested (Bishop & Nasrabadi 2006), where

$$\text{sigmoid: } f(x) = \frac{1}{1 + e^{-x}} \tag{13}$$

$$\tanh: f(x) = \frac{e^{x} - e^{-x}}{e^{x} + e^{-x}} \tag{14}$$

$$\text{Relu: } f(x) = \max\{0,\ x\} \tag{15}$$

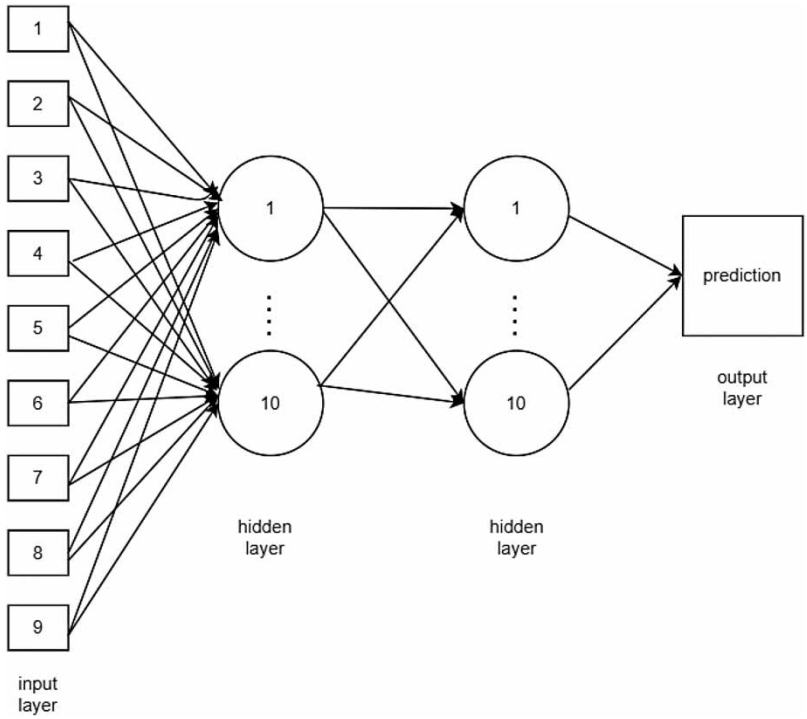


Figure 3 | Schematic diagram of the fully connected feed-forward neural network.

3.2. Two-stage forecast procedure

The single-stage forecast procedure with the input of the nine independent variables at time t and the output of CO_2 emissions at time $t+h$ is illustrated in Figure 4. The functional relationship between input and output is fitted by four ML models, respectively. That means each of the four ML models is trained on lagged data sets.

The single-stage procedure takes the time gap, h into account. To reduce the forecast errors caused by the time gap, this study designs a novel two-stage forecast procedure, as shown in Figure 5.

In the first stage of Figure 5, nine individual SVR models are applied to forecast nine independent variables at time $t+h$ by using their respective historical values up to time t. This procedure is completed through the sliding window method (Brownlee 2017), which is utilized to build a frame of supervised learning. That means we use previous time steps as input variables and use the next time step as the output variable, so that the chronological order is preserved. We set the window width as 3 from the experience of previous research (Faruque *et al.* 2022) and reorganize the time sequence of each independent variable as shown in Figure 6. Take $h=1$ for example (h can be set to any positive integer), the number of columns of the input matrix is 3, and each value in the output is at the next time step of the last input element at the same row. Finally, we use values at $t-2$, $t-1$, and t to forecast the value at $t+1$.

Next, the outputs of the first stage, i.e. the forecasted values of nine independent variables at time $t+h$, serve as the inputs of the second stage in Figure 5. The output of the second stage is the forecasted value of CO_2 emissions at time $t+h$, which is the same as the output of the single-stage procedure in Figure 4. This indicates that models used in the second stage need to fit the relationship between nine forecasted independent variables at time $t+h$ and CO_2 emissions at the same time domain.

Briefly, the first stage of the two-stage procedure updates the value of the independent variables at time t to time $t+h$, and the second stage forecasts CO_2 emissions at time $t+h$ through forecasted independent variables at time $t+h$.

To assess the generalization ability to unseen data of a particular model, the TimeSeriesSplit method (Pedregosa *et al.* 2011) is applied to cross-validate the four hybrids of ML models in the two-stage procedure and the four individual ML models in the single-stage procedure. It is a variation of k-fold cross-validation, which returns the first k folds as the training set and the $(k+1)$th fold as the testing set. In this experiment, the data set is split into eight groups of the training set and the testing set as

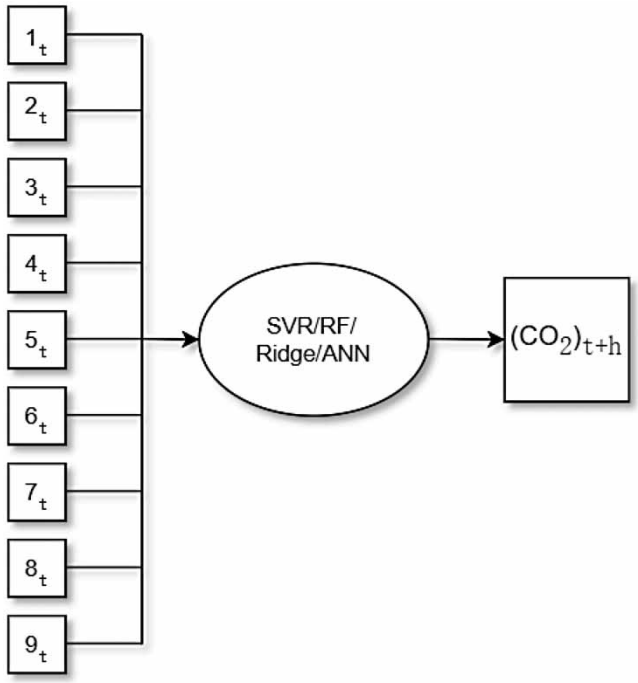


Figure 4 | The single-stage forecast procedure for forecasting CO_2 emissions *h* periods ahead.

shown in Figure 7. The yellow rectangles denote the training sets, followed by the blank rectangles that represent the testing sets. This indicates that testing indices are higher than training indices in each split, and successive training sets are supersets of those that come before them. The sample size of each testing set is fixed as 12, and the indices of each training set and testing set are also annotated in Figure 7.

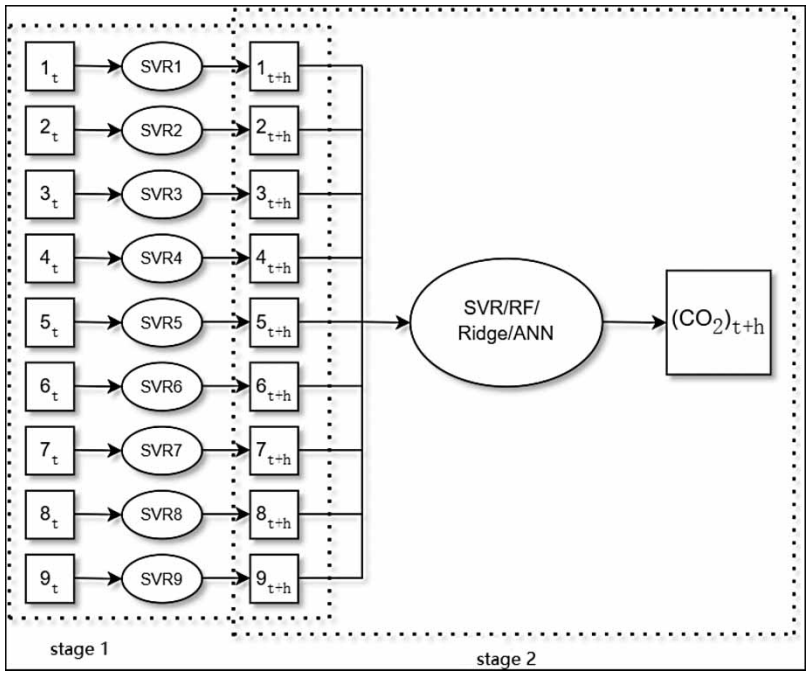


Figure 5 | The two-stage forecast procedure for forecasting CO_2 emissions *h* periods ahead.

input			output
1	2	3	4
2	3	4	5
3	4	5	6
.....			
t-3	t-2	t-1	t

Figure 6 | Sliding window method for forecasting time series. The number in the table denotes the time index of each individual independent variable.

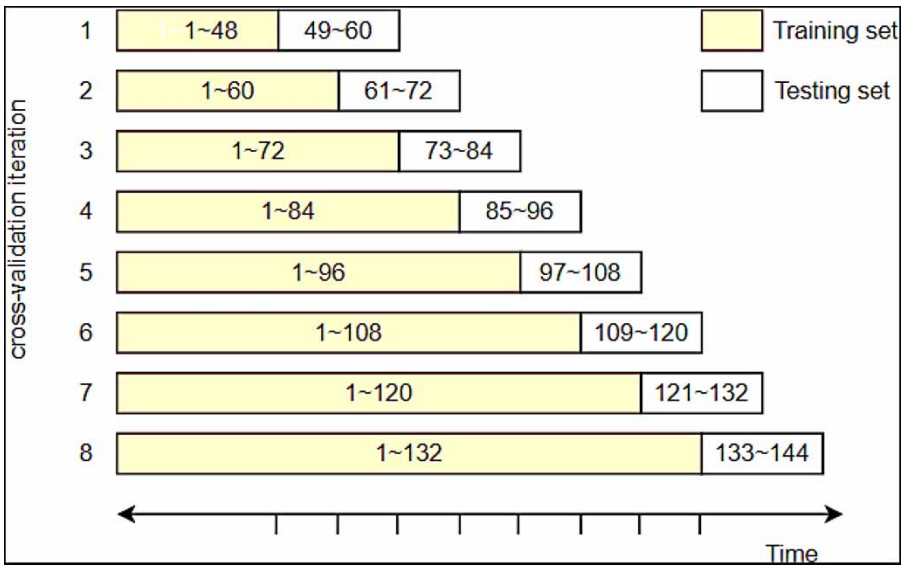


Figure 7 | TimeSeriesSplit method for time-based cross-validation. The data set is split into eight groups of training set and testing set. The yellow rectangles are training sets and the white rectangles are testing sets. The number represents the time index of each training set and testing set. Please refer to the online version of this paper to see this figure in colour: http://dx.doi.org/10.2166/wcc.2023.331.

We build a model on each training set and test it on the corresponding testing set, respectively. The final forecast error is averaged over the forecast errors of the eight testing sets.

4. RESULTS AND DISCUSSION

The single-stage forecast procedure tries four models: SVR, RF, Ridge, and ANN. The first stage of the two-stage forecast procedure uses SVR, and the second stage tries the same four models as in the single-stage version. So, we denote the two-stage models as SVR–SVR, SVR–RF, SVR–Ridge, and SVR–ANN. Tables 5 and 6 show the comparison of the average forecast error of the two-stage forecast procedure with that of the single-stage forecast procedure as h increases from 1 to 8. Figures 8 and 9 present the contents of Tables 4 and 5 in line chart form. In each figure, the horizontal axis is the number of periods h, and the vertical axis is the RMSE or MAE.

It is clear that SVR–RF vs. RF with $h = 2$ is the only scenario where the two-stage procedure shows a slightly larger average forecast error than the single-stage procedure. In all other cases, the average forecast errors of the two-stage procedure are smaller than that of the single-stage procedure.

Table 5 | Comparison of the average forecast error ($h = 1, 2, 3, 4$)

Model	RMSE ($h = 1$)	MAE ($h = 1$)	RMSE ($h = 2$)	MAE ($h = 2$)	RMSE ($h = 3$)	MAE ($h = 3$)	RMSE ($h = 4$)	MAE ($h = 4$)
SVR	0.0266	0.0233	0.0291	0.0252	0.0329	0.0286	0.0369	0.0323
SVR–SVR	0.0149	0.0123	0.0194	0.0169	0.0214	0.0188	0.021	0.0184
RF	0.0934	0.086	0.0871	0.0791	0.0932	0.0855	0.0928	0.0849
SVR–RF	0.0868	0.0789	0.088	0.0799	0.0884	0.0805	0.0865	0.078
Ridge	0.0243	0.0209	0.0236	0.0203	0.0291	0.0248	0.0341	0.0299
SVR–Ridge	0.0162	0.014	0.0177	0.0149	0.0146	0.0125	0.0162	0.0141
ANN	0.0133	0.0109	0.0138	0.0122	0.0164	0.0143	0.0136	0.0116
SVR–ANN	0.0099	0.0087	0.0106	0.0091	0.0124	0.0101	0.0122	0.0103

Table 6 | Comparison of the average forecast error ($h = 5, 6, 7, 8$)

Model	RMSE ($h = 5$)	MAE ($h = 5$)	RMSE ($h = 6$)	MAE ($h = 6$)	RMSE ($h = 7$)	MAE ($h = 7$)	RMSE ($h = 8$)	MAE ($h = 8$)
SVR	0.0414	0.0361	0.0451	0.0391	0.0497	0.0424	0.0522	0.0426
SVR–SVR	0.0282	0.0242	0.0304	0.026	0.0321	0.0271	0.039	0.0352
RF	0.0906	0.0824	0.0888	0.0807	0.0927	0.0833	0.0894	0.0803
SVR–RF	0.0867	0.0781	0.0821	0.0734	0.0813	0.0722	0.0853	0.0762
Ridge	0.036	0.0312	0.0383	0.0325	0.0403	0.0339	0.0428	0.0357
SVR–Ridge	0.017	0.014	0.0189	0.0161	0.0207	0.0171	0.0251	0.0217
ANN	0.016	0.0139	0.0153	0.0131	0.0164	0.0138	0.0173	0.0157
SVR–ANN	0.0154	0.013	0.0148	0.0124	0.0148	0.0128	0.016	0.0137

Figures 8 and 9 reveal that except for the RF and SVR–RF scenarios, the average forecast errors of all models display a slowly increasing trend as h moves up. But there is an apparent reduction in forecast error in the two-stage procedure compared with the single-stage procedure, which indicates that the two-stage forecast procedure proposed in this study indeed improves the forecasting accuracy of CO_2 emissions.

In addition, among these four two-stage models, SVR–ANN exhibits the smallest forecast error, whereas SVR–RF gives the greatest forecast error. The forecast errors of RF and SVR–RF show significant fluctuation as h increases. This can be explained by the characteristics of the RF algorithm, which consists of decision trees. According to Figures 1 and 2, four out of nine independent variables resemble the increasing trend of CO_2 emissions. Therefore, the decision tree may be randomly fitted by a single independent variable during training, which is insufficient for the forecasting task. This also implies that models involving decision trees may not be suitable for the data in this paper.

The percentage of the forecast error reduction of the four two-stage models relative to their single-stage versions is also calculated. The results are illustrated in Figure 10. It can be seen that the error reduction rates of SVR–ANN and SVR–RF are smaller than that of SVR–SVR and SVR–Ridge. Comparing the value of RMSE and MAE of SVR, RF, Ridge, and ANN in Tables 5 and 6, we find that both RMSE and MAE of ANN are the lowest for all h ranging from 1 to 8. It means that ANN already does a good job in the single-stage procedure, so there is little room left for improvement in the two-stage procedure. As for RF and SVR–RF, they both exhibit high forecast error, once again indicating models involving decision trees are not appropriate for the data we considered, either in the single-stage or two-stage procedure.

Finally, error reduction rates for the four two-stage models relative to their corresponding single-stage models ranging from 1 to 8 are averaged, respectively, resulting in Table 7. It indicates that the performance of SVR–Ridge and SVR–SVR is significantly improved than Ridge and SVR, respectively. The performance of SVR–ANN is moderately improved than ANN and SVR–RF is slightly improved than RF. This demonstrates the advantages of the proposed two-stage procedure.

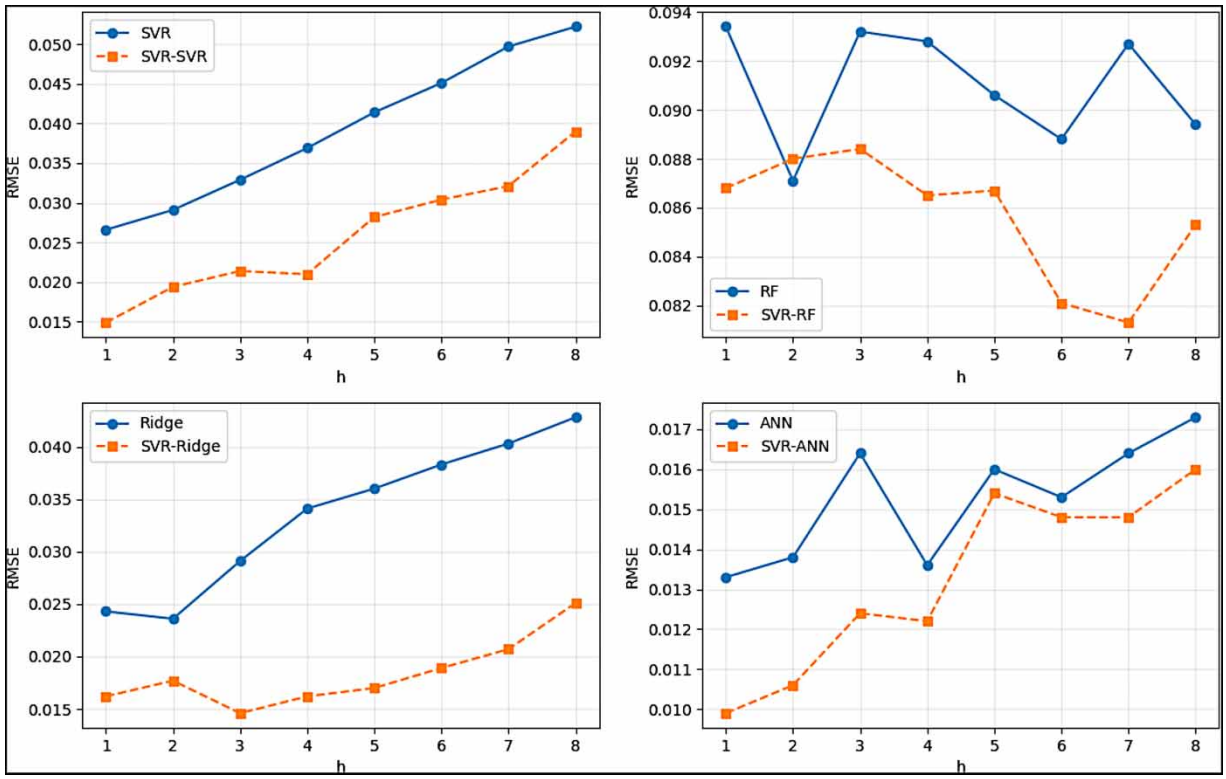


Figure 8 | Comparison of RMSE.

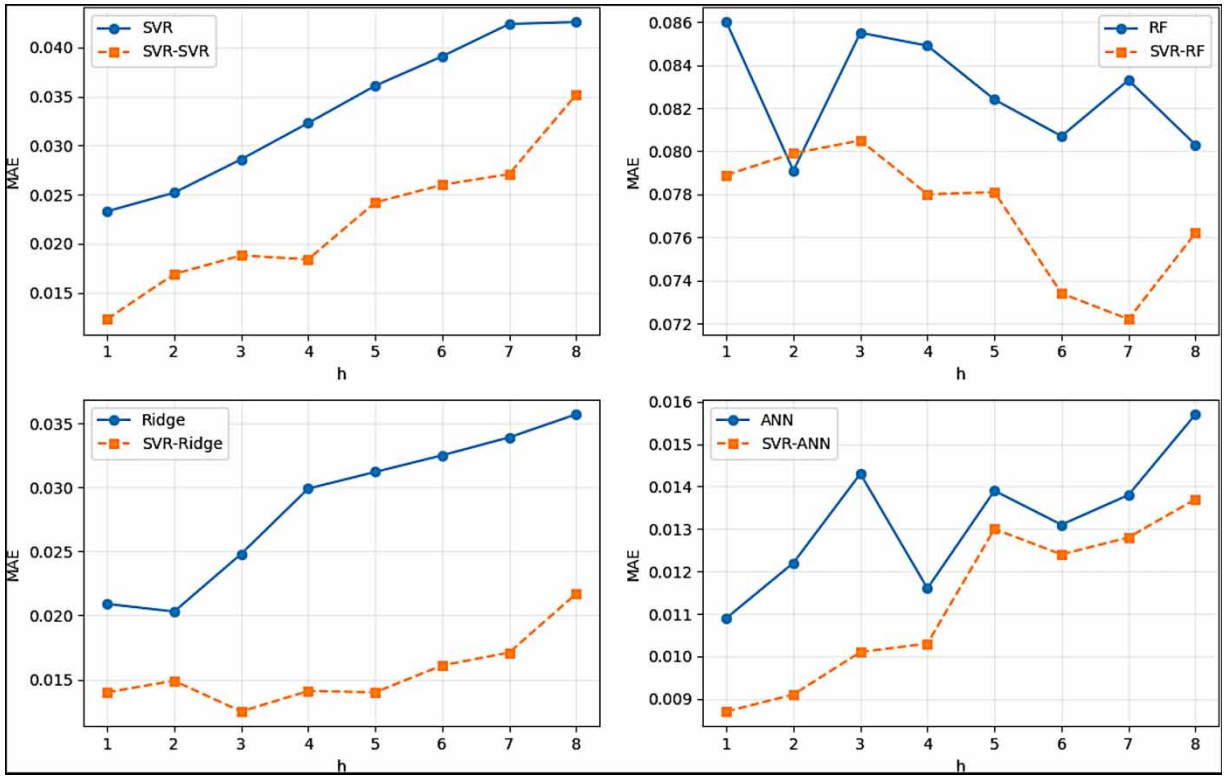


Figure 9 | Comparison of MAE.

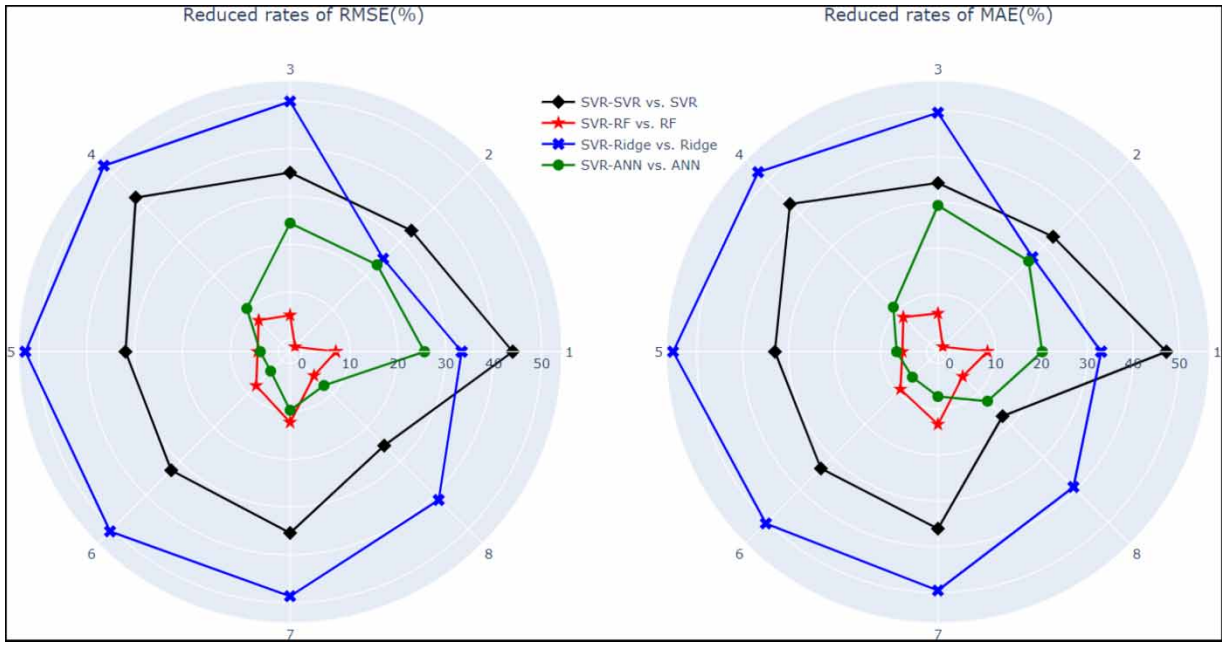


Figure 10 | Comparison of the error rate reduced by the two-stage procedure relative to the corresponding single-stage procedure for *h* from 1 to 8.

Table 7 | Average reduced rates of the four two-stage approaches vs. the four single-stage approaches

Models	RMSE	MAE
SVR–SVR vs. SVR (%)	36.06	36.06
SVR–RF vs. RF (%)	5.98	6.91
SVR–Ridge vs. Ridge (%)	43.05	43.27
SVR–ANN vs. ANN (%)	14.81	15.35

In conclusion, SVR–Ridge gives the best performance improvement than Ridge, but SVR–ANN has the lowest forecast error among all models. Therefore, it is appropriate to forecast China's CO_2 emissions via SVR–ANN.

In addition to the hybrids of SVR with the four ML models, alternative hybrids, i.e. RF–SVR, RF–RF, RF–Ridge, RF–ANN; Ridge–SVR, Ridge–RF, Ridge–Ridge, Ridge–ANN; and ANN–SVR, ANN–RF, ANN–Ridge, ANN–ANN, are also tested in the study. The resulting figures of RMSE, MAE, and reduced rates are provided in the Supplementary Material. It is observed that the two-stage models all have better prediction performance than their corresponding single-stage models.

5. CONCLUSION

Nine relevant independent variables have been adopted to forecast production-based CO_2 emissions in China. In the literature, it is found that some studies may not train their models on lagged data sets, and some train models on lagged data sets but ignore the forecast errors caused by the time gap, *h*, in the dependent variable. We develop a novel two-stage procedure based on ML models in order to reduce the forecast errors identified. The hybrids of ML models in the two-stage procedure are compared against the corresponding individual models in the single-stage procedure.

Our experiments point out that the average forecast errors of the four two-stage methods, SVR–SVR, SVR–RF, SVR–Ridge, and SVR–ANN, are almost consistently smaller than their single-stage versions. Other combinations of SVR, RF, Ridge, and ANN are also compared against the corresponding single-stage procedures. It is observed that our two-stage procedure can improve forecasting performance. An accurate forecast of CO_2 emissions can help policymakers to schedule carbon emission reduction plans more efficiently and promote the realization of the carbon peaking and carbon neutrality goals.

The window width used in the first stage of the two-stage procedure is fixed as 3. It may be worth trying more possibilities to give more accurate forecasts. Since ML models have the advantages of handling big data with many inputs, another direction for future study is to consider more independent variables that affect CO_2 emissions, such as consumer price index and environmental policy stringency (Ahmed & Ahmed 2018).

Finally, the two-stage procedure is also suitable for other forecast tasks. For example, it is possible to forecast the consumption of different types of fossil fuels and renewable energy, which are listed before.

DATA AVAILABILITY STATEMENT

All relevant data are available from an online repository or repositories.

CONFLICT OF INTEREST

The authors declare there is no conflict.

REFERENCES

Acheampong, A. O. & Boateng, E. B. 2019 Modelling carbon emission intensity: application of artificial neural network. *Journal of Cleaner Production* **225**, 833–856.

Ahmed, K. & Ahmed, S. 2018 A predictive analysis of CO_2 emissions, environmental policy stringency, and economic growth in China. *Environmental Science and Pollution Research* **25** (16), 16091–16100.

Al-Mulali, U., Ozturk, I. & Solarin, S. A. 2016 Investigating the environmental Kuznets curve hypothesis in seven regions: the role of renewable energy. *Ecological Indicators* **67**, 267–282.

Arashi, M., Saleh, A. M. E. & Kibria, B. G. 2019 Theory of Ridge Regression Estimation with *Applications*. John Wiley & Sons, Ames, IA.

Belgiu, M. & Drăguţ, L. 2016 Random forest in remote sensing: a review of applications and future directions. *ISPRS Journal of Photogrammetry and Remote Sensing* **114**, 24–31.

Bhatt, R. & Hossain, A. 2019 Concept and consequence of evapotranspiration for sustainable crop production in the era of climate change. *Advanced Evapotranspiration Methods and Application* **1**, 1–13.

Bishop, C. M. & Nasrabadi, N. M. 2006 *Pattern Recognition and Machine Learning*, Vol. 4(4). Springer, New York, p. 738.

Boulesteix, A. L., Janitza, S., Kruppa, J. & König, I. R. 2012 Overview of random forest methodology and practical guidance with emphasis on computational biology and bioinformatics. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery* **2** (6), 493–507.

Brownlee, J. 2017 *Introduction to Time Series Forecasting with Python: How to Prepare Data and Develop Models to Predict the Future*. Machine Learning Mastery. Available from: https://machinelearningmastery.com/introduction-to-time-series-forecasting-with-python/.

Chow, G. C. & Lin, A. L. 1971 Best linear unbiased interpolation, distribution, and extrapolation of time series by related series. *The Review of Economics and Statistics* **53** (4), 372–375.

Djaafari, A., Ibrahim, A., Bailek, N., Bouchouicha, K., Hassan, M. A., Kuriqi, A., Al-Ansar, N. & El-Kenawy, E. S. M. 2022 Hourly predictions of direct normal irradiation using an innovative hybrid LSTM model for concentrating solar power projects in hyper-arid regions. *Energy Reports* **8**, 15548–15562.

Faruque, M. O., Rabby, M. A. J., Hossain, M. A., Islam, M. R., Rashid, M. M. U. & Muyeen, S. M. 2022 A comparative analysis to forecast carbon dioxide emissions. *Energy Reports* **8**, 8046–8060.

Farzin, S., Anaraki, M. V., Naeimi, M. & Zandifar, S. 2022 Prediction of groundwater table and drought analysis; a new hybridization strategy based on bi-directional long short-term model and the Harris hawk optimization algorithm. *Journal of Water and Climate Change* **13** (5), 2233–2254.

Friedlingstein, P., Jones, M. W., O'Sullivan, M., Andrew, R. M., Bakker, D. C., Hauck, J., Le Quéré, C., Peters, G. P., Peters, W., Pongratz, J. & Sitch, S. 2022 Global carbon budget 2021. *Earth System Science Data* **14** (4), 1917–2005.

Gallo, C., Conto, F. & Fiore, M. 2014 A neural network model for forecasting CO_2 emission. *AGRIS on-Line Papers in Economics and Informatics* **6**, 31–36.

Geevaretnam, J. L., Zainuddin, N. M. M., Kamaruddin, N., Rusli, H., Maarop, N. & Hassan, W. A. W. 2022 Predicting the carbon dioxide emissions using machine learning. *International Journal of Innovative Computing* **12** (2), 17–23.

Han, M., Ding, L., Zhao, X. & Kang, W. 2019 Forecasting carbon prices in the Shenzhen market, China: the role of mixed-frequency factors. *Energy* **171**, 69–76.

Hou, Y., Wang, Q. & Tan, T. 2022 Prediction of carbon dioxide emissions in China using shallow learning with cross validation. *Energies* **15** (22), 8642.

Kuriqi, A., Pinheiro, A. N., Sordo-Ward, A. & Garrote, L. 2019 Influence of hydrologically based environmental flow methods on flow alteration and energy production in a run-of-river hydropower plant. *Journal of Cleaner Production* **232**, 1028–1042.

Kuriqi, A., Pinheiro, A. N., Sordo-Ward, A. & Garrote, L. 2020 Water-energy-ecosystem nexus: balancing competing interests at a run-of-river hydropower plant coupling a hydrologic–ecohydraulic approach. *Energy Conversion and Management* **223**, 113267.

Kuriqi, A., Pinheiro, A. N., Sordo-Ward, A., Bejarano, M. D. & Garrote, L. 2021 Ecological impacts of run-of-river hydropower plants – current status and future prospects on the brink of energy transition. *Renewable and Sustainable Energy Reviews* **142**, 110833.
Li, Y. 2020 Forecasting Chinese carbon emissions based on a novel time series prediction method. *Energy Science & Engineering* **8** (7), 2274–2285.
Lin, C. S., Liou, F. M. & Huang, C. P. 2011 Grey forecasting model for CO_2 emissions: a Taiwan study. *Applied Energy* **88** (11), 3816–3820.
Liu, Z., Ciais, P., Deng, Z., Davis, S. J., Zheng, B., Wang, Y., Cui, D., Zhu, B., Dou, X., Ke, P. & Sun, T. 2020 Carbon monitor, a near-real-time daily dataset of global CO_2 emission from fossil fuel and cement production. *Scientific Data* **7** (1), 1–12.
Loh, W. Y. 2011 Classification and Regression Trees. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery* **1** (1), 14–23.
Lotfalipour, M. R., Falahi, M. A. & Bastam, M. 2013 Prediction of CO_2 emissions in Iran using grey and ARIMA models. *International Journal of Energy Economics and Policy* **3** (3), 229–237.
Lu, I. J., Lewis, C. & Lin, S. J. 2009 The forecast of motor vehicle, energy demand and CO_2 emission from Taiwan's road transportation sector. *Energy Policy* **37** (8), 2952–2961.
Malka, L., Daci, A., Kuriqi, A., Bartocci, P. & Rrapaj, E. 2022 Energy storage benefits assessment using multiple-choice criteria: the case of Drini River Cascade, Albania. *Energies* **15** (11), 4032.
Mardani, A., Liao, H., Nilashi, M., Alrasheedi, M. & Cavallaro, F. 2020 A multi-stage method to predict carbon dioxide emissions using dimensionality reduction, clustering, and machine learning techniques. *Journal of Cleaner Production* **275**, 122942.
Morshed-Bozorgdel, A., Kadkhodazadeh, M., Valikhan Anaraki, M. & Farzin, S. 2022 A novel framework based on the stacking ensemble machine learning (SEML) method: application in wind speed modeling. *Atmosphere* **13** (5), 758.
Mselmi, N., Lahiani, A. & Hamza, T. 2017 Financial distress prediction: the case of French small and medium-sized firms. *International Review of Financial Analysis* **50**, 67–80.
Ning, L., Pei, L. & Li, F. 2021 Forecast of China's carbon emissions based on ARIMA method. *Discrete Dynamics in Nature and Society* **2021**, 1–12.
Nyoni, T. & Bonga, W. G. 2019 Prediction of CO_2 emissions in India using ARIMA models. *DRJ-Journal of Economics & Finance* **4** (2), 01–10.
Ouedraogo, I., Defourny, P. & Vanclooster, M. 2019 Application of random forest regression and comparison of its performance to multiple linear regression in modeling groundwater nitrate concentration at the African continent scale. *Hydrogeology Journal* **27** (3), 1081–1098.
Papouskova, M. & Hajek, P. 2019 Two-stage consumer credit risk modelling using heterogeneous ensemble learning. *Decision Support Systems* **118**, 33–45.
Patel, J., Shah, S., Thakkar, P. & Kotecha, K. 2015 Predicting stock market index using fusion of machine learning techniques. *Expert Systems with Applications* **42** (4), 2162–2172.
Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V. & Vanderplas, J. 2011 Scikit-learn: Machine learning in Python. *The Journal of Machine Learning Research* **12**, 2825–2830.
Platt, J. 1998 *Sequential Minimal Optimization: A Fast Algorithm for Training Support Vector Machines*. Technical Report MSR-TR-98-14, Microsoft. https://www.microsoft.com/en-us/research/publication/sequential-minimal-optimization-a-fast-algorithm-for-training-support-vector-machines/.
Rabaia, M. K. H., Abdelkareem, M. A., Sayed, E. T., Elsaid, K., Chae, K. J., Wilberforce, T. & Olabi, A. G. 2021 Environmental impacts of solar energy systems: a review. *Science of The Total Environment* **754**, 141989.
Ramasubramanian, K. & Singh, A. 2017 *Machine Learning Using R (No. 1)*. Apress, New Delhi, India.
Rehman, A., Ma, H., Ahmad, M., Ozturk, I. & Chishti, M. Z. 2021 How do climatic change, cereal crops and livestock production interact with carbon emissions? Updated evidence from China. *Environmental Science and Pollution Research* **28** (24), 30702–30713.
Sadorsky, P. 2021 Wind energy for sustainable development: driving factors and future outlook. *Journal of Cleaner Production* **289**, 125779.
Saleh, C., Dzakiyullah, N. R. & Nugroho, J. B. 2016 Carbon dioxide emission prediction using support vector machine. *IOP Conference Series: Materials Science and Engineering* **114** (1), 012148.
Smola, A. J. & Schölkopf, B. 2004 A tutorial on support vector regression. *Statistics and Computing* **14** (3), 199–222.
Stamenković, L. J., Antanasijević, D. Z., Ristić, M. Đ., Perić-Grujić, A. A. & Pocajt, V. V. 2015 Modeling of methane emissions using artificial neural network approach. *Journal of the Serbian Chemical Society* **80** (3), 421–433.
Valipour, M. 2017 Calibration of mass transfer-based models to predict reference crop evapotranspiration. *Applied Water Science* **7** (2), 625–635.
Vapnik, V. 1999 *The Nature of Statistical Learning Theory*. Springer Science & Business Media, Berlin/Heidelberg, Germany.
Wang, C. & Zhu, M. 2021 Scenario prediction of China's natural gas consumption and carbon emissions in the next ten years. *Frontiers of Economics in China* **16** (3), 569–587.
Wang, S., Zhao, Y. & Wiedmann, T. 2019 Carbon emissions embodied in China–Australia trade: a scenario analysis based on input–output analysis and panel regression models. *Journal of Cleaner Production* **220**, 721–731.
Yang, H. & O'Connell, J. F. 2020 Short-term carbon emissions forecast for aviation industry in Shanghai. *Journal of Cleaner Production* **275**, 122734.

First received 10 September 2022; accepted in revised form 21 January 2023. Available online 1 February 2023