Cleaner Chemical Engineering 5 (2023) 100095

Contents lists available at ScienceDirect

Cleaner Chemical Engineering

journal homepage: www.elsevier.com/locate/clce

# Forecasting and mitigation of global environmental carbon dioxide emission using machine learning techniques

Harsh Bhatt[a], Manan Davawala[a], Tanmay Joshi[a], Manan Shah[b,*], Ashish Unnarkat[b]

[a] Department of Computer Science and Engineering, Nirma University, Ahmedabad, Gujarat, India
[b] Department of Chemical Engineering, School of Energy Technology, Pandit Deendayal Energy University, Gandhinagar, Gujarat, India

ARTICLE INFO

*Keywords:*
Machine learning
$CO_2$ emission
Prediction

ABSTRACT

Carbon dioxide emission has emerged as a major concern in the 21st century. The rising global average temperature and its impact on climate change has a major impact on the socioeconomic affairs of the world. This rise, directly causes the melting of the polar ice caps which in turn, brings about many other issues including extinction of polar animals, flooding of coastal regions, exposure to ancient microbial life and bacteria frozen in the snow which pose a risk of many more global pandemics and unseen diseases. An urgent need to control this carbon emission is required. The initial step in this process is to accurately identify the threat levels and milestones. Certain thresholds need to be mapped that express the most critical levels of $CO_2$ such as – the risk point, point of no return, etc. This is the main issue that this paper aims to address. The paper also aims to suggest some methodologies to deal with the same issue. The flow of the experiment and paper is described in the following lines. Historical data was used to make a prediction for the year in which the earth will hit a particular threshold for carbon dioxide concentration in the atmosphere. This level must not be breached and is essential in the fight against climate change. Next, analysis and data are used to calculate the reduction needed in the emission levels in order to bring back the $CO_2$ concentrations into a safer range. The study concludes that the critical level of $CO_2$ - 500 ppm, will be achieved by the year 2047. This level is considered a point of no return. A reduction rate of 6.37% and reversal rate of 23.38% is required to bring the emissions back to safe levels. The study also concluded that various socioeconomic factors such as population, greenhouse gasses, combustion industries contribute the most to these emissions. The authors recommend that further research be carried out on this problem to ascertain further predictions on the point of no return so that an action plan can be developed accordingly. The authors also recommend that a shift to renewable energy sources be undertaken speedily and that carbon neutrality be a crucial goal of every organization.

## 1. Introduction

Climate change is a contemporary issue that has been the showcase of the times for a significant number of years. Greenhouse gasses contribute heavily to climate change and the main forerunner of this is carbon dioxide ($CO_2$). The additional release of $CO_2$ breaks the standard carbon cycle hence having adverse effects on global climate such as rise in temperatures. Our climate alters inextricably when the average temperature rises. Extreme weather occurrences, such as tropical storms, wildfires, severe droughts and heat waves are caused by this warming.

Carbon emissions also have a direct impact on humans, producing an increase in respiratory ailments due to smog and air pollution. In addition, carbon emissions can potentially wipe out particular animal species, disrupt crop yields, and destroy land.

Human activities are the main source of carbon emissions. Transportation, power production, industry, commercial and residential, agriculture, and land use and forestry are the six main contributors of greenhouse gasses, according to the Environmental Protection Agency.

Despite the decrease in 2020 due to the COVID-19 pandemic, global energy-related $CO_2$ emissions remained at 31.5 Gt, contributing to $CO_2$ reaching its highest ever average annual concentration in the atmosphere of 412.5 parts per million in 2020 – roughly 50% higher than when the industrial revolution began. Global energy-related $CO_2$ emissions are expected to rebound and increase by 4.8 percent in 2021, as demand for coal, oil, and gas recovers along with the economy.

The ultimate goal of every country should be to achieve carbon neutrality. Carbon neutrality implies that there is a state of net-zero carbon emissions. To achieve this, each country must have accurate and viable

* Corresponding author.
*E-mail address:* manan.shah@spt.pdpu.ac.in (M. Shah).

https://doi.org/10.1016/j.clce.2023.100095
Received 13 June 2022; Received in revised form 3 January 2023; Accepted 20 January 2023

data on its carbon footprint and the emissions from its various sectors. The Sustainable Development Goals (SDGs) and the Paris Climate Agreement of 2015 aimed to launch a global initiative, but most countries struggle to achieve the proposed carbon neutrality target which is essential for low $CO_2$ emissions (Shao et al., 2021). The main challenge in reducing carbon emissions is that the emissions can propagate through space and time, but can only be reduced at the time and location of occurrence of the emission themselves (Kriegler et al., 2013).

While technologies and methodologies to reduce carbon emissions and boost carbon neutrality exist [6–8], technologies to predict the emission rate are relatively scarce.

Certain researches and experiments have been conducted to map, analyze and predict the carbon emissions from various sectors and sources individually (Li et al., 2021; Liu et al., 2021; Nandi et al., 2021). These analysis and predictions can then help target the particular sources and sectors that can actively be changed to other sustainable forms. One such example of this is the energy industry that can be shifted to other natural sources and renewable energy.

However, there is a lack of research study that provides a holistic analysis of these emissions. This study aims to bridge this research gap by taking into account emissions from various sources. Predictions of $CO_2$ emissions from an exhaustive list of sources can help accurately demarcate milestones in time by which certain changes should be brought about in these levels.

This study provides a novelty by estimating the year by which the damage from $CO_2$ emissions to the environment would be irreversible, which helps establish one such crucial milestone. In addition, this study also provides accurate calculations of the reversal rate of $CO_2$ emissions required for attaining safe atmospheric carbon dioxide levels.

$CO_2$ predictions are mostly required in industries closely related to heavy energy consumption/production. The prediction of energy consumption and carbon emissions are essential to assist a company's energy improvement and carbon efficiency activities. For these purposes, generally time series data (TS data) is required (Modise and Mpofu, 2022). Statistical methods and semi hybrid methods such as ARIMA (Auto-Regressive Integrated Moving Average) (Khashei and Bijari, 2011) can be used for this purpose. However, predictions are best carried out by machine learning and deep learning techniques.

Machine Learning (ML) and Deep Learning (DL) techniques have come a long way since their introduction and are used in almost every industry in the world from healthcare to manufacturing to logistics and more. These techniques have also found their way into being utilized for climate change (Chantry et al., 2021; Rolnick et al., 2023). They provide an edge over basic statistical methods by utilizing multiple features and algorithms unlike traditional statistical models.

ML and DL techniques form various computerized and mathematical models which learn from the data provided and then predict future values. ML and DL algorithms can provide very accurate results provided that they are modeled accordingly.

Several well-established algorithms have been utilized for predicting $CO_2$ emission values.

Support Vector Machine algorithms are used for regression and classification purposes and are a type of supervised learning method utilizing a labeled dataset for training (Noble, 2006). They work on the principle of decision planes for determining the decision boundaries. This algorithm was utilized to predict carbon emission expenditure from input variables consisting of energy consumption from electrical energy and burning coal (Saleh et al., 2016).

K-Nearest Neighbors is a simple machine learning technique, also called a lazy-learning algorithm (Peterson, 2009). It is based on pattern recognition and can be used for regression and classification purposes. This algorithm classifies a new unidentified point by checking 'k' of its nearest neighbors. It was used to predict the amount of seafloor carbon (Lee et al., 2019).

Artificial Neural Networks are a type of deep learning technique. They are developed to replicate the working of neurons in human brains (Boger and Guterman, 1997) and provide excellent prediction results. A neural network model was developed and utilized for predicting carbon emissions in Apulia region of Southern Italy to study the impact of cultural methods on pollution reduction (Gallo et al., 2014).

### 1.1. Research objectives

Considering the above stated situation, this research aims to identify, study and find impactful results on the following issues:

- The year in which the $CO_2$ emissions will hit the critical level of 500 ppm
- The reduction and reversal required in the current emission levels to bring them down to the safe level of 316 ppm
- The relationship between various social and economic fields of a country and how they impact $CO_2$ emissions

This paper covers an experiment carried out on a dataset taken from the United States regarding its annual carbon emission over the years. The different features are stated and studied along with their relationship to carbon emission. The crucial factors are identified and highlighted. Machine learning models were developed, trained and used to predict future carbon emission values. These were then used to predict certain theoretical thresholds deemed crucial by field experts. The challenges and future scope regarding the field and methodologies have also been discussed.

## 2. Related works

Over time, several researchers have conducted experiments to predict the amount of carbon emissions.

Xu et al. (2021) conducted an experiment to forecast the carbon dioxide emissions in 53 countries and various other regions using a technique called non-equigap gray model. The experiment utilized energy consumption as input and carbon dioxide emissions as output.

Wang et al. (2020) used a Land Use Regression (LUR) model to predict air pollution levels from traffic. A dataset consisting of various features like time, coordinates, traffic, weather, and land use variables was utilized.

Hamrani et al. (2020) utilized nine different machine learning and deep learning models to predict agricultural soil greenhouse gas emissions. The LSTM model they developed gave the best performance for N2O and $CO_2$ prediction.

Magazzino et al. (2021) used advanced machine learning techniques such as Causal Direction from Dependency algorithm to predict and analyze the relationship between solar and wind energy production, coal consumption, GDP and $CO_2$ emissions.

Mardani et al. (2018) utilized multiple techniques from clustering, dimensionality reduction and machine learning to predict carbon emission. Techniques such as singular value decomposition (SVD), Fuzzy neural network.

Zhang et al. (2021) used machine learning techniques to predict urban blocks carbon emission (UBCE) based on environmental factors in Changxing city, China.

Jeff (2019a)) used machine learning techniques on time series data to predict $CO_2$ emissions in India. They utilized models such as Linear Regression, Random Forest Regression and LSTM.

Several works in this field used features that did not fully cover the spectrum of factors actively affecting $CO_2$ emissions. This study exhaustively takes into account several unutilized factors from various sectors including - industrial outlets such as powerhouses, power plants and factories; economic features such as GDP; energy utilization and combustion sources such as coal, oil and gas; emissions due to greenhouse gasses such as nitrous oxide and methane.

It was also observed that most of the similar studies of this field used a single machine learning technique thereby restricting the outcomes and results. This study utilizes dimensionality reduction and ensemble methods that combine multiple well-established machine learning techniques namely – bagging and boosting. These have been proven to provide better results. Using multiple techniques also ensures that the results have corroboration.

## 3. Dataset and methodology

The dataset that is used to build the machine learning model is taken from the United States of America containing the information regarding $CO_2$ emissions in the country itself. The dataset contains a total of 38 columns and 219 rows. The columns denote different features that are directly or indirectly responsible for calculation of $CO_2$ emissions in the USA (Newcomer et al., 2008). There are 219 rows showcasing the same number of different records that are the experimental values calculated for different years. The term energy-related $CO_2$ emissions, as used in these tables, refers to emissions released at the location where fossil fuels are combusted. If fuels are combusted in one state to generate electricity consumed in another state, we attribute the emissions to the state where the electricity was generated.

The dataset also contains features that are directly related to the economy (G.D.P. in this experiment). These features give clear signs that the population of the country and the economy of the country play a vital role in analyzing the $CO_2$ emissions in the country for past, present and future years. The dataset also contains the experimental values of different greenhouse gasses such as methane, nitrous oxide etc. This indicates that the emissions of such different greenhouse gasses are related to each other. The dataset is primarily about the change in $CO_2$ emissions in the past years. This may help to find a pattern and draw inferences about the future emissions. Eventually, this can help slow down and eventually stop the adverse effects that global warming can have if appropriate measures are not taken. Curbing of $CO_2$ emissions from factories and powerplants is one such step along with similar measures in different fields (Bjp et al., 1992).

### 3.1. Feature description

The dataset contains a total of 38 attributes (Blum and Langley, 1997). A brief description of these features is mentioned below:

1. **Year:** A very important feature of the dataset. Each tuple has experimental values for that particular year. It is used to extrapolate our predictions and predict that if the emissions increase at present rate, after what year will we not be able to achieve carbon neutrality.
2. **$CO_2$:** This is the target label of our machine learning model. The unit of the values is million-tones.
3. **Related to $CO_2$ consumption:** The consumption of $CO_2$ by different powerhouses, power plants and factories in the country.
4. **Related to $CO_2$ emission trading:** Carbon emissions trading is a scheme that permits businesses to buy and sell government-issued carbon dioxide production allotments. Carbon taxes or carbon emissions trading are used by 40 countries and 20 municipalities, according to the World Bank (Lin and Jia, 2019). This equates to 13% of global greenhouse gas emissions each year.
5. **Cumulative consumption and per-capita features:** These features have cumulative calculated values for above mentioned features. Some features are also related to per-capita indicating, they directly depend on the GDP and the population of the US.
6. **Energy_burning_sources**: This contains the $CO_2$ emissions when coal, oil and gas were used as sources of generating energy.
7. **Emissions of other greenhouse gasses:** The emissions of other greenhouse gasses such as Nitrous Oxide and Methane (Wunch et al., 2009).
8. **Population:** Population of the United States in a given year.
9. **GDP:** Gross Domestic Product of the United States for a given year.

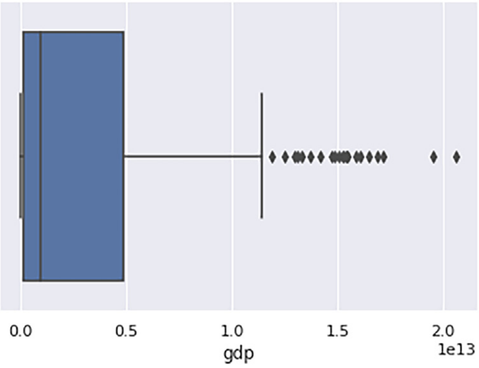

**Fig. 1.** Boxplot diagram for 'gdp' attribute.

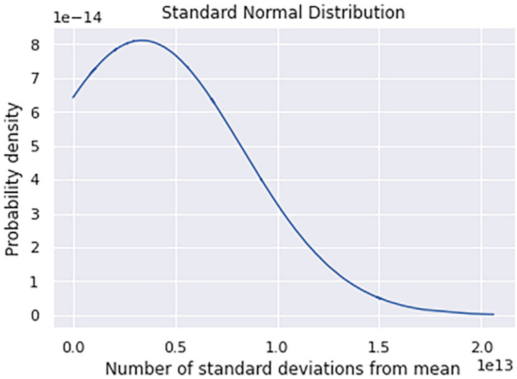

**Fig. 2.** Standard Normal Distribution for 'gdp' attribute.

### 3.2. Exploratory data analysis

I **Missing values**

Missing data is defined as values that aren't stored or absent for one or more variables in a dataset. During implementation it was found that missing values were present in 28 features out of 36. The missing values were for both categorical features and numerical features. Thresholding techniques were used to handle such data. The threshold was 80 percent. The features with the least missing values were replaced by zero. The remaining attributes were filled with the median values, as these features had many outliers the data was positively skewed. The box plot and the normal distribution plot of the 'gdp' feature is shown below in Fig. 1 and Fig. 2 respectively:

The columns with missing values greater than the threshold were dropped. 15 features were dropped.

II **Outliers**

Some of the features had outliers but this number was unsubstantial. Hence, the outlier values were not handled in order to maintain the variance of data

III **Feature engineering**

In machine learning, feature engineering is a critical stage. The technique of designing artificial features into an algorithm is known as feature engineering (Heaton, 2016). These developed traits are then used by the algorithm to increase its performance, or to achieve better outcomes. In the dataset there are 21 features. First a covariance matrix was plotted to get an intuition about interdependency of the features and which features should be removed. The covariance matrix is shown below in Fig. 3:

The features having negative correlation with the target feature - $CO_2$ - were then extracted. Further analysis was conducted by plotting the scatter plots of both the features which revealed no concrete evidence to eliminate these features and hence they were retained.

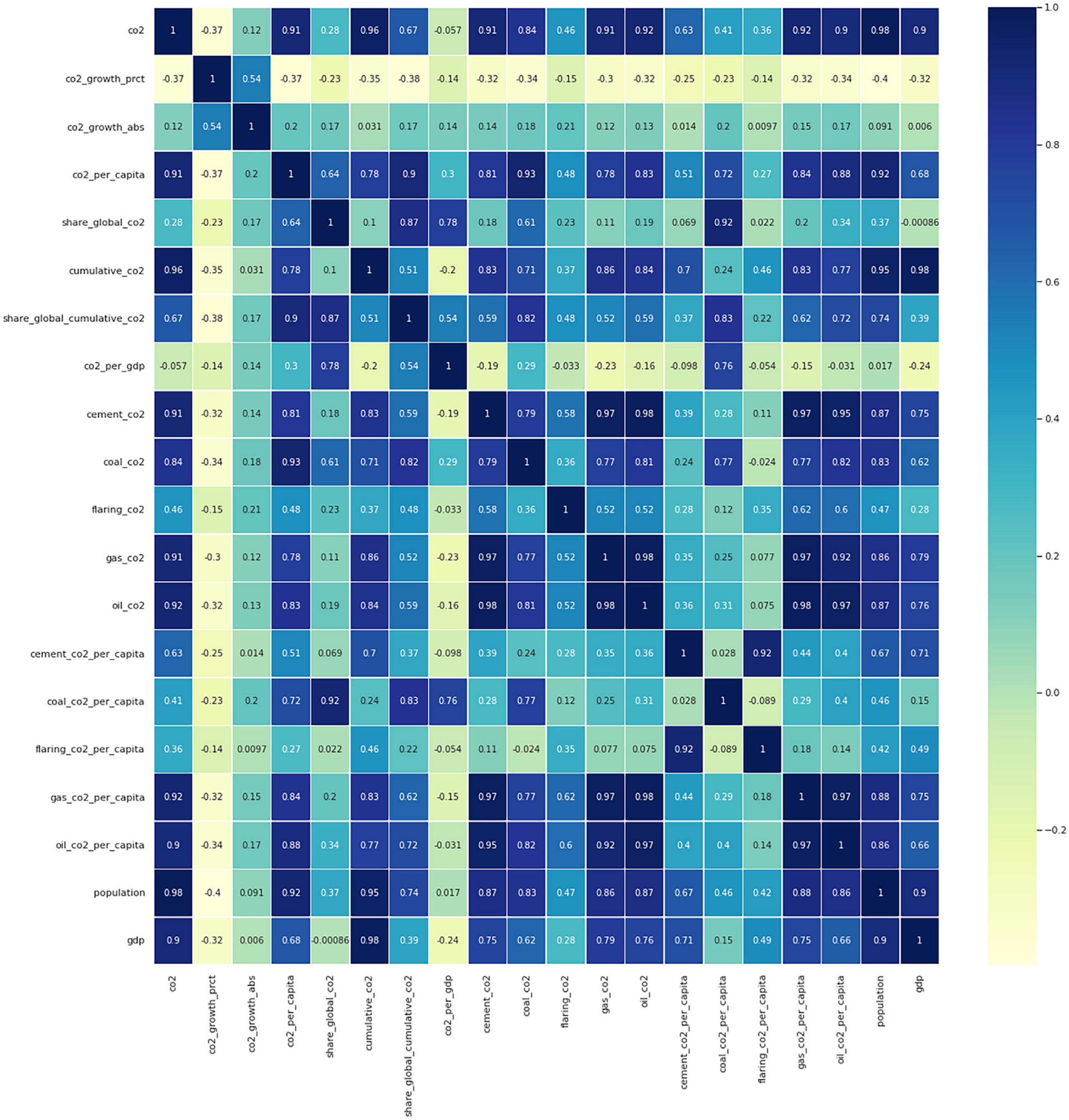

**Fig. 3.** Covariance matrix.

Following this, highly correlated features were extracted. The threshold was 90 percent and above. It was found that 8 features had a very high correlation with our target variable. For further analysis the heatmap was plotted for these highly correlated features.

On analyzing the above heatmap, it was clearly observed that the features of cement, gas and oil had a very high correlation of above 98 percent. Dimensionality reduction was done using Principal Component Analysis technique (Abdi and Williams, 2010).

### 3.3. Machine learning models

Machine Learning models like linear regression, ridge regression, lasso regression, k-nearest neighbor (KNN) regression, polynomial regression, forest regression, decision tree regression, gradient boosting regression, support vector regression and various neural network models were used for the predictions. These state-of-the-art models were chosen as they are highly effective, efficient, relevant and computationally simple.

### 3.4. Linear regression

Linear Regression (LR) is the most rudimentary supervised learning method for making predictions. The algorithm works by using the input parameters and output parameters in the training set and finding the optimized set of coefficients. This set of coefficients are formally known as parameters. The Gradient Descent techniques on Mean Squared Loss functions are used in this study for the yield of the optimized vector of Parameters (Liang and Song, 2009).

$$\hat{y} = \theta_0 + \theta_1 x_1 + \theta_2 x_2 + \ldots. + \theta_m x_m \quad (1)$$

Linear Regression prediction function

$$J(\theta) = \sum_{i=1}^{n} (y - \hat{y})^2/2n \qquad (2)$$

MSE cost function

In the above equations (Eqs. (1), (2)), $\theta$ represents the corresponding parameters and J ($\theta$) is the loss function.

### 3.5. Ridge regression

Ridge Regression has an analogous hypothesis to that for linear regression (LR) with added L2-Norm Squared Regulation, which does not apply when evaluating performance in a test set or predicting a real sample. Hyper-parameter controls how much the model will be controlled. When the hyper-parameters are huge, all parameters tend to be zero and eventually become a horizontal line over the average of the data (Liang and Song, 2009).

$$J(\theta) = \sum_{i=1}^{n} (y - \hat{y})^2/2n + \alpha \sum_{i=1}^{m} \theta_i^2 \qquad (3)$$

Regularized cost function - Ridge Regression

In the above equation (Eq. (3)), y represents truth value, $\hat{y}$ represents predictions, $\theta$ the corresponding parameters, $\alpha$ is the penalty and J ($\theta$) is the loss function.

### 3.6. Lasso regression

Lasso Regression also has a similar hypothesis to that for linear regression (LR) with added L1-Norm Absolute Regulation which does not apply when evaluating performance in a test set or predicting an actual sample. Hyper-parameters control how much the model will be controlled. When the hyper-parameters are huge, all parameters tend to be zero and eventually become a horizontal line over the average of the data (Liang and Song, 2009).

$$J(\theta) = \sum_{i=1}^{n} (y - \hat{y})^2/2n + \alpha \sum_{i=1}^{m} |\theta_i| \qquad (4)$$

Regularized cost function - Lasso Regression

In the above equation (Eq. (4)), y represents truth value, $\hat{y}$ represents predictions, $\theta$ the corresponding parameters, $\alpha$ is the penalty and J ($\theta$) is the loss function.

### 3.7. K-Nearest neighbor regression

k-nearest neighbor (KNN) algorithm, although commonly used for classification, can also yield accurate predictions on an optimized k value for large datasets. A standard KNN algorithm consists of assigning similar weightage to all training data points, calculating Euclidean distances and using the average of a data space consisting of predefined k data points to record observations for new values.

$$d = \sqrt{(x - x_i)^2 + (y - y_i)^2 + .... + (z - z_i)^2} \qquad (5)$$

Euclidean distance function

In the above equation (Eq. (5)), x, y, z represents new points and $x_i$, $y_i$, $z_i$ represents existing points.

### 3.8. Random forest regression

Using a random forest (RF) is a method of constructing a regression tree ensemble to decrease the variations of individual trees. Decision trees gather to build a forest, by using a concept known as "bootstrap aggregation" (bagging) to create many alike datasets sampled from the same source dataset. Bagging is a method of merging a trained basic model for training data. Because of its small bias and huge variance, the tree is prone to overfit. The random forest (RF) algorithm brings together extra randomness when growing trees; instead of searching for the very best feature when splitting a node, therein reduces instability. However, the downside of the decision tree is that it tends to overfit training data which can be regulated (Liang and Song, 2009).

### 3.9. Decision tree regression

Decision Tree, normally used for classification, is a node-based method with parameters like maximum depth of trees. Each node predicts a value after going through a series of true/false "branches". The Classification and Regression Tree (CART) training algorithm is used for bifurcation of the values at the node, thus reducing the mean-squared error for both nodes. The cost function for a decision tree node can be written as,

$$J(\theta) = m_L MSE_L + m_R MSE_R/m \qquad (6)$$

Decision Tree cost function

$$MSE_{node} = \sum_{i \in node} (y - \hat{y})^2/2m_{node} \qquad (7)$$

MSE loss function

In the above equations (Eqs. (6), (7)), $MSE_L$ means the mean squared error L branches, y represents truth value, $\hat{y}$ represents predictions, $\theta$ the corresponding parameters, $\alpha$ is the penalty and J ($\theta$) is the loss function.

Although it is good for handling both categorical and numerical data and is simpler to understand, careful examination is required as trees are vulnerable to overfitting as a result of orthogonal boundaries and instabilities that come with it. A trivial change can result in completely different tree structures and consequently different predictions (Liang and Song, 2009).

### 3.10. Model description

#### 3.10.1. Dimensionality reduction using principal component analysis

The 3 features - oil_co2, gas_co2 and cement_co2 had a very high correlation. Principal Component Analysis was applied and these three features were merged into a single feature called secondary_co2 and the three features were dropped (Fig. 4).

### 3.11. Model input parameters

The initial dataset comprised of 37 input parameters but not all of them were important for the prediction of $CO_2$ emission. In the feature engineering step, features with very high correlation and with semantic similarities were merged together. Also, the columns having more than 40% values missing were dropped from the dataset. The final dataset contained 17 input parameters which are described below:

1. **Secondary_co2:** Principal Component Analysis was applied on cement_co2, gas_co2 and oil_co2 as these parameters showed a very high correlation among themselves. After applying PCA, the new feature formed was Secondary_co2 (in million tones). Here cement_co2, gas_co2 and oil_co2 are the annual emissions of $CO_2$ from cement, gas and oil respectively.
2. **GDP:** Gross Domestic Product of the United States for a given year (in US dollars).
3. **Population:** Population of the United States in a given year.
4. **Oil_co2_per_capita:** Annual $CO_2$ emissions from oil (per capita) (in tonnes per person).
5. **Gas_co2_per_capita:** Annual $CO_2$ emissions from gas (per capita) (in tonnes per person).
6. **Flaring_co2_per_capita**: Annual $CO_2$ emissions from industrial flaring (per capita) (in tonnes per person).
7. **Coal_co2_per_capita**: Annual $CO_2$ emissions from coal (per capita) (in tonnes per person).
8. **Cement_co2_per_capita**: Annual $CO_2$ emissions from cement (per capita) (in tonnes per person).

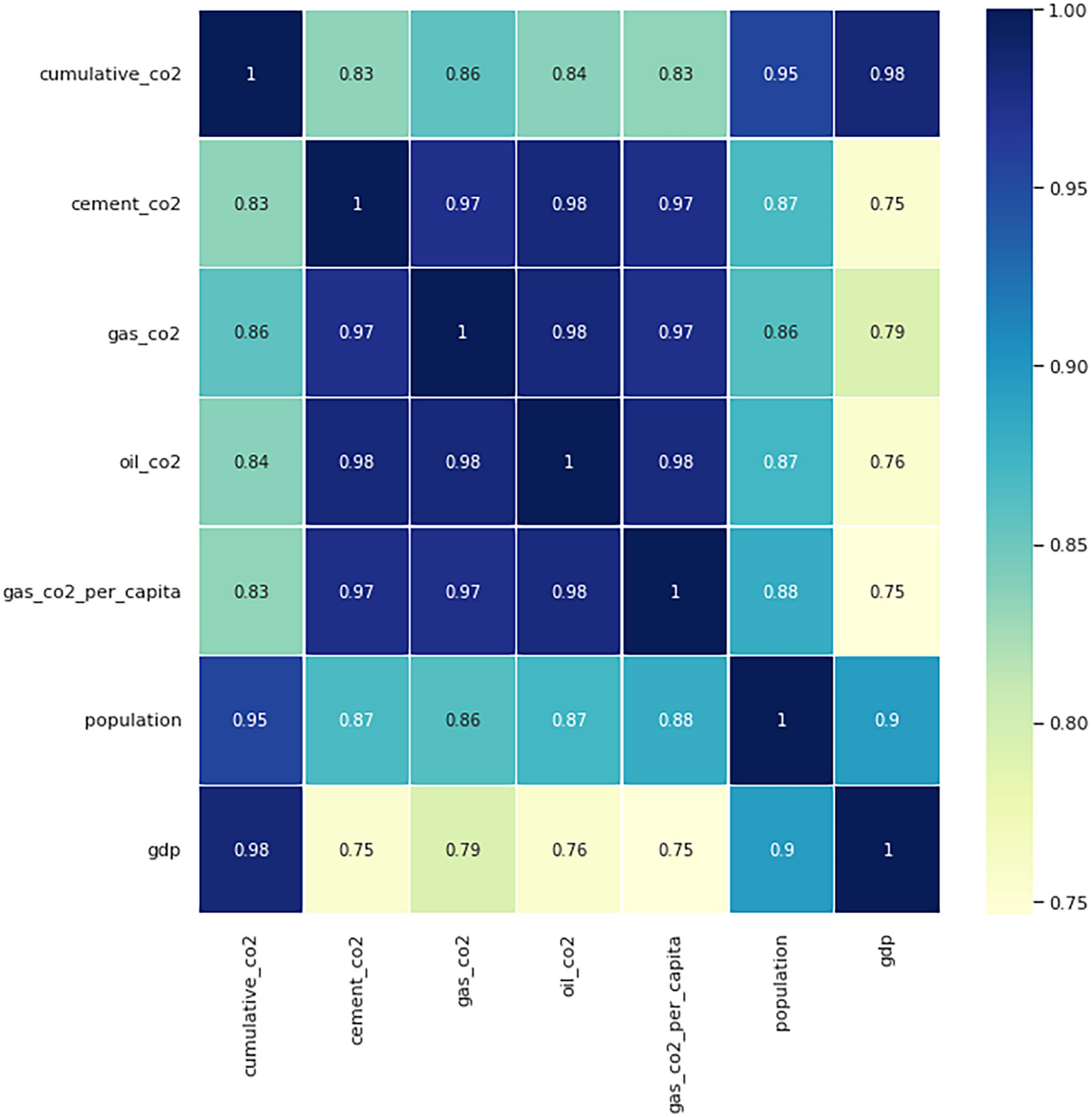

**Fig. 4.** Heatmap for highly correlated features.

9. **Flaring_co2:** Annual $CO_2$ emissions from industrial flaring (in million tonnes).
10. **Coal_co2:** Annual $CO_2$ emissions from coal (in million tonnes).
11. **Co2_per_gdp:** Annual $CO_2$ emissions per GDP (in kilograms per dollar).
12. **Share_global_cumulative_co2**: Share of global cumulative $CO_2$ emissions (in percentage).
13. **Cumulative_co2:** Cumulative $CO_2$ emissions (in million tonnes).
14. **Share_global_co2:** Share of global annual $CO_2$ emissions (in percentage).
15. **Co2_per_capita:** Annual $CO_2$ emissions (per capita) (in tonnes per person).
16. **Co2_growth_abs:** Annual change in product based $CO_2$ emissions (in million tonnes).
17. **Co2_growth_prct:** Annual percentage change in product based $CO_2$ emissions (in percentage).

A curve was then plotted that denotes the importance of each and every important feature and how important it is. The plot is shown below in Fig. 5:

### 3.12. Model output parameter

Annual $CO_2$ emissions in the United States of America is the output parameter for the ML model used. It gives the $CO_2$ emission measured in million tonnes.

### 3.13. Specific model description

The ML model that gave the best prediction accuracy on the dataset was the Decision Tree regressor. The minimum number of samples needed for splitting an internal node was kept as 2. The maximum depth of the tree was selected as 21. The loss function used was “Mean Squared

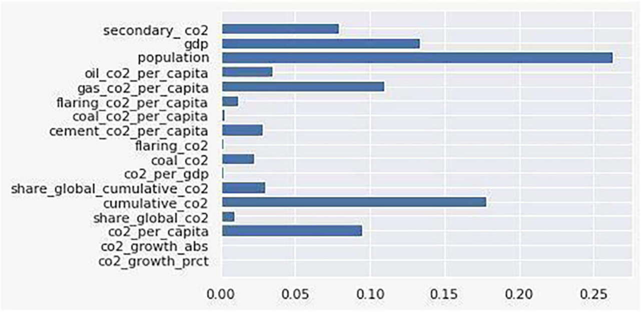

**Fig. 5.** Importance of every feature.

Error Loss". All the above parameters were specified after employing GridSearchCV to determine the optimal values.

## 4. Results and discussions

The globe is already changing as a result of the warming, with glaciers melting, coral reefs bleaching, heat waves and storms strengthening, among other things. Carbon dioxide levels above 450 parts per million are expected to lock in serious and irreversible climatic changes.

Keeping in mind the above inferences, it is safe to say that concentration of Carbon Dioxide above 500 ppm would cause irreversible damage to the atmosphere. This paper predicts the year in which global Carbon Dioxide concentration would surpass 500 ppm based on the past trends indicated by the historical data. The prediction given by the model was the year **2047**. Observational studies have shown that an increase of 0.75 ppm in $CO_2$ results in a 0.05 °C increase. Extrapolating this information, an increase of **5.6 °C** in global temperatures is expected by the year **2047**. Thus, the worrying conditions mount up as at the current rate the earth would approach irreversibly damaging conditions in less than 3 decades. Steps must be taken to prevent reaching this level at all costs. The next part of our study deals with this problem.

The experiment also predicted the reduction rate of carbon dioxide emission in order to bring its concentration back to what it was in 1991 – an estimated 316 ppm – which can be considered acceptable in today's world. Subsequently, cooling down the earth after repeated year-on-year recording breaking temperature rises. The target year we set is the year **2047**, which was the year predicted to result in irreversible harm to the earth's atmosphere. The reduction rate calculated was **6.37%** in order to achieve the aforementioned target. Although keeping in mind the average rate at which carbon dioxide emission is currently increasing, which is **12.59%**, the reversal rate of emission comes out to **23.38%**. To effectively bring about changes, it is important to identify the factors that are crucial.

Our study identified these pivotal features affecting $CO_2$ emissions the most. These features are **population** and emissions from **greenhouse gasses.** The average carbon footprint for an individual in the United States is 16tons. The United States population growth rate is currently increasing annually by 0.4%. This increase in population alone results in an increase in emission of 21.12 million tons every year.

All the above predictions were made by the model which was trained on historical data and gave predictions with around 99% accuracy. The data available pertains only to the United States. However, it must be noted that the United States consistently had around 15–18% share of the total global carbon dioxide emission. Hence, all the data was fitted by keeping this correlation in mind. The metric indicating the amount of carbon dioxide toxic to earth was parts per million (ppm). The data on which the model was trained measured the carbon dioxide quantity in million tones. Thus, a conversion formula of **1 ppm = 89 million tones**, was used to address this discrepancy in use of metrics.

All the above predictions were made by the model which was trained on historical data and gave predictions with around 99% accuracy. The RMSE achieved, methodologies used and the objectives are comparable to that of other works (Kadam et al., 2018a; Li et al., 2018):

**Table 1**
Model description.

| Model Name | Model Description |
|---|---|
| KNN Regressor | $K = 5$<br>Minkowski Distance Metric<br>Uniform weights<br>Mean Squared Error Loss |
| Random Forest Regressor | Number of estimators = 100<br>Mean Squared Error Loss |
| Gradient Boosting Regressor | Number of estimators = 100<br>Learning rate = 0.1<br>Mean Squared Error Loss |
| Decision Tree Regressor | Minimum Sample Split = 2<br>Maximum Depth = 21<br>Mean Squared Error Loss |
| Bagging Regressor | Number of estimators = 10<br>Mean Squared Error Loss |
| Extra Trees Regressor | Number of estimators = 100<br>Random State = 77<br>Mean Squared Error Loss |

**Table 2**
Comparison of techniques.

| | Experiments in our research | (Kadam et al., 2018a) | (Li et al., 2018) |
|---|---|---|---|
| Techniques Used | Linear Regression, KNN, Decision Tree, Random Forest Regression, SVM | Linear Regression | ELM, SVM, ELM-SVM |

**Table 3**
Comparison of RMSE using linear regression.

| | Experiment in our research | (Kadam et al., 2018b) |
|---|---|---|
| RMSE (Linear Regression) | 0.183 | 0.255 |

**Table 4**
Comparison of RMSE using SVM.

| | Experiment in our research | (Li et al., 2018) |
|---|---|---|
| RMSE(SVM) | 0.263 | 0.405 |

The best RMSE achieved in our experiment was **0.129** using Decision Tree Regression (Tables 1-4).

The most important feature identified by the Decision Tree regressor was that of population, thus, implying the significance of population on global Carbon Dioxide emissions. The simplest solution is to control the population explosion, specifically in the newly developing and industrializing countries. The other parameters that showed a high correlation with population were all different sources of $CO_2$ emissions such as $CO_2$ emissions from the oil industry, gas industry and the cement industry. Thus, for a comprehensive reduction in the levels of atmospheric carbon dioxide, a control in population is needed.

## 5. Challenges and future scope

Since 1958, scientists at the observatory have been measuring carbon dioxide levels in the atmosphere. However, scientists have data on levels dating back 800,000 years thanks to other types of investigation, such as those performed on ancient air bubbles trapped in ice cores.

The major challenge faced in Carbon Emission prediction is the availability and validity of the data available. The correct measurement of the carbon emission from industries is a tough task in itself. While large corporations and manufacturers may have technologies to keep an accurate track of their emissions, the same cannot be said for their smaller counterparts. The reliability of the data obtained from such industries must be questioned before it can be used for analysis.

Another major challenge is the high correlation between carbon emissions and various factors such as economic risk, financial risk, political risk and other socioeconomic factors (Hassan et al., 2022). This makes it tough to map the parameters which result in an increase or decrease in carbon emission levels. This in turn makes it tougher to measure what changes are required in which sectors to bring a positive effect on the emission levels.

Different industries have different sources or processes that result in emission. Hence, keeping an accurate track of the same in similar metrics so that they can be combined in an overall entity is an overwhelming task (Jeff, 2019b).

## 6. Conclusions

Every year, the earth's atmosphere contains around 3 ppm more $CO_2$. According to estimates, the planet will reach the penultimate level of 450 ppm in just over a decade. This study concludes that the final threshold of 500 ppm – the point of no return - will be achieved by 2047. To attain the safe levels of 316 ppm, the study identified the requirement of an emission reversal rate of 23.38%. The study also identified the factors that contribute the most to the emissions – population and greenhouse gasses. Various other features such as cement industry, combustion industries amongst others also contribute heavily to the emission spectrum. These conclusions are strengthened by the multiple machine learning techniques and the broad spectrum of factors utilized in this research, which effectively addresses certain drawbacks of other similar works.

It is evident that carbon dioxide emission reduction is the need of the hour. However, certain crucial milestones are yet to be established in this process. The authors recommend that extensive research be carried out speedily to determine these levels. Action plans must follow soon to effectively bring about changes before irreversible damage is done. A shift to renewable energy sources, and sustainable materials in industries will help mitigate these emissions. Achieving the carbon-neutrality mark should be the ultimate goal of every country.

## Authors contribution

All the authors make a substantial contribution to this manuscript. HB, MD, TS MS and AU participated in drafting the manuscript. HB, MD, TS and MS wrote the main manuscript. All the authors discussed the results and implication on the manuscript at all stages.

## Availability of data and material

All relevant data and material are presented in the main paper.

## Funding

Not applicable.

## Consent for publication

Not applicable.

## Ethics approval and consent to participate

Not applicable.

## Declaration of Competing Interest

The authors declare that they have no competing interests.

## Data availability

Data will be made available on request.

## Acknowledgements

The authors are grateful to Department of Computer Science and Engineering, Nirma University and Department of Chemical Engineering, School of Energy Technology, Pandit Deendayal Energy University for the permission to publish this research.

## References

Abdi, H., Williams, L.J., 2010. Principal component analysis. Interdiscip. Rev. Comput. Stat. 2, 433–459.

Bjp, M., Nicoletti, G., Jean-Marc, M.J.O., 1992. GREEN a multi-sector, multi-region general equilibrium for quantifying of curbing $CO_2$ emission. OECD Econ. Dep. Work Pap. 116. doi:10.1787/744101452772.

Blum, A.L., Langley, P., 1997. Selection of relevant features and examples in machine learning. Artif. Intell. 97, 245–271. doi:10.1016/S0004-3702(97)00063-5.

Boger, Z., Guterman, H., 1997. Knowledge extraction from artificial neural networks models. In: Proceedings of the IEEE International Conference on Systems, pp. 3030–3035.

Chantry, M., Christensen, H., Dueben, P., Palmer, T., 2021. Opportunities and challenges for machine learning in weather and climate modelling: hard, medium and soft AI. Philos. Trans. R Soc. A Math Phys. Eng. Sci. 379. doi:10.1098/rsta.2020.0083.

Gallo, C., Contò, F., Fiore, M., 2014. A neural network model for forecasting $CO_2$ emission. Agris On-Line Pap. Econ. Inf. 6, 31–36.

Hamrani, A., Akbarzadeh, A., 2020. Machine Learning For Predicting Greenhouse Gas Emissions from Agricultural Soils. Elsevier.

Hassan, T., Song, H., Kirikkaleli, D., 2022. International trade and consumption-based carbon emissions: evaluating the role of composite risk for RCEP economies. Environ. Sci. Pollut. Res. 29, 3417–3437. doi:10.1007/s11356-021-15617-4.

Heaton, J., 2016. An empirical analysis of feature engineering for predictive modeling. SoutheastCon 2016, 1–6.

Jeff, M., 2019a. Why agriculture's greenhouse gas emissions are almost always underestimatedle. Green Tech. Forbes.

Jeff, M., 2019b. Why agriculture's greenhouse gas emissions are almost always underestimatedle. Green Tech. Green Tech.

Kadam, P., 2018a. Prediction Model: $CO_2$ Emission Using Machine Learning. Ieeexplore.Ieee.Org.

Kadam, P., 2018b. Prediction Model: $CO_2$ Emission Using Machine Learning. Ieeexplore.Ieee.Org.

Khashei, M., Bijari, M., 2011. A new hybrid methodology for nonlinear time series forecasting. Model Simul. Eng. 2011. doi:10.1155/2011/379121.

Kriegler, E., Edenhofer, O., Reuster, L., Luderer, G., Klein, D., 2013. Is atmospheric carbon dioxide removal a game changer for climate change mitigation? Clim. Change 118 (1), 45–57. doi:10.1007/s10584-012-0681-4.

Lee, T.R., Wood, W.T., Phrampus, B.J., 2019. A machine learning (kNN) approach to predicting global seafloor total organic carbon. Glob. Biogeochem. Cycles 33, 37–46. doi:10.1029/2018GB005992.

Li, M., Wang, W., De, G., Ji, X., Tan, Z., 2018. Forecasting carbon emissions related to energy consumption in beijing-tianjin-hebei region based on grey prediction theory and extreme learning machine optimized by support vector machine algorithm. Energies 11 (9), 2475. doi:10.3390/EN11092475, 2018Vol. 11, Page 2475.

Li, S., Siu, Y.W., Zhao, G., 2021. Driving factors of $CO_2$ emissions: further study based on machine learning. Front. Environ. Sci. 9, 323. doi:10.3389/FENVS.2021.721517/BIBTEX.

Liang, H., Song, W., 2009. Improved estimation in multiple linear regression models with measurement error and general constraint. J. Multivar. Anal. 100, 726–741. doi:10.1016/j.jmva.2008.08.003.

Lin, B., Jia, Z., 2019. Impacts of carbon price level in carbon emission trading market. Appl. Energy 239, 157–170. doi:10.1016/j.apenergy.2019.01.194.

Liu, J., Murshed, M., Chen, F., Shahbaz, M., Kirikkaleli, D., Khan, Z., 2021. An empirical analysis of the household consumption-induced carbon emissions in China. Sustain. Prod. Consum. 26, 943–957. doi:10.1016/j.spc.2021.01.006.

Magazzino, C., Mele, M., 2021. A Machine Learning Approach On the Relationship Among Solar and Wind Energy production, Coal consumption, GDP, and $CO_2$ Emissions. Elsevier.

Mardani, A., Streimikiene, D., Nilashi, M., Arias Aranda, D., Loganathan, N., Jusoh, A., 2018. Energy Consumption, Economic Growth, and $CO_2$ Emissions in G20 Countries: Application of Adaptive Neuro-Fuzzy Inference System. Mdpi.Com doi:10.3390/en11102771.

Modise, R., Mpofu, K., 2022. Energy and Carbon Emission Efficiency Prediction: Applications in Future Transport Manufacturing n.d.. Mdpi.Com Retrieved July 13from.

Nandi, S., Sarkis, J., Hervani, A.A., Helms, M.M., 2021. Redesigning supply chains using blockchain-enabled circular economy and COVID-19 experiences. Sustain. Prod. 27, 10–22.

Newcomer, A., Blumsack, S.A., Apt, J., Lave, L.B., Morgan, M.G., 2008. Short run effects of a price on carbon dioxide emissions from U.S. electric generators. Environ. Sci. Technol. 42 (9), 3139–3144. doi:10.1021/es071749d.

Noble, W.S., 2006. What is a support vector machine. Nature 24, 1565–1567.

Peterson, L., 2009. K-nearest neighbor. Scholarpedia 4, 1883. doi:10.4249/scholarpedia.1883.

Rolnick, D., Donti, P.L., Kaack, L.H., Kochanski, K., Lacoste, A., Sankaran, K., Ross, A.S., Milojevic-Dupont, N., Jaques, N., Waldman-Brown, A., Luccioni, A.S., Maharaj, T., Sherwin, E.D., Mukkavilli, S.K., Kording, K.P., Gomes, C.P., Ng, A.Y., Hassabis, D., Platt, J.C., Bengio, Y., 2023. Tackling climate change with machine learning. ACM Comput. Surv. 55 (2). doi:10.1145/3485128.

Saleh, C., 2016. Carbon Dioxide Emission Prediction Using Support Vector Machine. Iopscience.Iop.Org doi:10.1088/1757-899X/114/1/012148.

Shao, X., Zhong, Y., Liu, W., 2021. Modeling the Effect of Green Technology Innovation and Renewable Energy On Carbon Neutrality in N-11 countries? Evidence from Advance Panel Estimations. Elsevier.

Wang, A., Xu, J., Tu, R., Saleh, M., Hatzopoulou, M., 2020. Potential of machine learning for prediction of traffic related air pollution. Transp. Res. Part D Transp. Environ. 88 (10259), 9. doi:10.1016/j.trd.2020.102599.

Wunch, D., Wennberg, P.O., Toon, G.C., Keppel-Aleks, G., Yavin, Y.G., 2009. Emissions of greenhouse gases from a North American megacity. Geophys. Res. Lett. 36 (15). doi:10.1029/2009GL039825.

Xu, Z., Liu, L., Wu, L., 2021. Forecasting the carbon dioxide emissions in 53 countries and regions using a non-equigap grey model. Environ. Sci. Pollut. Res. 28, 15659–15672. doi:10.1007/s11356-020-11638-7.

Zhang, X., Yan, F., Liu, H., Qiao, Z., 2021. Towards low carbon cities: a machine learning method for predicting urban blocks carbon emissions (UBCE) based on built environment factors (BEF) in Changxing City, China. Sustain. Cities Soc. 69 (10287), 5. doi:10.1016/j.scs.2021.102875.