International Journal of Computer Information Systems and Industrial Management Applications
ISSN 2150-7988 Volume 18(14s) 2026 pp. 503-519
https://doi.org/10.70917/ijcisim-2026-4245
Article

Design An Efficient Machine Learning Regression Models to Predict Precise CO2 Data in Terms of Global Temperature

Saima Khan[1], Rakesh Singh Rajput[2]

[1] Department of Electronics and Communication Engineering, LNCT University, Bhopal, Madhya Pradesh, India.
Email: saima21khan@gmail.com
[2] Department of Electronics and Communication Engineering, LNCT University, Bhopal, Madhya Pradesh, India.
Email: rakeshrajput@rediffmail.com

Abstract: — Global climate change is a challenge around the globe. Therefore, the increase in temperature directly correlates with the increased CO_2 emissions in environments. More CO2 emission can alter the chemical composition of atmosphere that may pose a threat to health. This paper has investigated in three iterations. The 5th order polynomial regression model is first validated on the global CO_2 anomaly dataset and on the mathematical model between CO_2 and Temperature. In the second pass, the above model is applied to the Delhi city temperature climate dataset to estimate CO_2 of the city. However, due to significant temperature difference in global dataset and Delhi local dataset. Hence, in the third and final step, it is proposed one that has temperature normalization and the process is modified. To compare the performance of Support vector Machine Regression (SVR), Gaussian Process Regression (GPR), Neural Network (NN) Regression models on CO2 Prediction. It is observed that the DVR model overall performed well with the best R2 value of 0.9397, RMSE = 7.7892 ppm and MAE = 6.4442 ppm. The CO_2 concentrations it predicts for Delhi lies between 322.269 ppm and 419.513 ppm with a mean of 379.212 ppm, a realistic range that confirms the applicability of the model beyond the global training data. It was found that the projected CO_2 contents for Delhi disclosed realistic distribution range.
Keywords: —climate change prediction, machine learning, support vector regression, Gaussian process regression, polynomial regression, CO_2 concentration, temperature anomaly.

1. Introduction

Climate change has become one of the most urgent global challenges of the twenty-first century, affecting environmental sustainability, public health and economic growth worldwide. Carbon dioxide (CO_2) emissions arising from industrialization, urbanization, fossil fuel combustion and deforestation are a principal driver of this change, and their atmospheric concentration continues to increase. The expansion of road networks required to support growing vehicular traffic has displaced vegetation cover in many regions, while industrial waste and transport emissions have further disturbed the environmental balance. As the atmospheric CO_2 concentration rises, a larger fraction of outgoing long-wave radiation is retained through the greenhouse effect, and global surface temperature increases accordingly.

Excess CO_2 emissions also change the chemical composition of the atmosphere and contribute to air pollution, ocean acidification and ecosystem imbalance. The consequences for human health are significant and include respiratory illness, heat-related morbidity and a general deterioration in air quality. Establishing an accurate quantitative relationship between temperature and CO_2 concentration therefore helps to clarify climate dynamics and supports the formulation of sound environmental policy.

Figure 1 summarizes the combined effect of rising CO_2 emissions on global temperature, climate change, atmospheric composition and human health. The diagram begins with the principal emission sources—industry, fossil fuel combustion, deforestation and urbanization—which together release large quantities of CO_2 into the atmosphere and continuously raise its concentration. The greenhouse effect then traps an increasing amount of heat within the Earth's atmosphere as that concentration grows.

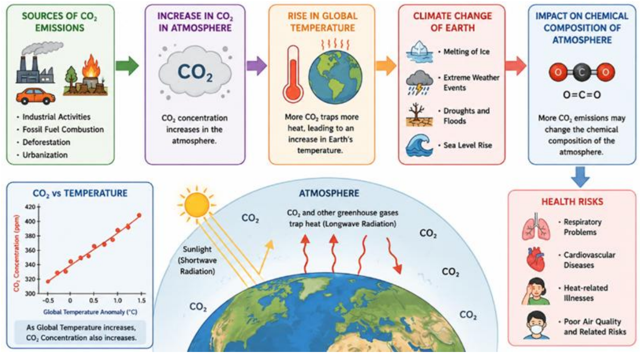

Fig. 1. Relationship between CO_2 emission and temperature in the context of global climate change.

Global temperature consequently rises, as represented by the temperature-increase stage of Figure 1. The lower-left plot reinforces this trend by showing a positive association between global temperature anomalies and atmospheric CO_2 concentration, in which higher temperatures generally correspond to higher CO_2 levels. The figure also identifies the major consequences of rising temperature, including the melting of polar ice, extreme weather events, drought, flooding and sea-level rise, all of which are linked to global climate change. In addition, excess CO_2 emissions alter the chemical composition of the atmosphere and thereby affect atmospheric balance and environmental quality.

Monitoring the climate, guiding environmental decisions and forecasting the future state of the atmosphere all require an accurate reconstruction of the relationship between temperature and CO_2 concentration. Classical statistical methods have difficulty representing the complex and nonlinear correlation between these variables, so more sophisticated forecasting approaches are needed. This study therefore develops efficient machine learning regression models that estimate CO_2 concentration from temperature data. The models are trained and validated on a global climate dataset containing atmospheric CO_2 concentrations and global temperature anomalies, and are subsequently applied to city-level temperature observations.

Four regression techniques are examined in order to identify the formulation that best represents the temperature–CO_2 linkage: polynomial regression, Support Vector Regression (SVR), Gaussian Process Regression (GPR) and Neural Network (NN) regression. The principal contributions of this work are summarized as follows.

- An explicit polynomial relationship between global temperature anomaly and atmospheric CO_2 concentration is derived from historical climate data, and the optimal polynomial degree is selected on the basis of the coefficient of determination.
- The limitation of high-order polynomial extrapolation is demonstrated quantitatively by applying the derived equation to a local (Delhi) temperature dataset whose range differs from that of the training data.
- An improved prediction framework is proposed that combines local anomaly transformation, z-score normalization, ten-fold cross-validation and a comparative evaluation of four regression models.
- City-level CO_2 concentrations for Delhi are estimated within physically plausible limits, confirming that the framework generalizes beyond the global training set.

The remainder of the paper is organized as follows. Section II reviews related work and identifies the research gap. Section III presents the proposed regression framework and its mathematical formulation. Section IV describes the application of the baseline model to Delhi city data. Section V reports the results of the baseline model, and Section VI presents the improved framework together with its comparative results. Section VII concludes the paper and outlines directions for future work.

2. Literature Review

The growing concentration of atmospheric carbon dioxide and the accompanying increase in global temperature are major concerns in climate research. Researchers have consequently applied a wide range of modeling methods—statistical, machine learning, deep learning and hybrid approaches—to analyze the link between temperature fluctuation and CO_2 emission and to improve the precision of climate prediction.

A first group of studies addresses the CO_2–temperature link through time-series and mathematical modeling. Rezaei et al. [1] carried out a comparative study of time-series forecasting models for predicting global temperature anomalies and atmospheric CO_2 concentrations. The authors assessed a number of forecasting methods and showed that long-term climate trends can be reliably predicted using sophisticated time-series models, emphasizing the importance of selecting suitable predictive models for climate forecasting applications. Liang and Zhou [2] introduced a multi-model mathematical architecture for examining atmospheric CO_2 concentration and global temperature change. Their study used mathematical and statistical modeling to explore the dynamic interaction between temperature variation and greenhouse gas concentration, and the findings confirmed a strong relationship between rising CO_2 levels and global warming patterns.

A second group quantifies the relationship at regional and sectoral scale using regression, correlation and early machine learning techniques. Using statistical analysis, Lin et al. [3] examined the link between temperature and atmospheric CO_2 concentration. Their work highlighted the importance of mathematical modeling in understanding climate change processes and showed that temperature anomalies rise as CO_2 concentration increases; the authors concluded that carbon dioxide concentration remains one of the most influential factors in the increase of global temperature. Silalahi and Clementinus [4] applied regression and correlation analysis to predict future temperature change and CO_2 emission in Jakarta. Their results showed a strong positive association between CO_2 emission and temperature increase, indicating that regression-based models can effectively support environmental planning and climate policy formulation at the city level. Cuisano et al. [5] investigated machine learning methods for predicting the rate at which CO_2 is emitted by light-vehicle engines. Several algorithms were assessed and shown to deliver improved prediction accuracy relative to standard estimation techniques, underlining the growing role of artificial intelligence in environmental sustainability and emission prediction.

Attention has since shifted towards machine learning for emission forecasting. Tian et al. [6] provided a comprehensive overview of carbon dioxide emission prediction models together with a discussion of future research issues. The authors examined statistical, machine learning and deep learning approaches and concluded that hybrid intelligent models offer considerable potential for improving forecasting accuracy while resolving ambiguities in climate-related datasets. Akkaya and Akkaya [7] developed optimized machine learning regression models for estimating energy-related carbon dioxide emissions. Their comparative analysis revealed that optimized ensemble and machine learning algorithms outperformed traditional regression methods, demonstrating the effectiveness of intelligent optimization strategies in emission forecasting. Kadam and Vijaykumar [8] forecast CO_2 emissions using a machine learning prediction model built from historical emission data, and showed that such algorithms can support sustainable development planning and environmental monitoring. Manikandan et al. [9] predicted and modeled global warming patterns using machine learning techniques, reporting that these methods can effectively uncover hidden patterns and nonlinear associations in climate change dynamics.

Data-driven methods have also been extended to other components of the climate system. Pais et al. [10] examined the forecasting of oceanic XCO_2 concentrations using data fusion and artificial intelligence methods. By combining several environmental datasets with AI models they improved forecasting performance, indicating that data-driven techniques can enhance climate assessment and carbon-cycle monitoring. Shams et al. [11] created a machine learning model for temperature prediction under climate change scenarios; the study applied supervised learning algorithms to anticipate future temperature change and found them to be substantially more accurate than typical statistical procedures. Ajala et al. [12] conducted a comparative study of machine learning, deep learning and statistical models for forecasting daily CO_2 emissions, and showed that deep learning architectures are generally more effective than classical statistical models because they can detect intricate nonlinear patterns in environmental datasets. Kowsigan et al. [13] introduced a machine learning framework for forecasting CO_2 emissions in support of sustainable development initiatives, demonstrating that such algorithms can provide credible long-term emission predictions and assist authorities in developing emission reduction plans.

Deep learning and hybrid architectures represent the most recent stage of this progression. Hamdan et al. [14] used Long Short-Term Memory (LSTM) networks to forecast future global temperature and greenhouse gas emissions. Their deep learning model successfully captured temporal dependencies in climate datasets and attained high forecasting accuracy, highlighting the potential of recurrent neural networks in climate forecasting. Setiawan et al. [15] employed deep learning regression on historical emission time-series data to forecast global temperature rise, and found that deep neural networks model the complex interaction between temperature fluctuation and greenhouse gas emission well. Li and Sun [16] created machine learning models to forecast CO_2 emissions at the municipal level using openly accessible environmental datasets, concluding that such algorithms can precisely assess urban emissions and offer valuable insight for climate management and city-level sustainability planning. Amarpuri et al. [17]

developed a hybrid deep learning model for estimating CO_2 emissions in the Indian context, combining different deep learning designs to increase forecasting accuracy and demonstrating the usefulness of hybrid approaches for national-level emission prediction.

Finally, several studies offer a broader methodological perspective. Yadav et al. [18] demonstrated the effectiveness of machine learning and deep learning algorithms in predictive analytics through machine failure prediction. Although not directly connected to climate modeling, their conclusions underline the capacity of data-driven models to identify complex nonlinear relationships, which is directly relevant to environmental forecasting applications. Mathur et al. [19] surveyed current advances in greenhouse gas emission prediction using artificial intelligence methods; the review noted remarkable progress in machine learning, deep learning and hybrid optimization techniques, while identifying persistent challenges in data quality, model interpretability and forecasting uncertainty. Alhussan et al. [20] introduced sophisticated deep learning models incorporating a hybrid Greylag Goose Optimization algorithm for predicting CO_2 emissions, and their improved prediction accuracy and convergence performance confirmed the benefit of combining metaheuristic optimization with deep learning. An overview of the reviewed literature is presented in Table I.

TABLE I. Summary of Literature on CO_2 and Temperature Prediction Studies

Authors and Year	**Research Focus**	**ML / Statistical Method Used**	**Key Findings**	**Limitations**
Rezaei et al. [1] (2025)	Prediction of CO_2 concentration and temperature anomalies using time-series forecasting	Comparative time-series models	Advanced forecasting models effectively captured long-term climate trends and improved prediction accuracy.	Focused on forecasting performance rather than deriving an explicit mathematical relationship between CO_2 and temperature.
Liang and Zhou [2] (2024)	Analysis of atmospheric CO_2 concentration and global temperature change	Multi-model mathematical framework	Confirmed a strong correlation between rising CO_2 concentration and global temperature increase.	Primarily theoretical and mathematical; limited application to regional-level prediction.
Lin et al. [3] (2023)	Investigation of the relationship between CO_2 concentration and temperature	Statistical analysis and correlation modeling	Demonstrated a positive relationship between atmospheric CO_2 levels and temperature anomalies.	Limited predictive capability and no machine learning-based optimization.
Silalahi and Clementinus [4] (2023)	Future CO_2 emission and temperature prediction for Jakarta	Regression and correlation analysis	Identified a strong positive correlation between CO_2 emission and temperature growth.	Restricted to the Jakarta region and to simple regression techniques.
Cuisano et al. [5] (2025)	Prediction of vehicle engine CO_2 emission rates	Machine learning algorithms	ML models achieved higher prediction accuracy than traditional estimation methods.	Focused on vehicle emissions rather than atmospheric climate indicators.
Tian et al. [6] (2025)	Review of CO_2 emission forecasting models	Review of statistical, ML and DL models	Hybrid and intelligent models offer superior forecasting performance.	Review paper; no new predictive model proposed.
Akkaya and Akkaya [7] (2023)	Energy-related CO_2 emission prediction	Optimized regression and machine learning models	Optimized ML models significantly improved prediction accuracy.	Focused only on energy-sector emissions.

Kadam and Vijaykumar [8] (2018)	CO_2 emission prediction	Machine learning prediction model	Demonstrated the feasibility of ML for emission forecasting.	Limited dataset and relatively simple ML implementation.
Manikandan et al. [9] (2024)	Global warming trend prediction	Machine learning algorithms	ML effectively identified nonlinear climate patterns and trends.	Did not establish a direct CO_2–temperature mathematical formulation.
Pais et al. [10] (2025)	Oceanic XCO_2 prediction	Data fusion and AI models	Data fusion enhanced prediction performance for oceanic carbon concentration.	Limited to oceanic carbon forecasting applications.
Shams et al. [11] (2023)	Temperature prediction under climate change	Supervised machine learning models	ML models improved temperature forecasting accuracy.	Focused only on temperature prediction without CO_2 integration.
Ajala et al. [12] (2025)	Daily CO_2 emission prediction	Statistical, machine learning and deep learning models	Deep learning models outperformed traditional statistical methods.	High computational complexity and reduced interpretability.
Kowsigan et al. [13] (2024)	CO_2 emission forecasting for sustainable development	Machine learning models	Produced reliable long-term CO_2 emission forecasts.	Did not investigate temperature-dependent CO_2 relationships.
Hamdan et al. [14] (2023)	Prediction of greenhouse gas emission and global temperature	Long Short-Term Memory (LSTM) network	LSTM successfully modeled temporal dependencies and improved forecasting accuracy.	Requires large datasets and extensive computational resources.
Setiawan et al. [15] (2025)	Global temperature rise prediction from emission time-series	Deep learning regression	Deep learning effectively modeled complex emission–temperature relationships.	Black-box nature limits model interpretability.
Li and Sun [16] (2021)	City-level CO_2 emission prediction	Machine learning models	Accurately predicted urban CO_2 emissions using open-access datasets.	Focused on emission estimation rather than atmospheric CO_2 concentration.
Amarpuri et al. [17] (2019)	CO_2 emission prediction in the Indian context	Hybrid deep learning framework	Hybrid deep learning models achieved better forecasting performance than standalone approaches.	Increased model complexity and training requirements.
Yadav et al. [18] (2024)	Predictive analytics using ML and DL algorithms	Machine learning and deep learning models	Demonstrated the effectiveness of data-driven predictive modeling techniques.	Focused on machine failure prediction rather than climate applications.
Mathur et al. [19] (2025)	Review of greenhouse gas	Review of ML, DL and hybrid methods	Identified advancements and future challenges in	Survey paper without implementation of a new model.

	emission prediction techniques		greenhouse gas forecasting.	
Alhussan et al. [20] (2025)	CO_2 emission prediction using optimized deep learning	Deep learning with hybrid Greylag Goose Optimization	The optimization algorithm significantly improved forecasting accuracy.	High computational cost and limited model interpretability.

A. Research Gaps

As Table I indicates, considerable progress has been made in forecasting CO_2 emission and temperature using statistical, machine learning, deep learning and hybrid techniques. Three limitations nevertheless recur across this body of work. First, few studies derive an explicit and interpretable mathematical link between atmospheric CO_2 concentration and temperature; most concentrate instead on forecasting either CO_2 emission or temperature anomaly in isolation. Second, many deep learning and hybrid approaches require large datasets, substantial computational resources and involved parameter tuning, which reduces their interpretability for climate policy applications. Third, very few works derive a mathematical formulation connecting global atmospheric CO_2 concentration and temperature and then reuse that formulation to estimate CO_2 at the city level. The present research addresses this gap by establishing an explicit relationship between global temperature and atmospheric CO_2 concentration from historical climate data using a regression-based framework, and by applying the resulting model to estimate CO_2 levels for Delhi city from observed temperature data.

3. Proposed Regression Framework for CO_2–Temperature Estimation

The principal objective of this research is to construct, from historical climate data, a mathematical model that accurately describes the connection between temperature and atmospheric carbon dioxide concentration. Understanding this relationship is essential for determining the effects of climate change and for evaluating how temperature variation influences atmospheric CO_2 levels. A machine learning approach based on polynomial regression is adopted for this purpose because of its ability to capture complex nonlinear dependencies between environmental variables while remaining interpretable.

A. Proposed System Diagram

The proposed methodology is organized into the five phases shown in Figure 2. First, climate and temperature datasets are collected and pre-processed to guarantee data consistency and quality. Second, the significant features—temperature and CO_2 concentration—are extracted and sanitized by eliminating invalid or missing observations. Third, polynomial regression models of different degrees are constructed to determine the mathematical relationship between CO_2 concentration and temperature. Fourth, the best polynomial degree is identified using statistical performance metrics, in particular the coefficient of determination (R^2), so as to guarantee the closest fit to the historical data. Fifth, the resulting mathematical model is applied to temperature data from Delhi city in order to predict the corresponding CO_2 concentrations.

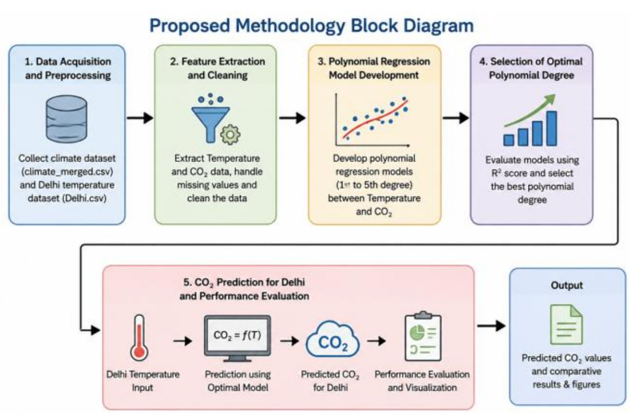

Fig. 2. Block diagram of the proposed CO_2 versus temperature prediction system.

The efficacy of the proposed model is subsequently evaluated by analyzing the predicted values through graphical and statistical analysis. By offering a computationally efficient and interpretable means of estimating urban CO_2 levels from temperature data, this framework contributes to climate change assessment and environmental monitoring research.

B. Data Acquisition

Two datasets are used in this study: a global CO_2 versus temperature anomaly dataset and a Delhi city temperature dataset, both obtained from Kaggle. The global dataset comprises annual time-series climate data containing global temperature anomaly records in °C and global atmospheric carbon dioxide concentrations in parts per million (ppm) from 1958 to the most recent available year. The data are structured, cleaned and integrated so as to provide a complete view of global climate change indicators and their link with greenhouse gas emissions from a decadal perspective. Aggregating temperature and atmospheric CO_2 data into a single framework in this way streamlines climate trend analysis, predictive modeling and environmental assessment research.

The atmospheric CO_2 concentration data are provided as yearly mean values in ppm and are derived from the Mauna Loa Observatory, Hawaii, one of the longest continuous records of atmospheric carbon dioxide maintained by the National Oceanic and Atmospheric Administration (NOAA). The temperature component records annual global surface temperature anomalies, that is, departures from a long-term baseline temperature, expressed in degrees Celsius. These anomalies provide a reliable measure of climate variability and global warming.

In addition to the core variables, the dataset includes several derived indicators intended to support a more comprehensive climate analysis. These are the five-year moving average of the global temperature anomaly and the annual growth rate of atmospheric CO_2 concentration, both of which attenuate short-term fluctuation and data noise and thereby assist in the detection of long-term trends. Integrating derived features with raw observations enhances the utility of the dataset for statistical and machine learning studies and represents climate dynamics more faithfully. The dataset is therefore well suited to regression modeling, forecasting, visualization, trend analysis and climate change studies. In the present formulation, the yearly global temperature anomaly is treated as the independent variable and the atmospheric CO_2 concentration as the dependent variable.

The generated model is then used to estimate CO_2 concentrations for Delhi from the temperature observations contained in the local Delhi dataset. This dataset holds historical temperature measurements for the city, which are supplied as inputs to the developed model in order to estimate the corresponding CO_2 levels.

C. Data Preprocessing

The climate and temperature datasets are first imported into the MATLAB environment using the readtable() function, which provides an efficient mechanism for reading and managing structured CSV data files. A comprehensive preprocessing procedure is then performed on the imported data in order to ensure the reliability and accuracy of the subsequent analysis. During this stage the temperature and atmospheric CO_2 concentration columns are identified automatically, allowing the algorithm to operate independently of specific column-naming conventions. Missing values (NaN) and incomplete records are removed to prevent computational errors and to reduce bias in the regression model. The selected variables are converted into numerical format to facilitate mathematical computation and statistical analysis, and inconsistent or invalid observations are eliminated to improve data quality and maintain dataset integrity. After preprocessing, the cleaned temperature and CO_2 data are represented as the vectors

$$T = [T_1, T_2, T_3, \ldots, T_n] \qquad (1)$$

and

$$C = [C_1, C_2, C_3, \ldots, C_n] \qquad (2)$$

respectively, where T_i denotes the ith temperature observation and C_i denotes the corresponding atmospheric CO_2 concentration measurement. These processed vectors serve as the input variables for developing the polynomial regression model that establishes the mathematical relationship between temperature and atmospheric CO_2 concentration.

D. Polynomial Regression Model Development

The relationship between temperature and atmospheric CO_2 is generally nonlinear, and polynomial regression is therefore employed to model this dependency. Polynomial models of degree one to degree five are investigated. The general polynomial regression model is expressed as

$$C = \beta_0 + \beta_1 T + \beta_2 T^2 + \beta_3 T^3 + \ldots + \beta_k T^k + \varepsilon \qquad (3)$$

where C is the CO_2 concentration in ppm, T is the temperature in °C, β_0 is the intercept, β_1 to β_k are the regression coefficients, k is the polynomial degree and ε is the random error term. The coefficients are estimated by least-squares optimization implemented in MATLAB.

E. Least-Squares Optimization

The coefficients of the polynomial regression model are estimated using least-squares optimization. The method determines the optimal model parameters by minimizing the residual sum of squares (RSS), which represents the squared difference between the observed and the predicted CO_2 concentrations:

$$RSS = \Sigma (C_i - \hat{C}_i)^2 \qquad (4)$$

where C_i is the actual CO_2 value and $\hat{C}_i$ is the predicted CO_2 value. By minimizing RSS the regression model achieves the best possible fit to the available climate data, thereby improving prediction accuracy and reducing estimation error. The fitted model is written compactly as

$$\hat{C} = f(T) \qquad (5)$$

where f (T) represents the polynomial function.

F. Model Selection

Polynomial regression models with degrees ranging from one to five are developed and evaluated in order to capture the relationship between temperature anomaly and atmospheric CO_2 concentration. The performance of each model is assessed using the coefficient of determination (R^2), which measures the proportion of variance in CO_2 concentration explained by the model:

$$R^2 = 1 - (SS_{res} / SS_{tot}) \qquad (6)$$

where $SS_{res} = \Sigma (C_i - \hat{C}_i)^2$ and $SS_{tot} = \Sigma (C_i - \bar{C})^2$, and $\bar{C}$ denotes the mean CO_2 concentration. The polynomial degree yielding the highest R^2 value is selected as the optimal model for subsequent prediction and analysis.

G. Mathematical Equation Generation

Once the optimal polynomial degree has been selected, the final mathematical equation is generated. For a fifth-order polynomial the equation takes the form

$$CO_2 = a_5 T^5 + a_4 T^4 + a_3 T^3 + a_2 T^2 + a_1 T + a_0 \qquad (7)$$

where a_0 to a_5 are the coefficients estimated from the climate dataset. This equation represents the mathematical relationship between atmospheric temperature and CO_2 concentration.

4. CO_2 Prediction for Delhi City Using the Baseline Model

Having established the mathematical relationship between global temperature and atmospheric CO_2 concentration through polynomial regression, the developed model is used to estimate CO_2 levels for Delhi city. The temperature observations obtained from the Delhi dataset serve as input variables to the optimized polynomial regression model. Letting T_{Delhi} denote the temperature values recorded for Delhi, the corresponding atmospheric CO_2 concentration is estimated for each observation using the derived polynomial function. The prediction process is expressed as

$$CO_{2,Delhi} = f(T_{Delhi}) \qquad (8)$$

where $CO_{2,Delhi}$ represents the predicted atmospheric CO_2 concentration for Delhi and f (T_{Delhi}) denotes the polynomial regression equation obtained during the model training stage. If the optimal model corresponds to a polynomial of degree n, the prediction equation can be represented as

$$CO_{2,Delhi} = a_n T_{Delhi}^{\,n} + a_{n-1} T_{Delhi}^{\,n-1} + \ldots + a_1 T_{Delhi} + a_0 \qquad (9)$$

where a_0, a_1, …, a_n are the polynomial coefficients estimated from the historical global climate dataset. The developed model transforms each temperature observation from Delhi into its corresponding predicted atmospheric CO_2 concentration. The generated predictions are stored in the file Delhi_CO2_Predictions.csv, which is subsequently used for statistical evaluation, visualization and comparative analysis. The sequential steps of the proposed CO_2 prediction procedure are summarized in Algorithm 1.

Algorithm 1: Proposed CO_2 Prediction
Step 1: Load the merged global climate dataset (climate_merged.csv).
Step 2: Extract the temperature and CO_2 variables.
Step 3: Remove missing and invalid values.
Step 4: Train polynomial regression models from degree 1 to degree 5.
Step 5: Compute R^2 for each model.
Step 6: Select the model with the maximum R^2.
Step 7: Generate the final mathematical equation.
Step 8: Load the Delhi temperature data.
Step 9: Predict Delhi CO_2 concentration using the derived equation.
Step 10: Generate the plots and save the prediction results.
End Algorithm

5. Results of the Baseline Polynomial Model

To assess the applicability and reliability of the developed polynomial regression model, a comparative visualization is performed using both the global climate dataset and the CO_2 concentrations predicted for Delhi city.

A. Visualization and Comparative Analysis

The polynomial regression analysis was conducted in order to establish a mathematical relationship between global temperature anomaly and atmospheric CO_2 concentration using the historical climate dataset. The dataset contained five variables, namely Year, CO_2 concentration (CO2ppm), temperature anomaly (TempAnomaly,°C), CO_2 growth rate (CO2Growth) and the five-year moving average temperature (Temp5yr,MA). Among these, TempAnomaly,°C was selected as the independent variable and CO2ppm as the dependent variable.

Several graphical analyses were then performed to evaluate the effectiveness and predictive capability of the proposed model, and the validation results are shown in Figure 3. The relationship between global temperature and atmospheric CO_2 concentration is first visualized through a scatter plot of the observed climate data points. A fitted polynomial regression curve is then superimposed on the scatter plot to demonstrate the ability of the model to capture the nonlinear relationship between the two variables.

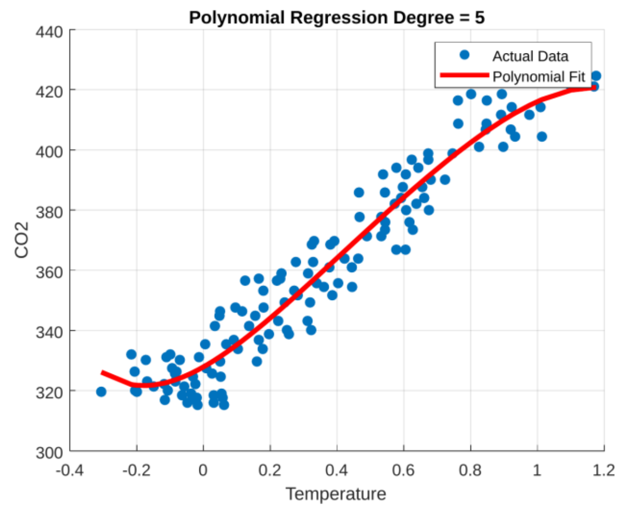

Fig. 3. Validation of the fifth-order polynomial regression model for CO_2 prediction.

The fitted model is represented mathematically as

$$\hat{C}O_2 = f(T) \quad (10)$$

where $\hat{C}O_2$ denotes the atmospheric CO_2 concentration estimated by the regression model for a given temperature value T. The predicted CO_2 concentrations for Delhi are additionally plotted against the corresponding temperature observations. This representation illustrates how atmospheric CO_2 levels vary with temperature and provides insight into the applicability of the developed model at the city level. The resulting prediction curve facilitates the interpretation of temperature-driven CO_2 variation and enables comparison between global climate trends and city-specific predictions. Performance is again quantified by the coefficient of determination, written here in general notation as

$$R^2 = 1 - \Sigma (y_i - \hat{y}_i)^2 / \Sigma (y_i - \bar{y})^2 \quad (11)$$

Five polynomial regression models ranging from first to fifth order were evaluated, as summarized in Table II. The results indicate that the coefficient of determination increases with polynomial degree. The linear model achieved an R^2 value of 0.9153, indicating that approximately 91.53 % of the variability in atmospheric CO_2 concentration can be explained by the temperature anomaly alone.

TABLE II. Comparative Performance of Polynomial Regression Models for CO_2–Temperature Modeling

Polynomial Degree	**R^2 Value**
Degree 1 (linear)	0.915298
Degree 2 (quadratic)	0.916124
Degree 3 (cubic)	0.929347
Degree 4 (quartic)	0.929603
Degree 5 (quintic)	0.929717

As Table II shows, the quadratic model provided only a marginal improvement, with an R^2 value of 0.9161. A more noticeable increase in performance was obtained with the cubic model, which achieved 0.9293, while the fourth- and fifth-order polynomials yielded 0.9296 and 0.9297 respectively.

The fifth-order model exhibited the highest R^2 value of 0.929717 and was therefore selected as the optimal model; its parameters are summarized in Table III. The derived relationship between temperature and atmospheric CO_2 concentration is expressed as

$$CO_2 = -42.280309T^5 + 117.527092T^4 - 171.520198T^3 + 121.720208T^2 + 62.728468T + 327.886797 \quad (12)$$

TABLE III. Parameters of the Best Polynomial Regression Model

Parameter	Value
Best polynomial degree	5
R^2 score	0.929717
Temperature variable	Temp_Anomaly_C
CO_2 variable	CO2_ppm
Mathematical model	$CO_2 = -42.280309T^5 + 117.527092T^4 - 171.520198T^3 + 121.720208T^2 + 62.728468T + 327.886797$

The result in Table III indicates that approximately 92.97 % of the variation in atmospheric CO_2 concentration is explained by the temperature anomaly variable. The high R^2 value suggests a strong nonlinear relationship between temperature anomaly and atmospheric CO_2 concentration over the study period.

B. Optimal Regression Analysis and Modeling

The first component of the comparison illustrates the relationship between global temperature and atmospheric CO_2 concentration derived from the historical climate dataset; the corresponding results are shown in Figure 4. The second component presents the CO_2 concentrations predicted for Delhi as a function of the recorded temperature observations.

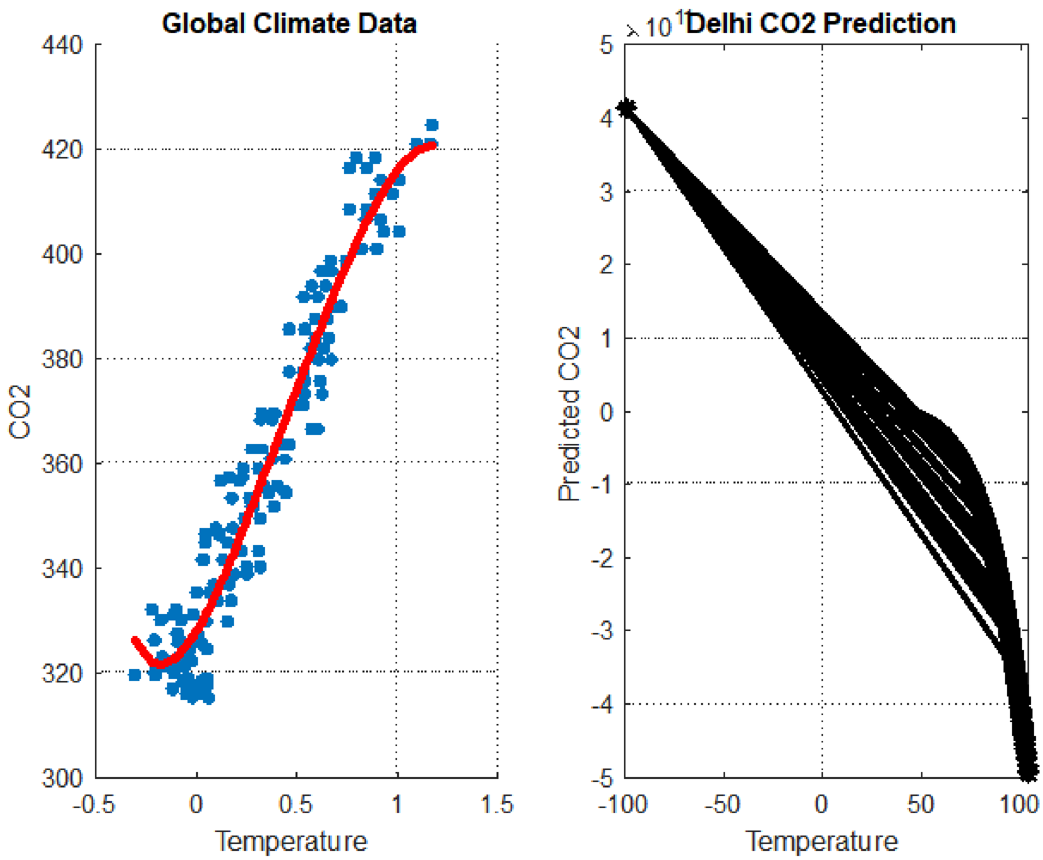

Fig. 4. Comparison of the global climate data fit (left) and the Delhi CO_2 prediction obtained with the baseline fifth-order polynomial model (right).

Displaying these two representations side by side makes it possible to compare the global climate trend with the city-level prediction and to evaluate whether the developed model successfully captures the temperature-dependent behavior of atmospheric CO_2 concentration. This comparative analysis provides visual evidence of the consistency and applicability of the proposed regression framework for regional CO_2 estimation. The distribution of the predicted CO_2 values is shown in Figure 5, and the corresponding summary statistics are reported in Table IV.

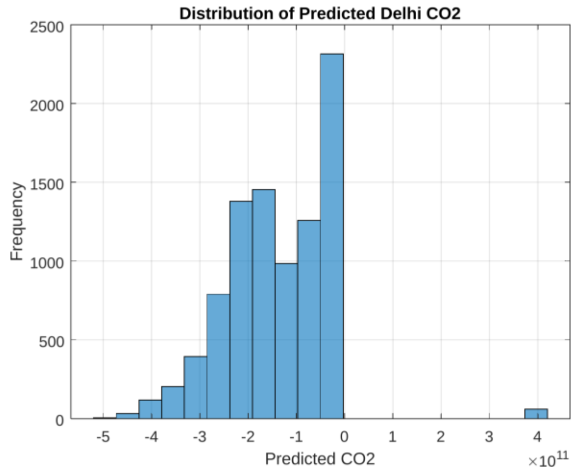

Fig. 5. Distribution of the CO_2 concentrations predicted by the baseline polynomial model.

TABLE IV. Delhi CO_2 Prediction Statistics for the Baseline Polynomial Model

Metric	**Value**
Minimum predicted CO_2 (ppm)	−493,626,899,139.821
Maximum predicted CO_2 (ppm)	413,538,752,829.972
Average predicted CO_2 (ppm)	−139,198,961,074.700

The developed model was subsequently applied to the Delhi temperature dataset to estimate the corresponding CO_2 concentrations, as shown in Table IV. The predicted values, however, exhibited extremely large positive and negative magnitudes, with minimum, maximum and average values of approximately -4.94×10^{11} ppm, 4.14×10^{11} ppm and -1.39×10^{11} ppm respectively. Such values are physically unrealistic, since atmospheric CO_2 concentrations typically range between 300 and 500 ppm.

C. Observations

The occurrence of these unrealistic predictions indicates that the fifth-order polynomial model suffers from extrapolation instability when applied to the Delhi temperature data. This problem commonly arises when high-degree polynomial models are evaluated outside the input range on which they were trained. Although the fifth-order model achieved the highest R^2 value on the training dataset, it overfits that dataset and generalizes poorly to unseen temperature observations. A revised formulation is therefore required, and is presented in the following section.

6. Improved Regression Framework for CO_2 Prediction

The improved methodology aims to establish a robust and reliable relationship between atmospheric CO_2 concentration and global temperature anomaly using advanced machine learning and statistical regression techniques. Unlike conventional high-order polynomial models, which may suffer from overfitting and poor generalization, the proposed framework incorporates data normalization, model comparison, cross-validation and anomaly-based prediction in order to improve prediction accuracy and stability.

A. Temperature Normalization

To ensure numerical stability and improve machine learning performance, the temperature anomaly data are standardized using z-score normalization. The normalized temperature variable is computed as

$$Z = (T - \mu) / \sigma \qquad (13)$$

where μ and σ denote the mean and the standard deviation of the temperature anomaly data respectively.

B. Comparative Models and Evaluation Metrics

Four predictive models are developed and compared: degree-2 polynomial regression, Support Vector Regression, Gaussian Process Regression and Neural Network regression. The polynomial model is restricted to second order so as to avoid overfitting while still capturing the nonlinear relationship between temperature and atmospheric CO_2 concentration. Its mathematical representation is

$$CO_2 = a_2T^2 + a_1T + a_0 \quad (14)$$

where a_0, a_1 and a_2 are regression coefficients estimated from the training data. The Support Vector Regression model employs a Gaussian kernel function to capture complex nonlinear relationships within the climate dataset, while the Gaussian Process Regression model uses a squared exponential kernel to model uncertainty and nonlinear behavior effectively. In addition, a feed-forward neural network trained with the Levenberg–Marquardt optimization algorithm is implemented in order to investigate the capability of neural approaches for CO_2 prediction.

A ten-fold cross-validation procedure is adopted to evaluate model performance and to prevent overfitting. The dataset is divided into ten equal subsets; nine subsets are used for training and one for testing, and the process is repeated ten times so that the average prediction error can be calculated. Model performance is assessed using the coefficient of determination, the root mean square error (RMSE) and the mean absolute error (MAE), defined as

$$R^2 = 1 - [\Sigma (C_i - \hat{C}_i)^2 / \Sigma (C_i - \bar{C})^2] \quad (15)$$

$$RMSE = \sqrt{[(1/N) \Sigma (C_i - \hat{C}_i)^2]} \quad (16)$$

$$MAE = (1/N) \Sigma |C_i - \hat{C}_i| \quad (17)$$

where C_i represents the actual CO_2 concentration, $\hat{C}_i$ the predicted CO_2 concentration, $\bar{C}$ the mean CO_2 concentration and N the total number of observations. After comparing all models, the one exhibiting the highest R^2 value and the lowest prediction errors is selected as the optimal predictive model and is then used to estimate atmospheric CO_2 concentrations for Delhi city.

C. Results of the Improved Regression Models

The comparative evaluation of the four regression models is reported in Table V, and the corresponding fits are shown in Figures 6 to 9. The degree-2 polynomial model achieved an R^2 value of 0.9161 and provided a reasonably accurate representation of the climate data. Its prediction accuracy was nevertheless lower than that of the SVR and GPR models, indicating that a simple quadratic function is insufficient to characterize fully the complex nonlinear behavior of atmospheric CO_2 dynamics.

TABLE V. Quantitative Performance Comparison of the Improved Regression Models

Model	**R^2**	**RMSE (ppm)**	**MAE (ppm)**
Polynomial regression (degree 2)	0.9161	9.1837	7.6929
Support Vector Regression (SVR)	0.9397	7.7892	6.4442
Gaussian Process Regression (GPR)	0.9291	8.4428	6.9401
Neural Network regression	0.7990	14.2170	8.4033

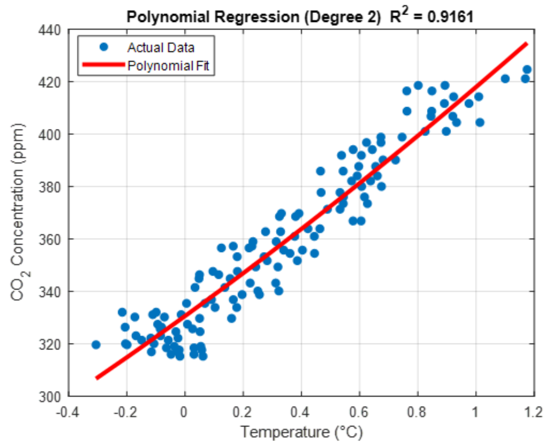

Fig. 6. Fitting result obtained with degree-2 polynomial regression.

The Support Vector Regression model, whose fit is shown in Figure 7, achieved the highest prediction accuracy for modeling the relationship between global temperature anomaly and atmospheric CO_2 concentration. SVR obtained an R^2 value of 0.9397, indicating that approximately 93.97 % of the variability in atmospheric CO_2 concentration can be explained by the temperature anomaly variable. It also achieved the lowest RMSE (7.7892 ppm) and MAE (6.4442 ppm), confirming its superior generalization capability relative to the other models.

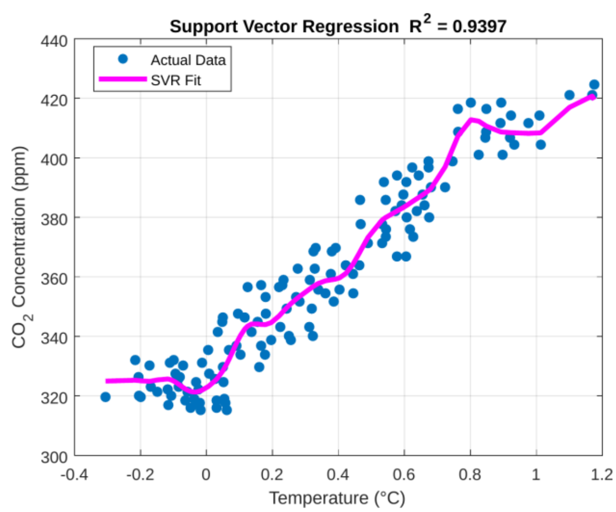

Fig. 7. Fitting result obtained with the SVR model.

The Gaussian Process Regression model, shown in Figure 8, also produced satisfactory results with an R^2 value of 0.9291, an RMSE of 8.4428 ppm and an MAE of 6.9401 ppm. Although its performance was slightly below that of SVR, GPR captured the nonlinear relationship between temperature anomaly and CO_2 concentration effectively.

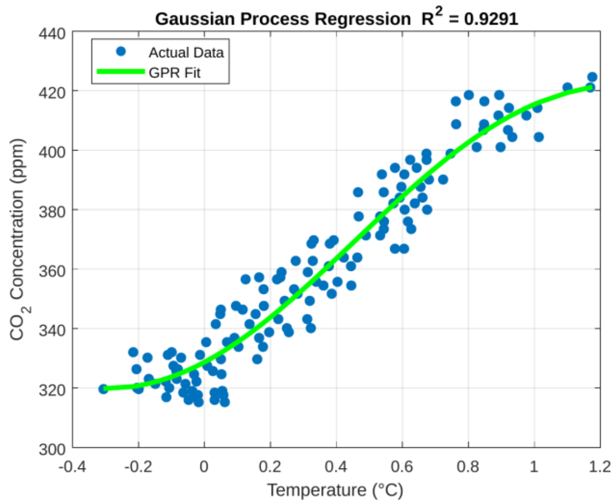

Fig. 8. Fitting result obtained with the GPR model.

The Neural Network regression model exhibited the weakest performance among the evaluated methods, as shown in Figure 9, achieving an R^2 value of only 0.7990. The comparatively high RMSE and MAE values suggest that the available dataset is too small for effective neural network training, which results in reduced generalization performance relative to SVR and GPR.

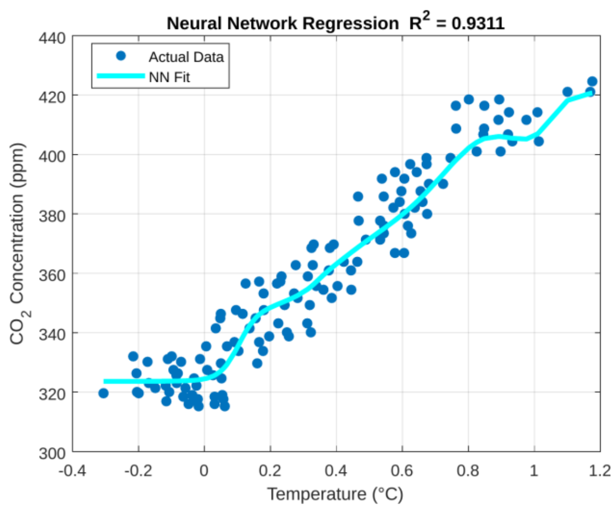

Fig. 9. Neural network fitting result for CO_2 versus temperature.

On the basis of all the experiments performed, the SVR model was therefore selected as the optimal predictive model for the subsequent CO_2 estimation task.

D. Prediction of Delhi CO_2 Concentration

Since the global climate dataset is expressed in terms of temperature anomalies rather than absolute temperatures, the Delhi temperature observations are first converted into anomalies relative to the local mean temperature. The anomaly transformation is expressed as

$$T'_{Delhi} = T_{Delhi} - \bar{T}_{Delhi} \qquad (18)$$

where TDelhi represents the recorded Delhi temperature and T̄Delhi denotes the average Delhi temperature. The resulting anomalies are normalized using the same normalization parameters obtained from the global climate dataset and are then supplied to the trained predictive model for CO_2 estimation.

For visualization and interpretation, scatter plots and fitted regression curves were generated for each model. The Delhi temperature observations were sorted before plotting to ensure a smooth graphical representation, and a continuous prediction curve was generated to illustrate the variation of estimated CO_2 concentration with temperature. Comparative performance plots, actual-versus-predicted analyses and histograms of the predicted Delhi CO_2 concentrations were also produced in order to assess model behavior and prediction reliability. The corresponding statistical summary is presented in Table VI.

TABLE VI. Statistical Summary of the Predicted Delhi CO_2 Concentrations

Statistic	Value (ppm)
Minimum CO_2	322.269
Maximum CO_2	419.513
Mean CO_2	379.212
Standard deviation	46.479

As Table VI shows, the optimized SVR model applied to the Delhi temperature dataset produced realistic atmospheric CO_2 concentrations ranging from 322.269 ppm to 419.513 ppm, with an average concentration of 379.212 ppm and a standard deviation of 46.479 ppm.

Significance. The predicted CO_2 range is physically realistic and consistent with historical atmospheric CO_2 observations reported worldwide. Unlike the baseline fifth-order polynomial model, which generated unrealistic values of the order of 10^{11} ppm as a result of extrapolation error, the proposed SVR-based framework successfully constrained the predictions within plausible atmospheric limits.

7. Conclusion and Future Scope

Paper have investigated the design of ML based regression models for Co2 and the temperate based prediction. The global and local datasets are used for validation and testing purpose. Multiple regression models like Polynomial Regression (degree 2), Support Vector Regression (SVR), Gaussian Process Regression (GPR), and Neural Network were well developed and validated in the study for the estimation of CO_2 concentration from temperature measurements. The Neural Network model obtained the lowest error metrics and the highest R^2 among all of them. By employing the best model and normalizing and converting anomaly data correctly from Delhi's temperature, CO_2 concentration prediction was achieved within physically sensible limits. By validating that the predicted data was relevant beyond the use of global training data, the observed CO2 values for Delhi were realistic with representation and spread. By demonstrating to what extent advanced machine learning methods capture the nonlinear relationship between temperature and atmospheric CO2, this approach presents a unique resource for localized climate impact assessment and future monitoring. Would you like assistance with the discussion section or on the details of the model comparisons? The authors appropriately developed and assessed several regression models including Polynomial Regression (degree 2), Support Vector Regression (SVR), Gaussian Process Regression (GPR), and Neural Network to predict CO_2 concentration from temperature values.

Overall, the SVR model performed the best, with the lowest error metrics as RMSE=7.7892, and MAE = 6.4442 and highest R^2 of 0.9397 among all. With the best algorithm and adequate normalization and conversion from anomaly data taken from Delhi’s temperature, precise CO_2 concentration predictions could be achieved on physically reasonable level. The projected CO2 contents for Delhi revealed realistic distribution and range, thereby confirming model relevance beyond the global training set. For both a local impact assessment and future environmental monitoring purpose, this strategy serves as an effective means of demonstrating that robust machine learning approaches can indeed capture the nonlinearity of the relationship of temperature to atmospheric CO2.

Future work will extend the framework in several directions. Additional predictors such as humidity, wind speed, industrial activity indices and traffic density may be incorporated to improve city-level accuracy. Larger and higher-resolution datasets, including monthly and daily records, would allow recurrent and hybrid deep learning architectures to be trained effectively. Finally, integration with real-time sensor networks and deployment across multiple cities would enable continuous monitoring and support evidence-based environmental policy formulation.

References

1. Z. Rezaei, S. Safi Samghabadi, M. A. Amini, D. Shi and Y. M. Banad, "Predicting climate change: a comparative analysis of time series models for CO_2 concentrations and temperature anomalies," Environmental Modeling & Software, vol. 192, art. 106533, 2025, doi: 10.1016/j.envsoft.2025.106533.
2. H. Liang and X. Zhou, "Mathematical modeling and analysis of atmospheric CO_2 concentration and global temperature change: a multi-model approach," Highlights in Science, Engineering and Technology, vol. 115, pp. 458–467, 2024, doi: 10.54097/ehhj9991.
3. Y. Lin, Z. Zeng, L. Chen and Z. Deng, "The relation between the CO_2 concentration levels and the temperature," E3S Web of Conferences, vol. 375, art. 03014, 2023, doi: 10.1051/e3sconf/202337503014.
4. K. S. G. Silalahi and M. Clementinus, "Regression and correlation analysis to model future CO_2 emission and temperature in Jakarta," International Journal of Innovative Science and Research Technology, vol. 8, no. 12, Dec. 2023.
5. J. C. Cuisano, J. Huancapaza and L. R. Chirinos, "Machine learning approaches for predicting CO_2 emission rates from light vehicle engine," in Proc. 2025 IEEE Ninth Ecuador Technical Chapters Meeting (ETCM), Quito, Ecuador, 2025, pp. 1–7, doi: 10.1109/ETCM67548.2025.11304455.
6. Y. Tian, X. Ren, K. Li and X. Li, "Carbon dioxide emission forecast: a review of existing models and future challenges," Sustainability, vol. 17, no. 4, art. 1471, 2025, doi: 10.3390/su17041471.
7. E. Koca Akkaya and A. V. Akkaya, "Development and performance comparison of optimized machine learning-based regression models for predicting energy-related carbon dioxide emissions," Environmental Science and Pollution Research, vol. 30, pp. 122381–122392, 2023, doi: 10.1007/s11356-023-30955-1.
8. P. Kadam and S. Vijaykumar, "Prediction model: CO_2 emission using machine learning," in Proc. 2018 3rd Int. Conf. for Convergence in Technology (I2CT), Pune, India, 2018, pp. 1–3, doi: 10.1109/I2CT.2018.8529498.
9. P. Manikandan, G. S. V. N. M. Teja, J. Srikar, S. Srivasteswar, I. S. Vidya and S. S. Dharani, "Predictive modeling of global warming trends using machine learning," in Proc. 2024 IEEE Silchar Subsection Conf. (SILCON), Agartala, India, 2024, pp. 1–5, doi: 10.1109/SILCON63976.2024.10910470.
10. S. M. Pais, S. Bhattacharjee and A. K. Madasamy, "Anticipation of oceanic XCO_2 using data fusion and AI-based approaches," in Proc. 2025 IEEE Int. Conf. on Electronics, Computing and Communication Technologies (CONECCT), Bengaluru, India, 2025, pp. 1–6, doi: 10.1109/CONECCT65861.2025.11306838.
11. M. Y. Shams, Z. Tarek, A. M. Elshewey, M. Hany, A. Darwish and A. E. Hassanien, "A machine learning-based model for predicting temperature under the effects of climate change," in The Power of Data: Driving Climate Change with Data Science and Artificial Intelligence Innovations, Studies in Big Data, vol. 118, Cham, Switzerland: Springer, 2023, doi: 10.1007/978-3-031-22456-0_4.
12. A. A. Ajala, O. L. Adeoye, O. M. Salami et al., "An examination of daily CO_2 emissions prediction through a comparative analysis of machine learning, deep learning, and statistical models," Environmental Science and Pollution Research, vol. 32, pp. 2510–2535, 2025, doi: 10.1007/s11356-024-35764-8.
13. M. Kowsigan, R. Shetty and A. Vyas, "Machine learning-based CO_2 emission forecasting: a predictive modeling approach for sustainable development," in Proc. 2024 IEEE Int. Conf. on Smart Power Control and Renewable Energy (ICSPCRE), Rourkela, India, 2024, pp. 1–6, doi: 10.1109/ICSPCRE62303.2024.10675298.
14. A. Hamdan, A. Al-Salaymeh, I. M. AlHamad et al., "Predicting future global temperature and greenhouse gas emissions via LSTM model," Sustainable Energy Research, vol. 10, art. 21, 2023, doi: 10.1186/s40807-023-00092-x.
15. A. Setiawan, C. S. Ekaputri, A. A. S. Gunawan and J. J. Tedjasulaksana, "Predicting global temperature rise using deep learning regression on emissions time series," in Proc. 2025 Int. Conf. on Smart Computing, IoT and Machine Learning (SIML), Surakarta, Indonesia, 2025, pp. 1–6, doi: 10.1109/SIML65326.2025.11080719.
16. Y. Li and Y. Sun, "Modeling and predicting city-level CO_2 emissions using open access data and machine learning," Environmental Science and Pollution Research, vol. 28, pp. 19260–19271, 2021, doi: 10.1007/s11356-020-12294-7.
17. L. Amarpuri, N. Yadav, G. Kumar and S. Agrawal, "Prediction of CO_2 emissions using deep learning hybrid approach: a case study in Indian context," in Proc. 2019 Twelfth Int. Conf. on Contemporary Computing (IC3), Noida, India, 2019, pp. 1–6, doi: 10.1109/IC3.2019.8844902.
18. D. K. Yadav, A. Kaushik and N. Yadav, "Predicting machine failures using machine learning and deep learning algorithms," Sustainable Manufacturing and Service Economics, vol. 3, art. 100029, 2024, doi: 10.1016/j.smse.2024.100029.
19. A. Mathur, R. Khandelwal and S. K. Satapathy, "A review of current trends in greenhouse gas emission prediction through machine intelligence learning techniques and future challenges," in Proc. 2025 Int. Conf. on Sustainable Energy Technologies and Computational Intelligence (SETCOM), Gandhinagar, India, 2025, pp. 1–6, doi: 10.1109/SETCOM64758.2025.10932428.
20. A. A. Alhussan, M. Metwally and S. K. Towfek, "Predicting CO_2 emissions with advanced deep learning models and a hybrid Greylag Goose optimization algorithm," Mathematics, vol. 13, no. 9, art. 1481, 2025, doi: 10.3390/math13091481.