Transactions on Computer Science and Intelligent Systems Research
ISSN: 2960-1800, eISSN: 2960-2238 | Volume 9
AIDML 2025

A Review of Machine Learning and Deep Learning Approaches for Predicting Greenhouse Gas Emissions

Xiaowen Chen *

Geography Department, Social Science & Public Policy Faculty, King's College London, London, United Kingdom

* Corresponding Author Email: xiaowenchenyaoxu@outlook.com

Abstract. Accurate prediction of greenhouse gas (GHG) emissions is critical in guiding informed climatic policymaking but has traditionally suffered from spatial biases, reporting lags, and inconsistent data. Machine learning (ML) and deep learning (DL) have in recent times become viable options for dealing with the complex, non-linear relationships in emissions data. These data-oriented methodologies have been used widely across sectors, showing high accuracy in prediction and flexibility. Here, this paper reviews recent advances in the application of ML and DL for the prediction of GHG emissions through their theoretical underpinnings, empirical performance using public datasets, and the use of explainable artificial intelligence (AI). A contrastive evaluation of exemplary studies identifies patterns in model performance, interpretability, and dependence on data. Special attention is paid to model interpretability and the contribution of explainable AI in increasing the policy applicability of methodologies. This paper contributes to an open, scalable, and actionable approach to data-driven decision support in climatic management, by linking technical advances to policy utility.

Keywords: Greenhouse Gas Prediction; Machine Learning; Deep Learning; Explainable AI; Emission Modelling.

1. Introduction

Climate change is among the most vital global issues today, and its driving factor is the increase in greenhouse gas (GHG) emissions. Carbon dioxide (CO_2), methane (CH_4), and nitrous oxide (N_2O) have been steadily rising as a consequence of human activities like fossil fuel combustion, industrialization, deforestation, and agriculture [1] [2]. These emissions are responsible for increasing global warming, climatic instability, and general socio-environmental disruption. Meanwhile, the Intergovernmental Panel on Climate Change (IPCC) highlights having to achieve the 1.5°C of Paris Agreement warming target through emission reductions at a rapid and sustained pace [3]. Importantly, realising such targets largely hinges on the capacity to monitor accurately, project precisely, as well as control GHG emissions across regions and sectors.

Two typical traditional methods for predicting GHG emissions are bottom-up inventory modeling and top-down atmospheric measuring [4]. These two methods, however, are generally marred by time-lagged reporting, low resolution and high uncertainty in the presence of rapidly changing city centers as well as in the case of developing economies [5]. Accordingly, Machine learning (ML) and deep learning (DL) then arrived as potent solutions to such limitations. These data-driven methods, as noted by [6] and [7], can extract complex nonlinear correlations between variables, process large and diverse datasets and make timely and spatial-specific predictions. For this reason, the flexibility of ML and DL methods also allows for aggregating multiple sources of data including satellite photos, weather patterns, land use activities and industrial production for delivering an adaptive and flexible solution for the estimation of the GHG emissions [2].

Furthermore, recent studies have increasingly employed ML and DL techniques to project GHG emissions from specific industries and regions. Support vector machines and random forest have been reported to have strong performance in modeling structured emission datasets both at the national scale and at the city scale [2][8][9]. Multilayer perceptrons and recurrent networks have instead been found to have advantages in handling temporal patterns as well as high-dimensional input [10-12].

The availability of public datasets from the likes of the World Bank and EUCoM has also facilitated replicable and scalable modeling attempts [13][14]. Challenges aside, however, lie in the fact that deep models are not necessarily interpretable, comparability of data is impaired due to fragmentation, and evaluation measures differ largely between studies. This review redresses these deficits through a systematic examination of the most recent ML and DL uses in the context of the forecast of GHG emissions, comparison of their empirical performance, scrutiny of public data utilization, and exploration of room for future integration of methods and applicability at the policy level.

2. Theoretical and Technical Foundations

2.1. Greenhouse Gas Emissions and Conventional Modelling Approaches

GHG, e.g., carbon dioxide, methane, and nitrous oxide, are released through many anthropogenic activities including fossil fuel combustion, agriculture, and industrial processes. Proper estimation of these emissions is important for the formulation of successful mitigation strategies. Typical two traditional methods for GHG emission prediction, as noted in [15], are bottom-up and top-down methods. The former method, mainly relies on compiling detailed inventory using information on activities and emission factors. As such, this method is advantageous in source specificity for predicting GHG emissions. However, the bottom-up method shows restrictions in applicability in terms of real time and spatial detail [15]. The later method, as discussed by [16], relies upon using atmospheric measurements and inverse modeling. While acknowledging that the top-down method is good at a greater degree of coverage, it shows weaker source specificity.

Some studies especially criticise the limitations of these two methods, notifying the estimation inconsistencies in predicting GHG emissions [17], [18]. For instance, the study of [17] emphasizes the great value of integrating both methods to establish the integrated frameworks, so as to fill gaps and augmenting accuracy in the estimation. Similarly, the study of [18] argues that such integration of both methods brings together the higher detailed source information of bottom-up techniques and the vast spatial information of top-down methods. Accordingly, the combined framework of bottom-up and top-down methods can yield a more sound and credible estimation of emissions.

2.2. ML and DL Approaches to GHG Modelling

Consider the complex and multifaceted nature of GHG emissions, the increasing demand for modelling techniques that can capture nonlinear relationships and time-dependent variability. In keeping with this requirement, machine learning (ML) and deep learning (DL) have become influential substitutes for traditional statistics [19], [20]. For instance, the study of [14] notifies that Support Vector Machines and Random Forests have been employed as means for modeling multifaceted relationships between environmental and socio-economic factors impacting emissions. Also, the work of [19] discusses that ML techniques surpass classical statistical methods for estimating agricultural soils' CO_2 emissions in terms of flexibility and predicting accuracy under changing conditions.

Furthermore, the DL techniques—most notably neural network architectures—have been found to possess significant potential in terms of mapping sophisticated emission structures [6], [12]. For instance, in the research of [12], they presented the integration of Dual-Path Recurrent Neural Networks (DPRNNs) and the Ninja Optimization Algorithm (NiOA) for the forecasting of CO_2 emissions in their DL framework. As implied from the empirical results, the DL methods not only show high prediction accuracy but also indicate evidence of effective abilities in modeling short-term variations and distant temporal relationships. As such, in comparison to the existing conventional statistical methods, the DL model had higher performance in replicating dynamic emission dynamics.

While ML and DL methods have made significant advances in modeling GHG emissions, their application is still hamstrung by some open issues. Data problems, such as lack of coverage, uneven quality, and the disaggregated nature of sources on hand, are still constraining model reliability [20].

In addition, high-capacity model complexity tends to render them intractable to interpret in many cases, generating concerns about the use of these instruments to guide environmental policy. And as the work of [2] suggests, stakeholders' trust does not just rely upon prediction accuracy, but on the transparency and accountability of the modeling process as well. Addressing these issues would demand more than algorithmic optimization. Meanwhile, it would also demand improved curating of datasets, enhanced data-harvesting practices, and more fully integrated domain expertise to help guarantee model outputs as both reliable and interpretable.

3. Empirical Applications of ML and DL in GHG Prediction

3.1. Machine Learning Approaches

ML methods have been extensively used in the modeling of GHG emissions in the last few years due to the ability of such methods to identify complicated relationships between diverse environmental and socio-economic datasets [13], [19]. For instance, typical ML models, e.g. Random Forests, Gradient Boosting, and Support Vector Machines, are among the popular algorithms used due to their differing capacities for dealing with nonlinearity, high dimensionality, and heterogeneous input variables. Accordingly, this section summarises some of the chosen empirical works which have utilised such ML methods to forecast GHG emissions, specifically with an interest in comparing their performance across sectors, conditions of the data, and modelling purposes.

The empirical study of [9] tested the application of ML methods for urban-level carbon emissions modelling on 254 Chinese cities from 2011 to 2020, based on six tree-based ML models: Extra-Trees, XGBoost, Bagging, Random Forest, Decision Trees, and Boosting. Of these, the best performance was reported by Extra-Trees with a MAPE of 0.21164, an MSE of 0.00049, and an R^2 of 0.98121, all of which significantly outperformed the baseline of the standard OLS model (MAPE = 0.55279; R2 = 0.93711). Adaptive Lasso regression was used to increase model interpretability and also to filter down the feature subset from 19 to 14 of the most essential variables [9]. These features included energy consumption as an internal driver as well as foreign direct investment as the main external influences and economic policy and other uncertainties. Additionally, partial dependence plots (PDPs) were used to visualise the non-linear relationships between predictors and emissions, highlighting the dominant impact of energy consumption and the emerging importance of digital finance and uncertainty measures.

On a broader scale, the work of [14] applies ML models to forecast long-term CO_2 emissions from the building sector across eight global regions using historical data from 1971 to 2014. They compare the performance of multiple model families—Linear Regression, ARIMA, Shallow Neural Networks (SNN), and Deep Neural Networks (DNN)—using both univariate and multivariate modeling with lagged values or polynomial transformations as feature engineering techniques. Their evaluation metric is MAPE, assessed on held-out test data (2005–2014). For example, the SNN with multivariate lagged features achieved a MAPE of 2.02% on the World Average dataset and 7.19% on the European Union dataset, significantly better than the polynomial transformation variants, which yielded higher MAPE values across most regions [14]. Only models fulfilling both the overfitting and naïve benchmark test were retained for projecting in the long run to 2050 as an indicator of the strength of multivariable-lagged SNNs in forecasting.

While both studies are based on ML methods, their modeling purposes and methodologies vary. The study of [9] centers on explanation power through the application of ensemble models to rank variable importance and visualise marginal effects through PDPs. This approach allows for nuanced interpretation for the purposes of urban policy formulation. By contrast, the work of [14] takes a forecasting stance, with the objective of high prediction accuracy on long horizons, shallow neural networks providing an optimal trade-off between accuracy and interpretability. The former combines cross-sectional heterogeneity and dynamic modeling of drivers, while the latter focuses on temporal coherence as well as cross-regional generalizability.

3.2. Deep Learning Approaches

DL models have become more prominent in the prediction of GHG emissions because of their better ability to capture non-linear relationships and long-term time-series dynamics in large environmental datasets. In contrast to conventional ML techniques with the need for heavy feature engineering tasks, DL models have the ability to learn automatically through representation learning processes, as discussed by [2] and [21], which is of particular importance in dealing with high-dimensional or time-series environmental variables.

The work of [12] suggests a hybrid deep learning methodology integrating Dual-Path Recurrent Neural Networks (DPRNNs) with the new metaheuristic optimization approach Ninja Optimizer Algorithm (NiOA) to predict CO_2 emissions from cement production. An extensive data preprocessing pipeline with Principal Component Analysis (PCA) and Blind Source Separation (BSS) was applied to separate noise and enhance feature selection [12]. DPRNNs successfully capture short- and long-term relationships in the emission time series, while dynamic adjustment of the related hyperparameters by NiOA prevents them from getting trapped in local optima. Empirical results from [12] indicated the NiOA-DPRNNs model performs better than other competing methodologies like JAYA and HHO and also outperforms SCA. It recorded the highest R^2 of 0.9736 and the lowest error rates in all cases. Robustness was also verified using ANOVA and Wilcoxon statistical testing. The research also highlights real-time CO_2 monitoring potential and the ease of model adaptation to other GHGs and the suitability of the model as a future environmental control system.

In another field of application, the work of [1] introduces CarbonMLP, a high-performing yet lightweight deep learning model for estimating vehicle CO_2 emissions using real-world Canadian data from the national open data platform. The model is an optimized multilayer perceptron (MLP) trained on an extensive dataset of 7,385 records of vehicle specifications including engine size, fuel type, transmission type, and city/highway fuel consumption. The best-performing model had an R^2 of 0.9938 and a Mean Squared Error (MSE) of 0.0002 compared to standard ML baselines. To deal with the black box problem typically found with deep models, the work of [1] added Explainable AI (XAI) techniques, namely SHapley Additive exPlanations (SHAP), which allowed input features to be ranked according to their contribution to predicted emissions. Their research indicates engine displacement and city fuel consumption as the strongest predictors. The paper of [1] also illustrates how deep learning in combination with XAI can be used to inform not just technical optimization but decision-making at policy and at the level of the consumer.

A comparison of the two studies emphasizes the varied capabilities of DL architectures in meeting various GHG modeling issues. A focus on optimizing time-series performance via recurrent structures and evolutionary tuning techniques is utilised in the case of [12], who do particularly well in situations where emission trends change across time. On the other hand, the work of [1] opts for efficiency in terms of interpretability and real-world applicability via efficient feedforward networks and clear-cut XAI techniques, demonstrating high accuracy whilst maintaining transparency via lightweight models. Although the approach of [12] is well tailored to industrial or process-based emission modeling with changing input-output behavior, the CarbonMLP from [1] best fits snapshot-based prediction in the transportation domain where categorical and numerical inputs are combined.

3.3. Empirical Applications Based on Public GHG Datasets

Publicly available datasets have become a key enabler of reproducible and scalable GHG emissions research. Studies using structured sources such as UNFCCC reports, national vehicle registries, or city-level spatial inventories have successfully applied ML methods to diverse prediction tasks, as summarised in Table 1. For example, the work of [8] used a range of ML and DL classifiers, including graph-based neural networks, to classify and fill missing emissions reports in UN and ClimateTRACE databases. Also, the study of [22] employed decision tree models on UK WLTP vehicle emissions test data, achieving high predictive accuracy with a MAPE of 1.69%.

In the work of [13], however, they utilized the application of XGBoost to forecast CO_2 emissions for over 92,000 European cities ($R^2 = 0.88$) via satellite proxies and city attributes to develop an inter-regionally transferable modeling pipeline. As implied from the findings of [13], the ML approach contributes better to emphasize interpretability, cross-regional reproducibility, as well as suitability for real-world applications like online emission calculators.

On the other hand, deep learning (DL) models are also shown to perform well on more nuanced and multimodal datasets in prior empirical works. For instance, the work of [10] created EcoPerceiver, a Transformer model trained on MODIS satellite and FLUXNET carbon flux observations, which performed superior to tree-based baselines in several biomes with NSE at 0.8482. In this regard, the model of [10] also sets the state of the art for carbon forecasting on the ecosystem scale using standardised open data for empirically applying the DL models. Table 1 shows the empirical applications based on public GHG datasets

Table 1. Empirical applications based on public GHG datasets.

Author (Year)	Model Type	Use of Public Dataset	Prediction Target	Model(s) Used	Best Performance (Metric)	Key Strength / Contribution
[13]	ML	EUCoM, ODIAC, spatial proxies for 92,636 European cities	City-level CO_2 emissions and trend classification	XGBoost with bootstrapped intervals	$R^2 = 0.88$, MAPE ≈ 8%	Predicts emissions for non-reporting cities with confidence intervals
[22]	ML	UK WLTP vehicle emission tests (VCA, 2020–2023)	CO_2 emissions from light-duty vehicles	Linear, Ridge, Lasso, SVR, RF, Tree	MAPE = 1.69%, MAE = 2.20	Interactive web tool for real-time vehicle emissions estimation
[8]	ML + DL	UNFCCC, ClimateTRACE, EcoInvent	Missing GHG emitter classification	Graph ML, RF, MLP, ResNet, GCN	Multi-level accuracy; best models vary per scope	Developed a multi-tier classification framework and benchmarked 18 models
[10]	DL	CarbonSense (MODIS + FLUXNET global tower network)	Carbon fluxes: NEE, GPP, RECO	EcoPerceiver (Transformer)	NSE = 0.8482 (best); RMSE lower than XGBoost	Cross-biome DL benchmark with standardised environmental data
[2]	ML + DL	World Bank (101 countries, 1991–2021)	National-level GHG emissions prediction	SVR (best), DL models, longitudinal regression	SVR model had lowest MAE and RMSE	Combines prediction and regression; identifies key emission drivers

Another notable contribution is provided by [2], who assessed the predictive accuracy of both ML and DL methods using 31 years of GHG emissions data from 101 countries, sourced from the World Bank. Their results show that support vector regression (SVR) outperformed DL models on key metrics such as MSE and RMSE. Also from their regression analysis was the confirmation of cereal yield and cropland expansion contributing to emissions while forest cover and renewable energy consumption decreasing them, again demonstrating the policy utility of public data in explanatory GHG modelling.

4. Conclusions

Within the literature review of this paper, the evolving role of ML and DL in projecting GHG emissions has been examined specifically in relation to their usage on an empirical basis, modeling paradigms, and application using public datasets. From city-level emissions forecasting to sectoral CO2 modeling across various fields of application, ML methods like ensemble tree models have performed well with high interpretability and flexibility using structured data. On the other hand, DL

architectures like recurrent and transformer networks have outperformed in dealing with high-dimensional and non-linear emission patterns.

While ML and DL possess much potential for augmenting the accuracy and scalability of modeling GHG emissions, it is emphasized in this review that this full potential is contingent on their integration into larger research environments. Mapped against new sets of empirical uses and contrasted with performance across research fields in this study is a systematic overview of existing directions of methodology. The gaps in interpretability, availability of data, and comparability in modeling are also highlighted. Rather than proposing a single optimal approach, this review emphasizes the need to align model choice with the research objective, data type, and policy requirements. In this sense, it can offer a foundation for both future research and the practical deployment of data-driven GHG prediction strategies.

References

[1] Zheng Xueyan, Streimikiene Dalia, Balezentis Tomas, Mardani Abbas, Cavallaro Federico, Liao Hua. A Review of Greenhouse Gas Emission Profiles, Dynamics, and Climate Change Mitigation Efforts Across the Key Climate Change Players. Journal of Cleaner Production, 2019, 234: 1113–1133.

[2] Yazd S D, Gharib N, Derakhshandeh J F. Investigations on Machine Learning, Deep Learning, and Longitudinal Regression Methods for Global Greenhouse Gases Predictions. International Journal of Environmental Science and Technology, 2025, 22(7): 5249–5262.

[3] IPCC. Sixth Assessment Report—IPCC [Online]. Available at: https://www.ipcc.ch/assessment-report/ar6/ [Accessed in: 15 May, 2025].

[4] Zheng Bo, Geng Geng, Ciais Philippe, Davis Steven J., Martin Randall V., Meng Jing, Zhang Qiang. Satellite-Based Estimates of Decline and Rebound in China's CO_2 Emissions During COVID-19 Pandemic. Science Advances, 2020, 6(49): eabd4998.

[5] Li Xiaoyan, Zhang Xiaoxing. A Comparative Study of Statistical and Machine Learning Models on Carbon Dioxide Emissions Prediction of China. Environmental Science and Pollution Research, 2023, 30(55): 117485–117502.

[6] Altikat Selim. Prediction of CO_2 Emission from Greenhouse to Atmosphere with Artificial Neural Networks and Deep Learning Neural Networks. International Journal of Environmental Science and Technology, 2021, 18(10): 3169–3178.

[7] Behvandi Omid, Ghorbani Hamed. Predicting Future Climate Scenarios: A Machine Learning Perspective on Greenhouse Gas Emissions in Agrifood Systems. Frontiers in Environmental Science, 2024, 12: 1471599.

[8] Cullen L., Marinoni A., Cullen J. Machine Learning for Gap-Filling in Greenhouse Gas Emissions Databases. Journal of Industrial Ecology, 2024, 28(4): 636–647.

[9] Yu Wen, Xia Lin, Cao Qiang. A Machine Learning Algorithm to Explore the Drivers of Carbon Emissions in Chinese Cities. Scientific Reports, 2024, 14(1): 23609.

[10] Fortier Mathieu, Richter M. L., Sonnentag Oliver, Pal C. CarbonSense: A Multimodal Dataset and Baseline for Carbon Flux Modelling. arXiv preprint, 2024: arXiv:2406.04940.

[11] Alam G M I, Arfin Tanim S, Sarker S K, Watanobe Y, Islam R, Mridha M F, Nur K. Deep Learning Model Based Prediction of Vehicle CO_2 Emissions with Explainable AI Integration for Sustainable Environment. Scientific Reports, 2025, 15(1): 3655.

[12] Ghorbal A B, Grine A, Elbatal I, Almetwally E M, Eid M M, El-Kenawy E S M. Predicting Carbon Dioxide Emissions Using Deep Learning and Ninja Metaheuristic Optimization Algorithm. Scientific Reports, 2025, 15(1): 4021.

[13] Hsu Angel, Wang Xin, Tan Jingru, Toh Wei, Goyal Nikhil. Predicting European Cities' Climate Mitigation Performance Using Machine Learning. Nature Communications, 2022, 13(1): 7487.

[14] Giannelos S, Bellizio F, Strbac G, Zhang Tong. Machine Learning Approaches for Predictions of CO_2 Emissions in the Building Sector. Electric Power Systems Research, 2024, 235: 110735.

[15] Reay D S, Davidson E A, Smith K A, Smith P, Melillo J M, Dentener F, Crutzen P J. Global Agriculture and Nitrous Oxide Emissions. Nature Climate Change, 2012, 2(6): 410–416.

[16] Quéré Sylvie, Levasseur A. Bottom-Up Versus Top-Down Hands-On Options for Measuring GHG and Pollutants in Smart Cities. WIT Transactions on Ecology and the Environment, 2020, 249(12): 285–293.

[17] Timilsina G, Pang J, Yang X. Linking Top-Down and Bottom-Up Models for Climate Policy Analysis: The Case of China. Unpublished, 2021.

[18] Saunois M, Martinez A, Poulter B, Zhang Z, Raymond P A, Regnier P, Zhuang Q. Global Methane Budget 2000–2020. Earth System Science Data, 2025, 17(5): 1873–1958.

[19] Hamrani A, Akbarzadeh A, Madramootoo C A. Machine Learning for Predicting Greenhouse Gas Emissions from Agricultural Soils. Science of The Total Environment, 2020, 741: 140338.

[20] Wu Lijuan, Huang Zhen, Zhang Xinyu, Wang Yiyang. Harmonizing Existing Climate Change Mitigation Policy Datasets with a Hybrid Machine Learning Approach. Scientific Data, 2024, 11(1): 580.

[21] Ajala A A, Adeoye O L, Salami O M, Jimoh A Y. 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, 2025, 32: 1–26.

[22] Udoh J, Lu Jing, Xu Qiang. Application of Machine Learning to Predict CO_2 Emissions in Light-Duty Vehicles. Sensors, 2024, 24(24): 8219.