A Review of Machine Learning Applications in Evaluating Carbon Policy

Ethan Kim[1]

Received September 19, 2025
Accepted July 25, 2026
Electronic access October 15, 2026

This literature review focuses on the application of machine learning (ML) methods for evaluating carbon emissions and the effectiveness of climate policies. As decarbonization efforts become increasingly urgent, the role of computational tools such as ML methods becomes increasingly significant. By drawing from 15 recent empirical and modeling studies, this literature review aims to evaluate the use of ML methods, and identify gaps in current research and opportunities to improve ML implementation for a more robust climate policy analysis. Deep Learning and Neural Networks have shown gains in predictive accuracy in emissions forecasting, Natural Language Processing (NLP) methods have proven useful in policy classification, and hybrid ML–econometric frameworks extend the reach of carbon policy analysis in terms of scale and applicability, though only a small share of policies (about 4%, or 63 of 1,500 in one study) were found to achieve substantial emissions reductions. While ML methods have shown potential in emission prediction, policy analysis and policy tracking, they face limitations such as a lack of causal credibility, interpretability issues, sectoral and regional biases, and limited consideration of outside constraints.

Keywords: Carbon emissions, climate policy, machine learning, carbon pricing, policy evaluation, emissions reduction

Introduction

In 2015, governments around the world agreed via the Paris Agreement, which seeks to limit the temperature increase compared to pre-industrial times 'well under' 2 degrees Celsius, to take action on climate change and reduce their emissions[1]. In recent years, investment into renewable energy sources, such as solar, wind and nuclear energy, and the demand on renewable energy sources have been increasing and there have been further efforts to reduce carbon emissions[2]. To achieve this, many countries have introduced a wide variety of policy changes, including encouraging the use of electric vehicles, and in some countries, carbon taxes.

Despite policymakers' attempts to decrease carbon emissions, these efforts to mitigate carbon emissions and its effects often fall short of proposed goals to reverse or halt climate change and its effects. A previous study that studied 1500 climate change policies from an OECD (Organization for Economic Cooperation and Development) database concluded that only about 63 policies brought about meaningful change in carbon emissions, or about 4% which combined reduces emissions by 0.6 to 1.8 gigatons, albeit not adequate to mitigating the effects of climate change. The study cites a disparity in ambition between goals and outcomes of policies as the reason that these policies have not meaningfully decreased emissions[3].

Among the most common policy changes which are most pertinent to carbon emissions are plans related to renewable energy, such as subsidies, expansion planning, auctions, as well as plans to phase out fossil fuels and introduce a carbon/fuel tax[4]. One such policy change that has been attempted at a nationwide level and has found some degree of success is the carbon tax, though the degree of its success is contested[4].

The ability to track and forecast carbon emissions allows policymakers to better understand the impact of their policies and set more achievable carbon reduction targets[5]. Current efforts to analyze carbon emissions often consist of statistical models, machine learning (ML) approaches, or a hybrid approach. Forecasting carbon emissions is particularly important in terms of setting targets and informing policymakers. These types of models can take on many forms, and each have their areas of strengths and weaknesses. For instance, statistical models are highly effective in terms of interpretability and for handling linear data but present difficulties in relationships that are not linear and complex, including policy changes.

ML methods are highly effective for complex and multifaceted relationships, but sometimes struggle with overfitting, where the model fits too closely to the provided dataset that it cannot accurately predict future data. Various ML has been

[1] *Singapore American School, Singapore*

utilized, from natural language processing (NLP) for identifying patterns in certain plans, to neural networks, deep learning techniques, and various regression methods. ML techniques have been utilized somewhat frequently in recent years to assess the impact of policies. In the context of modeling carbon emissions and analyzing carbon policy, it has been implemented in studies because ML techniques, compared to traditional statistical methods, offer a wide degree of flexibility and the ability to handle nonlinear and complex relationships.

This review examines the application of machine learning to carbon emissions forecasting and policy evaluation, with the aim of identifying methodological strengths, limitations, and directions for future research. Here, carbon policy is defined as any policy implementation that aims to reduce carbon emissions. Policy evaluation refers to an assessment of the design, implementation, impact, or optimization of carbon policy where effectiveness is defined by emissions reduced at acceptable economic and social cost.

Methods

To gather sources that apply ML to predict carbon emissions or analyze carbon policy, multiple different academic databases such as Google Scholar, SSRN, ScienceDirect, SpringerLink, ResearchGate, and Nature were used. Keywords such as "machine learning" AND "emissions", "carbon policy" AND "evaluation", "carbon tax", "subsidies" OR "climate policy" AND "evaluation", "emissions" AND "regression" were applied.

The exclusion criteria includes studies which do not mention carbon emissions or policy changes. Studies that lack empirical data were also excluded. Studies that used ML to analyze emissions or policies were included. To ensure relevance and that recent publications are prioritized, a filtering criterion was set to only select literature published after 2020. Relevant sources that were referenced in the selected literature were also used. However, due to this criteria, the scope of this paper will be limited to quantitative data, excluding qualitative insights. This paper will mostly exclude conventional econometrical and statistical methods, except for the purpose of providing background information. In addition, only studies written in English were included. More importantly, this paper focuses on covering policies specifically focused on decreasing emissions, such as carbon taxes and subsidies. To ensure that a source fits this criteria, shortlisted literature based on their keywords were then assessed based on their abstract.

An initial database search across Google Scholar, SSRN, and ScienceDirect using the keywords yielded a large volume of results. As such, 59 of the most relevant studies (ie. top results) were selected for title and abstract screening. From these 59 studies, 33 were selected for full text evaluation according to their relevance to this review. Of these 33 studies, 15 were selected to be included in this review according to the inclusion and exclusion criteria. From these studies, details such as ML method used, model architecture and key hyperparameters, data sources and sample characteristics, validation approach, performance metrics, causal design (where applicable), geographic and sectoral focus, and primary limitations were extracted. Findings were synthesized to evaluate the extent to which ML methods advance carbon policy analysis and to identify methodological gaps in the literature.

Literature Review

The use of machine learning in climate and carbon policy research has expanded over the past decade. As governments implement complex policy instruments such as carbon pricing, renewable energy subsidies, and regulatory standards, there is an increasing need for tools capable of capturing high-dimensional and nonlinear relationships.

This section reviews existing research across four categories: (1) traditional econometric and simulation-based approaches, (2) ML-based emissions and energy prediction, (3) ML-based policy evaluation, and (4) natural language processing (NLP) applications for climate policy analysis. Table 1 has been provided to summarise the literature assessed in this review.

i) Non-ML Approaches to Carbon Policy Evaluation

Besides the use of ML methods, carbon policy evaluation is conducted using econometric and simulation-based frameworks. Computable general equilibrium (CGE) models are used to simulate the macroeconomic impacts of carbon pricing. For instance, Akkemik, Borges, and Dang (2024) use a dynamic CGE model to assess the potential macroeconomic and emissions impacts of a carbon tax in Vietnam, finding GDP reductions of 1.2–2.7% depending on the tax level[6]. Alongside CGE models, integrated assessment models (IAMs) link economic and climate systems to evaluate long-run mitigation pathways and compute metrics such as the social cost of carbon[7].

Traditional econometric methods complement these simulation approaches by using panel or time-series data to support causal inference through quasi-experimental designs. For example, Rafaty, Dolphin, and Pretis (2025) employ a synthetic control factor model across a panel of 39 countries to estimate the emissions response to carbon pricing, finding reductions in emissions growth of 1–2% on average following policy introduction[8]. Although these methods provide valuable policy insights, they face limitations. CGE and IAM frameworks have limited capacity to handle nonlinearity[9,10], while quasi-experimental econometric methods depend on assumptions such as parallel trends that may not always hold in practice[11].

ii) Machine Learning for Emissions and Energy Prediction

A substantial portion of the literature applies machine learning to forecasting carbon emissions and energy system dynamics, framing the problem as a supervised learning task. These studies are important for carbon policy analysis as they inform target-setting and planning.

Mujeeb and Javaid (2023) introduce a sophisticated deep learning framework for forecasting emissions in power systems[12]. Their model integrates Spearman Correlation Analysis (SCA) for feature selection and Improved Shallow Denoising Autoencoders (ISDAE) to filter out noise from raw data. Then, a Deep Neural Network (DNN) was used, after its hyperparameters were tuned via Improved Particle Swarm Optimization (IPSO). This architecture demonstrates superior accuracy in capturing the non-linear relationship between renewable energy integration and emissions compared to traditional multiple linear regression (MLR) and support vector machines (SVM). The proposed model boasts a normalized root mean square error (NRMSE) of 0.1730 compared to that of 0.3427 for SVM and 0.3814 for MLR, indicating strong predictive accuracy. However, although this study uses Spearman correlation for feature selection, causal effects are not established. In addition, data were collected only from New England, whose power system uses less renewable energy (with $\sim$7% wind and $\sim$0.6% solar), meaning the model may not apply to other areas or regions. Its scope of features is also limited, as it excludes policy variables, economic activity, and behavioral changes.

Similarly, Wang et al. (2023) propose a hybrid two-stage forecasting model for carbon emissions in China[13]. By combining support vector regression (SVR) with artificial neural networks (ANN) and random forests, they demonstrate that two-stage hybrid models consistently outperform single-stage versions across various time horizons, with the SVR-ANN configuration achieving the lowest RMSE. However, this study does not establish causal relationships, purely using machine learning methods to minimize RMSE/MAE. Its dataset is also limited in size, consisting of only 36 actual annual observations which are then split to 144 data points, statistically generated via the Chow-Lin method. In addition, the model's performance on error metrics (RMSE/MAE) may be overstated as the Chow-Lin method involves smoothing data. This paper serves as a good example of how machine learning can optimize predictions for carbon emissions, though its ability to generate meaningful insights about the drivers of emissions or inform policy decisions is limited. Wei and Xu (2023) address carbon emission prediction using a hybrid TCN-LSTM architecture, drawing on data from EIA, EPA, EEA, IEA, though sample sizes, time periods, and geographic granularity are unspecified[14]. The model uses a 70/30 train-test split and reported, and is benchmarked against baseline studies using statistical methods plus studies on the TCN and LSTM components, achieving the highest reported accuracy (97.66%) and F-score (94.56%). This study highlights the predictive accuracy that ML can bring for carbon emissions forecasting. However, instead of using regression metrics conventionally used, this study uses classification metrics, making these metrics somewhat difficult to interpret. In addition, no causal design is attempted.

Wang, Hu, Bai, Chang, and Yu (2026) develop a two-stage hybrid framework to forecast China's national carbon allowance cap for 2026-2035 using path analysis and supervised machine learning. First, path analysis identifies key predictors from 19 candidate variables, including lagged carbon allowances, energy consumption, carbon emissions, GDP, and environmental regulation intensity. These variables are then used to train five ML models (decision tree, backpropagation neural network, random forest, support vector machine, and Gaussian process regression) on data from China's eight carbon-trading pilot regions. Gaussian Process Regression (GPR) performs best across all error metrics, with Diebold-Mariano tests to confirm. For projections through 2035, the authors adjust the GPR baseline using emission reduction targets from a Goldman Sachs report and China's 2035 NDCs, showing the result as a policy constrained scenario rather than a statistical forecast. They also emphasize that path analysis identifies structural associations rather than causal relationships. However, the study remains geographically limited to China, and its long-term projections rely partly on a non peer-reviewed financial report rather than a government or academic source[15].

The literature zin emissions forecasting showcases a progression toward deep learning and hybrid architectures that achieve gains in predictive accuracy over traditional statistical baselines. However, these methods remain characterized by a trade-off between predictive power and causal transparency. While these frameworks excel at capturing non-linear patterns in high-dimensional datasets, their "black box" nature and the absence of formal causal identification limit their utility in explaining the structural drivers of emissions.

iii) Machine Learning for Carbon Policy Evaluation

Recent research has begun applying ML to the evaluation of climate policies, moving beyond pure prediction toward identifying effective policy portfolios.

D'Orazio and Pham (2025) examine how climate-related financial policy sequencing and bindingness relate to CO2 emissions and renewable energy production across 87 countries from 2000–2023[16]. After comparing 19 classical regressors via PyCaret, they select an Extra Trees Regressor (n_estimators=100, random_state=123), which achieves R^2 above 0.9 on a 30% validation split for both outcome variables. SHAP values are then used to decompose feature contributions by country group. The high R^2 indicates the model fits the observed relationships well, but this is a measure of association, not causal attribution, and the paper explicitly disclaims causal inference. A SHAP value showing that policy sequencing contributes to lower emissions in advanced economies does not establish that sequencing caused the reduction, since GDP, institutional readiness, and openness are also strong contributors and likely correlated with sequencing itself. The finding that emerging markets respond differently than advanced economies shows that results vary substantially by region, with Sub-Saharan Africa and South Asia showing strong responsiveness to policy in renewable energy production specifically, even as aggregate emissions outcomes lag advanced economies.

Abrell, Kosch, and Rausch (2022) evaluate the UK's Carbon Price Support (CPS), a carbon tax on all fossil-fired power plants introduced in 2013, using an approach that combines economic theory with LASSO regression to construct a counterfactual without a control group [17]. Because the CPS rate changes only annually, providing insufficient variation for model training, the authors use variation in the coal-to-gas fuel price ratio as a proxy for the carbon tax, as this affects plant output similarly. They find that between 2013 and 2016, the CPS reduced electricity sector emissions by 6.2 percent at an average cost of €18 per ton of CO_2, primarily through fuel-switching from coal to gas. They also find that the tax's effectiveness is driven more by relative fuel prices than by the tax rate itself, so a higher carbon tax does not necessarily yield greater emissions reductions. As this study constructs a counterfactual, it links the policy to the claimed emissions reduction. However, by focusing exclusively on the short run, it omits longer-term responses such as investment in renewables, energy efficiency improvements, and plant closures. Additionally, the validity of the fuel price ratio as a proxy depends on whether or not firms behave identically to a change in fuel price ratio or a change in the tax, somewhat limiting its reproducibility in other policy contexts.

Green and Knittel (2020) leverage LASSO with k-fold cross-validation to estimate household-level incidence of climate policies [18]. This study's use of LASSO allows the researchers to downscale to the census tract scale, enabling policy analysis at a fine geographic resolution. With this, Green and Knittel could demonstrate that the distributed effects vary significantly across income and geography. However, the ML component is fairly conservative as LASSO is essentially regularized linear regression, and the low model performance (measured by R^2) for consumer expenditure categories undermines confidence in the goods-and-services portion of the footprint estimates. In addition, it lacks causal inference as the researchers take a correlational approach.

In contrast to purely correlational designs, Stechemesser et al. (2024) employ a "data-driven causal impact assessment" by developing a ML extension of a difference-in-differences (DID) approach with indicator saturation, drawing on the OECD CAPMF policy database and EDGAR emissions data [3]. By employing break detection and synthetic control methods across 1,500 policies in 41 countries, they identified only 63 cases of large scale emission reductions. This study finds that the most successful interventions typically involve policy portfolios rather than isolated instruments. One limitation of this study is that it does not measure the impact of smaller-scale policies, as it targets policies with an effect size of at least 4.5%-13%, which means that this study may understate the impact of gradual changes.

Pretis (2022) evaluates the effectiveness of British Columbia's carbon tax using difference-in-differences, synthetic control, and a ML-based break-detection approach [19]. The break-detection method uses indicator saturation, a general-to-specific ML variable selection procedure, to identify structural breaks in emissions across a panel of Canadian provinces without requiring prior knowledge of treatment timing or assignment. This allows the author to test whether the carbon tax registered as a detectable intervention and what else moved emissions in the data. The results find a significant long-run reduction in transportation emissions of approximately 19%, but no statistically significant effect on aggregate CO_2 emissions, with the break-detection approach revealing that the largest detected reductions coincide not with carbon pricing but with industrial closures and efficiency improvements in electricity generation in untaxed provinces. Rather than evaluating a known policy intervention, the ML component provides an account of what actually drove emissions changes in the panel, which standard difference-in-differences and synthetic control approaches cannot offer. However, the analysis covers a single treated province over a short post-treatment window of nine years, which raises concerns about statistical power to detect aggregate effects, and the findings may not generalise beyond the specific context of British Columbia.

While Abrell et al. (2022) find the UK CPS reduced electricity emissions by 6.2% [17], Pretis finds no significant aggregate effect for the BC tax [19]. The contrast likely reflects available abatement mechanisms rather than tax design or stringency. The UK result came from fuel-switching between coal and gas, a substitution margin whose effectiveness depended more on relative fuel prices than the tax rate itself. BC lacked an equivalent low-cost substitution channel economy-wide, and Pretis's break-detection results suggest reductions previously credited to the tax may instead reflect industrial closures and efficiency gains already underway in untaxed provinces. Therefore, carbon tax effectiveness in this case appears conditional on accessible abatement mechanisms, not simply on whether a tax exists.

Li and Adriaens (2025) use causal machine learning to estimate the impact of U.S. municipal green bond issuance on county-level carbon emissions, addressing the limited evidence on municipal green finance compared with corporate green bonds. Using causal forests within a double/debiased machine learning (DML) framework, they estimate treatment effects across 468 county-year observations from 2009-2019, controlling for factors such as GDP, payroll per employee, education, and establishment size. They find that a 1% increase in green bond issuance is associated with a 0.039% reduction in CO_2 emissions two years later, corresponding to an estimated abatement cost of about $192 per ton, although the result is only marginally significant (p = 0.086). Emission reductions are concentrated in counties with lower payroll per employee and smaller establishments, suggesting that green finance may particularly benefit capital-constrained firms. The authors find no statistically significant difference between certified and self-labeled green bonds, but note that the small certified sample (N=67) makes this result inconclusive. Methodologically, the study differs from other reviewed approaches by directly combining causal forests with DML[20].

Ezenkwu, Cannon, and Ibeke (2024) develop a monitoring-based approach to carbon policy assessment by combining an LSTM model of UK per-capita CO_2 emissions from 1750–2021 with Statistical Process Control (Shewhart I-MR charts) to identify periods when emissions deviate from the model's expected baseline. They then compare these “out-of-control” periods with major UK carbon policies. Although the LSTM outperforms ARIMA, exponential smoothing, and feedforward ANN models, the approach remains fundamentally correlational. The authors acknowledge that linking detected anomalies to specific policies is difficult, particularly when flagged periods span multiple years and overlap several interventions. Unlike DiD and LASSO based methods, this LSTM-SPC framework treats policy identification as an anomaly detection problem rather than estimating a counterfactual. While this makes the approach more scalable, it cannot independently establish policy effects or support causal policy recommendations without additional qualitative analysis, limiting its usefulness for policy evaluation[21].

These evaluative studies are essential to carbon policy analysis because they show how policies work in complex, real-world environments. By using machine learning, researchers can identify the impact of specific policy combinations, helping policymakers make informed decisions regarding climate policy. However, many of these studies lack formal causal inference and cost-benefit analysis, limiting their ability to give insights into what causes policy success and the effects of carbon policy beyond emissions and also possibly indicating confounding factors.

iv) Natural Language Processing and Climate Policy Analysis

NLP has emerged as a tool for classifying and analyzing carbon policy. For instance, Sachdeva et al. (2022) analyzed 318 city-level climate action plans using keyword-based lexicons and logistic regression to determine which textual patterns predict ”ambitious” economy-wide net-zero targets[22]. After converting plain text into Term Frequency-Inverse Document Frequency (TF-IDF) features, the authors fit an L1-regularized logistic regression, validated through leave-one-out cross-validation averaged over 50 train-test splits, with coefficient significance assessed via chi-square tests. The paper finds that quantitative specificity and governance language (particularly mentions of the mayor) are the strongest predictors of ambitious target-setting.

Wu et al. (2024) developed the Global Climate Change Mitigation Policy Dataset (GCCMPD), a dataset covering 73,625 policies globally[23]. Using a semi-supervised hybrid ML approach that combines expert knowledge-based dictionary mapping with advanced NLP, they provide information on policy objectives, target sectors, and legal bindingness, providing a valuable foundation for empirical research. However, this classification is limited in accuracy in some areas. For instance, policies under the category “taxes” were classified with only 0.18 precision, meaning that only 18% of the policies labelled as taxes were correctly labelled.

Li et al. (2026) extend NLP-based policy analysis from text classification to causal policy evaluation. They apply three transformer models, ERNIE 3.0, BERT-base-Chinese, and RoBERTa-wwm-ext, to 1,310 Chinese national climate policies from 1992-2023, evaluating policy objectives, instruments, and stringency and combining these measures into a “Policy Intensity” index. ERNIE 3.0 achieves the highest accuracy across the six classification tasks. Unlike Sachdeva et al. (2022) and Wu et al. (2024), who use NLP primarily to classify policy text, Li et al. incorporate their NLP derived index into a fixed-effects panel regression of CO_2 emissions across 224 Chinese cities from 2003-2020. They further address endogeneity through two-stage least squares (2SLS), using an international climate-momentum index and the frequency of climate-related language in national leaders' speeches as instruments. The results support instrument validity, and the 2SLS estimates indicate that stronger policy intensity significantly reduces city-level emissions, with larger effects in less carbon-intensive, service-oriented cities. This creates a distinct hybrid approach in which NLP constructs the policy measure while econometric IV methods provide causal identification, rather than causal-ML techniques such as causal forests or DML. However, the study is limited by its focus on China and national-level policies, while its causal conclusions depend on the plausibility of the instruments' exclusion restrictions[24].

Hooper et al. (2024) develop a semi-automated pipeline for synthesizing causal claims from policy literature using a deep learning relation-extraction model (SCITE), SBERT semantic clustering, and graph analytics. Applied to 28 articles on emissions trading schemes (ETS), the pipeline extracted 154 causal sentences and 119 unique cause-effect pairs, which were combined into a causal map of 159 links. The map reproduced established ETS mechanisms, including fuel switching, cost pass-through, emissions leakage, and windfall profits, while also identifying a feedback loop in which declining emissions reduce allowance demand and weaken future price signals. Comparison with two manual reviews showed substantial overlap, while reducing extraction time by roughly 90%. However, the method synthesizes causal claims from existing literature rather than estimating causal effects, making it distinct from causal ML approaches such as causal forests and DML. Its 38% recall also means many relationships may be missed, while its inability to capture null findings or geographic and sectoral heterogeneity can obscure important conclusions. This study is relevant to the review as a distinct application of NLP in carbon policy analysis, using deep learning to automate the synthesis of causal evidence across a literature[25].

NLP can classify and compare policies, providing insights on how policymakers can structure policies. However, it is important to note that natural language processing does not directly measure the impact of policies enacted. In addition, NLP only measures policy presence, not policy effectiveness. As such, NLP's utility is primarily as a means of reading, classifying, and standardizing policies.

Discussion

This review examined how machine learning methods have been applied to carbon emissions forecasting and carbon policy evaluation, assessing the extent to which these methods advance beyond the limitations of traditional econometric and simulation-based approaches. The findings reveal that ML has expanded the analytical toolkit available to researchers and policymakers, but that methodological gaps persist, most critically around causal inference, interpretability, and regional biases.

i) Contributions

The reviewed literature demonstrates clear areas where ML offers advantages over traditional approaches. Deep learning architectures such as the TCN-LSTM of Wei and Xu (2023) and the DNN of Mujeeb and Javaid (2023) demonstrate superior capacity to capture nonlinear relationships between emissions and socioeconomic drivers, directly addressing the nonlinearity limitations of CGE and IAM frameworks identified[12,14]. At scale, Stechemesser et al. (2024) screen 1,500 policies across 41 countries using ML-extended synthetic control methods[3], and Wu et al. (2024) classify 73,625 climate policies using NLP[23]. These tasks would be infeasible using traditional methods. However, it is important to distinguish between ML as a tool for prediction and classification and ML as a tool for causal policy evaluation. Deep learning achieves strong predictive accuracy but offers no causal purchase; hybrid ML–econometric frameworks sacrifice predictive precision for identification rigor; and NLP spans multiple roles, from classifying policy text (Sachdeva et al. 2022; Wu et al. 2024) to constructing a policy measure that feeds causal identification via 2SLS (Li et al. 2026) to synthesizing causal claims from existing literature (Hooper et al. 2024) (Table 2).

ii) Causal Inference

The most significant limitation in the reviewed literature is a lack of causal identification. Of the fifteen core papers, Abrell et al. (2022)[17], Stechemesser et al. (2024)[3] and Pretis (2022)[19] make serious attempts at causal attribution. In addition, Li and Adriaens (2025) and Li et al. (2026) utilize the ML component itself to perform the causal identification[20,24]. The remaining nine papers either disclaim causal inference or take a mostly correlational approach. This is a critical limitation for policy relevance, because a model that cannot attribute changes to a specific policy instrument provides limited actionable guidance to policymakers. These studies' emphasis on predictive accuracy, classification performance, or anomaly detection, while internally rigorous, does not satisfy the causal attribution that policy evaluation demands.

iii) Interpretability and Regional/sectoral Bias

A second limitation is the tradeoff between predictive accuracy and interpretability. The highest-performing forecasting models, Mujeeb and Javaid (2023), Wei and Xu (2023), and Wang et al. (2023), and Wang, Hu, Bai, Chang & Yu (2026) are black boxes whose internal logic cannot be directly interpreted for policy insights[12–15]. LASSO-based methods used by Abrell et al. (2022) and Green and Knittel (2020) are more interpretable but sacrifice the nonlinear flexibility that motivates the use of ML[17,18]. D'Orazio and Pham (2025) attempt to bridge this gap through SHAP values, which is promising but describes variable importance rather than causal mechanisms[16]. There also appears to be regional and sectoral bias. The studies cluster around the power sector, though many are economy wide aggregates or cover multiple sectors, and a small number of national contexts, including the US, UK, and China. D'Orazio and Pham (2025) explicitly

Table 1 Summary of the 15 machine learning studies included in this literature review.

Author / year	Policy type	ML technique	Definition of policy effectiveness	Main findings	Geography
Mujeeb and Javaid (2023)[12]	N/A — emissions forecasting	DNN with SCA feature selection, ISDAE noise filtering, IPSO hyperparameter tuning	Predictive accuracy (NRMSE, MAE, MSE) — not a policy effectiveness measure	NRMSE of 0.1730; ∼55% improvement over MLR and SVM baselines	New England, USA
Wang *et al.* (2023)[13]	N/A — emissions forecasting	Two-stage hybrid SVR-ANN and Random Forest	N/A	SVR-ANN achieves lowest RMSE; two-stage hybrid outperforms single-stage models	China
Wei and Xu (2023)[14]	N/A — emissions forecasting	TCN-LSTM with attention mechanism	Classification accuracy (accuracy, F-score, AUC) against binarized emissions outcomes — threshold unspecified	97.66% accuracy and 94.56% F-score, outperforming six baselines (predominantly grey-system statistical models)	Not specified
Abrell, Kosch and Rausch (2022)[17]	Carbon tax (UK Carbon Price Support)	LASSO with k-fold cross-validation; theory-grounded counterfactual simulation	Cost-effectiveness: emissions reduced (% and Mt CO_2) and cost per ton of CO_2 abated (€/tCO_2)	CPS reduced electricity sector emissions by 6.2% at avg. cost of €18/tCO_2 (2013–2016); effectiveness driven by relative fuel prices, not tax rate alone	UK (electricity sector)
Green and Knittel (2020)[18]	Climate policy household incidence	LASSO with k-fold cross-validation	Distributional cost-benefit: net household cost/benefit ($) by income quintile and geography across 12 policy scenarios	Distributed effects vary significantly across income and geography; dividend schemes produce net benefits for low-income households	USA (census tract)
Stechemesser *et al.* (2024)[3]	1,500 carbon policies (mixed)	ML-extended DiD with break detection and synthetic control	Causal emissions reduction: statistically significant structural break exceeding a 4.5–13% effect-size threshold	Only 63 of 1,500 policies (∼4%) produced large-scale emissions reductions; successful interventions typically involve policy portfolios	41 countries
Pretis (2022)[19]	Carbon tax (British Columbia)	DiD, synthetic control, indicator saturation (agnostic break detection)	Causal emissions reduction: statistically significant structural break in provincial emissions attributable to the tax	∼19% long-run reduction in transportation emissions; no statistically significant aggregate CO_2 effect; largest detected reductions linked to industrial closures and efficiency gains in untaxed provinces	Canada (British Columbia)
D'Orazio and Pham (2025)[16]	Climate-related financial policies	Extra Trees Regressor with SHAP values; policy sequencing score	Predictive association: CO_2 emissions levels and renewable energy production as outcome variables; no causal identification	Strategic policy sequencing reduces emissions in advanced economies; EMDEs show strong responsiveness in renewable energy production despite institutional constraints	87 countries
Sachdeva *et al.* (2022)[22]	City-level climate action plans	Keyword-based lexicons with logistic regression	N/A — classifies plan ambition (economy-wide net-zero target vs. not); does not measure emissions outcomes	Ambitious plans emphasize quantitative metrics and high-emitting sectors; energy focus dominates at the expense of land-use and climate impacts	318 cities (global, predominantly Global North)
Wu *et al.* (2024)[23]	Global mitigation policies	Semi-supervised hybrid NLP (ClimateBERT, dictionary mapping, BM25, BERTopic)	N/A — classifies policy characteristics; does not measure emissions outcomes	GCCMPD dataset covering 73,625 policies; macro F1 of 0.91 for sector classification; tax category precision 0.18 due to classification standard mismatch	Global (216 entities)
Wang, Hu, Bai, Chang & Yu (2026)[15]	N/A — carbon allowance cap forecasting	Path analysis for predictor selection; five ML models compared (decision tree, BPNN, random forest, SVM, Gaussian Process Regression)	N/A — forecasting task; path analysis identifies structural associations, not causal relationships	GPR outperforms all other models across error metrics; 2026–2035 projections adjusted using external policy targets (Goldman Sachs report, China's 2035 NDCs)	China (8 carbon-trading pilot regions)
Li and Adriaens (2025)[20]	Municipal green bond issuance	Causal forests within double/debiased machine learning (DML)	Emissions reduction: heterogeneous treatment effect of green bond issuance on CO_2 emissions, controlling for GDP, payroll, education, establishment size	1% increase in issuance associated with 0.039% reduction in CO_2 two years later (∼$192/ton abatement cost); marginally significant ($p = 0.086$); effects concentrated in lower-payroll, smaller-establishment counties	USA (468 county-year observations, 2009–2019)
Ezenkwu, Cannon & Ibeke (2024)[21]	UK carbon policies (mixed, retrospective)	LSTM emissions model + Statistical Process Control (Shewhart I-MR charts) for anomaly detection	N/A — correlational monitoring approach; flags "out-of-control" emissions periods for qualitative comparison against known policy timelines.	LSTM outperforms ARIMA, exponential smoothing, and feedforward ANN baselines; detected anomaly periods overlap with major UK policies (Energy Conservation Act, Climate Change Levy, EU ETS, Climate Change Act) but cannot isolate individual policy effects	UK (1750–2021)
Li *et al.* (2026)[24]	1,310 Chinese national climate policies	Transformer-based NLP (ERNIE 3.0, BERT-base-Chinese, RoBERTa-wwm-ext) to construct "Policy Intensity" index; fixed-effects panel regression with 2SLS	Causal emissions reduction: 2SLS estimate of policy intensity's effect on city-level CO_2, instrumented via international climate-momentum index and leader speech frequency	ERNIE 3.0 achieves highest classification accuracy; stronger policy intensity significantly reduces emissions, with larger effects in less carbon intensive, service oriented cities	China (224 cities, 2003–2020)
Hooper *et al.* (2024)[25]	Emissions trading schemes (ETS), literature derived	Deep learning relation-extraction (SCITE), SBERT semantic clustering, graph analytics	N/A — synthesizes existing causal claims from literature into a causal map rather than estimating a new causal effect	Extracted 154 causal sentences, 119 cause-effect pairs, 159-link causal map; reproduced known ETS mechanisms (fuel switching, cost pass-through, leakage) plus a feedback loop; substantial overlap with 2 manual reviews at ∼90% less extraction time; 38% recall	Global (28 articles on ETS)

Table 2 Comparison of ML families by task suitability, key strengths and limitations, and causal credibility.

ML family	Best suited for	Key strengths	Key limitations	Causal credibility
Regularized regression (e.g. LASSO, Ridge)	High-dimensional feature selection; distributional and incidence analysis	Interpretable coefficients; compatible with economic theory; handles many predictors without overfitting	Assumes linear relationships; limited flexibility for complex nonlinear patterns	Low to moderate. Correlational by default; causal credibility only achieved when embedded in a structural economic model
Deep learning (e.g. DNN, TCN-LSTM)	Time-series emissions forecasting	Highest predictive accuracy; captures complex nonlinear and temporal dependencies	Black box; no causal identification; results sensitive to undocumented implementation choices	None
Tree ensembles (e.g. Random Forest, Extra Trees)	Multi-country policy impact prediction	Robust to noise; handles high-dimensional data; compatible with post-hoc interpretability tools (e.g. SHAP)	Underperforms on monotonically trending data; SHAP describes variable importance, not causation	None
Hybrid ML–econometric (e.g. ML-extended DiD, indicator saturation)	Large-scale causal policy evaluation	Combines ML scalability with econometric identification; agnostic break detection without pre-specified treatment timing	Requires strong parallel trends assumptions; computationally intensive at scale	Strong, only family achieving credible causal attribution in the reviewed literature
NLP (e.g. BERT, TF-IDF logistic regression)	Policy text classification and dataset construction	Enables analysis at scales infeasible through manual coding; captures semantic patterns across thousands of documents	Measures policy presence, not effectiveness; cannot construct counterfactuals; precision varies by category (Sachdeva *et al.* 2022[22]; Wu *et al.* 2024[23]); credible when paired with econometric identification, e.g. 2SLS (Li *et al.* 2026[24]), though credibility derives from the IV strategy rather than the NLP step	Not applicable when used for classification alone

find that policy effectiveness varies significantly between advanced and emerging economies[16], but no reviewed paper develops frameworks specifically suited to low-institutional-capacity contexts. Across the 15 reviewed papers, 9 focus on a single country or region, 5 span across multiple regions, and 1 (Wei and Xu 2023) does not specify its geographic scope. Of the single region papers, 6 are focused on developed countries, and 3 are focused on a developing country, which is China. Of the papers that are focused on multiple regions, Sachdeva et al. (2022) more heavily focuses on developed countries, and the other 3 include both developing and developed nations, though Wu et al. (2024) and Sachdeva et al. (2022) note data availability issues for developing regions[22,23].

iv) Economic and Implementation Dimensions

The reviewed literature focuses somewhat narrowly on emissions outcomes, neglecting the broader economic and implementation dimensions of policy effectiveness. Only Abrell et al. (2022), Green and Knittel (2020), and Li and Adriaens (2025) address economic impacts: the first through a cost-effectiveness estimate of €18 per ton of CO_2, the second through distributional household incidence analysis, and the third through an estimated abatement cost of $\sim$\$192 per ton of CO_2 from municipal green bond issuance, though this last estimate is only marginally significant[17,18,20]. No reviewed paper thoroughly examines employment effects, GDP impacts, or industry competitiveness. Similarly, while Stechemesser et al. (2024) implicitly signal widespread policy underperformance by identifying only 63 successful interventions out of 1,500, no reviewed paper investigates the mechanisms behind policy failure, such as enforcement gaps, political economy constraints, and behavioral factors that cause carbon policies to underperform[3]. This narrow scope limits the actionability of the reviewed literature for policymakers who must weigh emissions outcomes against economic and social costs.

Implementation barriers receive little attention across the reviewed literature. D'Orazio and Pham (2025) provide the most substantive treatment, finding that weak enforcement mechanisms, absent green taxonomies, and structural reliance on carbon-intensive industries can cause well-designed policies to produce limited or counterproductive outcomes[16]. Stechemesser et al. (2024) and Green and Knittel (2020) gesture toward political economy constraints, the former by finding that successful interventions require policy portfolios rather than isolated instruments, the latter by flagging that carbon taxes' distributional burdens represent a significant political obstacle, but neither investigates these dynamics systematically[3,18]. Most of the remaining studies largely assume implementation away. Forecasting models treat policy as an input variable without considering enforcement, while NLP studies measure policy presence rather than compliance. Future research should incorporate enforcement quality, political feasibility, and behavioral response as variables in carbon policy evaluation.

v) Replicability of studies

Replicability varies substantially across the reviewed literature. Stechemesser et al. (2024) and Pretis (2022) are the most replicable, as both use publicly archived emissions data, formally specified econometric methods, and openly available code, with Stechemesser et al. providing a full replication archive on Zenodo including a dedicated R package[3,19]. For Abrell et al. (2022) and Green and Knittel (2020), while methodology is sufficiently documented for reconstruction, practical replication is constrained by restricted commercial data access and the absence of formal replication packages[17,18]. Li and Adriaens (2025) and Wang, Hu, Bai, Chang and Yu (2026) both fully specify their methods, but rely on proprietary data (Bloomberg; CNRDS) with no code or data repository[15,20]. D'Orazio and Pham (2025), Sachdeva et al. (2022) and Wu et al. (2024) demonstrate replicability, as D'Orazio and Pham (2025) and Sachdeva et al. (2022) both use publicly available data sources and open-source tooling, with Sachdeva et al. providing code on GitHub and Wu et al. (2024) depositing the full GCCMPD dataset on Figshare[16,22,23]. Ezenkwu, Cannon and Ibeke (2024) is similarly replicable at the data level, using the public "Our World in Data" UK CO_2 series, though no code repository is cited[21]. Hooper et al. (2024) uses open source tools and a listed source corpus, but the reported causal map depends on manual verification and an "iteratively tuned" clustering threshold, limiting reproducibility[25]. The weakest cases are the deep learning forecasting studies. Mujeeb and Javaid (2023) and Wei and Xu (2023) present novel architectures involving non-standard components that were developed by the authors themselves, with no code repositories, and underspecified data extraction procedures[12,14]. Li et al. (2026) use standard pretrained transformers and publish full scoring criteria, but the manually annotated dataset, partly drawn from a subscription legal database, is not deposited and neither is the code[24]. As the aim is to inform real-world policy decisions, replicability limitations represent a meaningful constraint on the weight that can be placed on the forecasting literature.

vi) Hybrid Approaches and Future Directions

The papers that most credibly advance carbon policy analysis are those that combine ML with causal econometric frameworks. Abrell et al. (2022) use LASSO to construct a theory-grounded counterfactual, Stechemesser et al. (2024) extend DiD with ML-based break detection, and Li, Shen, Lin and Liu (2026) combine NLP-derived measures with 2SLS instrumental variables, each achieving causal credibility through a different hybrid design[3,17,24]. Future research should prioritise developing causally grounded hybrid ML frameworks, expanding geographic and sectoral coverage, incorporating economic and distributional dimensions of policy effectiveness, and advancing interpretable ML methods suited to policy analysis contexts. In addition, methods such as reinforcement learning, graph neural networks, and Bayesian approaches remain underrepresented in this literature, despite their theoretical fit for carbon policy analysis. Reinforcement learning could simulate how economic agents adapt their behavior in response to evolving policy, graph neural networks could capture spillover effects and interdependencies between countries and sectors, and Bayesian methods could provide formal uncertainty quantification for emissions forecasts and policy impact estimates.

vii) Limitations of This Review

This review adopts a narrative rather than a systematic synthesis approach, introducing potential selection bias in included papers. Including only studies published after 2020 ensures methodological currency but excludes earlier foundational work. The regional and sectoral biases of the reviewed literature are inherited by this review, meaning conclusions may reflect developed-country and power-sector contexts more than the global picture. In addition, the methodological heterogeneity of included studies makes direct cross-study comparison difficult, and findings should be interpreted as thematic observations rather than definitive conclusions about specific methods.

Conclusion

This review has examined how machine learning methods have been applied to carbon emissions forecasting and carbon policy evaluation across fifteen empirical studies. At scale, NLP methods have enabled policy classification and synthesis that would be infeasible through manual methods. In forecasting, deep learning architectures have demonstrated meaningful accuracy gains over traditional statistical baselines. And in causal policy evaluation, hybrid ML–econometric frameworks and causal ML methods have shown that ML can extend the reach of credible causal inference to policy screening.

However, the most significant limitation is that the majority of papers prioritize predictive accuracy over causal identification, limiting their utility for policy guidance. Regional and sectoral coverage remains heavily skewed toward developed economies and the power sector, with agriculture, buildings, and industry underrepresented, and the Global South underrepresented in both single country and multi country studies. Replicability is a further concern, concentrated in the deep learning forecasting papers, where novel architectures and absent code repositories make independent verification difficult. In future studies, hybrid ML–econometric approaches and causal ML should be extended to developing country contexts and underrepresented sectors. Replication archives and preregistered model specifications should become standard practice, particularly for studies making policy relevant causal claims. Future research should focus on expanding geographic and sectoral coverage, including economic and behavioral impacts in their analyses, and including implementation barriers in their analyses. ML's potential contribution to carbon policy analysis is substantial, but realizing it requires methodological standards that the reviewed literature has not yet consistently met.

Glossary

Attention mechanism — A neural network component that assigns learned weights to different parts of an input sequence, allowing a model to prioritize the most relevant information rather than treating all inputs equally.

Computable General Equilibrium (CGE) model — A simulation-based economic model that represents how markets across an entire economy adjust to a policy change, used to estimate macroeconomic and emissions impacts.

Causal forest — A tree based ensemble method that estimates how a treatment's effect varies across individual observations by using tree-splitting rules designed to maximize differences in estimated treatment effects between subgroups.

Difference-in-differences (DiD) — A causal inference method that compares changes in an outcome over time between a treated group and a control group, isolating the effect of an intervention from general trends affecting both groups.

Double/debiased machine learning (DML) — A framework that uses machine learning to estimate nuisance functions (such as the outcome model and treatment assignment model) while applying a statistical correction so that the resulting treatment effect estimate remains valid even if the ML components are imperfectly specified.

Endogeneity — A situation in which an explanatory variable is correlated with a regression model's error term, biasing coefficient estimates.

Extra Trees Regressor — An ensemble method similar to Random Forest that introduces additional randomness by selecting both random feature subsets and random split thresholds at each decision point, further reducing variance.

Improved Shallow Denoising Autoencoder (ISDAE) — A neural network architecture that learns to reconstruct clean data from noisy input, used here to filter noise from raw power system data before forecasting.

Integrated Assessment Model (IAM) — A model that links economic and climate systems to evaluate long-run mitigation pathways and compute metrics such as the social cost of carbon.

K-fold cross-validation — A model validation technique that splits data into k subsets, training on $k-1$ of them and testing on the remaining subset, repeating this process to estimate prediction error.

LASSO (Least Absolute Shrinkage and Selection Operator) — A regularized regression technique that penalizes the size of coefficients, shrinking some to exactly zero, which performs feature selection and prediction simultaneously while reducing overfitting.

Logistic regression — A statistical model used to predict a binary outcome based on one or more predictor variables, expressed as the log-odds of the outcome occurring.

Long Short-Term Memory (LSTM) — A type of recurrent neural network designed to retain information over long sequences using internal gating mechanisms, addressing limitations of simpler recurrent models in capturing long-range dependencies.

NRMSE (Normalized Root Mean Square Error) — RMSE scaled relative to the range or mean of the observed data, allowing comparison of prediction error across datasets with different scales.

Quasi-experimental design — A research design that approximates the conditions of a randomized experiment using naturally occurring variation, rather than random assignment, to estimate causal effects.

R^2 (coefficient of determination) — A measure of how much variance in the outcome variable is explained by a model, ranging from 0 to 1 (or negative, for very poor fits), with higher values indicating better fit.

Random Forest (RF) — An ensemble learning method that combines predictions from many decision trees trained on random subsets of data and features, reducing variance and overfitting relative to a single tree.

Regularization — A general technique that penalizes model complexity during training to prevent overfitting and improve out-of-sample prediction.

RMSE / MAE (Root Mean Squared Error / Mean Absolute Error) — Standard metrics for measuring the average difference between a model's predicted values and actual observed values, with lower values indicating better predictive accuracy.

Semi-supervised learning — A machine learning approach that combines a small set of manually labeled data with a much larger set of unlabeled data to train a classification model at scale.

SHAP (Shapley Additive Explanations) — An interpretability method derived from cooperative game theory that attributes a model's prediction to the individual contribution of each input feature, enabling post-hoc explanation of complex models.

Spearman Correlation Analysis (SCA) — A feature selection method that ranks variables by the strength of their monotonic (not necessarily linear) relationship with the target variable.

Support Vector Machine (SVM) / Support Vector Regression (SVR) — A machine learning method that finds an optimal boundary or function separating or fitting data, effective for nonlinear relationships in small-to-medium datasets.

Synthetic control — A method that constructs an artificial comparison unit by weighting a combination of untreated units to closely match the treated unit's pre-intervention trajectory, providing a counterfactual when no single natural control group exists.

Temporal Convolutional Network (TCN) — A neural network architecture that applies convolutional filters across time steps to capture temporal patterns in sequential data.

Transformer model — A neural network architecture that processes sequential data, such as text, using self-attention rather than recurrence, allowing it to model relationships between all elements of a sequence simultaneously.

References

1 J. Rogelj, M. d. Elzen, N. Höhne, T. Fransen, H. Fekete, H. Winkler, R. Schaeffer, F. Sha, K. Riahi, M. Meinshausen. Paris Agreement climate proposals need a boost to keep warming well below 2 °C. *Nature*. Vol. 534, pg. 631–639, 2016. [https://doi.org/10.1038/nature18307.

2 Frankfurt School–UNEP Collaborating Centre for Climate & Sustainable Energy Finance, BloombergNEF. Global trends in renewable energy investment 2020. 2020. https://wedocs.unep.org/handle/20.500.11822/32700.

3 A. Stechemesser, N. Koch, E. Mark, A. Krueger, G. Shapiro, H. Geissler, S. Brandt, J. Reckin, C. Gornott, M. Jakob. Climate policies that achieved major emission reductions: Global evidence from two decades. *Science*. Vol. 385, pg. 884–892, 2024. [https://doi.org/10.1126/science.adl6547.

4 E. Levine. A case study of British Columbia's carbon tax. *CLOSUP Student Working Paper Series*. No. 70, University of Michigan, 2021. [https://closup.umich.edu/research/student-working-papers/case-study-british-columbias-carbon-tax.

5 Y. Han, L. Cao, Z. Geng, W. Ping, X. Zuo, J. Fan, J. Wan, G. Lu. Novel economy and carbon emissions prediction model of different countries or regions in the world for energy optimization using improved residual neural network. *Science of The Total Environment*. Vol. 860, pg. 160410, 2023. [https://doi.org/10.1016/j.scitotenv.2022.160410.

6 K. A. Akkemik, J. T. Borges, P. T. Dang. Assessing carbon tax using a CGE model with firm heterogeneity: An application to Vietnam. *Journal of Environmental Management*. Vol. 365, pg. 121585, 2024. [https://doi.org/10.1016/j.jenvman.2024.121585.

7 J. Weyant. Some contributions of integrated assessment models of global climate change. *Review of Environmental Economics and Policy*. Vol. 11, pg. 115–137, 2017. [https://doi.org/10.1093/reep/rew018.

8 R. Rafaty, G. Dolphin, F. Pretis. Carbon pricing and the elasticity of CO2 emissions. *Energy Economics*. Vol. 144, pg. 108298, 2025. [https://doi.org/10.1016/j.eneco.2025.108298.

9 S. S. Scrieciu. The inherent dangers of using computable general equilibrium models as a single integrated modelling framework for sustainability impact assessment: A critical note on Böhringer and Löschel (2006). *Ecological Economics*. Vol. 60, pg. 660–670, 2007. [https://doi.org/10.1016/j.ecolecon.2006.07.032.

10 S. Asefi-Najafabady, L. Villegas-Ortiz, J. Morgan. The failure of integrated assessment models as a response to "climate emergency" and ecological breakdown: The Emperor has no clothes. *Globalizations*. Vol. 18, pg. 1178–1188, 2021. [https://doi.org/10.1080/14747731.2020.1853958.

11 J. Roth, P. H. C. Sant'Anna, A. Bilinski, J. Poe. What's trending in difference-in-differences? A synthesis of the recent econometrics literature. *Journal of Econometrics*. Vol. 235, pg. 2218–2244, 2023. [https://doi.org/10.1016/j.jeconom.2023.03.008.

12 S. Mujeeb, N. Javaid. Deep learning based carbon emissions forecasting and renewable energy's impact quantification. *IET Renewable Power Generation*. Vol. 17, pg. 873–884, 2023. [https://doi.org/10.1049/rpg2.12641.

13 C. Wang, M. Li, J. Yan. Forecasting carbon dioxide emissions: application of a novel two-stage procedure based on machine learning models. *Journal of Water and Climate Change*. Vol. 14, pg. 477–493, 2023. [https://doi.org/10.2166/wcc.2023.331.

14 X. Wei, Y. Xu. Research on carbon emission prediction and economic policy based on TCN-LSTM combined with attention mechanism. *Frontiers in Ecology and Evolution*. Vol. 11, pg. 1270248, 2023. [https://doi.org/10.3389/fevo.2023.1270248.

15 X. Wang, W. Hu, L. Bai, W. Chang, X. Yu. Forecasting the total carbon allowance cap under emission-reduction targets using a hybrid path analysis and supervised machine learning framework. *Frontiers in Environmental Science*. Vol. 14, pg. 1757914, 2026. [https://doi.org/10.3389/fenvs.2026.1757914.

16 P. D'Orazio, A. D. Pham. Evaluating climate-related financial policies' impact on decarbonization with machine learning methods. *Scientific Reports*. Vol. 15, pg. 1694, 2025. [https://doi.org/10.1038/s41598-025-85127-7.

17 J. Abrell, M. Kosch, S. Rausch. How effective is carbon pricing? – A machine learning approach to policy evaluation. *Journal of Environmental Economics and Management*. Vol. 112, pg. 102589, 2022. [https://doi.org/10.1016/j.jeem.2021.102589.

18 T. W. Green, C. R. Knittel. Distributed effects of climate policy: a machine learning approach. *The Roosevelt Project Working Paper Series*, 2020. [https://ceepr.mit.edu/wp-content/uploads/2021/09/The-Roosevelt-Project-WP-3.pdf.

19 F. Pretis. Does a carbon tax reduce CO2 emissions? Evidence from British Columbia. *Environmental and Resource Economics*. Vol. 83, pg. 115–144, 2022. [https://doi.org/10.1007/s10640-022-00679-w.

20 D. Li, P. Adriaens. Green bond issuance and carbon emissions: can causal machine learning inform forward-looking policy decisions? *Environmental Science Technology*. Vol. 59, pg. 24672–24682, 2025. [https://doi.org/10.1021/acs.est.5c04966.

21 C.P. Ezenkwu, S. Cannon, and E. Ibeke. Monitoring carbon emissions using deep learning and statistical process control: a strategy for impact assessment of governments' carbon reduction policies. *Environmental Monitoring and Assessment*. Vol. 196, pg. 231, 2024. [https://doi.org/10.1007/s10661-024-12388-6.

22 S. Sachdeva, A. Hsu, I. French, E. Lim. A computational approach to analyzing climate strategies of cities pledging net zero. *npj Urban Sustainability*. Vol. 2, pg. 21, 2022. [https://doi.org/10.1038/s42949-022-00065-x.

23 L. Wu, Z. Huang, X. Zhang, Y. Wang. Harmonizing existing climate change mitigation policy datasets with a hybrid machine learning approach. *Scientific Data*. Vol. 11, pg. 580, 2024. [https://doi.org/10.1038/s41597-024-03411-z.

24 Y. Li, J. Shen, Y. Lin, and X. Liu. The intensity of China's climate

change policy from 1992 to 2023: Measurement, evolution and consequence. *Habitat International*. Vol. 167, pg. 103667, 2026. [https://doi.org/10.1016/j.habitatint.2025.103667.

25 R. Hooper, N. Goyal, K. Blok, L. Scholten. A semi-automated approach to policy-relevant evidence synthesis: combining natural language processing, causal mapping, and graph analytics for public policy. *Policy Sciences*. Vol. 57, pg. 875–900, 2024. [https://doi.org/10.1007/s11077-024-09548-3.

