Array 29 (2026) 100639

Contents lists available at ScienceDirect

Array

journal homepage: www.sciencedirect.com/journal/array

An explainable comparative study of statistical, machine learning, deep learning, and hybrid models for CO_2 emissions forecasting in Australia

Safa Ghannam

School of Professional Practice and Leadership Faculty of Engineering & IT, Australian Artificial Intelligence Institute, University of Technology Sydney, Australia

ARTICLE INFO

Keywords:
CO_2 emissions forecasting
Statistical models
Machine learning
Deep learning
Explainable AI
Net-zero 2050
Climate policy

ABSTRACT

Accurate forecasting of national CO_2 emissions is critical for evidence-based climate policy and for meeting commitments such as Australia's 2050 net-zero target and the United Nations Sustainable Development Goal 13 (Climate Action). This study implements and evaluates thirteen forecasting approaches, including statistical models (ARIMA), machine learning methods (random forest, XGBoost, SVR), kernel methods (GPR), hybrid approaches (ELM, ISSA-ELM), deep learning networks (MLP, LSTM, GRU, RNN), and two ensemble models (stacking regressor and enhanced stacking regressor), using annual Australian data from 1982 to 2022 within a reproducible pipeline. Thirty random seeds ensured robustness for stochastic learners. Ensemble tree methods delivered the most accurate and stable predictions: Random Forest achieved mean cross-validation $R^2 \approx 0.989 \pm 0.003$ and RMSE $\approx 0.018 \pm 0.002$ and generalized well to unseen 2016–2022 data ($R^2 \approx 0.96$; RMSE ≈ 2.43 Mt CO_2). Pairwise significance testing confirmed that Random Forest and stacking significantly outperformed most individual learners ($p < 0.01$). SHAP analysis identified energy productivity, total GHG excluding land-use change, total energy consumption, and population as dominant drivers. Scenario experiments show that deterministic adjustments yield only modest 2050 reductions (−0.49 % to −2.68 %), with population shifts treated as exogenous sensitivities, underscoring the need for system-level action to achieve net-zero. Limitations include reliance on annual data and exclusion of policy and trade factors. Future work could extend this framework through causal inference and hybrid physics-informed machine learning. Building on global advances in emissions forecasting, this study contributes a localized, interpretable comparative framework tailored to Australia's emissions profile, addressing a notable gap in national-level forecasting research. This transparent and reproducible approach provides evidence-based guidance for model selection and supports policy-relevant discussions on national CO_2 forecasting.

1. Introduction

Carbon dioxide (CO_2) emissions, which account for approximately 63 % of Australia's total

greenhouse gas emissions as reported in the quarterly update of Australia's National.

Greenhouse Gas Inventory March 2024, remain a principal driver of global climate change.

Accurate and interpretable forecasting of CO_2 emissions is crucial to support evidence-based policymaking, environmental planning, and sustainability strategies. Australia, with one of the highest per capita CO_2 emission rates globally, faces unique challenges due to its reliance on fossil fuels, expansive urban development, and diverse climatic zones [1]. In line with international commitments such as the Paris Agreement, Australia has pledged to achieve net-zero emissions by 2050, aligning with the United Nations Sustainable Development Goal 13 (Climate Action). This context underscores the importance of developing robust forecasting frameworks that can inform national decarbonisation pathways.

Machine learning (ML) methods have emerged as powerful tools for forecasting emissions, demonstrating superior capabilities over traditional statistical models in capturing nonlinear relationships and complex feature interactions [2,3]. Recent studies have expanded ML applications beyond national aggregates to urban contexts, identifying critical drivers such as transportation activity [4], GDP and energy consumption [5], urban density and population growth [7], industrial output [17] and climatic variables [18]. The integration of diverse datasets, including socio-economic indicators and sensor-based traffic data, has further enhanced the granularity and relevance of emissions prediction models [7].

E-mail address: safa.ghannam@uts.edu.au.

https://doi.org/10.1016/j.array.2025.100639
Received 22 September 2025; Received in revised form 7 December 2025; Accepted 8 December 2025
Available online 11 December 2025

Of the various ML models, ensemble approaches such as random forest, gradient boosting, and XGBoost have consistently outperformed traditional statistical techniques, particularly in terms of predictive accuracy and robustness against overfitting [6,8]. However, while deep learning models like long short-term memory networks (LSTM) and convolutional neural networks (CNNs) have shown strong performance in high-frequency, large-volume datasets, their effectiveness diminishes when applied to smaller, coarse-grained datasets such as annual national CO_2 emissions, as observed in this study and supported by prior research [9,10].

Increasingly, model transparency has become as critical as predictive performance, particularly for policy-relevant applications. Tools such as SHAP (Shapley Additive Explanations) are widely employed to interpret complex ML models and reveal the relative influence of input variables [4,8]. Feature importance methods based on ensemble trees and interpretability frameworks ensure that models provide actionable insights into the underlying factors driving emissions, beyond mere predictive accuracy.

Despite recent advancements, few studies provide a unified, explainable framework comparing a broad spectrum of forecasting models from traditional statistical approaches to advanced deep learning methods within the Australian context. Most existing studies either focus on a single class of models or prioritize predictive accuracy without addressing interpretability. Although there has been significant global progress, the application of interpretable, machine-learning-driven CO_2 forecasting specific to Australia remains largely unexplored. Much of the existing literature concentrates on countries such as the United States or China, with limited attention given to Australia's unique economic, geographic, and policy environment [11].

This study addresses this gap by systematically evaluating thirteen forecasting models for Australia's CO_2 emissions from 1982 to 2022, using national-level datasets from Our World in Data and the Australian Energy Statistics. The analysis emphasizes both predictive performance and explainability, ensuring that the results provide not only accurate forecasts but also actionable insights for policymakers and environmental planners. By integrating statistical, machine learning, deep learning, and hybrid approaches within a reproducible pipeline, this work contributes a transparent comparative framework that supports Australia's decarbonisation strategies and aligns with Australia's 2050 net-zero commitments.

2. Literature review and related work

The accurate forecasting of CO_2 emissions has been a critical area of research in both environmental science and machine learning. Early forecasting approaches relied heavily on traditional statistical models such as grey models, ARIMA, and SARIMAX, offering interpretability but often struggling to capture the nonlinear, multivariate dynamics inherent in emissions data [2,6]. Comparative studies have consistently demonstrated that ML models, particularly ensemble methods, outperform statistical models in both predictive accuracy and adaptability [3,8].

Across the literature, common influencing factors used for emissions modelling include GDP, total energy consumption, industrial activity, transportation metrics, urban density, population, and weather-related variables [5,12]. Open-access socio-economic and environmental datasets have become increasingly valuable, enabling robust feature selection across diverse geographic scales [4,7]. While urban-focused studies highlight real-time, sensor-based traffic emissions as critical drivers [4], national-level research similarly confirms that economic and energy-related factors remain dominant predictors.

Ensemble learning models, such as random forest, gradient boosting machines, and XGBoost, have consistently demonstrated strong performance by effectively modelling nonlinear relationships and complex feature interactions without significant overfitting [3,6]. Deep learning models, including multilayer perceptrons, CNNs, and LSTM networks, have shown promise particularly in handling large, high-frequency datasets [4,5]. However, several studies emphasize that deep models require rich temporal or spatial resolution to outperform simpler ML techniques, and their advantage diminishes when applied to smaller, annual datasets [9,10].

Pre-processing techniques such as normalization, standardization, outlier removal, and logarithmic transformation have been widely adopted to enhance model performance across multiple studies [5,13,14,18]. In parallel, feature selection methods, including tree-based importance rankings, recursive feature elimination, and ReliefF have been instrumental in improving model efficiency and mitigating overfitting [8,15]. Evaluation frameworks and integrated multi-factor approaches, such as hybrid decomposition models and methodological guides for reproducible ML workflows, further enhance interpretability and robustness in emissions forecasting [19,20]. Beyond feature engineering, several studies have emphasized the importance of robust evaluation frameworks. Optimized regression-based ML models for energy-related CO_2 emissions have been benchmarked using widely adopted performance metrics such as MSE, RMSE, R^2, MAE, and MAPE [21]. Similarly, daily carbon emission prediction studies have employed multi-stage feature selection combined with extreme learning machines to improve accuracy and reliability [22]. Machine learning has also been applied to national-level CO_2 forecasting in the United States, further demonstrating the global relevance of data-driven approaches and the consistency of evaluation practices [23].

Recent trends in the literature highlight the growing importance of model interpretability, particularly through SHAP values. Studies applying SHAP to ensemble models, such as random forest and XGBoost, provide a clearer understanding of feature contributions and foster greater trust in model outputs [4,8]. More broadly, explainable artificial intelligence (XAI) has been recognized as a powerful tool in renewable energy systems, enhancing transparency, accountability, and overall model efficacy [24]. The integration of renewable sources into urban energy systems has likewise been identified as a critical component of sustainable development and long-term emission-reduction strategies [25]. In maritime transportation, SHAP and LIME have also been applied to predict fuel consumption and identify key operational drivers [26]. Collectively, these studies underscore the growing importance of explainability across energy and environmental modelling domains. These examples highlight the broader relevance of explainable AI in energy and environmental modelling.

In the context of CO_2 forecasting, SHAP analyses often reveal that demographic, economic, and energy-related factors such as GDP [1,5], total energy use [5,13], population [7], and industrial output [17] are among the most influential factors. While considerable advances have been made globally, research specifically tailored to Australia's emissions forecasting remains scarce. Most studies continue to prioritize major economies or broader regional analyses, often overlooking Australia's unique emission patterns, policy frameworks, and urban infrastructure dynamics [11,16]. Given Australia's combination of high per capita emissions and fossil fuel dependence, localized and interpretable modelling approaches are urgently needed.

Global forecasting efforts have also explored metaheuristic algorithms, such as in India's greenhouse gas trajectory modelling [27], highlighting methodological diversity but reinforcing the scarcity of localized, interpretable approaches tailored to Australia. Recent studies have adopted multi-method feature selection techniques to enhance the accuracy and interpretability of CO_2 emissions forecasting. For instance, Spearman correlation and mutual information have been employed to detect both linear and nonlinear associations between emissions and predictors, while machine-learning-based approaches such as random forest and XGBoost provide feature importance scores that reflect complex interactions. To further improve model transparency, explainable AI tools such as SHAP have been integrated into these workflows to quantify the contribution of each input variable [19,20]. The combined use of these statistical and machine learning methods offers a comprehensive strategy for identifying influencing factors and improving model performance in emissions modelling.

This study addresses key gaps in carbon emissions forecasting by conducting one of the few comprehensive, data-driven comparisons of statistical, machine learning, deep learning, and hybrid models using national-level annual data from 1982 to 2022 in the Australian context. It prioritizes not only predictive accuracy but also model interpretability by integrating explainable AI techniques, particularly SHAP analysis, into random forest and XGBoost models. This dual focus reveals the relative importance of influential factors such as population, energy supply, and fossil fuel use, and provides actionable insights for environmental policy and planning. The study also highlights a critical limitation: deep learning models underperform on coarse-grained annual datasets, reinforcing the value of ensemble methods like random forest and stacking regressors as more transparent and reliable alternatives for medium-sized datasets.

3. Methodology

This study adopts a structured and rigorous methodology to evaluate and compare a diverse set of forecasting models for national CO_2 emissions in Australia. Drawing on annual data from 1982 to 2022, the research follows a multi-phase process comprising data collection, feature selection, exploratory analysis, data preprocessing, model development, hyperparameter tuning, and performance evaluation. Thirteen models spanning statistical, machine learning, deep learning, and hybrid approaches were implemented, offering a comprehensive basis for comparison. The models were trained and validated using historical data, with a dedicated test set used to assess generalization on unseen records. Emphasis was placed not only on predictive accuracy but also on model interpretability, using SHAP analysis to explain the contribution of individual factors in tree-based models. All experiments were conducted in Python using standard libraries and executed in the Google Colab environment, ensuring reproducibility and transparency.

3.1. Data Overview

This study draws on national-level data for Australia spanning the years 1982–2022. Two reputable and publicly available sources were used: Our World in Data (OWID) for greenhouse gas emissions and related environmental metrics, and the Australian Energy Statistics for detailed energy usage and production indicators. These datasets offer comprehensive, longitudinal coverage of key factors influencing CO_2 emissions across economic, energy, and environmental domains.

A total of 22 factors were selected based on their consistent appearance in prior emissions forecasting studies and their documented relevance to national CO_2 dynamics in Australia and other high emitting countries. These include metrics related to energy consumption, fossil fuel use, economic output, electricity generation, and population dynamics. Table 1 provides a summary of these variables, along with their units and brief descriptions. Together, they offer a multi-dimensional view of the drivers behind Australia's CO_2 emissions and serve as input features for all forecasting models developed in this study.

Table 1
Summary of influencing factors with corresponding units and descriptions.

Influencing Factors	Unit	Description
CO_2 emissions	Mt CO_2	Total carbon dioxide emissions
Total greenhouse gas emissions excluding land use change	Mt CO_2-e	Total GHG emissions excluding emissions from land use, land-use change, and forestry
Gross Domestic Product (GDP)	Billion AUD	Total market value of goods and services produced in Australia
Consumption (Total Energy Consumption)	PJ	Total energy used in the Australian economy across all energy types (oil, gas, coal, renewables)
Electricity supply	PJ	Total electricity supplied
Coal consumption	PJ	Energy from coal consumption
Gas consumption	PJ	Energy from natural gas consumption
Oil consumption	PJ	Energy from oil consumption
Transport energy consumption	PJ	Energy used specifically for transport purposes
Energy growth in Queensland	PJ	Annual growth of energy consumption in Queensland
Energy growth in the rest of Australia	PJ	Annual growth of energy consumption in the rest of Australia (excluding QLD and NT)
Energy growth in Northern Territory	PJ	Annual growth of energy consumption in Northern Territory
Total generation	PJ	Total energy generation from all sources
Residential energy consumption	PJ	Energy consumed by the residential sector
Commercial energy consumption	PJ	Energy consumed by the commercial sector
Renewable energy	PJ	Total renewable energy produced (including electricity and direct uses like firewood and solar hot water)
Renewable energy generation	GWh	Renewable energy used for electricity generation only
Net energy exports	PJ	Energy exports minus imports
Population growth	Million people	Change in population over time
Energy intensity	GJ per million AUD	Energy consumed per million AUD of GDP
Energy productivity	GDP per PJ	Economic output produced per unit of energy input
Land use change CO_2 emissions	Mt CO_2-e	CO_2 emissions from land use, land-use change, and forestry activities
Energy Consumption	TWh	Total electricity consumption in the national electricity market

3.2. Exploratory data analysis

Exploratory data analysis has played a foundational role in examining the dynamics between CO_2 emissions and their influencing factors. A comprehensive analysis was conducted to explore data patterns, distributional properties, and relationships among key factors.

Table 2 presents a detailed statistical summary of Australia's CO_2 emissions from 1982 to 2022, offering valuable insights into historical patterns and variability. The mean emission level over the period was 339.36 million tonnes, while the median was higher at 362.54 million tonnes, indicating a left-skewed distribution likely influenced by lower values in the earlier years. The standard deviation of 69.47 million tonnes reflects notable fluctuations across the decades, particularly in recent years, potentially shaped by economic transitions and environmental policy changes. A skewness value of −0.59 suggests a prolonged period of increasing emissions followed by more recent stabilization or decline. Furthermore, the negative kurtosis (−1.12) indicates a relatively flat distribution, with few extreme values or abrupt changes. Emissions ranged from a minimum of 207.65 to a maximum of 415.77 million tonnes, representing a significant rise over the 40-year period. The standard error of 10.85 million tonnes reinforces the reliability of

Table 2
Descriptive statistics for CO_2 emissions from 1982 to 2022.

Statistic	Value
Mean	339.362
Median	362.537
Standard Deviation	69.474
Standard Error	10.850
Kurtosis	−1.121
Skewness	−0.586
Minimum	207.645
Maximum	415.770

the mean as a central estimate. This statistical profile, as shown in Table 2, provides a robust foundation for interpreting long-term emission trends and their potential drivers in the Australian context.

Time-series visualizations in Fig. 1 reveal persistent upward trends in emissions, energy use, and GDP, with notable downturns corresponding to global events such as the 2008 financial crisis and the COVID-19 pandemic. Fig. 2, which displays distribution plots, highlights skewness in several influencing factors, particularly fossil fuel consumption and industrial energy use, underscoring the importance of normalization techniques. Fig. 3 presents pairwise scatter plots that illustrate both linear and nonlinear relationships between CO_2 emissions and influencing factors such as energy intensity and transport emissions. These findings support the application of both linear and nonlinear modelling approaches in subsequent analyses. This combination of statistical and visual exploration confirms that economic activity and energy-related factors are strongly associated with CO_2 emission levels, a relationship well-documented in the environmental forecasting literature.

3.3. Data preprocessing

Preprocessing was conducted in Python using Google Colab. Several key steps were taken to ensure data quality.

- **Handling missing values**: Backward fill was applied, especially effective for temporally ordered data.
- **Renaming columns**: Simplified factor names were adopted for clarity.
- **Log transformation**: Applied to factors such as CO_2 emissions, GDP, and population to reduce skewness and stabilize variance. As shown in Fig. 4, this transformation normalized factor distributions, which helps enhance model performance and training stability.

3.4. Model development and evaluation

To build a reliable predictive framework capable of accurately modelling CO_2 emissions in Australia, this study implemented and evaluated a suite of thirteen models spanning machine learning, hybrid, statistical, and neural network approaches. The selected models represent a diverse array of architectures from tree-based ensembles to biologically inspired neural networks and classical time series baselines allowing for a comprehensive assessment of the forecasting models.

Model development and evaluation were based on historical CO_2 emissions data from 1982 to 2022. To ensure robust validation and minimize the risk of data leakage, the dataset was split into two parts: 80 % for training and 20 % for validation using data from 1982 to 2015. Each model was run 30 times with different random seeds, and the results were averaged to provide a stable and reproducible performance evaluation. A diverse set of machine learning, statistical, deep learning and hybrid models was explored during this phase, and their performance was assessed using established evaluation metrics. Following this rigorous evaluation, the best-performing model was selected to forecast CO_2 emissions for the short-term period from 2016 to 2022 data that was deliberately set aside to serve as unseen input to test the model's forecasting capability. This evaluation phase helped identify the most effective model and revealed areas for improving current approaches to reduce forecasting discrepancies. The two-phase design also ensured a fair assessment and offered a realistic measure of the model's ability to generalize to future CO_2 emission patterns [4]. All experiments were implemented in Python within the Google Colab environment, making use of libraries such as scikit-learn, XGBoost, Keras, TensorFlow, and statsmodels. The following models were implemented, trained, and evaluated.

1. Random forest regressor (RF)
2. XGBoost regressor
3. Stacking regressor combining the best RF and best XGBoost with linear regression as final estimator
4. Enhanced stacking regressor combining RF, XGBoost, and support vector regressor (SVR)
5. SVR
6. ARIMA
7. Extreme learning machine (ELM)
8. ISSA-ELM, simulated version using a different hidden layer configuration
9. Backpropagation neural network (MLP)
10. Gaussian process regression (GPR)
11. Long short-term memory (LSTM)
12. Recurrent neural network (RNN)
13. Gated recurrent unit (GRU)

All models were tuned to optimize performance and ensure generalizability. For tree-based models (RF and XGBoost), hyperparameters such as the number of estimators, maximum tree depth, and learning rate were explored using grid search. SVR was optimized for kernel type, regularization parameter (C), and epsilon parameters. Stacking models combined optimized base learners (RF, XGBoost, SVR) with a linear regression meta-estimator. MLP were tuned for hidden layer sizes and maximum iterations, with fixed learning rate and activation functions. Sequence models, including LSTM, GRU, and RNN, were trained using early stopping and learning rate reduction to prevent overfitting, with fixed layer sizes, batch sizes, and epochs. This systematic approach ensured robust and high-performing models across all algorithms.

3.5. Model hyperparameters and structural Characteristics

To ensure transparency and reproducibility, the hyperparameters and structural configurations of all thirteen models are summarized in Table 3. The settings were chosen to balance model expressiveness with the limitations imposed by the modest sample size of annual national data. For ensemble methods such as RF and XGBoost, the number of estimators and maximum tree depth were selected to avoid overfitting while preserving predictive power. Stacking models combined these optimized base learners with a linear regression meta-estimator. Neural models, including MLP, LSTM, GRU, and RNN, were restricted to relatively compact architectures with fixed layer sizes and controlled epochs, leveraging early stopping and learning rate adjustments where applicable. These architectures were deliberately kept compact to avoid overfitting given the modest annual dataset size. Layer sizes were fixed based on best practices for small datasets, and training employed early stopping rather than extensive tuning to control epochs. The hyperparameter choices ensure fair and reproducible comparisons across model classes.

This table consolidates the hyperparameters used in the repeated 30-run analysis. The settings reflect a balance between model expressiveness and stability, given the modest sample size of annual national emissions data.

3.6. Evaluation metrics

All the developed models were evaluated using a set of widely adopted performance metrics to ensure a comprehensive assessment of forecasting accuracy, namely mean squared error (MSE), root mean squared error (RMSE), R-squared (R^2), mean absolute error (MAE), mean absolute percentage error (MAPE), mean squared log error (MSLE), and median absolute error (MedAE). These metrics are commonly used in the literature and provide complementary perspectives on model performance [14,21,23].

3.7. Assessment of model uncertainty via repeated trials

To account for the inherent randomness in some of the models, each

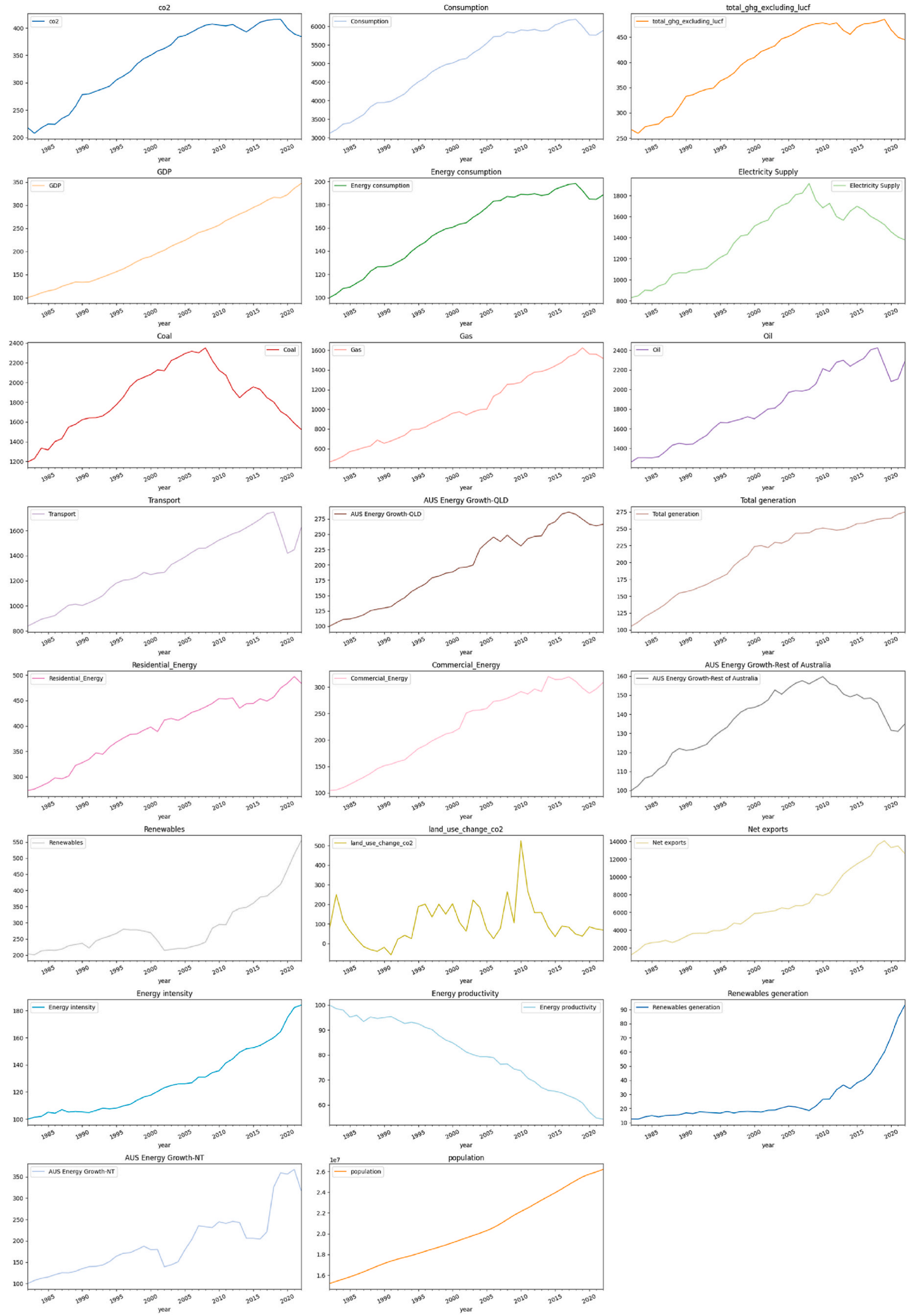

Fig. 1. Time series of 22 economic, demographic, and energy-related variables influencing Australia's CO_2 emissions (1982–2022). Panels show trends across GDP, population, CO_2 emissions, energy consumption, electricity supply, transport, fossil fuels, renewables, and regional energy growth, providing a multidimensional view of Australia's energy-emissions dynamics for the forecasting models.

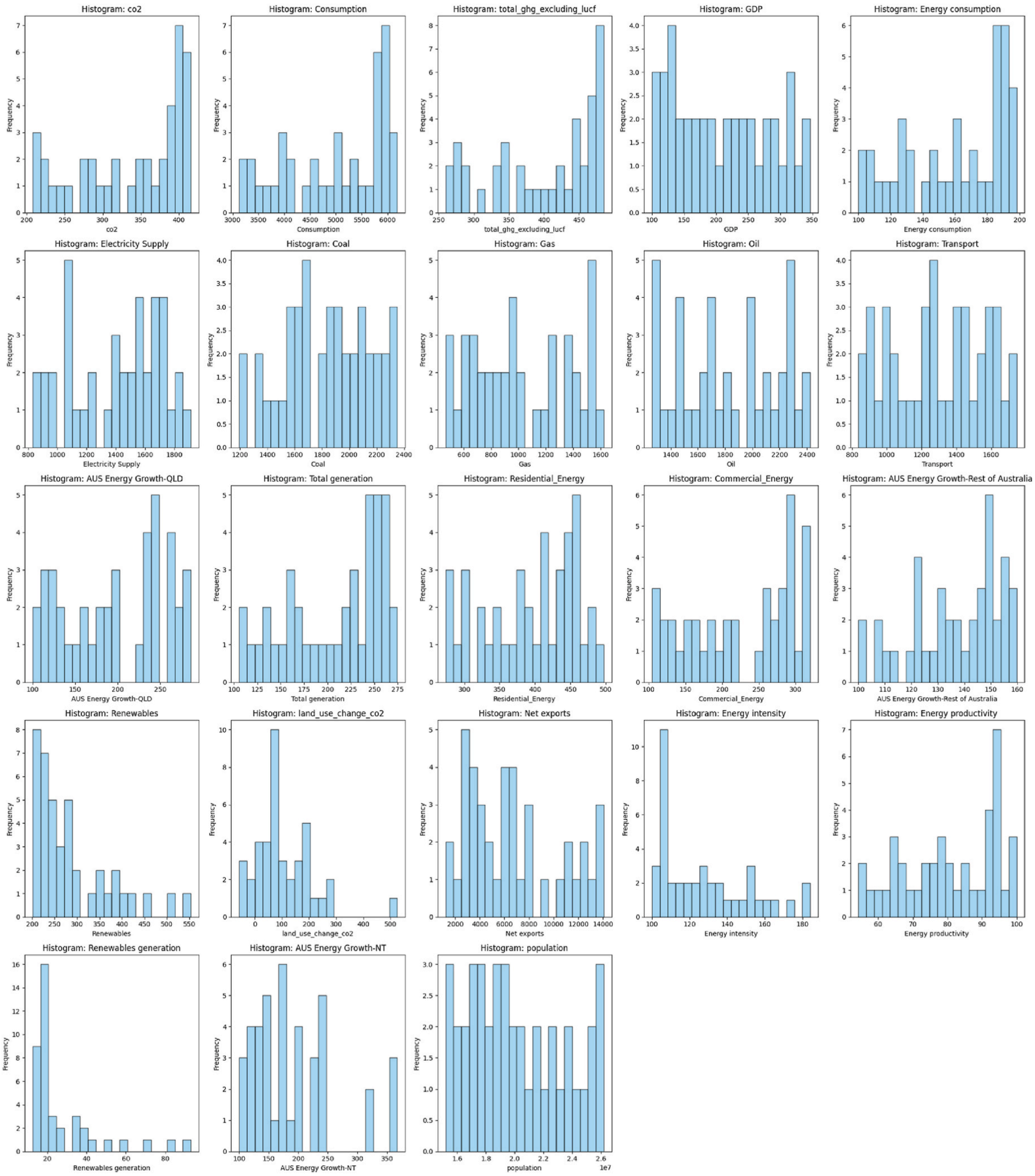

Fig. 2. Distributional profiles of 22 economic, demographic, and energy-related variables used in the forecasting framework. Each histogram illustrates the frequency and spread of values across indicators such as GDP, population, CO_2 emissions, energy consumption, electricity supply, transport, fossil fuels, renewables, and regional energy growth. These visualizations support exploratory analysis and provide insight into feature behaviour prior to model development.

algorithm was trained and evaluated 30 times using different random seeds. Reporting the mean and 95 % confidence interval for each metric across these runs helps reduce the influence of chance and provides a more reliable estimate of model performance.

All repeated trials were conducted within the chronological window 1982–2015. Random seeds affected only model initialization and resampling, so the temporal order of the data was preserved while still capturing stochastic variability. This approach ensures that the results reflect consistent performance patterns rather than isolated outcomes, strengthening the reliability and interpretability of the comparisons across models.

3.8. Statistical significance testing

To assess whether observed differences in predictive performance are statistically meaningful, pairwise comparisons were conducted using two complementary approaches.

- **Paired *t*-test**: A parametric test suitable when the differences between paired observations are approximately normally distributed.
- **Wilcoxon signed-rank test**: A non-parametric alternative that does not assume normality of differences.

Both tests were applied to the RMSE values obtained from the 30 independent runs. This dual approach ensures that the conclusions are

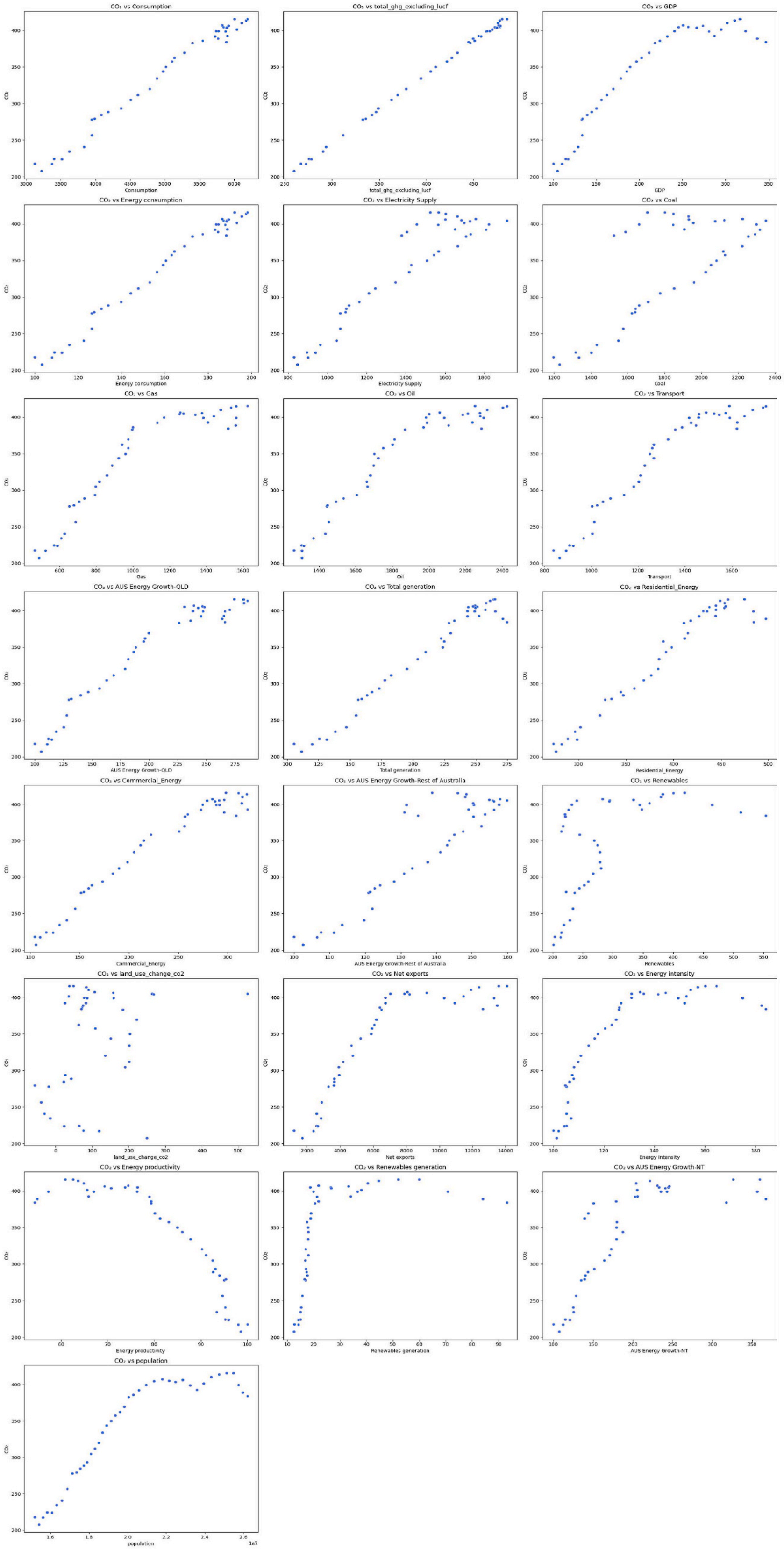

Fig. 3. Scatter plots illustrating the relationships between CO_2 emissions and selected economic, demographic, and energy-related variables used in the forecasting models. Each panel shows the association between emissions and a key feature, with fitted regression lines highlighting direction and strength of correlation, supporting feature relevance assessment and interpretability.

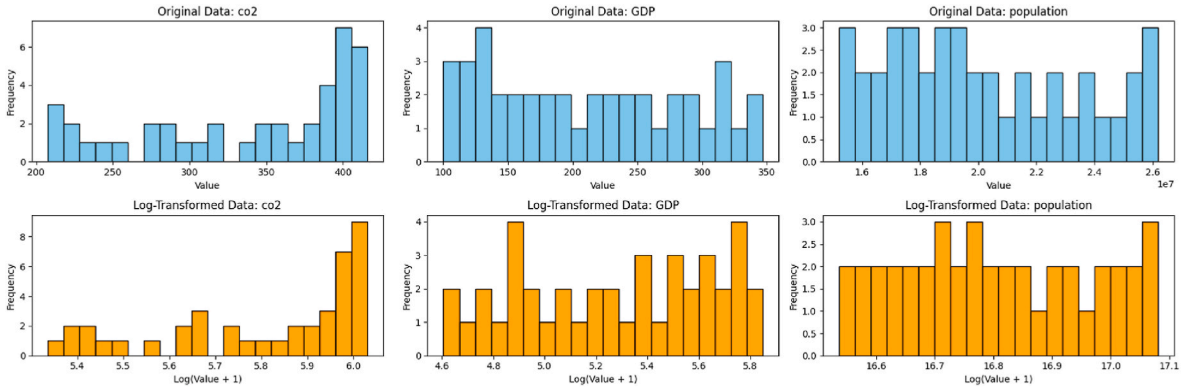

Fig. 4. Histograms of selected factors before and after log transformation, illustrating the effect on distribution normalization.

Table 3
Machine learning model architecture and key hyperparameter Configuration.

Model Category	Model Name	Core Architecture	Key Hyperparameters	Parameter Description	Regularization Mechanism
Ensemble/ Tree	Random Forest	Bagging + Random Feature Subsets	n_estimators = 300, max_depth = 10, min_samples_split = 2, max_features = ’sqrt’	Number of trees, tree depth, min samples to split, feature sampling	Ensemble Averaging
Ensemble/ Tree	XGBoost	Second-Order Gradient Boosting	n_estimators = 100, max_depth = 3, learning_rate = 0.05, subsample = 1.0	Learning rate, tree depth, number of boosting rounds, subsample fraction	L2 Regularization + Early Stopping
Ensemble	Stacking	RF + XGB → Linear Regression	"Stacking [RF + XGB] with LinearRegression final estimator"	Base learners + meta learner	Ensemble Averaging
Ensemble	Enhanced Stacking	RF + XGB + SVR → Linear Regression	"Stacking [RF + XGB + SVR] with LinearRegression final estimator"	Base learners + meta learner	Ensemble Averaging
Kernel/ Linear	SVR	Support Vector Regression (RBF)	C = 10, epsilon = 0.01, kernel = ’rbf’	Regularization strength, tube size, kernel type	L2 Regularization
Time Series	ARIMA	Autoregressive Integrated MA	order=(5,1,0)	AR, differencing, MA terms	None
Neural Network	MLP	Multi-Layer Perceptron	hidden_layer_sizes=(50,50), max_iter = 500	Number of hidden neurons per layer, max training iterations	Weight decay (implicitly via solver)
Kernel/ Bayesian	GPR	Gaussian Process Regression	kernel = 3.41**2 * RBF(length_scale = 9.12), n_restarts_optimizer = 5	Kernel function, optimizer restarts	None
Neural Network	ELM	Extreme Learning Machine	n_hidden = 100, activation = ’tanh’	Number of hidden neurons, activation function	None
Neural Network	ISSA-ELM	Extreme Learning Machine + ISSA	n_hidden = 200, activation = ’tanh’	Number of hidden neurons, activation function	None
Neural Network	LSTM	Long Short-Term Memory Network	layers = [50,50], optimizer = Adam, loss = MSE, early_stopping = True	Layer sizes, optimizer, loss function, early stopping	Implicit via Early Stopping & Adam
Neural Network	RNN	Simple RNN	layers = [50,50], optimizer = Adam, loss = MSE, early_stopping = True	Layer sizes, optimizer, loss function, early stopping	Implicit via Early Stopping & Adam
Neural Network	GRU	Gated Recurrent Unit	layers = [50,50], optimizer = Adam, loss = MSE, early_stopping = True	Layer sizes, optimizer, loss function, early stopping	Implicit via Early Stopping & Adam

robust regardless of distributional assumptions. The detailed outcomes of these tests are reported in subsection 4.1.5, where statistically significant and non-significant differences between model performances are highlighted, providing a rigorous assessment of relative predictive accuracy.

4. Comparative analysis and critical discussion

This section presents the results from all the implemented and evaluated forecasting models using both quantitative performance metrics and qualitative analyses such as actual vs. predicted

Table 4
Mean ± 95 % confidence interval of key performance metrics for all forecasting models across 30 independent runs.

Model	MSE	RMSE	R^2	MAE	MAPE	MedAE	MSLE
RF	0.0004 ± 0.0001	0.0182 ± 0.0023	0.9890 ± 0.0032	0.0142 ± 0.0017	0.2507 ± 0.0308	0.0118 ± 0.0018	0.0000 ± 0.0000
XGBoost	0.0011 ± 0.0003	0.0315 ± 0.0037	0.9613 ± 0.0149	0.0253 ± 0.0026	0.4425 ± 0.0468	0.0187 ± 0.0024	0.0000 ± 0.0000
Stacking	0.0003 ± 0.0001	0.0159 ± 0.0021	0.9913 ± 0.0026	0.0123 ± 0.0015	0.2160 ± 0.0281	0.0095 ± 0.0014	0.0000 ± 0.0000
Enhanced Stacking	0.0003 ± 0.0001	0.0166 ± 0.0023	0.9910 ± 0.0023	0.0130 ± 0.0018	0.2303 ± 0.0324	0.0100 ± 0.0016	0.0000 ± 0.0000
SVR	0.0012 ± 0.0005	0.0289 ± 0.0066	0.9748 ± 0.0078	0.0198 ± 0.0037	0.3542 ± 0.0694	0.0133 ± 0.0026	0.0000 ± 0.0000
ARIMA	0.0639 ± 0.0089	0.2477 ± 0.0184	−1.1670 ± 0.6909	0.2189 ± 0.0170	4.7259 ± 0.5190	0.2138 ± 0.0234	0.0014 ± 0.0002
ELM	5.4131 ± 4.9047	1.3434 ± 0.6914	−111.9307 ± 89.0143	0.7344 ± 0.3558	12.9326 ± 6.1750	0.2651 ± 0.1194	0.0329 ± 0.0218
ISSA-ELM	1.4215 ± 2.2025	0.5515 ± 0.3847	−22.1546 ± 30.2725	0.2883 ± 0.1661	5.1782 ± 3.0627	0.1149 ± 0.0462	0.0218 ± 0.0212
MLP	0.3803 ± 0.2285	0.4243 ± 0.1629	−11.0970 ± 6.8079	0.4110 ± 0.1648	7.1203 ± 2.8487	0.4091 ± 0.1662	0.0071 ± 0.0041
GPR	0.0003 ± 0.0002	0.0146 ± 0.0031	0.9907 ± 0.0051	0.0117 ± 0.0024	0.2057 ± 0.0427	0.0091 ± 0.0016	0.0000 ± 0.0000
LSTM	0.0718 ± 0.0168	0.2558 ± 0.0290	−2.0081 ± 1.2453	0.2313 ± 0.0333	4.0580 ± 0.5699	0.2163 ± 0.0431	0.0015 ± 0.0003
RNN	0.4444 ± 0.7415	0.3499 ± 0.2065	−6.5571 ± 10.5953	0.2468 ± 0.0848	4.3979 ± 1.5728	0.1839 ± 0.0329	0.0015 ± 0.0003
GRU	0.5663 ± 0.1134	0.7219 ± 0.0774	−20.9966 ± 10.7309	0.6615 ± 0.0734	11.3746 ± 1.2422	0.6972 ± 0.0788	0.0116 ± 0.0024

comparisons, feature importance interpretation, and learning curve diagnostics.

4.1. Model evaluation and multi-run performance comparison

The predictive performance of all models was evaluated over 30 independent runs to account for stochastic variability. Each run involved training and testing on randomly sampled train–test splits. Key metrics MSE, RMSE, R^2, MAE, MAPE, MedAE, and MSLE were recorded. Table 4 summarizes the mean values along with 95 % confidence intervals, providing a clear assessment of both accuracy and stability.

Tree-based ensembles (random forest [RF], stacking, enhanced stacking) and Gaussian process regression (GPR) consistently achieved the lowest errors and highest R^2 values, with narrow confidence intervals across runs. Deep learning architectures (LSTM, GRU, RNN, and MLP) exhibited higher variability and occasional extreme predictions, reflecting their sensitivity to stochastic effects and the constraints of annual data. Single-layer biologically inspired models, such as ELM, showed moderate variability, while ISSA-ELM demonstrated improved stability.

These results indicate that ensemble and kernel-based methods provide robust predictions under repeated trials, whereas neural and single-layer models require careful multi-run evaluation to achieve reliable performance estimates. Cross-validation metrics in Table 4 were computed on log-transformed CO_2 emissions to stabilize variance, so all reported values are therefore in log units.

4.1.1. Ensemble and hybrid models

RF demonstrated the highest accuracy, achieving the lowest RMSE and MSE values. Stacking, which combines RF and XGBoost with a linear regression meta-learner, offered comparable accuracy and slightly improved stability. Enhanced stacking, which incorporates SVR, maintained strong performance, though confidence intervals were marginally wider. Ensemble models proved most reliable and reproducible across 30 runs.

4.1.2. Individual learning models

XGBoost performed well but slightly below the ensemble models. SVR showed moderate accuracy with higher variability. ARIMA underperformed, confirming its limitations in multivariate emission forecasting.

4.1.3. Neural and biologically inspired models

ELM exhibited moderate accuracy with notable variability. ISSA-ELM improved stability, while GPR consistently achieved reliable performance with narrow confidence intervals.

4.1.4. Deep learning architectures

LSTM, RNN, GRU, and MLP underperformed consistently, reflecting limitations of annual data for capturing temporal dependencies. Their predictions were less reliable despite the theoretical capability for complex sequence modelling.

4.1.5. Statistical significance of model performance

Pairwise statistical tests (paired *t*-test and Wilcoxon signed-rank) on RMSE values across 30 runs were conducted to assess the robustness of observed performance differences. Table 5 summarizes selected comparisons among top models.

Key findings.

- RF significantly outperforms most individual learning and deep learning models ($p < 0.01$).
- Differences between RF and stacking or enhanced stacking are smaller; only RF vs Stacking reached statistical significance.
- XGBoost, while strong, was generally surpassed by ensemble models.
- GPR showed robust performance with occasional significant differences relative to RF and ensembles.

Table 5
Pairwise Statistical Significance of RMSE Across Top-Performing Models (30 Independent Runs) Using Paired *t*-Test and Wilcoxon Signed-Rank Test.

Model Comparison	*t*-test RMSE	p-value	Wilcoxon	p-value	Significance
RF vs XGBoost	−7.132	0.0000	4.000	0.0000	Significant
RF vs Stacking	2.775	0.0096	115.000	0.0145	Significant
RF vs Enhanced Stacking	1.942	0.0620	147.000	0.0803	Not significant
RF vs GPR	2.097	0.0449	112.000	0.0120	Significant
RF vs LSTM	−15.733	0.0000	0.000	0.0000	Significant
RF vs GRU	−17.691	0.0000	0.000	0.0000	Significant
XGBoost vs Stacking	7.990	0.0000	0.000	0.0000	Significant
XGBoost vs Enhanced Stacking	7.473	0.0000	1.000	0.0000	Significant
GPR vs LSTM	−15.712	0.0000	0.000	0.0000	Significant
GPR vs GRU	−18.081	0.0000	0.000	0.0000	Significant

These results statistically confirm the conclusions drawn from the multi-run performance metrics, reinforcing that tree-based ensembles particularly RF and hybrid stacking approaches are the most reliable forecasting models.

Overall, ensemble and hybrid models, particularly RF and stacking variants, demonstrated the highest predictive accuracy and stability across repeated runs. Kernel-based methods such as GPR also provided reliable performance. Neural and single-layer biologically inspired models exhibited greater variability, while deep learning architectures (LSTM, RNN, GRU, and MLP) consistently underperformed, reflecting the constraints of annual data frequency for capturing temporal dependencies.

4.2. Computational resources, runtime, and memory profiling

All experiments were conducted using Google Colab's free-tier environment, which provides a cloud-based virtual machine with approximately 12 GB of RAM and a lightweight multi-core CPU (typically two vCPUs). GPU acceleration was employed only for deep learning models, while classical and tree-based approaches were executed on the CPU. As Colab's resource allocation is dynamic and may vary across sessions, the reported runtime and memory results should be interpreted as representative of the environment during our runs rather than fixed hardware specifications. The computational efficiency of all evaluated models was assessed in terms of runtime and peak memory usage over 30 repeated runs. Table 6 summarizes the mean ± 95 %

Table 6
Runtime and peak memory usage of all models (mean ± 95 % CI). Memory values < 0.001 MB are reported as measured, reflecting precise consumption even when negligible for practical deployment.

Model	Runtime (s)	Memory (MB)
RF	0.35 ± 0.03	0.04 ± 0.02
XGBoost	0.11 ± 0.04	0.53 ± 0.33
Stacking	2.65 ± 0.24	0.40 ± 0.17
Enhanced Stacking	2.47 ± 0.17	0.03 ± 0.03
SVR	0.006 ± 0.001	0.00 ± 0.00
ARIMA	0.33 ± 0.11	0.00026 ± 0.00051
ELM	0.0018 ± 0.0003	0.021 ± 0.014
ISSA-ELM	0.0025 ± 0.0003	0.023 ± 0.017
MLP	0.039 ± 0.008	0.020 ± 0.016
GPR	0.086 ± 0.013	0.015 ± 0.024
LSTM	6.51 ± 0.18	32.64 ± 8.27
RNN	5.55 ± 0.31	12.39 ± 11.66
GRU	7.63 ± 0.34	32.90 ± 19.16

confidence interval (CI) for these metrics.

Simpler models such as RF, XGBoost, SVR, and ELM-based methods (ELM and ISSA-ELM) demonstrated low runtime (0.002–0.35 s) and minimal memory consumption (0–0.53 MB). These models are computationally lightweight and can be efficiently deployed for rapid predictions on standard hardware.

The ARIMA model, despite being a classical time series approach, exhibited a moderate runtime of 0.33 ± 0.11 s with a negligible memory footprint ~0 M. Although computationally inexpensive, ARIMA struggled to capture the complex dynamics of the dataset, highlighting its limitations for CO_2 emissions forecasting.

Stacking and enhanced stacking models required longer runtimes (2.47–2.65 s) with moderate memory usage (0.03–0.40 MB), reflecting the additional overhead of combining multiple base learners. Similarly, feedforward and recurrent neural networks (MLP, LSTM, RNN, GRU) incurred substantially higher computational costs. LSTM and GRU networks required over 6–7 s per run with peak memory exceeding 32 MB, while RNNs used slightly less memory (~12 MB). This highlights the trade-off between the flexibility of deep learning models and their computational demands.

As shown in Table 6, simpler models such as RF, XGBoost, SVR, and ELM-based approaches exhibit extremely low runtime and minimal memory usage, with some values approaching or below 0.001 MB. Although negligible for practical deployment, these values are reported precisely to reflect the measured performance. In contrast, deep learning models, particularly LSTM and GRU, incur substantially higher computational and memory demands, highlighting the trade-off between predictive flexibility and resource requirements.

Overall, the profiling reveals a clear trend that simpler models and classical approaches are efficient in terms of runtime and memory but may underperform in capturing complex dynamics, whereas advanced neural architectures offer enhanced predictive capability with higher computational demands. Ensemble methods strike a favourable balance between predictive accuracy and computational cost, making them practical for integration into policy-oriented forecasting systems where transparency and scalability are essential. These findings provide guidance for researchers seeking to balance accuracy and deployment efficiency when selecting appropriate forecasting models.

4.3. Model behaviour and interpretability

In addition to quantitative metrics, we evaluated each model's ability to replicate observed CO_2 emission patterns through visual diagnostics. To assess robustness and consistency, predictions were

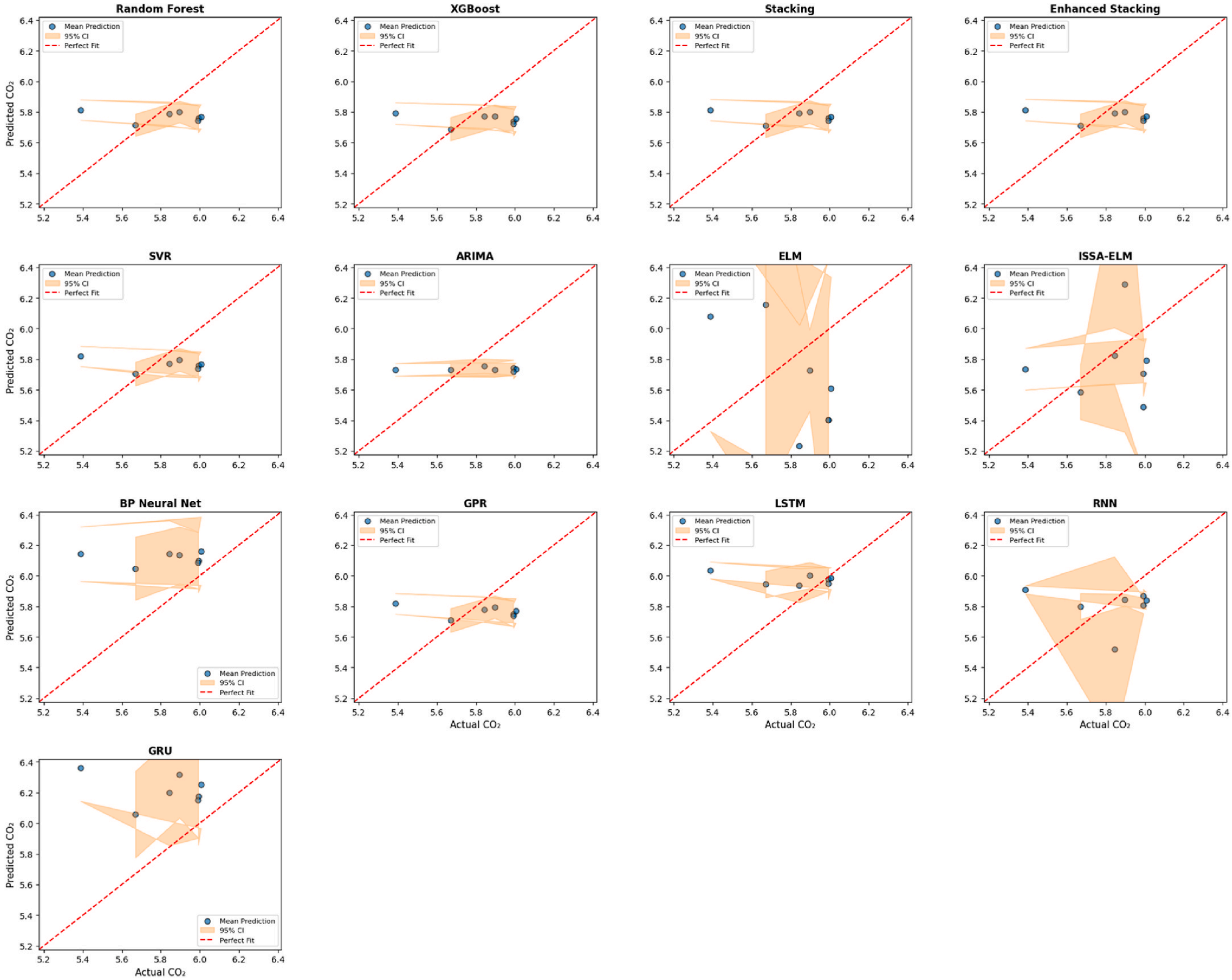

Fig. 5. Scatter plots of actual versus predicted CO_2 emissions in Australia (1982–2015) across 30 independent runs for all forecasting models. Orange shading denotes the 95 % confidence interval across seeds.

generated across 30 independent runs using different random seeds, with results aggregated to reflect mean behaviour and variability.

4.3.1. *Actual vs. predicted comparisons*

Fig. 5 presents scatter plots comparing predicted and actual CO_2 emissions for all forecasting models for the period 1982–2015. Each plot displays mean predictions (blue dots) and actual values (black dots), with orange shading denoting the 95 % confidence interval across 30 seeds. The red diagonal line represents perfect prediction alignment.

- RF and stacking exhibit strong agreement with actual emissions. Predictions cluster tightly around the diagonal, and narrow confidence intervals indicate low variability across seeds.
- Enhanced stacking, ELM, and GPR also track observed emissions reliably during stable periods. Enhanced stacking and GPR remain near the diagonal with slightly wider confidence intervals, suggesting moderate sensitivity to data splits, whereas ELM exhibits unstable behaviour and poor fit, consistent with its negative R^2 and large RMSE.
- XGBoost performs well but is generally surpassed by ensemble methods; prediction points show more scatter and wider confidence intervals than RF or stacking.
- SVR, ARIMA, and neural networks (MLP, LSTM, GRU, RNN) deviate more substantially from actual values. Neural networks occasionally flatten or introduce artificial volatility, while SVR and ARIMA struggle to capture both trend and fluctuation dynamics. Broader confidence intervals reflect increased variability and reduced robustness under limited data.

These visual diagnostics are consistent with quantitative metrics: RF and stacking achieved the lowest RMSE (≈0.016–0.018) and highest R^2 (≈0.989–0.991), XGBoost and SVR performed moderately well, and ARIMA and ELM showed poor fit (negative R^2). Together, the diagnostics and metrics confirm that tree-based ensembles, particularly RF and stacking, offer the most stable and interpretable performance for annual CO_2 emission forecasting. The use of multi-run predictions enhances confidence in model reliability, supporting their suitability for policy-oriented applications.

4.4. *Learning curve insights*

To complement the multi-run evaluation and the interpretability analysis, learning curves were generated, as shown in Fig. 6, for the RF and XGBoost models to examine how their performance changes with increasing training data. These curves offer an additional perspective on model behaviour by showing not only how well each model fits the data but also how efficiently they learn from the limited annual observations available for Australia's emissions.

Across the 30-seed runs, the RF model showed a clear and steady convergence between its training and validation RMSE values. As the training set expanded, both curves stabilised at low error levels, accompanied by consistently high validation R^2 scores. This pattern indicates that the model captures underlying relationships in a balanced manner, learning relevant nonlinear interactions without overfitting, even when trained on relatively small subsets of data. The narrowing gap between training and validation performance reinforces earlier findings that RF is both accurate and structurally well-suited to annual CO_2 emissions data.

XGBoost, in contrast, displayed a different learning profile. While it ultimately reached competitive performance, the early stages of its learning curve showed a more noticeable separation between training and validation RMSE, alongside greater variability across seeds. This behaviour suggests mild overfitting, reflecting the model's sensitivity to boosting depth and learning rate, particularly when applied to datasets with limited temporal granularity. Although XGBoost remained a strong individual learner, its learning trajectory was less stable than that of RF, consistent with the wider confidence intervals observed in the multi-run evaluation.

Taken together, the learning curves offer a broader view of model behaviour beyond point-estimate performance metrics. They confirm that RF not only achieves higher accuracy overall but also learns more consistently and reliably from small annual datasets. For policy-oriented forecasting where transparency, stability, and reproducibility are essential, this learning behaviour strengthens the case for RF as a dependable foundation for national-level CO_2 modelling. The findings also highlight an important implication for future work: analyses using higher-resolution or more granular data (such as quarterly emissions or sector-level series) may provide the conditions under which boosting-based models can fully realise their potential.

4.5. *Feature importance analysis*

To assess the relative influence of input variables on model predictions, classical feature importances were computed for both RF and XGBoost and averaged across 30 random seeds, reducing the stochastic variability introduced by different initialisations, complementing the learning-curve insights shown in Fig. 6, and enhancing reproducibility. Fig. 7 presents the aggregated feature importances, with error bars representing standard deviations across seeds, quantifying attribution uncertainty.

For RF, importance scores were broadly distributed, reflecting the model's ability to integrate signals from both macroeconomic and

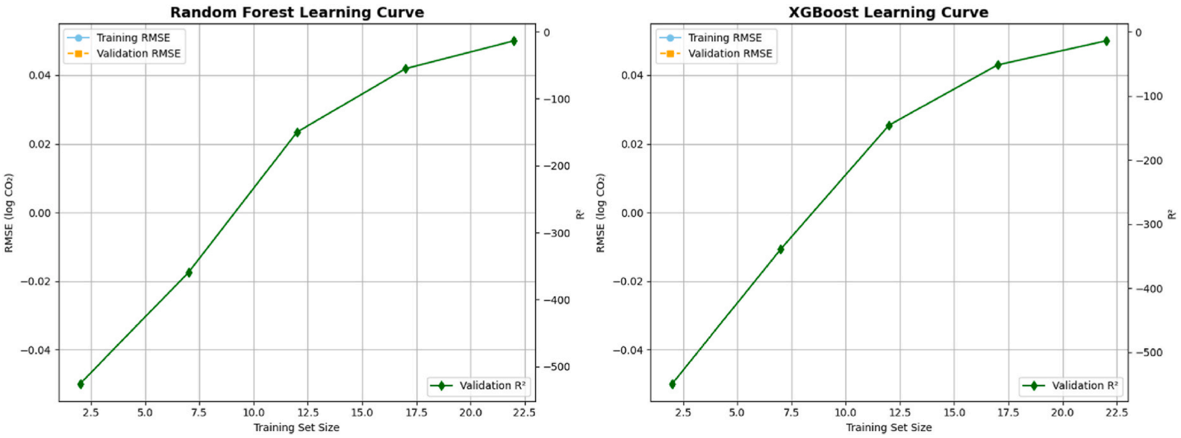

Fig. 6. Learning curves for random forest and XGBoost models, showing training and validation RMSE and validation R^2 across increasing training set sizes, averaged over 30 seeds.

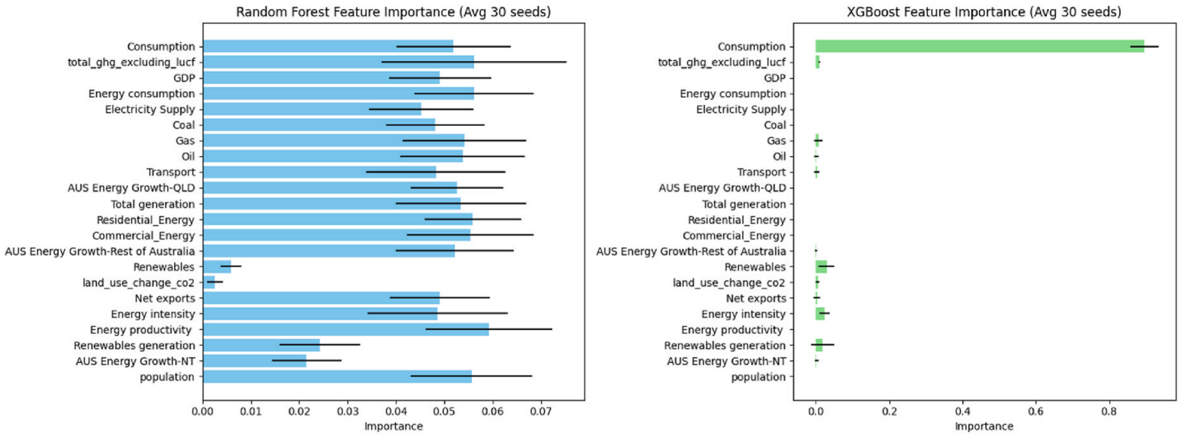

Fig. 7. Aggregated feature importances for RF and XGBoost across 30 seeds. Error bars show standard deviations, indicating attribution uncertainty. RF distributes importance across multiple drivers, while XGBoost concentrates importance on a narrower subset.

sector-level energy variables. Energy productivity (0.0593 ± 0.0131), total_ghg_excluding_LUCF (0.0562 ± 0.0191), energy consumption (0.0562 ± 0.0124), population (0.0557 ± 0.0126), and residential energy (0.0559 ± 0.0099) were the most influential features. Gas consumption (0.0542 ± 0.0128) and total electricity generation (0.0535 ± 0.0135) also contributed meaningfully. The moderate SDs across seeds indicate consistent RF attribution, aligning with the stable learning behaviour described in Section 4.3.

In contrast, XGBoost exhibited a selective profile: total_ghg_excluding_LUCF (0.00996 ± 0.00240) and Gas (0.00680 ± 0.01128) were the most influential features, while several others had standard deviations larger than their means, reflecting variability in attribution across seeds. This sparsity reflects XGBoost's strong regularization under coarse-grained annual data and mirrors the more variable learning trajectory observed in Section 4.3. Although XGBoost remained competitive in predictive accuracy, its attribution behaviour indicates a narrower emissions-driver representation compared with RF.

4.6. SHAP analysis of feature contributions

To complement classical importance metrics, SHAP values were computed for each model and averaged across seeds to capture the magnitude and direction of feature effects. SHAP enables a transparent interpretation by quantifying each variable's contribution to individual predictions across the dataset.

Fig. 8 presents the mean absolute SHAP values. For the RF model, the SHAP profile closely mirrors the classical importance results. Energy productivity (0.0125), total greenhouse gas emissions excluding LUCF (0.0115), population (0.0112), commercial energy consumption (0.0111), and residential energy consumption (0.0110) emerged as key contributors. Several other energy-related variables also contributed meaningfully, reflecting RF's ability to integrate signals across both macroeconomic and sectoral drivers. Error bars (standard deviations across 30 seeds) were relatively narrow, underscoring attribution consistency and providing a direct measure of uncertainty. This stability reinforces the reproducibility of RF's interpretability under coarse-grained annual data, consistent with the stable learning behaviour observed in Section 4.3.

For the XGBoost model, SHAP values reveal a sparse attribution profile. Total_ghg_excluding_LUCF (0.00573) and Gas (0.00026) were the main contributors, while most other features were effectively zero, with several exhibiting standard deviations larger than their mean SHAP values, reflecting variability across seeds. Despite this, the overall sparsity pattern remained consistent when averaged across 30 seeds, reinforcing the reproducibility of XGBoost's selective attribution behaviour. This concentration of importance on a narrow set of drivers reflects XGBoost's regularization under coarse-grained annual data and aligns with the more variable learning trajectory observed in Section 4.3.

Fig. 9 shows the SHAP summary for the first-seed RF model. Although feature rankings vary slightly across seeds, the top drivers closely align with the 30-seed average shown in Fig. 8, indicating that the first seed provides a representative view of the model's feature attributions. AUS Energy Growth-QLD is top ranked, followed by Energy

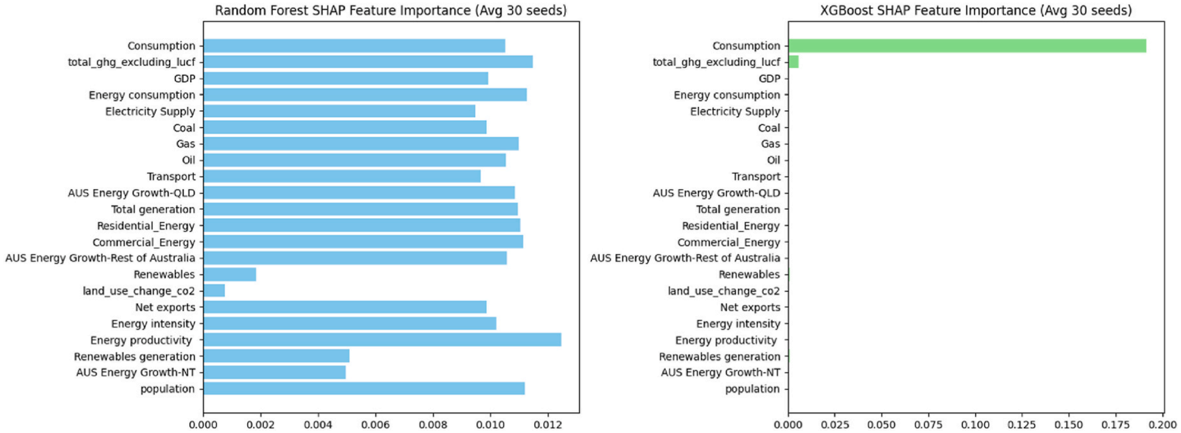

Fig. 8. Mean absolute SHAP values for RF and XGBoost across 30 seeds. Error bars show standard deviations, reflecting attribution uncertainty. RF exhibits distributed and stable attributions, while XGBoost relies on a smaller set of dominant drivers.

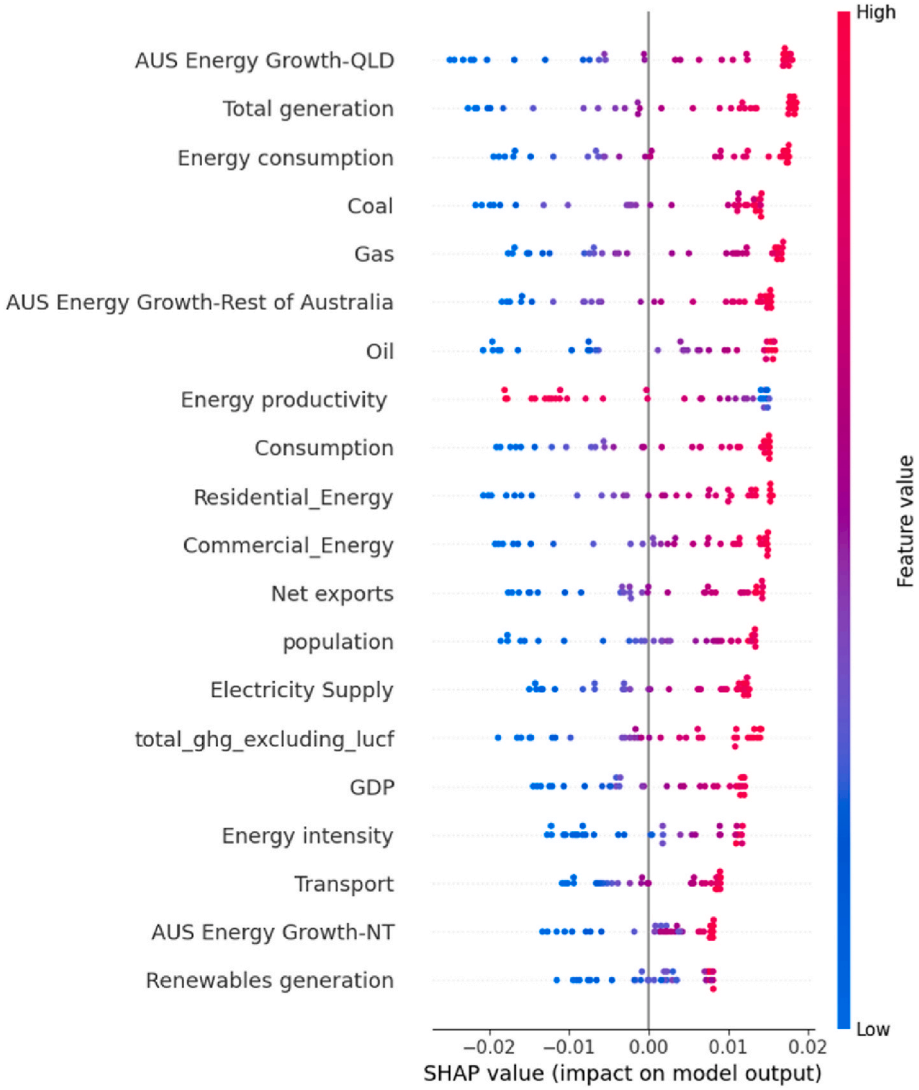

Fig. 9. RF SHAP summary dot plot for a single seed. Dot colour indicates feature value (blue: low, red: high); x-axis reflects SHAP impact. Multiple features show directional influence on predictions, highlighting the integrated effect of macroeconomic and energy-sector drivers.

productivity, Commercial_Energy, Residential_Energy, Consumption, and total_ghg_excluding_LUCF, all with broad SHAP ranges indicating distributed contributions across instances. Higher feature values (red dots) generally correspond to positive SHAP impacts, particularly for Energy productivity and Residential_Energy. Although individual seed rankings vary slightly, these patterns are consistent across the multi-seed framework and align with the averaged SHAP values in Fig. 8, reinforcing RF's reproducible attributions under coarse-grained annual data. These results illustrate the integrated influence of multiple economic and energy-sector variables on national emission trajectories.

For completeness, the XGBoost SHAP summary was also examined, as shown in Fig. 10. Consumption and total_ghg_excluding_LUCF dominate the predictions, showing high SHAP magnitudes and wide dispersion, while most other features cluster near zero regardless of value. These patterns confirm the selective attribution observed in both the averaged SHAP values and classical feature importance metrics, highlighting that XGBoost relies on a narrow set of drivers compared with RF's more integrative structure.

Overall, these SHAP analyses reinforce the attribution patterns observed in the classical feature importance metrics and learning curves, emphasizing RF's integrative representation of multiple emission drivers and XGBoost's reliance on a smaller subset of dominant predictors. These interpretability results provide the foundation for the scenario analysis in Section 6, where the practical implications of these drivers are explored under alternative trajectories.

5. Forecasting CO_2 emissions on unseen data using a multi-seed ensemble approach

To evaluate generalization capacity, forecasting experiments were conducted on unseen test data spanning 2016–2022 using a 30-seed ensemble framework. Models were trained exclusively on 1982–2015 data, and predictions were inverse-transformed to Mt CO_2 prior to metric calculation. This design ensures that performance results reflect predictive ability beyond the training samples, providing a robust assessment of model reliability. The RF ensemble achieved high predictive accuracy (MSE = 5.89, RMSE = 2.43 Mt CO_2, $R^2 = 0.96$), with modest errors (MAE = 1.81, MedAE = 1.16, MAPE <0.5 %) and a near-zero MSLE (0.000037), confirming stability across emission scales. As shown in Fig. 11, scatter points represent mean predictions, the shaded band indicates ±1 standard deviation across seeds, and the dashed line denotes the 1:1 reference. These results indicate that the ensemble captured meaningful and reproducible relationships in the data rather than overfitting, making the model both statistically sound and practically valuable. Beyond statistical performance, the ensemble's stability supports applications in policy analysis, national emissions forecasting,

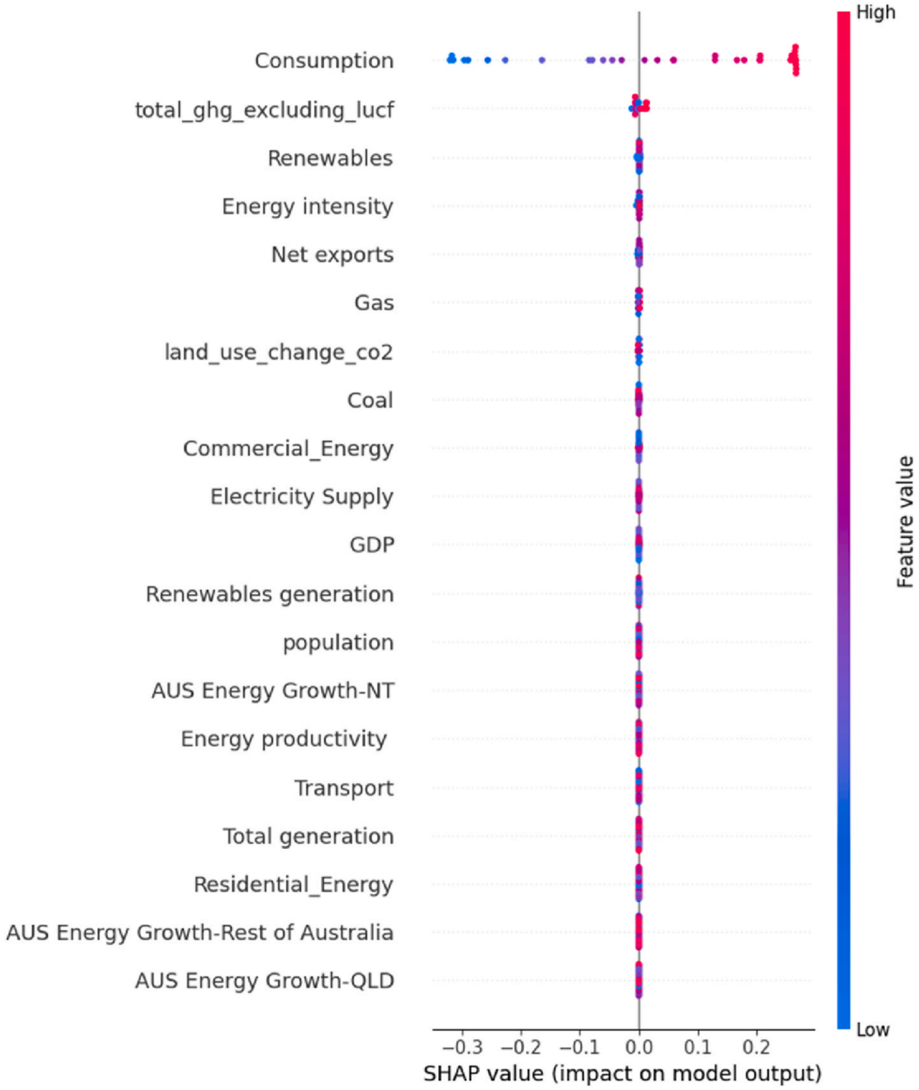

Fig. 10. SHAP summary plot for XGBoost using the first seed. Consumption and total GHG excluding LUCF dominate the model's predictions, while most other features exert minimal influence.

and integration into decision-making frameworks.

6. Scenario analysis and policy implications

6.1. Scenario analysis

To examine how shifts in key drivers may influence Australia's 2050 emissions outcomes, a set of illustrative decarbonisation scenarios was constructed using RF feature importance and SHAP analysis. Deterministic adjustments were applied to the most influential variables: energy productivity, total energy consumption, and population. These scenarios are illustrative rather than predictive; values are deterministic and intended to highlight the potential impacts of driver adjustments on long-term emissions outcomes**.**

As shown in Fig. 12, the baseline scenario reaches 410 Mt CO_2 by 2050. Incremental adjustments deliver consistently modest reductions: a 5 % improvement in energy productivity lowers emissions to 406.7 Mt (−0.80 %), while a 10 % improvement achieves 403.4 Mt (−1.61 %). Demand-side efficiency produces similar results, with a 5 % reduction in consumption yielding 408.0 Mt (−0.49 %) and a 10 % reduction 406.0 Mt (−0.98 %). Population sensitivity tests confirm exogenous effects: ±5 % shifts translate to ±0.5 % changes (412.1 Mt and 407.9 Mt, respectively), reinforcing that population is a contextual driver rather than a policy lever.

A combined adjustment of productivity and consumption achieves 404.0 Mt (−1.46 %), while an aggressive decarbonisation package delivers 399.0 Mt (−2.68 %). Only the Net-Zero 2050 scenario achieves complete decarbonisation (0 Mt, −100 %), highlighting the scale of systemic transformation required beyond incremental measures. Across all tested adjustments, the sensitivity range is −1.61 % to +0.51 % relative to baseline, confirming that incremental measures even when combined yield consistently modest changes.

6.2. Policy implications

The scenario outcomes and model interpretability converge on a clear message: coordinated, system-level action is necessary to support meaningful decarbonisation. Efficiency-oriented measures, such as improving energy productivity or reducing consumption, deliver only marginal gains when applied in isolation, and lowering consumption without changing the energy mix limits overall progress. Population dynamics, while influential, remain exogenous and outside the scope of direct policy intervention. Comprehensive changes across energy supply and demand, technology adoption, and infrastructure are required to reshape the energy–economy system and generate reinforcing effects. SHAP analysis reinforces this conclusion: efficiency improvements,

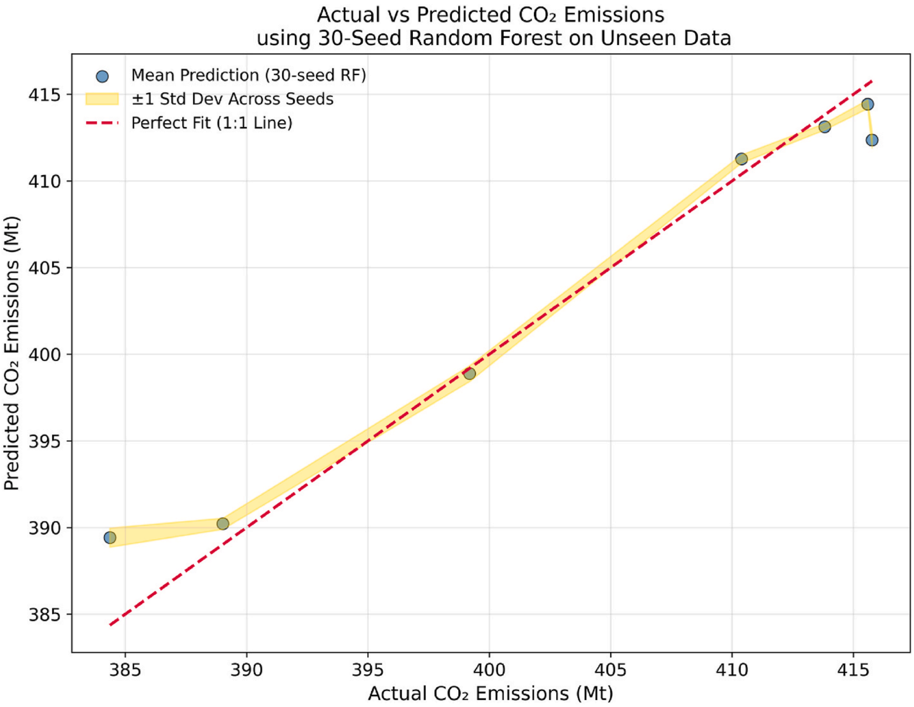

Fig. 11. Actual versus predicted CO_2 emissions for the 30-seed RF ensemble on unseen data (2016–2022). Scatter points show mean predictions, the shaded band represents ±1 standard deviation across seeds, and the dashed line indicates the 1:1 perfect fit.

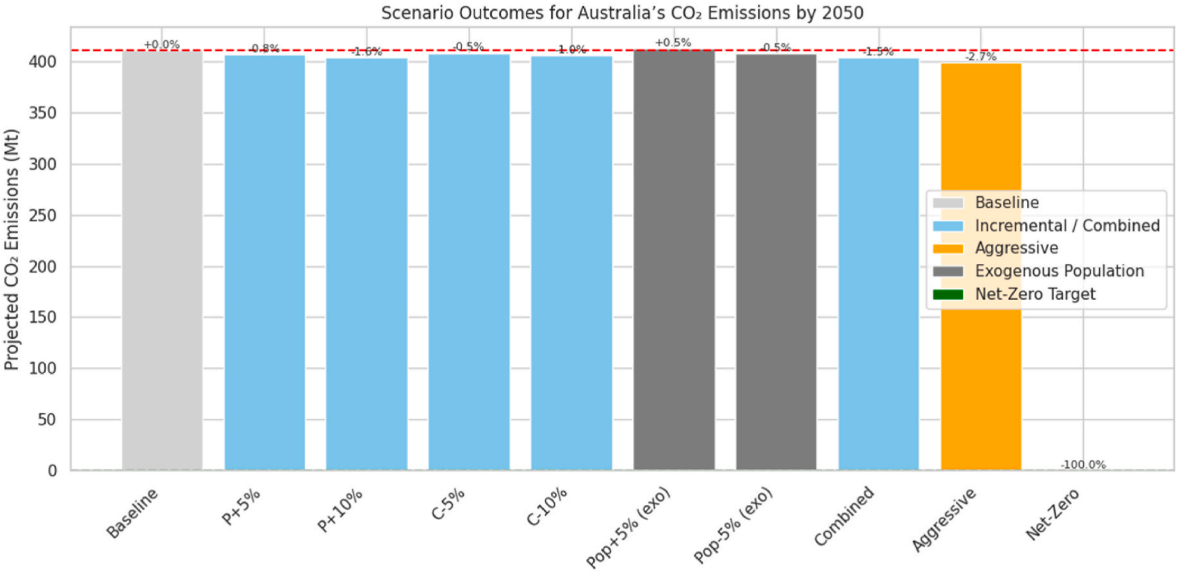

Fig. 12. Illustrative 2050 CO_2 emissions scenarios for Australia. Bars show projected emissions (Mt) relative to the 410 Mt baseline. Incremental adjustments are shown in light blue, combined/aggressive measures in dark blue, population sensitivity in grey (exogenous, not policy levers), and the legislated net-zero target in green. The "Aggressive decarbonisation" scenario is illustrative and not derived from a specific policy package.

renewable expansion, and demand reduction act synergistically rather than as substitutes. Policy frameworks that treat these levers as complementary for example, coupling renewable targets with demand-side efficiency initiatives are more likely to achieve meaningful decarbonisation within this modelling framework.

The RF framework clarifies the drivers behind emissions changes, enhancing transparency, supporting evidence-based interpretation, and improving communication of results. These findings provide actionable insights for national climate strategy, emphasizing multi-sector coordination and alignment between federal and state initiatives. Incremental reforms alone are insufficient; coordinated, system-level approaches are required to meet Australia's climate commitments. Ensemble outputs, including feature importance scores and SHAP explanations, further enhance transparency and contribute to the growing role of explainable AI in climate and energy policy discussions. By combining scenario outcomes with SHAP interpretability, the analysis demonstrates how explainable AI can bridge quantitative forecasts with actionable policy insights, reinforcing the case for coordinated, system-level strategies to achieve national climate targets.

7. Limitations and future work

This study provides valuable insights but has several limitations. The reliance on annual national-level data constrains the effectiveness of deep learning models, which are better suited to higher-frequency temporal patterns. Additionally, the exclusion of policy instruments such as carbon pricing, renewable energy targets, and regulatory interventions as well as international trade factors limits the ability to capture broader economic and policy interactions.

While a broad set of economic, energy, and demographic variables was included (GDP, electricity supply, coal, gas, oil, transport, renewables, energy intensity, energy productivity, and population), national-level aggregation from 1982 to 2022 can mask subnational or sector-specific dynamics and limits the ability of models to capture short-term fluctuations or shocks. Incorporating higher-resolution data, such as quarterly or sectoral series, could enhance model performance and provide a more nuanced understanding of emissions drivers.

Model uncertainty is an important consideration. RF predictions exhibited minor variability across random seeds, highlighting the importance of reproducibility checks and uncertainty quantification. While the ensemble approach helped stabilize results, scenario experiments consistently showed that deterministic adjustments produced only modest reductions by 2050, ranging from −0.49 % for incremental measures to −2.68 % for aggressive changes, with the combined productivity and consumption adjustment reaching −1.46 % relative to the baseline. Population shifts were treated as exogenous sensitivities rather than policy levers.

Several avenues for future research emerge from this study. These include leveraging higher-frequency, sectoral, and regional data to improve temporal resolution and capture detailed dynamics. Incorporating policy and trade variables would enhance projection realism. Methodologically, integrating machine learning with causal inference or physics-informed hybrid models could ensure that forecasts reflect structural drivers rather than purely statistical correlations. Probabilistic forecasting approaches (e.g., Monte Carlo simulations) would provide policymakers with clearer guidance on risks and confidence intervals. By addressing these limitations, future studies can build on this work's contribution: a transparent, interpretable comparison of forecasting approaches for national CO_2 emissions. Such extensions would provide more robust insights into Australia's legislated 2030 target and net-zero 2050 pathway, strengthen scientific rigor, and enhance policy relevance by demonstrating how transparent, reproducible modelling can inform Australia's transition pathways under SDG 13.

8. Conclusion

This study systematically compared statistical, machine learning, hybrid, and deep learning approaches for forecasting Australia's CO_2 emissions. Scenario analyses indicate that incremental adjustments yield only modest reductions by 2050, underscoring the limited impact of isolated measures and reinforcing the importance of coordinated, system-level approaches to align with net-zero targets.

Methodologically, the study contributes a transparent and reproducible forecasting pipeline with interpretable outputs, including feature importance rankings and SHAP explanations. By balancing rigor and transparency, the framework enhances trust in model outputs, supports evidence-based discussions of Australia's climate transition, and contributes to broader global climate objectives under SDG 13.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Data availability

All datasets and code used in this study are publicly available. Data were obtained from Our World in Data (https://ourworldindata.org/co2-and-greenhouse-gas-emissions) and the Australian Energy Update 2024 (https://www.energy.gov.au/publications/australian-energy-update-2024). The code for model training, evaluation, and multi-run analyses is available at https://github.com/safaghannam/CO2-Emissions-Forecasting-in-Australia.

References

[1] Shahbaz M, Bhattacharya M, Ahmed K. CO_2 emissions in Australia: economic and non-economic drivers in the long run. Appl Econ 2017;49(13):1273–86. https://doi.org/10.1080/00036846.2016.1217306.

[2] Li X, Zhang X. A comparative study of statistical and machine learning models on carbon dioxide emissions prediction of China. Environ Sci Pollut Control Ser 2023;30(55):117485–502. https://doi.org/10.1007/s11356-023-30428-5.

[3] Ajala AA, Adeoye OL, Salami OM, Jimoh AY. An examination of daily CO_2 emissions prediction through a comparative analysis of machine learning, deep learning, and statistical models. Environ Sci Pollut Control Ser 2025. https://doi.org/10.1007/s11356-024-35764-8.

[4] Alam GMI, Tanim SA, Sarker SK, Watanobe Y, Islam R, Mridha MF, Nur K. Deep learning model-based prediction of vehicle CO_2 emissions with explainable AI integration for sustainable environment. Sci Rep 2025. https://doi.org/10.1038/s41598-025-87233-y.

[5] Bhatt H, Davawala M, Joshi T, Shah M, Unnarkat A. Forecasting and mitigation of global environmental carbon dioxide emission using machine learning techniques. Cleaner Chemical Engineering 2023;4:100095. https://doi.org/10.1016/j.clechem.2023.100095.

[6] Jin Y, Sharifi A, Li Z, Chen S, Zeng S, Zhao S. Carbon emission prediction models: a review. Sci Total Environ 2024;927:172319. https://doi.org/10.1016/j.scitotenv.2024.172319.

[7] Li Y, Sun Y. Modeling and predicting city-level CO_2 emissions using open access data and machine learning. Environ Sci Pollut Control Ser 2021;28(15):19260–71. https://doi.org/10.1007/s11356-020-12294-7.

[8] Hong S, Fu T, Dai M. Machine learning-based carbon emission predictions and customized reduction strategies for 30 Chinese provinces. Sustainability 2025;17(5):1786. https://doi.org/10.3390/su17051786.

[9] Yao X, Zhang H, Wang X, Jiang Y, Zhang Y, Na X. Which model is more efficient in carbon emission prediction research? A comparative study of deep learning models, machine learning models, and econometric models. Environ Sci Pollut Control Ser 2024;31(13):19500–15. https://doi.org/10.1007/s11356-024-32083-w.

[10] Kumari S, Singh SK. Machine learning-based time series models for effective CO_2 emission prediction in India. Environ Sci Pollut Control Ser 2022;30(2023):116601–16. https://doi.org/10.1007/s11356-022-21723-8.

[11] Foong LK, Blazek V, Prokop L, Misak S, Atamurotov F, Khalilpoor N. Improve carbon dioxide emission prediction in the Asia and oceania (OECD): nature-inspired optimisation algorithms versus conventional machine learning. Eng Appl Comput Fluid Mech 2024;18(1). https://doi.org/10.1080/19942060.2024.2391988.

[12] Shen J, Zheng F, Ma Y, Deng W, Zhang Z. Urban travel carbon emission mitigation approach using deep reinforcement learning. Sci Rep 2024;14:27778. https://doi.org/10.1038/s41598-024-79142-3.

[13] Kahia M, Moulahi T, Mahfoudhi S, Boubaker S, Omri A. A machine learning process for examining the linkage among disaggregated energy consumption, economic growth, and environmental degradation. Resour Policy 2022;79:103104. https://doi.org/10.1016/j.resourpol.2022.103104.

[14] Ghannam S, Hussain F. Investigating the influence of Australia day and christmas day on water demand in the greater Sydney region. Water Supply 2024;24(10):3540–67.

[15] Ran Q, Bu F, Razzaq A, Ge W, Peng J, Yang X, Xu Y. When will China's industrial carbon emissions peak? Evidence from machine learning. Environ Sci Pollut Control Ser 2023;30:57960–74. https://doi.org/10.1007/s11356-023-26333-6.

[16] Marques AC, Fuinhas JA, Leal PA. The impact of economic growth on CO_2 emissions in Australia: the environmental kuznets curve and the decoupling index. Environ Sci Pollut Control Ser 2018;25(27):27283–96. https://doi.org/10.1007/s11356-018-2768-6.

[17] Aras S, Van MH. An interpretable forecasting framework for energy consumption and CO_2 emissions. Appl Energy 2022;328:120163. https://doi.org/10.1016/j.apenergy.2022.120163.

[18] Li S, Siu YW, Zhao G. Driving factors of CO_2 emissions: further study based on machine learning. Front Environ Sci 2021;9:721517. https://doi.org/10.3389/fenvs.2021.721517.

[19] Zhao L-T, Miao J, Qu S, Chen X-H. A multi-factor integrated model for carbon price forecasting: market interaction promoting carbon emission reduction. Sci Total Environ 2021;796:149110. https://doi.org/10.1016/j.scitotenv.2021.149110.

[20] Brownlee J. Machine learning mastery. Machine Learning Mastery; 2022.

[21] Akkaya EK, Akkaya AV. Development and performance comparison of optimized machine learning-based regression models for predicting energy-related carbon

dioxide emissions. Environ Sci Pollut Control Ser 2023;30:122381–92. https://doi.org/10.1007/s11356-023-30955-1.

[22] Kong F, Song J, Yang Z. A daily carbon emission prediction model combining two-stage feature selection and optimized extreme learning machine. Environ Sci Pollut Control Ser 2022;29:87983–97. https://doi.org/10.1007/s11356-022-21277-9.

[23] Chukwunonso BP, Al-Wesabi I, Shixiang L, AlSharabi K, Al-Shamma`a AA, Hussein Farh HM, Saeed F, Kandil T, Al-Shaalan AM. Predicting carbon dioxide emissions in the United States of America using machine learning algorithms. Environ Sci Pollut Control Ser 2024;31:33685–707. https://doi.org/10.1007/s11356-024-33460-1.

[24] Nguyen VN, Tarełko W, Sharma P, El-Shafay AS, Chen WH, Nguyen PQP, Hoang AT. Potential of explainable artificial intelligence in advancing renewable energy: challenges and prospects. Energy & Fuels 2024;38(3):1692–712.

[25] Hoang AT, Nguyen XP. Integrating renewable sources into energy system for smart city as a sagacious strategy towards clean and sustainable process. J Clean Prod 2021;305:127161.

[26] Hoang AT, Bui TAE, Nguyen XP, Bui VH, Nguyen QC, Truong TH, Chung N. Explainable machine learning-based prediction of fuel consumption in ship main engines using operational data. Brodogradnja: An International Journal of Naval Architecture and Ocean Engineering for Research and Development 2025;76(4): 1–24.

[27] Bakır H, Ağbulut Ü, Gürel AE, Yıldız G, Güvenç U, Soudagar MEM, Afzal A. Forecasting of future greenhouse gas emission trajectory for India using energy and economic indexes with various metaheuristic algorithms. J Clean Prod 2022;360: 131946.