Research Article

A Comparative Analysis of Machine Learning Models on Predicting GDP Based on Greenhouse Gas Emissions

Khanh Pham

Hanoi – Amsterdam High School for the Gifted, 1 Hoang Minh Giam St., Hanoi 100000, Vietnam

ABSTRACT

As climate change intensifies, it is increasingly important to understand how greenhouse gas (GHG) emissions affect economic performance. Using machine learning, this study examined the potential of using GHG emission data to predict gross domestic product (GDP) across 161 nations. Linear Regression, Decision Tree, Random Forest, and Multi-layer Perceptron (MLP) were trained and optimized individually per country, with their hyperparameters tuned by Optuna. The results show that MLP significantly reduced its mean MAPE from 27.583% to 5.265% and mean RMSE from 154.221 to 34.051 billion USD, while also significantly suppressing outliers (with its maximum error dropping from 117.955% to 22.661%). Random Forest also showed strong improvement, with mean MAPE decreasing to 5.477% and a notable reduction in large errors. Decision Tree showed some small-scale improvement, while Linear Regression was relatively poor with a mean MAPE of 11.244%, which highlighted the non-linearity of GHG-GDP relationship. Statistically significant improvements were found in all analyses ($p<0.001$) using Wilcoxon signed-rank tests, indicating consistent gains across different countries. Findings indicate that GHG emissions can serve as a source of meaningful predictive signal for non-linear models, making them valuable tools for estimating economic trends in areas with limited data or policy constraints, where emissions data may be more readily available than GDP reports. Hyperparameter optimization is a key factor in improving model accuracy and reliability, as highlighted in this study. Future work should expand the feature set, incorporate time series forecasting techniques, and improve model interpretability to support real-world policy applications.

Keywords: Machine Learning; Greenhouse Gas Emissions; Gross Domestic Product; Hyperparameter Optimization; Economic-Environmental Modeling; Time Series Forecasting; Environmental Economics

INTRODUCTION

Greenhouse gases (GHGs) are atmospheric gases that absorb some of the heat that occurs when sunlight hits the Earth's surface. Three main GHGs include carbon dioxide, methane, and nitrous oxide, all of which contribute to the greenhouse effect (1). GHGs are integral to the climate system; however, human activities, including industrial production, transportation, and agriculture, have significantly increased their concentration in the

Corresponding author: Khanh Pham, E-mail: khanh.phamnam310108@gmail.com.

Received July 14, 2025**; Accepted** August 10, 2025
https://doi.org/10.70251/HYJR2348.34263269

atmosphere, intensifying the greenhouse effect and affecting global warming (2). As global temperature increases, the climate system becomes increasingly unstable, causing extreme weather, rising sea level, and disrupting the ecosystem (3). These changes also pose significant challenges for modern cities. Therefore, understanding GHGs is crucial to ensuring a sustainable future.

Gross domestic product (GDP) is one of the most widely used indicators to assess a nation's economic performance and overall standard of living. It is the total market value of all goods and services produced within a country in a certain period of time (4). Governments and organizations often use GDP to measure economic growth and living standards. However, the increasing pursuit of GDP growth often means increased energy consumption. Therefore, understanding the relationship between GDP and GHG is consequential, which presents challenges for policymakers to balance environmental changes and long-term economic sustainability.

LITERATURE REVIEW

There is a significant overlap in the relationship between GHGs and economic growth. One such example is to predict how economies might be affected by climate change to prioritize actions and mitigate risks. This relationship was further explored in the paper written by Romero and Gramkow in 2021, where the researchers used data for 67 countries between 1967 and 2012, describing the Economic Complexity Index from MIT's Observatory of Economic Complexity and GHG emissions (in kilotons of CO2 equivalent, CO2e) from the World Development Indicator database (5). The paper found that an increase of 0.1 in the economic complexity index generates a 2% decrease in the next period's emissions of kilotons of CO2e per billion dollars of output as well as in emissions per capita (5). While this paper is profound in its results and implications, it does not complete the story of the relationship between economic activity and GHG emissions. This is because the paper does not include land use, land-use change and forestry. Given the lack of inclusion of these details, this study seeks to provide a more comprehensive analysis.

METHODS AND MATERIALS

Dataset

The dataset was sourced from Our World in Data, including a comprehensive collection of data from 161 countries over the period of 1991 to 2022. Although the original dataset contains multiple socioeconomic and environmental variables, a subset of key features was selected for this study. Specifically, GDP data from the University of Groningen GGDC's Maddison Project Database (6), expressed in US dollars, was utilized, as well as greenhouse gas emissions data, namely CO_2, CH_4, and NO_2, all measured in tons of carbon dioxide-equivalents and adjusted for land-use change (7).

Crucially, all modeling was conducted at the country level: each nation's data was isolated, resulting in 161 independent datasets. For each country, the data was then split into a training set (80% of the years) and a test set (20% of the years). All models were then trained, validated, and evaluated separately for each country using only that nation's historical GHG emissions to predict its GDP.

In addition, the StandardScaler of the scikit-learn library was used to standardize all the features to guarantee a fine level of consistency and comparison. Standardization was a very important preprocessing step as the features of the dataset, in this case, GDP and GHG emissions metrics, were on a different scale and unit. By centering the data around a mean of zero and scaling it to have a unit variance, the StandardScaler reduces potential biases in the model caused by differences in feature magnitudes. Primarily, it would also ensure that the features contributed to the predictions in equal measure, hence improving the reliability of the outcome.

Methodologies

This study used four different machine learning algorithms, including Linear Regression, Decision Tree, Random Forest, and Multi-layer Perceptron (MLP), to perform a thorough comparative analysis of machine learning models on predicting GDP based on GHG emissions. Since each model has its own advantages and methods for identifying connections in data, the analysis was used to provide a strong foundation for assessing how well they predict outcomes and comprehending the underlying trends between environmental effects and economic production.

Linear Regression. Linear Regression is a basic machine learning model to describe the linear relation between a dependent variable (GDP) and one or more independent variables (GHG emissions). The concept is to find the best-fit straight line that minimizes the sum of the squared difference between observed and predicted values.

It is a simple baseline for comparison because of its

simplicity, and it is simple to see how variables relate to each other. However, Linear Regression might not perform well on complex, non-linear data since it can be influenced by outliers and it assumes a linear relationship. That said, this study will apply Linear Regression to find a basic understanding of the direct linear relation of GHG emissions to GDP.

Decision Tree. Decision Trees are non-parametric supervised learning algorithms that partition the data into subsets depending on the feature values to construct a tree-based model of decisions and possible results. This is the procedure of recursive division of the data set into increasingly homogeneous subsets. In the context of this study, each internal node represents a test or decision on a GHG emission attribute (CO_2, CH_4, or N_2O levels), each branch represents an outcome of that decision, and each leaf node represents a predicted GDP value. The algorithm determines the optimal split points by minimizing a criterion such as Mean Average Percentage Error (MAPE).

Decision Trees are intuitive and can capture non-linear relationships up to a certain degree by visualizing the respective decision rules. Nevertheless, individual Decision Trees are prone to overfitting, especially if they are deep, and can be unstable, meaning slight modifications in the dataset may lead to substantially different tree structures.

Random Forest. Random Forest is an ensemble learning technique that extends the idea of Decision Trees to mitigate their weaknesses, particularly overfitting and instability. It does so by training a multitude of Decision Trees at training and outputting the average prediction of the individual trees. This ensemble approach takes advantage of two main methods: bagging (bootstrap aggregating), in which each tree is trained on a different bootstrap sample of the training data; and the random subspace method, where each tree considers only a random subset of features when determining the best split.

By aggregating the predictions of numerous diverse trees, Random Forest significantly reduces variance and improves predictive accuracy. This model is suitable for handling complex, non-linear relationships and high-dimensional data, making it a hypothetically effective tool for predicting GDP based on various GHG emissions.

Multi-layer Perceptron (MLP). The Multi-layer Perceptron (MLP) is a class of feedforward artificial neural networks, including at least three layers of nodes: an input layer, one or more hidden layers, and an output layer. Every node (or neuron) in a layer is linked with all nodes in the subsequent layer, with each connection carrying a weight that specifies the strength of the link. Information moves unidirectionally from the input layer, through the hidden layers, to the output layer. The hidden layers contain non-linear activation functions so that the MLP can learn and model complicated, non-linear relationships between inputs (GHG emissions) and outputs (GDP). The network is trained using a method called backpropagation, in which the error between the predicted output and the actual output is passed backwards through the network, modifying the weights to reduce this error.

Although MLPs are highly predictive and capable of learning complex patterns, they are typically hard to interpret due to their complex internal structure. They also need larger training sets and careful hyperparameter tuning to function at their best.

Optimization

In addition to providing baseline performance, this study also emphasizes optimization of the models to achieve maximum predictive performance for all countries. That is, the hyperparameters of Decision Tree, Random Forest, and MLP were optimized for each country. Through this, it is ensured that the models are optimized for the specific characteristics of individual country data, making the predictions more accurate.

For Decision Tree, the maximum depth of the tree, the minimum number of samples required to split an internal node, the minimum number of samples required to be at a leaf node, the number of features to consider when looking for the best split, and the Minimal Cost-Complexity Pruning parameter were tuned.

For Random Forest, the number of trees in the forest, the maximum depth of the tree, the minimum number of samples required to split an internal node, the minimum number of samples required to be at a leaf node, the number of features to consider when looking for the best split, and whether bootstrap samples are used when building trees were tuned.

For the MLP Regressor, the number and size of hidden layers, the activation function for the hidden layers, the solver for weight optimization, the strength of the L2 regularization term, the learning rate, the maximum number of iterations, and early stopping parameters were tuned.

For effective and efficient hyperparameter tuning, Optuna (8), an open-source library for hyperparameter tuning, was utilized. Optuna determines the best hyperparameters by specifying a search space for every parameter and employing various sampling algorithms (for example, Tree-structured Parzen Estimator) to

search the space intelligently. Optuna also employs pruning strategies (for example, median pruning) to automatically stop unpromising trials during training, saving computational resources. This dynamic process allows for effective hyperparameter combinations that would otherwise be overlooked using conventional grid search or random search methods, leading to improved model performance.

Model Evaluation

For each country, the performance of each machine learning model was evaluated using the Mean Absolute Percentage Error (MAPE) and Root Mean Squared Error (RMSE).

MAPE and RMSE are defined as:

$$MAPE = \frac{100\%}{n}\left|\frac{Y_i - \hat{Y}_i}{Y_i}\right|$$

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2}$$

where Y_i is the actual GDP, $\hat{Y}_i$ is the predicted GDP, and n is the number of samples.

Statistical Significance Testing

To determine whether hyperparameter optimization led to significant gains, the Wilcoxon signed-rank test was applied to analyze prediction errors before and after tuning. The MAPE and RMSE of the baseline and optimized versions of each model (Decision Tree, Random Forest, and MLP) were paired by country across all 161 nations.

Because MAPE and RMSE distributions are often skewed and not normally distributed, this non-parametric test evaluates whether the median difference in error (baseline vs. optimized) is significantly greater than or equal to zero.

RESULTS

This study evaluated the performance of four machine learning algorithms (Linear Regression, Decision Tree, Random Forest, and MLP) for predicting GDP using GHG emissions for 161 nations during the period 1991 – 2022. The available data for the analysis were all nations within the dataset with full annual observations over this full 32-year span, providing a balanced panel and no missing value imputation or case-wise deletion. This strict temporal alignment ensured that one period of time was used to train and test all models, removing bias from missing or interpolated data.

All features were extracted from the same period and geographical scope: total GHG emissions (CO_2, CH_4, N_2O, adjusted for land-use change) in tons of CO_2-equivalents, and GDP in US Dollars. Other data variables in the original dataset (for instance, population, energy consumption, and policy indicators) were omitted to isolate the predictive signal of GHG emissions. This specific reduction makes it possible to assess how well emissions data predict economic performance.

Figure 1 displays MAPE distribution of the 161 countries over all models, baseline and optimized. The x-axis scale ranges from 0% to over 100% MAPE, allowing visualization of outlier prediction errors in the graph. The y-axis shows how many countries fall into each range of errors. Baseline models showed significant variation: Linear Regression showed a moderately spread distribution with peak below 20% but extending past 40%, suggesting consistent but suboptimal performance. Decision Tree and Random Forest produced more compact central clustering, but both had high outliers above 50%. Most notably, MLP possessed the highest distribution of errors, with a peak between 20% and 60% and numerous countries exceeding 100% MAPE.

After optimization, all non-linear models showed improved reliability. Decision Tree's distribution shifted toward lower errors with most countries below 20% and fewer outliers. Random Forest achieved a highly peaked distribution, with the majority clustered in the 0-10% range and low tail spread. However, the most dramatic improvement was observed in MLP: its optimized version transformed from being the worst-performing model to one of the best, with nearly all countries below 15% MAPE and tight clustering below 10%. Visually, this was a suppression of large errors and a shift towards consistent, reliable predictions.

These trends were confirmed quantitatively in Table 1 and Table 2, which summarize minimum, maximum, and mean minimum/maximum MAPE (%) and RMSE (billion USD) for all countries. In baseline configuration, models performed as follows by mean MAPE: Random Forest (7.987%) > Decision Tree (10.991%) > Linear Regression (11.244%) > MLP (27.583%). Following tuning, the ranking shifted significantly: MLP achieved the lowest mean MAPE (5.265%) and a dramatic decrease in RMSE (from 154.221 billion USD to 34.051 billion USD), and it even outperformed tuned Random Forest (5.477%) and Decision Tree (7.206%). The highest MAPE of MLP also dropped from 117.955% to 22.661%.

To evaluate the statistical significance of these

improvements across countries, Wilcoxon signed-rank tests were applied to paired values of MAPE and RMSE (baseline vs. optimized) for every tunable model. All reductions in MAPE and RMSE were statistically significant ($p<0.001$), confirming that all performance gains were consistent and not due to random variation.

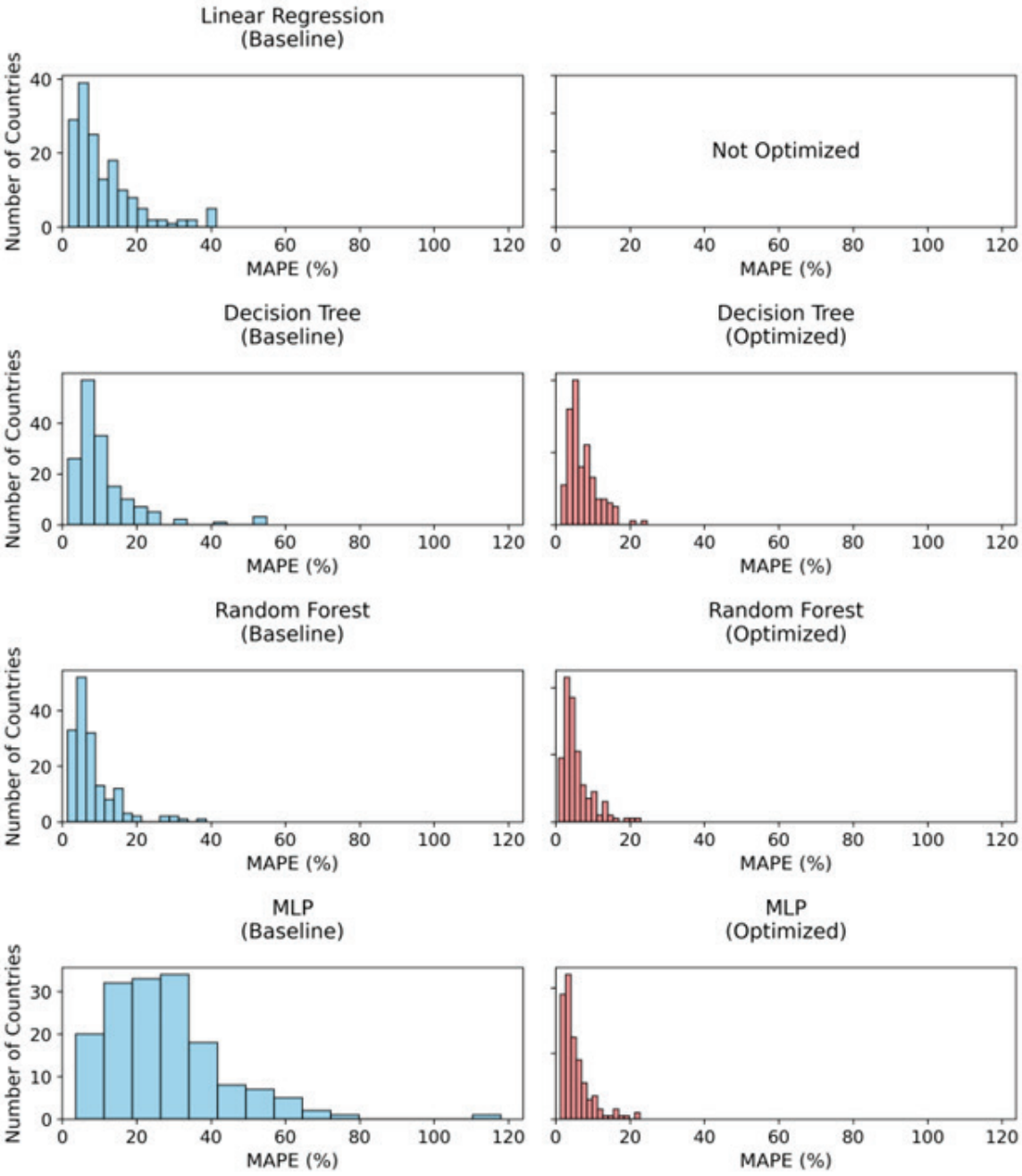

Figure 1. Distribution of MAPE across 161 countries for baseline and optimized models. Each row compares one algorithm: (left) baseline, (right) optimized. Models shown are (A) Linear Regression, (B) Decision Tree, (C) Random Forest, and (D) Multi-layer Perceptron. The optimized versions show tighter error distributions and fewer high-MAPE outliers.

Table 1. Baseline Model Results

	Min MAPE (%)	Max MAPE (%)	Mean MAPE (%)	Min RMSE (USD bn.)	Max RMSE (USD bn.)	Mean RMSE (USD bn.)
Linear Regression	1.678	41.51	11.244	0.020	1794.985	49.385
Decision Tree	1.499	54.958	10.991	0.025	1453.313	50.796
Random Forest	1.423	38.715	7.987	0.020	784.332	34.611
MLP	3.583	117.955	27.583	0.109	3924.333	154.221

Table 2. Optimized Model Results

	Min MAPE (%)	Max MAPE (%)	Mean MAPE (%)	Min RMSE (USD bn.)	Max RMSE (USD bn.)	Mean RMSE (USD bn.)
Linear Regression	N/A					
Decision Tree	1.440	24.566	7.206	0.058	972.356	38.761
Random Forest	0.761	22.797	5.477	0.020	779.749	28.889
MLP	1.205	22.661	5.265	0.013	1224.061	34.051

DISCUSSION

Model Performance

MLP's dramatic shift from the worst baseline performer to one of the best optimized models suggests that its predictive potential was limited by poor default hyperparameters rather than weak model fit. Once tuned, MLP improves its ability to learn complex, non-linear relationships between GHG emissions and GDP, enabling it to adapt to diverse national economic frameworks. Its error distribution becomes highly compact, with almost no large outliers, indicating stable and reliable predictions across countries.

Random Forest also achieves strong performance after tuning, with a tightly clustered error distribution and low mean MAPE (5.477%), indicating its ability to reduce variance through ensemble averaging and effective hyperparameter control. However, its improvement is less noticeable than MLP's, suggesting it was already close to optimal before tuning.

Decision Tree shows measurable gains, with more accurate predictions after optimization, but its wider error spread, suggesting vulnerability to high-error samples in several countries. Meanwhile, unoptimized Linear Regression displays moderate errors and long-tailed spread, highlighting the limitations of linear modeling in understanding the environmental-economic relationship.

Together, the results show that hyperparameter optimization is not only beneficial, but also transformative for models like MLP, allowing them to find meaningful predictive signals from GHG emissions alone.

Limitations & Future Work

This study has several methodological and practical limitations that should be taken into consideration. First, the models only rely on GHG emissions when predicting GDP. However, GDP is influenced by various economic and social factors, including labor force participation, capital investment, technological development, population growth, trade openness as well as institutional quality. These variables are not included in the models, so they do not accurately represent the full range of economic systems and should not be interpreted as causal GDP forecasting tools.

Moreover, the modeling technique is country-specific, as each nation's model is trained separately using its own time series and hyperparameters that are tuned separately. Although this allows for high customization and improved suitability for national emission-economic dynamics, it limits scalability and generalizability. The models cannot be applied directly to countries that are not included in the training set or which have limited historical data, such as small island states and areas of conflict. Moreover, as the patterns

are not uniform across countries, the approach misses opportunities to learn from cross-national insights.

MLP and Random Forest are complex models that offer high performance, but their interpretability is compromised. "Black boxes" in these models make it challenging to extract precise policy insights or understand the relationship between GDP and emissions at a mechanistic level. This poses challenges for policymakers seeking transparent, actionable information.

Based on historical data from 1991 to 2022, the models are trained to assume that the correlation between GHG emissions and GDP will persist. However, unpredictable events, such as global pandemics, wars or technological advancements in technology, can disrupt these, reducing predictive validity.

Future work should address these limitations in multiple ways. To improve model reliability, it is essential to expand the feature set by including more socioeconomic and environmental variables, such as energy consumption, population, urbanization, trade volume, and government policy indicators. Furthermore, alternative modeling frameworks such as panel data models or approaches that consider both country-specific adaptation and shared global patterns should be considered. In addition, time-series forecasting techniques like ARIMA, SARIMA, or LSTM networks could be utilized to better capture accurate long-term trends and temporal dependencies.

Moreover, future models should aim to have greater practical value by incorporating real-time emission data and working with policymakers to make outputs interpretable and decision relevant. Ultimately, incorporating machine learning into environmental-economic forecasting should not only improve accuracy but also help in developing sustainable and evidence-based policies.

CONCLUSION

This study investigated the effectiveness of machine learning models in predicting GDP from GHG emissions for 161 countries from 1991 to 2022, emphasizing the role of hyperparameter optimization. The performance of the models improved with hyperparameter tuning, with the tuned MLP and Random Forest models being the most accurate and reliable predictors on the GHG emissions data. This highlights the potential of machine learning to model complex economic-environmental relationships and provides a foundation for future forecasting and policymaking. Future studies should consider larger feature sets and advanced techniques to enhance accuracy and real-world applications.

ACKNOWLEDGEMENTS

I would like to thank my mentor Neven Vaduthala for his guidance and support throughout this research process.

CONFLICT OF INTERESTS

The author declares that there are no conflicts of interest regarding the publication of this article.

REFERENCES

1. Filonchyk M, Peterson MP, Zhang L, Hurynovich V & He Y. Greenhouse gases emissions and global climate change: Examining the influence of CO2, CH4, and N2O. S*cience of the Total Environment*. 2024; 935 (935): 173359–173359. https://doi.org/10.1016/j.scitotenv.2024.173359
2. Evseeva O, Evseeva S & Dudarenko T. The impact of human activity on the global warming. *E3S Web of Conferences*. 2021; 284 (11017): 11017. https://doi.org/10.1051/e3sconf/202128411017
3. Fontúrbel FE, Nespolo RF, Amico GC & Watson DM. Climate change can disrupt ecological interactions in mysterious ways: using ecological generalists to forecast community-wide effects. *Climate Change Ecology*. 2021; 2: 100044. https://doi.org/10.1016/j.ecochg.2021.100044
4. Landefeld JS, Seskin EP & Fraumeni BM. Taking the Pulse of the Economy: Measuring GDP. Journal of Economic Perspectives. 2008; 22 (2): 193–216. https://doi.org/10.1257/jep.22.2.193
5. Romero JP & Gramkow C. Economic complexity and greenhouse gas emissions. *World Development*. 2021; 139: 105317. https://doi.org/10.1016/j.worlddev.2020.105317
6. Bolt J & Luiten J. Maddison-style estimates of the evolution of the world economy: A new 2023 update. *Journal of Economic Surveys (Print)*. 2024. https://doi.org/10.1111/joes.12618
7. Jones MW, Peters GP, Gasser T, Andrew RM, *et al*. National contributions to climate change due to historical emissions of carbon dioxide, methane and nitrous oxide. *Zenodo*. 2024. https://doi.org/10.5281/zenodo.14054503
8. Akiba T, Sano S, Yanase T, Ohta T & Koyama M. Optuna. Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining. 2019. https://doi.org/10.1145/3292500.3330701