Planbureau voor de Leefomgeving

VU VRIJE UNIVERSITEIT AMSTERDAM

A Panel-Based Machine-Learning Framework for Benchmarking National Net-Zero Pledges

Predicting Cumulative Per-Capita Greenhouse-Gas Budgets and Implied Net-Zero Years

Master Project Business Analytics

Said Çetinkaya (2723453)

Author:	Said Çetinkaya
University:	Vrije Universiteit Amsterdam
Thesis internship:	PBL Netherlands Environmental Assessment Agency
External supervisors:	Michel G. J. den Elzen and Stefan Troost
University supervisor:	Rob van der Mei
Second reader:	Joost Berkhout
Date:	June 2026

Abstract

National net-zero pledges are increasingly used to explain the climate ambition of a country, yet their comparability still remains a challenge to evaluate since countries differs in their emissions profiles, energy systems, economic structures, natural-resource dependency, governance conditions, and the availability of data. The recent work of den Elzen et al. [1] have introduced a machine learning approach to the problem by modelling the cumulative per-capita greenhouse-gas emissions implied by national net-zero pledges which is constructed from a base-year emissions level to the pledged net-zero target year while passing through a 2030 NDC or current-policy emission benchmark. In this thesis, we try to answer how accurately and interpretably can a panel based machine learning model estimate these pledge derived cumulative emissions budgets and measure countries' announced net-zero target years against model implied years based on country level yearly indicators?

In our analysis, we use a country-year panel covering the period from 2000 to 2024 and combine emissions, energy, natural resources, economic, governance, demographic, and risk-related indicators. In our modelling pipeline, we include country level PCA embeddings, alternative imputation branches based on hierarchical median imputation and Multiple Imputation by Chained Equations (MICE), feature engineering using interaction terms, rolling statistics, and finally Recursive Feature Elimination (RFE) with an Extra Trees regressor for feature selection. Three different target scenarios (namely, linear, pessimistic and optimistic) are constructed to represent different assumptions about the shape of emission reductions towards reaching net-zero emissions by the pledged target year.

We have assessed 13 regression estimators by using nested cross-validation across four outer test years (2018–2021), two validation strategies, two imputation branches, and three target scenarios. For optimization of hyperparameters we have made use of Tree-structured Parzen Estimator search. The candidates, which are chosen by their average MAE scores, are then trained on the pre-2022 training data and evaluated on a final 2022 holdout partition. The 2022 holdout is reported in two different subset of the test set, namely, a strict subset, which consists of 109 country observations that satisfy the $< 30\%$ row-missingness rule, and a lenient subset, which has 123 target-valid country observations that are kept regardless of that feature-missingness threshold.

The results show that support vector regression is the best estimator in the experimental design. While SVR ranks first in all six primary combinations of validation strategy and target scenario, HistGradientBoosting, MLP, and CatBoost are sharing second and third places in the remaining winner candidates. On the strict 2022 holdout set, the best-performing configuration is the optimistic-scenario random-CV SVR with simple imputation, which achieves a test MAE of 0.153 and an R^2 of 0.956. Within the linear scenario, the random-CV SVR with simple imputation achieves a test MAE of 0.213 and an R^2 of 0.917, while the temporal-CV SVR with simple imputation achieves a test MAE of 0.181 and an R^2 of 0.942. SHAP-based interpretation is used to understand the main predictors of the chosen models, and the bootstrap resampling is used to evaluate the uncertainty in the final projections.

The thesis contributes to the literature by extending the cross-sectional (the countries at one time point) machine-learning framework of den Elzen et al. [1] into a panel-based (the countries over extended time points) setting, by comparing a broader set of regression models. The study is not intended to replace the existing integrated assessment models or detailed country specific policy analysis. Rather, it provides a complementary tool for the measurement of the national net-zero pledges against observable country characteristics.

Contents

List of Figures

List of Tables

List of Abbreviations

Abbreviation	Full term
AdaBoost	Adaptive Boosting
API	Application Programming Interface
CatBoost	Categorical Boosting
CH_4	Methane
CI	Confidence interval
CO_2	Carbon dioxide
CO_2e	Carbon dioxide equivalent
CV	Cross-validation
EC	European Commission
Extra Trees	Extremely Randomized Trees
FAO	Food and Agriculture Organization of the United Nations
GDP	Gross domestic product
Gg	Gigagram
GHG	Greenhouse gas
HFCs	Hydrofluorocarbons
HISTCR	Historical country-reported scenario
HistGradientBoosting	Histogram-based Gradient Boosting
IEA	International Energy Agency
INFORM	Index for Risk Management
IPCC	Intergovernmental Panel on Climate Change
IQR	Interquartile range
ISO	International Organization for Standardization
JRC	Joint Research Centre
L1	L1 norm or L1 regularization
L2	L2 norm or L2 regularization
LightGBM	Light Gradient Boosting Machine
LULUCF	Land use, land-use change, and forestry
MAE	Mean absolute error
MICE	Multiple Imputation by Chained Equations
MJ	Megajoule
MLP	Multilayer perceptron
Mt	Megatonne
N_2O	Nitrous oxide
NDC	Nationally Determined Contribution
NF_3	Nitrogen trifluoride
NZ	Net zero
OECD	Organisation for Economic Co-operation and Development
P05/P25/P75/P95	5th, 25th, 75th, and 95th percentiles
PBL	PBL Netherlands Environmental Assessment Agency
PCA	Principal Component Analysis
PFCs	Perfluorocarbons
PIK	Potsdam Institute for Climate Impact Research
PPP	Purchasing power parity
PRIMAP-hist	PRIMAP-hist national historical emissions time series
R^2	Coefficient of determination
RBF	Radial basis function
ReLU	Rectified linear unit
RFE	Recursive Feature Elimination
RMSE	Root mean squared error

Abbreviation	Full term
RSF	Reporters Without Borders
SF_6	Sulfur hexafluoride
SHAP	SHapley Additive exPlanations
SVR	Support vector regression
TFC	Total final consumption
TJ	Terajoule
TPE	Tree-structured Parzen Estimator
UN	United Nations
UN DESA	United Nations Department of Economic and Social Affairs
UNFCCC	United Nations Framework Convention on Climate Change
USD	United States dollar
V-Dem	Varieties of Democracy
WGI	Worldwide Governance Indicators
WPP	World Population Prospects
XGBoost	Extreme Gradient Boosting
YoY	Year-on-year

1 Introduction

1.1 Policy context and motivation

The Paris Agreement has established a global framework to keep the warming to below 2 °C above pre-industrial levels and trying to limit warming to 1.5 °C [2]. Within the framework, the national net-zero targets has become a central reference point in climate policy assessment, but the target year alone provides only limited information about countries' ambitions. A pledge to reach net zero by 2050, 2060, or 2070 implies a cumulative emissions pathway. The pathway depends on a country's current emissions level, energy structure, economic development, land-use conditions, and institutional capacity. For this reason, comparing the net-zero pledges requires more than listing announced target years. It requires a framework that ties those targets to observable country characteristics.

The thesis bases and builds on den Elzen et al. [1], who introduced a machine-learning approach for estimating the cumulative per-capita emissions associated with the national net-zero targets. Their study provides the closest predecessor to the present work, but it is primarily cross-sectional in which each country is represented once, using a single set of recent country characteristics to compare differences between countries at one point in time. The contribution of the thesis is to extend that framework further to a panel-based machine-learning design, in which countries are observed repeatedly over multiple years. The panel structure will allow the model to use both the differences between countries and the changes within the same country over time. It will also broaden the model comparison and evaluate the resulting predictions using stricter validation and holdout procedures.

The central research question is "how accurately and interpretably can a panel-based machine-learning framework estimate the pledge derived cumulative emissions budgets and benchmark countries' announced net-zero target years against model-implied years based on observable country characteristics?"

1.2 What shapes a net-zero pledge?

National net-zero target years are anchored in the international climate policy architecture that is created by the Paris Agreement. Article 4.1 sets the objective of reaching the global peak in greenhouse gas emissions as soon as possible. It also sets the objective of achieving a balance between emissions and removals in the second half of this century [2]. The IPCC Sixth Assessment Report also indicates that the pathways limiting the warming to 1.5 °C requires global greenhouse-gas emissions to reach net zero around mid-century, with rapid reductions before 2050 and net-zero CO_2 around 2050 [3]. In this global shared guideline, national circumstances, in particular the structure of the energy system, the composition of hard-to-abate sectors, the endowment of natural carbon sinks, the level of economic development, and the quality of governance institutions, jointly shape both the technical feasibility and the political credibility of any pledged trajectory [4–6]. Earlier empirical work using classical statistical methods has linked NDC ambition to indicators of democracy, climate vulnerability, fossil-fuel dependence, and macroeconomic structure. It also highlights that countries facing more severe domestic policy constraints tend to submit less ambitious NDCs and to adopt fewer policies to deliver them [5, 6]. A complementary literature has examined the timing of net zero from a fair-shares perspective, either by allocating remaining global budgets according to equity principles [7] or by mapping the systemic changes required to reach net zero in integrated-assessment-model pathways [4].

Despite the growing body of work, the use of machine learning to study national climate ambition is recent. A notable contribution has been made by von Dulong and Hagen [8], who used

machine-learning methods to identify the institutional drivers of climate-policy stringency. However, their target variable concerns the implementation of the existing policies rather than the long-term ambition embodied in a net-zero pledge. The most direct precursor to the present work has been conducted by den Elzen et al. [1], which is the first study to apply an machine learning algorithm, specifically XGBoost [9], to the problem of projecting and explaining national net-zero target years. Using a cross-sectional sample of 74 countries and an indicator set covering emissions, energy, land use, and institutional dimensions, the authors calculate a cumulative per-capita emission budget from a baseline year to the pledged net-zero year, train XGBoost and a comparison linear regression on this target, and interpret the fitted model via SHapley Additive exPlanations (SHAP) values [10]. The indicator set used by den Elzen et al. is summarised in Table 2. They identify that per-capita greenhouse-gas emissions, renewable energy share, hard-to-abate-sector emissions, energy intensity, oil and gas rents, and press freedom are the leading determinants of national ambition. Their results suggest that observable country characteristics contain meaningful information about the plausibility and relative ambition of national net-zero pledges. However, their design remains primarily cross-sectional which means that each country is a one observational entity once rather than as a repeated country-year series. The model comparison is also narrow in scope in which they make use of XGBoost and benchmarking it against a linear regression. The question remains open whether the same target construction logic remains robust when applied to a larger panel dataset (country-year) and evaluated across a broader set of model families.

A further difference is that the thesis does not reproduce the country-clustering step used in the earlier study. In the current panel setting, it is taught that clustering countries into discrete groups would substantially reduce the effective training sample within each group and would make nested cross-validation, imputation, and model comparison less stable. Instead, country heterogeneity is represented continuously through the observed emissions, energy, land-use, economic, and governance indicators, supplemented by country PCA embeddings and interaction features. The choice keeps all countries in a common benchmark model while still allowing structurally different country profiles to influence the predicted cumulative budget.

Table 2: Indicators used by den Elzen et al. to characterise determinants of net-zero targets

Dimension	**Indicators**
Emissions	Emissions per capita*, Non-CO_2 emissions per capita*, Emissions growth, Total hard-to-abate emissions
Energy systems	Renewable energy share, Renewable electricity share, Energy intensity, Oil and gas rents*, Coal rents*
Land use	Forest area, Annual deforestation, Agricultural land, Population density*, Average supply of protein of animal origin
Institutional and Governance	GDP per capita*, Electoral democracy, Control of corruption, Political stability and absence of violence, Regulatory quality, Voice and accountability, Press freedom index, Climate Change Risk Index

Note: Indicators marked with * are log-transformed. Source: adapted from den Elzen et al. [1], Table 1.

1.3 Open methodological questions

Den Elzen et al. [1] provides a methodological starting point, their framework also leaves open a set of methodological questions, some of which the authors themselves acknowledge. Three of these limitations motivate the thesis.

First, the dataset is small by the standards of contemporary supervised learning. With 74 countries entering the regression after filtering, the effective sample size per cross-validation fold is in the order of fifty observations, which limits the statistical power of the models. This also amplifies the risk of performance differences between XGBoost and the linear benchmark which may be sensitive to the particular validation split [1]. Cross-country panels are limited in temporal dimension because net-zero pledges are a recent phenomenon, still they can be substantially broadened by adding territories, regional aggregates, and additional countries for which data are not available for every selected indicator.

Second, the evaluation is restricted to a single tree-based ensemble and a single linear baseline. Although the cited XGBoost and linear comparison is informative, it does not address the question of which estimator family is best suited to the structural properties of the target variable. The methodological literature on cross-country regression has shown that the relative performance of different model families can vary substantially with sample size, dimensionality, feature scale, and the existence of nonlinear interactions [11]. A more decisive methodological conclusion requires a broader candidate set evaluated under the same cross-validation procedure in which the data is repeatedly split into training and test subsets, while checking whether a model performs well on the training data but poorly on unseen data.

Third, the earlier study uses a cross-sectional design, which means that each country appears only once in the regression dataset, and each country represented by the most recent available values of its explanatory indicators. This design compares differences between countries at a single point in time, but it does not make use of repeated country-year observations to learn from within-country changes over time. This choice simplifies the construction of the target variable, but it also discards the within-country variation over time. However, exploiting the temporal dimension is not easy. The cumulative per-capita budget to a distant net-zero year is, by construction, dominated by the integral of future emissions and is therefore relatively insensitive to small year-on-year changes in observed features. The relevance of the panel design must be shown rather than assumed. Also, the design itself must be matched with imputation, feature-engineering, and validation choices that are appropriate with a panel rather than a cross-section.

1.4 Research aims and contributions

The thesis revisits the prediction of national net-zero ambition through an extended panel-based machine-learning framework. The analysis is organised around four aims.

First, to test a broad set of machine-learning methods under a common evaluation framework. Thirteen candidate regression estimators are evaluated which are regularised linear models (Ridge, ElasticNet), tree-based ensembles (Random Forest, Extra Trees, AdaBoost, Histogram-based Gradient Boosting, LightGBM, XGBoost, CatBoost), kernel methods (support vector regression, kernel ridge), a multilayer perceptron, and a dummy mean baseline. Each candidate are making use of an identical preprocessing pipeline and tuned via Tree-structured Parzen Estimator search [12] inside a nested cross-validation design with separate inner-loop temporal and random folds. This allows the relative performance of different model families to be diagnosed rather than asserted.

Second, the improvement of the data transparency and the coverage. The underlying panel is expanded from 74 countries in the earlier study [1] to 282 entities from 2000 to 2024. This total

consists of 227 sovereign countries and dependent territories, plus 55 regional, income, institutional, and global aggregates. The supervised model evaluation uses the 123 countries with usable net-zero target pathways, while the final pledge benchmark is reported for 119 countries with an announced net-zero target . All preprocessing rules, domain-range constraints, outlier filters, and missingness thresholds are documented. Although the panel data extends to 2024, the main supervised modelling window ends in 2022. The year 2022 is used as the final holdout and baseline observation year (t_0) for reconstructing model-implied net-zero years. The reason is the less complete data for 2023 and 2024. The 2023–2024 rows are retained for documenting data availability and recent source coverage and they are not used in the primary baseline years for model evaluation or pledge benchmarking. Two parallel imputation branches are maintained throughout (a hierarchical median branch and a MICE branch [13]), so that the sensitivity of the downstream results to the imputation strategy can be quantified.

Third, to understand and improve the results in a way that is interpretable. The thesis treats policy interpretability as a central requirement. The analysis must produce projected net-zero years that can be explained to a policy audience. Accordingly, the SHAP interpretation focuses on substantively meaningful drivers, feature engineering is limited to transformations that can be interpreted easily rather than making purely statistical gains. The implied net-zero year is constructed through the model's predicted log-budget and with the individual country's projected 2030 emissions.

Fourth, to test the robustness of the results to the choices embedded in the framework. Three target variable scenarios are constructed by varying the power-law exponent that governs the post-2030 emission decay ($\alpha = 0.5$ for a pessimistic, slow early reduction; $\alpha = 1.0$ for a linear path; $\alpha = 2.0$ for an optimistic, rapid early reduction). A 200-iteration country-block bootstrap propagates training sample variability into the implied net-zero year which returns country-specific predictive intervals. Sensitivity analyses then compare results across CV strategies, imputation branches, and emission scenarios, so that the conclusions can be assessed against the methodological choices.

The thesis offers country-level implied net-zero years with bootstrap-based uncertainty predictive intervals for the countries with the announced pledges, together with a benchmark of these predictions against the commitments compiled by the Net Zero Tracker [14] and the Climate Pledge NDC tool of PBL [15]. Thus, the framework is best read as a pledge-benchmarking tool rather than as a precision forecasting system. Because the target variable is constructed from announced net-zero pledges, following the target-construction logic of den Elzen et al. [1], rather than from observed net-zero outcomes. The model does not predict when countries will actually reach net zero. Instead, it estimates the net-zero year implied by a country's characteristics relative to the relationships learned from other countries' pledge derived targets. The main policy relevant output is the gap between this model-implied net-zero year and the country's announced pledge. A negative, near-zero, or positive gap indicates whether the announced pledge is later than, close to, or earlier than the model-implied benchmark. This interpretation is discussed further in Section 6.4. Such relative pledge benchmarks can inform the broader assessment of national net-zero commitments in the Paris Agreement process [16, 17], but they should not be read as forecasts of actual policy delivery.

1.5 Roadmap

The remainder of the thesis proceeds as follows. Section 2 documents the seven public sources from which the panel is assembled and the construction of the three scenario-specific target variables, and sets out the inverse reconstruction of the implied net-zero year. Section 3 presents an exploratory analysis of the assembled dataset. Section 4 describes the preprocessing pipeline, the candidate

models, the nested cross-validation design, hyperparameter optimisation, overfitting diagnostics, winner selection, and the bootstrap uncertainty quantification scheme. Section 5 reports the empirical results, including model comparison, final 2022 holdout evaluation, sensitivity analysis, SHAP interpretation, predictive intervals, and the benchmark of model-implied net-zero years against announced pledges. Section 6 discusses the methodological implications, acknowledges the principal limitations, and concludes with directions for future research.

2 Data Collection

The empirical foundation of this study is a dataset covering 282 entities over the period from 2000 to 2024, constructing 7050 (282×25) entity–year observations before any filtering. Here, 'entities' refers to tracked sovereign countries and dependent territories, together with aggregate groups reported by the individual sources. Table 3 summarises this composition. The dataset is assembled from 9 public sources, each contributing a complementary dimension of entity level socioeconomic, environmental, governance, and energy characteristics. Data collection follows a standardised workflow comprising three stages, namely, acquisition of source materials, source-specific quality control and feature construction, and integration of all sources with the constructed target variables into a unified dataset.

Table 3: Composition of the 282-entity panel.

Entity type	**Count**	**Description**
Sovereign countries and dependent territories	227	Country-level records, including non-sovereign territories where source data use separate 3 letter codes, such as Bermuda, Cayman Islands, and Greenland.
Aggregate groups	55	Regional, income, institutional, and global aggregates, such as Arab World, Sub-Saharan Africa, High Income, and other group totals.

2.1 Data Sources

Table 4 provides an overview of the 9 data sources used in this study. Each source is described in detail below.

Table 4: Summary of data sources used in the study.

Source	**Provider**	**Coverage**	**Feature sets contributed**
PRIMAP-hist	PIK / Zenodo	2000–2024	GHG emissions (4 features)
World Bank database	World Bank	2000–2024	Economic & land features (13 features)
IEA Energy Balances	IEA / OECD	2000–2022	Energy mix & intensity (3 features)
INFORM Risk Index	JRC / EC	2015–2024	Disaster risk score
V-Dem	V-Dem Institute	2000–2024	Political regime type
RSF Press Freedom	RSF	2002–2024	Press freedom score
FAO / Data360	World Bank	2000–2022	Dietary protein supply
UN WPP 2024	UN DESA	2000–2100	Population projections (target construction)
NZ Targets Workbook	Custom / ClimateAnalytics	2023 data	Pledged net-zero years, 2030 NDC and current-policy emission projections

PRIMAP-hist. Greenhouse gas emission data are obtained from the PRIMAP-hist dataset [18], which provides entity level historical emissions derived from a combination of UNFCCC national

inventory reports and third-party data sources. For this study, the most recent release (version 2.7) of Zenodo at the time is used, and analysis is restricted to the HISTCR (historical country-reported) scenario. Four features are derived out of this source are, first, total GHG emissions per capita (t CO_2e/person) computed by summing all Kyoto gases under IPCC category 0 and dividing by country-level population, second, non-CO_2 emissions per capita (t CO_2e/person) which comprises of CH_4, N_2O, HFCs, PFCs, SF_6, and NF_3, third, year-on-year GHG growth rate (%), and lastly, hard-to-abate sector emissions per capita (t CO_2e/person) which is covering industrial processes, agriculture, transport, and waste (IPCC categories 1.A.3, 1.A.4, 2, and 3). Raw values are in Gg CO_2e and are converted to tonnes CO_2e per person.

World Bank 13 socioeconomic and environmental features are retrieved from the World Bank and the 2025 Worldwide Governance Indicators (WGI) workbook. Economic and land-use features are retrieved from the World Bank and span economic rents from fossil fuels (oil, natural gas, and coal rents as percentages of GDP), land use (agricultural land and forest area as percentages of total land area, population density), economic output (GDP per capita in constant 2015 USD and GDP at purchasing power parity (PPP) in constant 2021 international dollars), and annual deforestation rate. Governance quality indicators (Control of Corruption, Political Stability and Absence of Violence, Regulatory Quality, and Voice and Accountability) are sourced from the 2025 WGI workbook rather than the World Bank database to ensure access to the most current estimates. Total population is additionally retrieved for use in target variable construction but is excluded from model training features. All series are compiled as complete entity–year observations for all available entities.

IEA Energy Balances. Energy features are constructed from the International Energy Agency's world energy balances. The source data are restructured into a entity–year format, and three features are computed. First, renewable energy share of total final consumption (TFC, %) is calculated by combining modern renewables in electricity, heat, and direct TFC components. Second, renewable electricity share (%), is defined as the ratio of renewable electricity output to total electricity output. Lastly, energy intensity (MJ per USD PPP) is computed as total energy supply (TJ, converted to MJ) divided by GDP at purchasing power parity. Fourteen renewable product categories are recognised to compute renewable energy share and renewable electricity share. These 14 renewable product categories are hydro, wind, solar, geothermal, tidal, primary solid biofuels, liquid biofuels, biogases, renewable municipal waste, and charcoal.

INFORM Risk Index. The INFORM Risk Index [19] is measure of a country's risk of humanitarian crisis. It is produced by the European Commission's Joint Research Centre. It integrates three sub-dimensions, namely, hazard and exposure, vulnerability, and lack of coping capacity. The index ranges from 0 (lowest risk) to 10 (highest risk). In this study, the INFORM Risk Index is retrieved from the annual INFORM trend dataset covering reference years from 2015 to 2024. Prior to model year 2015, the INFORM index is unavailable, resulting in the relatively high missingness reported in Table 6.

V-Dem Political Regime. Regime type data are obtained from the Varieties of Democracy (V-Dem) project [20] which is a large-scale, expert coded political dataset. The main feature used is the V-Dem regime classification in which it assigns each entity–year observation to one of four categories. Those categories are closed autocracy (0), electoral autocracy (1), electoral democracy

(2), and liberal democracy (3). This ordinal measure is represented as a categorical predictor in the modelling stage.

RSF Press Freedom Index. The Reporters Without Borders (RSF) Press Freedom Index defines the degree of freedom available to journalists in each country on a scale from 0 to 100 scale in which 100 is being the best. Annual entity-level observations are collected from RSF for the years between 2002 and 2024. One important observation is that the year 2011 is unavailable and is imputed using the 2012 value. Another important point is that scores for 2002–2012 used an inverted scale in which lower was bette) and they are normalised using min-max scaling to align with the 2013–2024 convention.

FAO Dietary Protein Supply. FAO Food Security features is accessed through the World Bank Data360 platform which provides average dietary protein supply of animal origin (g/capita/day, three-year average). The feature reflects dietary quality and served as an indirect proxy for livestock activity.

Population Projections and Net-Zero Target Workbook. Two additional datasets are used exclusively in construction of the target variable (Section 2.3) and they are not used as model features. Population projections between 2024 and 2100 are drawn from the UN World Population Prospects 2024 workbook's medium variant [21], which are providing country-level annual population estimates that are used to normalise emission pathways to a per-capita basis. Net-zero commitment years and projected 2030 emission levels are retrieved from a curated dataset in which combines nationally pledged net-zero target years from the Net Zero Tracker [14] with 2030 emissions projections for both the NDC scenario and the 'current policies' scenario from den Elzen et al. (2025) and the PBL Climate Pledge NDC tool [15, 22].

2.2 Data Processing

Each data source obtained from databases and workbooks are harmonised before integration into a common panel. The harmonisation steps includes standardising country identifiers to ISO 3166-1 alpha-3 codes, validating numeric fields, and constructing derived features where it is required.

The harmonised sources are then integrated by entity and year. To ensure temporal and cross-entity comparability, the merged data are aligned to the full entity–year grid of 282 countries, territories, and aggregate entities over 25 years. The missing observations are preserved explicitly as missing values rather than removing through row deletion. The final panel is restricted between years 2000 and 2024, and to valid 3-letter ISO codes.

2.3 Target Variable Construction

The target variable for each scenario captures the log transformed cumulative per capita GHG budget available to a country between the observed year t and its pledged net-zero year T_{NZ}. The calculation proceeds in four steps. Appendix A provides a complete worked example for Iceland in which it shows both the emission pathway construction and the year by year calculation of the target variable.

Step 1: Emission pathway generation. For each country in the panel dataset with a pledged net-zero year and a projected 2030 emission level, an annual emission trajectory is calculated from the country's most recent observed total GHG emission level at 2024 (denoted E_{t_0}) through to the

net-zero year ($E_{T_{\mathrm{NZ}}} = 0$). The 2030 emission level is taken from the country's NDC unconditional target as the primary emission level, in case of its absence, the current policy 2030 median used as a fallback. The pathway is composed of two segments:

1. Linear segment (2024–2030): emissions decline linearly from E_{t_0} to the 2030 projection E_{2030}.
2. Power-law segment (2030–T_{NZ}): emissions follow a power-law decay,

$$E_t = E_{2030}\left(1 - \frac{t - 2030}{T_{\mathrm{NZ}} - 2030}\right)^{\alpha}, \quad t \in [2030, T_{\mathrm{NZ}}], \tag{1}$$

where the exponent α defines the shape of the trajectory and below are the three scenarios considered in this study:

- Optimistic ($\alpha = 2.0$) which assumes a rapid early reductions followed by a slower tail.
- Linear ($\alpha = 1.0$) which assumes a constant rate of decline.
- Pessimistic ($\alpha = 0.5$) which assumes a slow initial reductions with rapid acceleration near the net-zero year.

Step 2: Per-capita normalisation. Annual emission values are divided by entity level population projections from the UN WPP (World Population Prospects) 2024 medium variant by their corresponding year, which then converted to tonnes CO_2e per person per year. This is applied both to the future trajectories and, to historical emission levels that are obtained through PRIMAP total GHG emissions.

Step 3: Backwards cumulative aggregation. Starting from T_{NZ} and working backwards to $t = 2000$, per-capita emissions are summed to return the cumulative per-capita GHG budget at each historical year t:

$$B_t = \sum_{\tau=t}^{T_{\mathrm{NZ}}} e_{\tau}^{\mathrm{pc}}, \tag{2}$$

where e_{τ}^{pc} is the per-capita emission in the corresponding year τ. This amount equals to the total per-capita emission budget remaining from year t until net zero is achieved. Larger values of B_t shows that a country has a larger remaining emission relative to its population.

Step 4: Logarithmic transformation. To lower the right skewness and to make it comparable across countries, the natural logarithm is applied:

$$y_t = \ln(B_t), \quad B_t > 0. \tag{3}$$

Observations with $B_t \leq 0$ are treated as missing and removed from the model estimation since these are the countries with negative GHG emissions.

The effect of the logarithmic transformation is illustrated in Figure 1. The cumulative per-capita budgets are clearly right skewed, with a small number of entity–year observations shows a very large remaining budgets. After applying the natural logarithm, the distribution becomes more symmetric and reduces the influence of the extreme values.

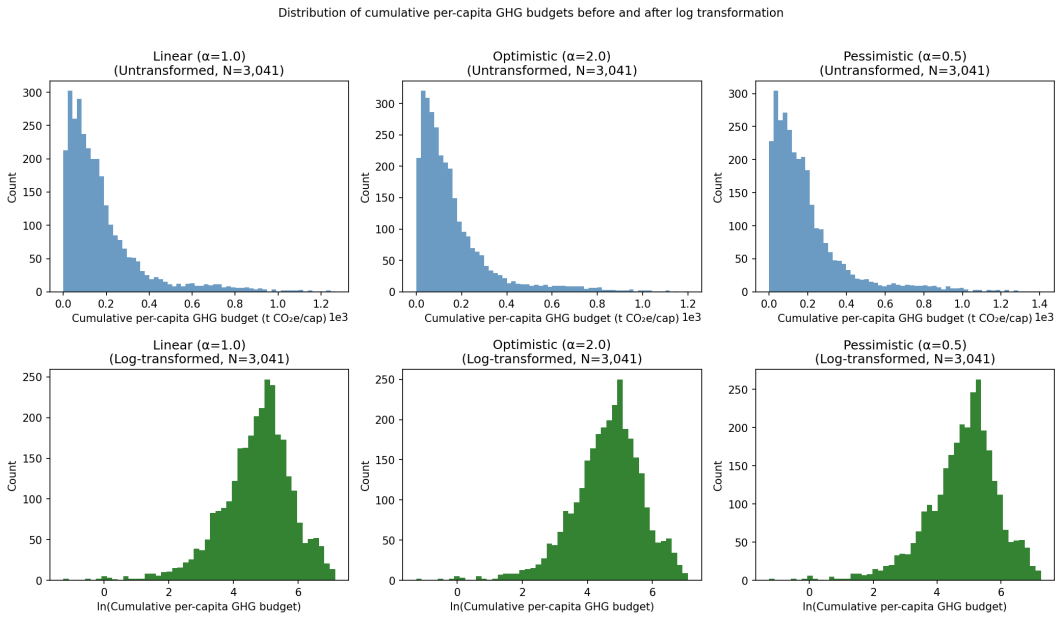

Figure 1: Distribution of cumulative per-capita GHG budgets before and after logarithmic transformation. The log scale reduces right-skewness and compresses extreme budget values.

2.4 Inverse Reconstruction of Net-Zero Years

Because the predictive models are trained on the log transformed cumulative per capita GHG budget, the model's predicted log budget (denoted $\hat{y}_{t_0}$) must be reconstructed back to an implied net-zero year T^*_{NZ}. Since the cumulative per capita budget function $B_{t_0}(T)$ is strictly monotonically increasing relative to the target net-zero year T , a unique root exists and can be reconstructed in four steps.

Step 1: Objective function formulation. For a given country and scenario, the continuous implied net-zero year T^* is defined as the value that satisfies the following mapping:

$$\ln\left(B_{t_0}(T^*)\right) = \hat{y}_{t_0}, \tag{4}$$

where $B_{t_0}(T^*)$ is the cumulative per capita budget starting from the baseline year t_0 up to an arbitrary target year T^* which is calculated using the identical pathway logic described in Section 2.3. This is solved by finding the root of the objective function:

$$f(T) = \ln\left(B_{t_0}(T)\right) - \hat{y}_{t_0} = 0. \tag{5}$$

Step 2: Boundary conditions and clamping. The numerical root-finding is bounded within a search interval $[T_{\text{lo}}, T_{\text{hi}}]$, where the lower bound is set to $T_{\text{lo}} = t_0 + 2$ years (2024) and the upper bound is set to a long-term ceiling of $T_{\text{hi}} = 2200$.

Step 3: Root optimization via Brent's method. For intermediate predictions where the boundary signs differ ($f(T = 2024) < 0 < f(T = 2200)$), the root T^* is solved numerically using Brent's method.

Step 4: Discrete year rounding. To get a discrete year that can be compared to national policy pledges, the continuous root T^* is rounded to the nearest integer year:

$$T^*_{\mathrm{NZ}} = \lfloor T^* + 0.5 \rfloor. \tag{6}$$

3 Data Analysis

This section presents an exploratory analysis of the panel dataset. It describes the entity-year panel structure, look into missing data patterns, summarises the distributional properties of each feature and the three target variables calculated at section 2.3, and analyse relationships among the features.

3.1 Panel Structure

The merged dataset is a balanced panel with 7050 entity–year observations with 282 entities over 25 periods (2000–2024). Because the panel grid is balanced by construction (each entity has a row for every year), the total observation count equals to 282×25. This panel is retained for descriptive feature analysis. However, the supervised model estimation makes only use of entities with a valid constructed target variable.

Only 124 entities have valid cumulative budget target pathways. The remaining 158 entities are removed from the supervised modelling because they fail at least one target-construction requirement. There are 65 individual entities with no announced net-zero target year. 81 entities are regional aggregates, territories, or microstates without the necessary independent target inputs. 18 entities are net-negative carbon sinks. Lastly, Mozambique (MOZ) is missing 2030 NDC or current policy target input. There is a slight overlap in categories, but together explain the reduction from 282 entities to 124. Examples of countries without announced net-zero targets include Afghanistan (AFG), Albania (ALB), Azerbaijan (AZE), Bosnia and Herzegovina (BIH), Belarus (BLR), and Botswana (BWA). Examples of excluded aggregates and territories include Arab World (ARB), High Income (HIC), Sub-Saharan Africa (SSF), Bermuda (BMU), Cayman Islands (CYM), and Greenland (GRL). Examples of net-negative carbon-sink countries include Bhutan (BTN), Gabon (GAB), Guyana (GUY), Madagascar (MDG), Panama (PAN), and Venezuela (VEN).

3.2 Feature Set

After merging, the dataset contains 24 raw input features, three target variables and the panel identifiers (country and year). The features span five thematic domains:

1. GHG emissions (PRIMAP-hist): total GHG per capita, non-CO_2 GHG per capita, year-on-year GHG growth , and hard-to-abate sector per capita emissions.
2. Energy (IEA): renewable energy share of TFC, renewable electricity share, and energy intensity.
3. Economy & resources (World Bank API): GDP per capita, GDP at PPP (constant 2021 international USD), population density, oil rents, natural gas rents, coal rents.
4. Land use & food (World Bank API, FAO): agricultural land, forest area , annual deforestation, and average dietary protein supply of animal origin.

5. Governance & institutions (World Bank WGI, INFORM, RSF, V-Dem): Control of Corruption, Political Stability, Regulatory Quality, Voice and Accountability, INFORM Risk Index, Press Freedom Index, and political regime type.

Table 5 presents descriptive statistics for all continuous features.

Table 5: Descriptive statistics of input features directly used in model training.

Feature	Unit	N	Mean	Std	Min	P25	Median	P75	Max
Per Capita GHG	t CO_2e/cap	4975	4.363	17.39	−189.04	1.904	4.230	8.380	105.06
Non-CO_2 GHG per capita	t CO_2e/cap	4975	0.384	0.807	0.001	0.047	0.145	0.390	10.46
Hard-to-Abate per capita	t CO_2e/cap	4975	1.240	1.746	0.000	0.383	0.753	1.406	20.50
GHG Growth	% YoY	4992	1.174	61.63	−2152	−2.423	0.973	4.607	2414
Renewable Energy Share	%	3522	30.19	28.39	0.000	7.042	20.53	47.59	98.34
Renewable Electricity Share	%	3600	33.30	32.12	0.000	5.152	22.29	57.22	100.00
Energy Intensity	MJ/USD PPP	3335	4.481	3.085	0.000	2.821	3.723	5.181	29.02
GDP per Capita	2015 USD	6176	14669	21664	233.03	1893	5583	18573	247170
Population Density	per km^2	6056	358.98	1733	0.136	33.81	76.13	180.36	21530
Oil Rents	% GDP	5173	3.811	9.000	0.000	0.000	0.084	2.464	82.78
Natural Gas Rents	% GDP	5196	0.622	2.452	0.000	0.000	0.007	0.332	55.01
Coal Rents	% GDP	5168	0.290	1.696	0.000	0.000	0.000	0.079	48.72
Agricultural Land	%	5945	37.29	21.04	0.436	20.00	38.24	50.15	86.48
Forest Area	%	6008	32.30	23.46	0.000	11.59	30.90	47.78	95.58
Annual Deforestation	%	216	0.553	1.234	−3.580	0.000	0.288	1.037	6.135
Protein Supply	g/cap/day	3709	38.45	21.63	2.500	19.20	36.00	56.30	111.90
Control of Corruption	index	4998	−0.003	0.996	−2.065	−0.779	−0.191	0.760	2.397
Political Stability	index	5050	0.008	0.991	−3.052	−0.629	0.087	0.864	1.760
Regulatory Quality	index	4971	0.016	0.995	−2.581	−0.726	−0.106	0.809	2.260
Voice & Accountability	index	5041	0.011	0.991	−2.361	−0.802	0.047	0.858	1.984
INFORM Risk Index	score	1719	3.813	1.710	0.600	2.500	3.500	4.900	8.800
Press Freedom Index	score	3981	67.65	20.44	0.000	56.18	70.46	82.90	100.00
Regime Type	class	4457	1.583	0.993	0.000	1.000	2.000	2.000	3.000

3.3 Missing Data

Table 6 reports the missing percentage for each feature. Missingness follows two broad patterns: structural gaps, where a series is unavailable for specific countries or periods (e.g., INFORM before 2015), and reporting gaps.

Table 6: Missing data by feature, sorted by missingness rate.

Feature	Present obs.	Missing (%)
Annual Deforestation	216	96.9
INFORM Risk Index	1719	75.6
Energy Intensity (MJ/USD PPP)	3335	52.7
Renewable Energy Share (%)	3522	50.0
Renewable Electricity Share (%)	3600	48.9
Protein Supply (g/cap/day)	3709	47.4
Press Freedom Index	3981	43.5
Regime Type (V-Dem)	4457	36.8
Regulatory Quality (WGI)	4971	29.5
Total GHG Emissions	4975	29.4

Feature	Present obs.	Missing (%)
Non-CO_2 Emissions	4975	29.4
Hard-to-Abate Emissions	4975	29.4
GHG Growth (% YoY)	4992	29.2
Control of Corruption (WGI)	4998	29.1
Voice & Accountability (WGI)	5041	28.5
Political Stability (WGI)	5050	28.4
Coal Rents (% GDP)	5168	26.7
Oil Rents (% GDP)	5173	26.6
Natural Gas Rents (% GDP)	5196	26.3
GDP at PPP	5934	15.8
Agricultural Land (%)	5945	15.7
Forest Area (%)	6008	14.8
Population Density (per km^2)	6056	14.1
GDP per Capita (2015 USD)	6176	12.4

The annual deforestation has the most missing (96.9%), because the feature is reported for only a limited number of countries. The INFORM Risk Index has 75.6% missingness since it is only available from 2015 onward. IEA energy features (49–53% missing) shows coverage gaps in the IEA extended energy balances for smaller and low income countries. Governance features have approximately 29% missingness, primarily in early years (pre-2002) and non-sovereign territories. The most complete features are GDP per capita (12.4%), population density (14.1%), forest area (14.8%), and agricultural land (15.7%).

3.4 Target Variable Analysis

The three target variables represent log-transformed cumulative per-capita GHG budgets under different emission reduction pathway assumptions. All three are calculated for the same set entity–year observations across countries with valid inputs.

Table 7 presents summary statistics and normality tests for each target. The distributions are moderately left skewed, with skewness values ranging from -0.800 (optimistic) to -0.926 (pessimistic). This indicates a modest excess of low-budget observations relative to a normal distribution. Excess kurtosis ranges from 1.907 to 2.102 over the three scenarios. This suggests heavier tails than the normal distribution. A Shapiro Wilk test rejects normality for all three targets ($p < 0.001$), consistent with the observed skewness and kurtosis.

Table 7: Summary statistics and normality tests for the three log-transformed target variables.

Scenario	Unit	N	Mean	Std	Min	Median	Max	Skewness	Ex. Kurtosis	Shapiro p
Linear	ln(t CO_2e/cap)	3041	4.710	1.101	−1.261	4.843	7.143	−0.873	2.035	< 0.001
Optimistic	ln(t CO_2e/cap)	3041	4.610	1.084	−1.261	4.725	7.090	−0.800	1.907	< 0.001
Pessimistic	ln(t CO_2e/cap)	3041	4.796	1.119	−1.261	4.948	7.249	−0.926	2.102	< 0.001

The three scenarios are highly correlated (Pearson $r > 0.99$), as they differ only in the shape of the 2030–NZ trajectory and share the identical 2024–2030 linear segment.

3.5 Feature–Target Correlations

Total GHG emissions per capita shows the strongest positive correlation with the targets ($r = 0.675$). The strongest negative correlates are the INFORM Risk Index ($r = -0.458$) and renewable energy

share ($r = -0.430$) which indicates that countries with higher disaster risk or higher renewable share tend to have smaller remaining cumulative budgets. Regulatory quality ($r = 0.428$), protein supply ($r = 0.407$), and control of corruption ($r = 0.343$) also show positive correlation which suggests that wealthier, better-governed countries tend to have higher remaining budgets.

Figure 2 presents the full pairwise correlation matrix among input features. There are several high-collinearity pairs. The four Worldwide Governance Indicators are strongly correlated ($r > 0.75$). Total GHG, non-CO_2, and hard-to-abate emissions are similarly collinear. GDP per capita and protein supply are strongly correlated ($r \approx 0.77$).

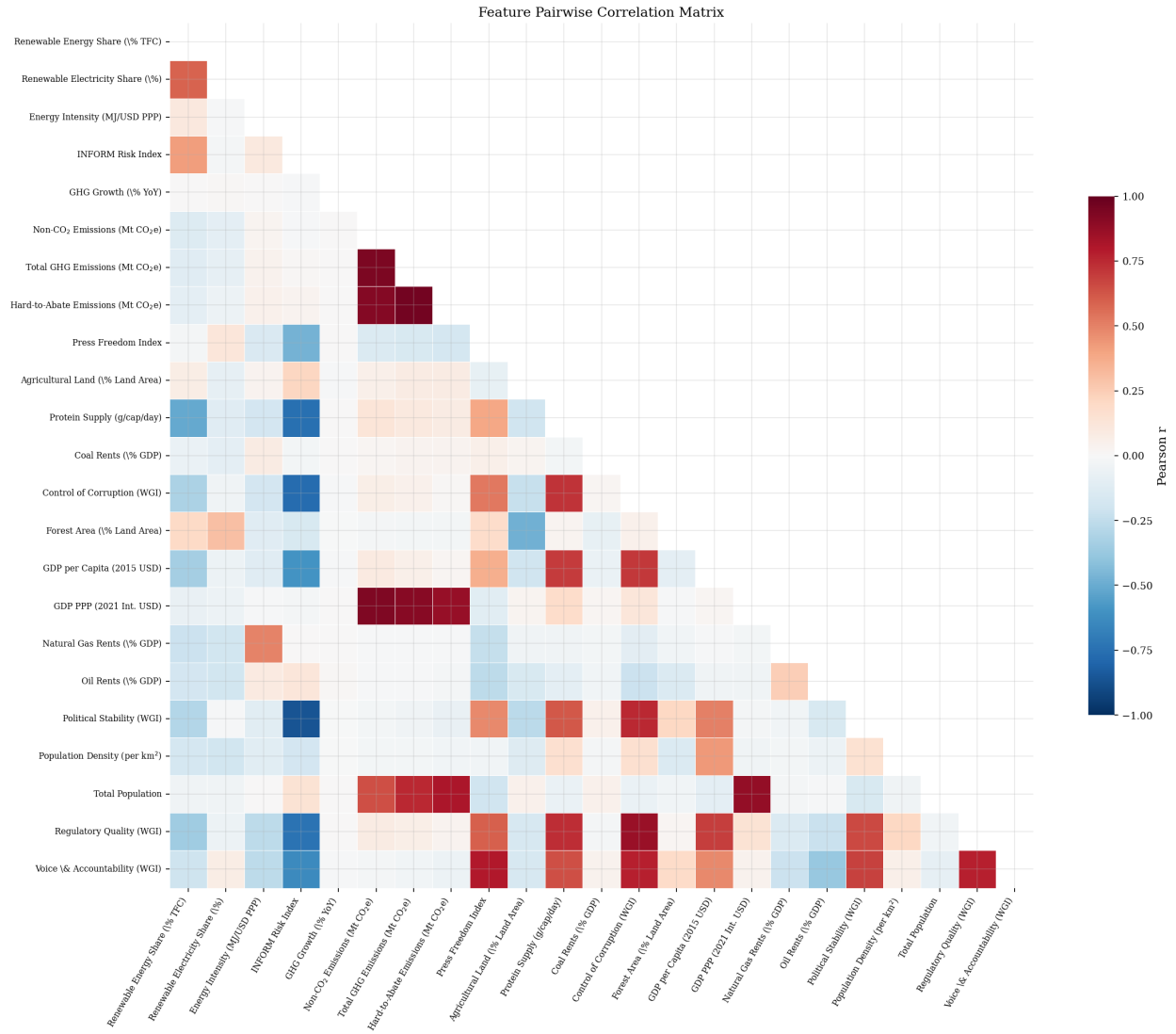

Figure 2: Pairwise Pearson correlation matrix for all continuous input features. Red indicates positive correlation; blue indicates negative correlation. The lower triangle only is shown.

3.6 Political Regime Distribution

Figure 3 shows the distribution of entity–year observations over the four V-Dem regime categories for the 4 457 observations with non-missing regime data. Electoral autocracies and electoral democracies each account for roughly 1 300–1 500 observations, while liberal democracies contribute ap-

proximately 1 000 observations and closed autocracies the fewest (around 700). The distribution reflects the global political landscape of the 2000–2024 period, during which a significant share of countries maintained competitive elections while falling short of full liberal democracy.

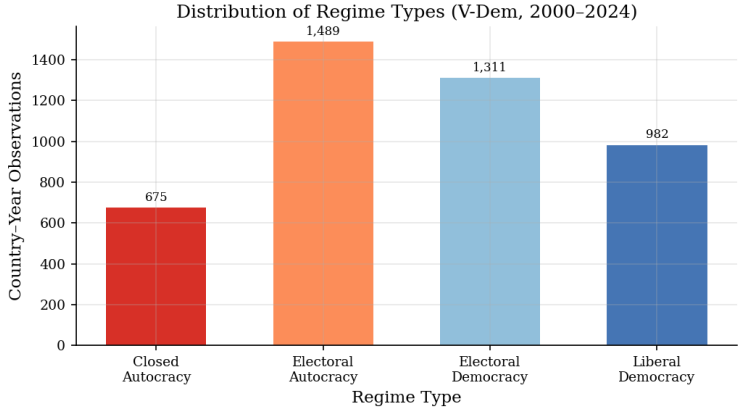

Figure 3: Distribution of country–year observations across the four V-Dem political regime types for the years 2000–2024.

3.7 Renewable Energy Trends by Regime

Figure 4 contains two median renewable energy deployment across regime categories. Panel (a) shows renewable energy as a share of total final consumption (TFC), while panel (b) shows renewable electricity as a share of total electricity output. The TFC panel changes gradually, indicating that renewables contributes final energy consumption more slowly because transport, heating, and industrial fuel use remain harder to decarbonise. The renewable-electricity panel shows an upward shift after approximately 2010.

Across the figure, electoral autocracies occupy the upper part for much of the period. This indicates relatively high median renewable shares compared with the other regime groups. The main change appears in the renewable electricity share during the last five to seven years, when liberal and electoral democracies move upward more and overtake the earlier electoral-autocracy share. Closed autocracies remain closer to the lower part of the figure overall, but they also show a significant increase in renewable electricity share during the last five years. This is an indication to the recent renewable power growth is not limited to more democratic regime types. The figure therefore suggests a common global increase in renewable deployment.

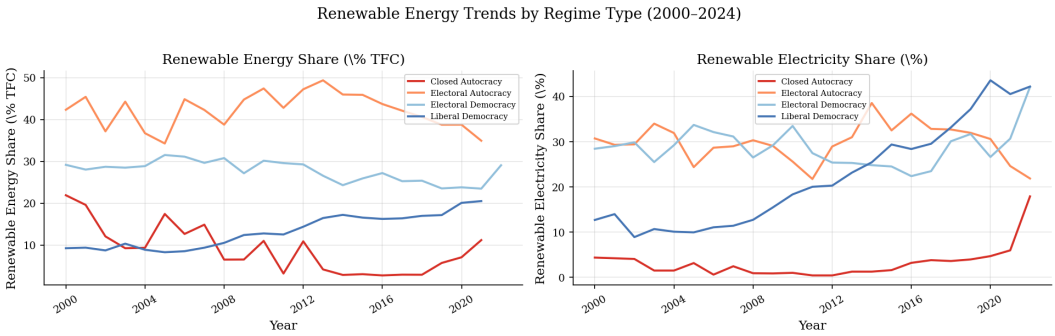

Figure 4: Median (a) renewable energy share of TFC and (b) renewable electricity share by V-Dem regime type, 2000–2024.

4 Methodology

This section describes the methodology employed to predict the log transformed cumulative per capita GHG budget introduced in Section 2.3. It consists of two principal stages. First, preprocessing pipeline that transforms the raw panel data into model-ready feature matrices. Second, a training pipeline that covers candidate model evaluation, hyperparameter optimisation, overfitting control, winner selection, and bootstrap uncertainty quantification. All preprocessing transformations are fitted only on training partitions to prevent information leakage from validation or test sets.

4.1 Preprocessing

The preprocessing pipeline transforms data into a set of model feature matrices through four sequential phases, namely, data cleaning, missing-value imputation, feature engineering, and feature selection. Two imputation branches (referred to as the simple branch and the MICE branch) are constructed for the purpose of evaluating the sensitivity of results to the imputation strategy.

4.1.1 Cleaning

In the cleaning phase, observations and features that are identified structurally inadequate for modelling are removed. It is applied independently to each training partition. Then, the cleaning is applied identically to validation and test partitions.

Domain-range enforcement. Each continuous feature is checked for range constraints informed by its physical or statistical definition. For example, renewable energy shares are bounded to $[0, 100]\%$, governance features to $[-4, 4]$, and total GHG per capita emissions to $(0, 500)$ Mt CO_2e. Values that resides outside these bounds are flagged missing rather than removal so that they enter the imputation stage. These domain constraints are the primary filter for structural country-profile variables, such as resource rents, GDP per capita, population density, land-use shares, renewable shares, and per-capita emissions. It is highly probablr that these valid extreme values often represent meaningful country signals rather than measurement error. In the assembled data, this rule flags one boundary violation. Albania's 2021 renewable electricity share slightly exceeds 100 % because of floating-point precision..

Outlier treatment. Outliers are identified using a feature aware and grouped interquartile range (IQR) procedure following Tukey [23]. Country-level IQR bounds are calculated within each country's historical time series. Similarly, year-level IQR bounds are calculated cross-sectionally within each year. An observation is considered IQR-extreme, if it is more than $1.5 \times$ IQR below the first quartile or above the third quartile of the relevant group.

The calculated IQR bounds are not applied uniformly across all features. For structural country features, year-level IQR flags are not used to coerce values to missing. This prevents legitimate country signals, such as high resource rents in hydrocarbon exporters, high population density in city states, or high GDP per capita in wealthy economies, from being mistaken for data errors. For these variables, domain-range enforcement provides the destructive check.

For volatile rate, index, and governance variables, IQR filtering is applied more conservatively. A value is coerced to missing only when it is flagged by both the country level and year level IQRs. Thus, an unusual value is not removed merely because a country is structurally different from others in the same year. It must also show a relatively inexplicable value to that country's own historical distribution. Values removed by the rule are set to missing rather than removing, preserving the row for subsequent imputation.

In the final training data with valid net-zero targets, this outlier treatment has limited effect on its scope. 103 cells out of 57 201 continuous-feature cells (approximately 0.18 %) are coerced to missing and affecting 103 rows out of 2487 training rows. The affected variables are concentrated in two features, namely, annual GHG growth (about 96.4 % of removals) and the RSF Press Freedom Index (about 3.6 %). Examples include extreme GHG-growth values for Sweden in 2012, Mongolia in 2012, Fiji in 2006, and Chile in 2003, as well as unusually low press-freedom values for Laos in 2003, Nepal in 2005, and Libya in 2005.

Row-level missingness filtering. Observations with more than 30 % of features are missing are excluded from model training since very sparse rows carry insufficient information for reliable feature engineering and imputation. Observations with missing target values are excluded. To preserve temporal continuity for the rolling-window computations described in Section 4.1.3, flagged rows are kept as ghost rows during the feature engineering phase and removed only after all temporal features have been computed.

Of the 3041 country-year observations with valid log-budget targets, 549 observations (18.1 %) are excluded from model training because more than 30 % of their candidate predictors are missing. The filter affects all 124 countries with valid net-zero targets. Each country loses at least two observations. The exclusions are not randomly distributed. In the early historical observations, especially 2001, observations are removed for 122 countries since governance and socioeconomic indices, including the INFORM Risk Index, the Reporters Without Borders Press Freedom Index, and WGI governance estimators, are not available for those years. In the most recent prediction years, observations are removed for 123 countries in 2023 and 102 countries in 2024. This reflects international database reporting lags from sources such as the World Bank and IEA. Features including renewable energy share, energy intensity, and fossil-fuel or resource rents are typically released with a two to three year delay. A small set of microstate or island cases is affected more severely. Andorra (AND), the Marshall Islands (MHL), and Tuvalu (TUV) are removed for all 25 years from 2000 to 2024 because data for these states are largely unavailable with average feature missingness above 50 %.

4.1.2 Imputation

Missing values are handled through two parallel imputation strategies. Maintaining two branches allows the training pipeline to assess the impact of imputation choice on prediction quality; separate winner models are selected for each branch.

Prior to imputation, a set of country embedding features is constructed to provide imputers with a compact representation of each country's profile. For each country, the per-feature median value and missingness rate are computed from the training data to create a per-country summary matrix. This matrix is standardised and subjected to Principal Component Analysis [24], with five components. The resulting components are added as permanent synthetic features and capture latent dimensions, such as income level, energy structure, and institutional capacity. Country embeddings allow models to identify the country-level factors not captured by the observed features alone. The year is additionally standardised to zero mean and unit variance and added both as a linear term $z_t = (t - \bar{t})/\sigma_t$ and as a quadratic term z_t^2 to enable models to capture nonlinear temporal trends.

Simple branch: hierarchical median imputation. The simple branch fills each missing value using a three-level hierarchical fallback. First, the training-set median for the respective country. If unavailable, second, the training-set median for the respective calendar year. Lastly, the global training-set median. This strategy is well suited to since it respects structural differences across countries while degrading to aggregate summaries when country-level information is insufficient [25]. For the categorical regime-type variable, mode rather than median imputation is applied at the same three levels.

MICE branch: iterative chained equations. The MICE branch employs Multiple Imputation by Chained Equations [13]. MICE cycles iteratively through a system of univariate regression models, one per feature with missing values, where each incomplete feature is predicted from the remaining columns in the imputation matrix. MICE captures multivariate dependencies among features, which is valuable when several correlated indicators are missing at the same time. Before imputation, synthetic ghost rows are added for missing years within each country's historical time series so that the imputer observes a temporally continuous panel. These rows support rolling window consistency.

The MICE design matrix has three groups of predictors. The first group has the raw continuous features after cleaning, including the original numerical indicators, country-level principal-component embedding features, and two standardised calendar-year features. The second group consists of panel-aware auxiliary predictors. These are created for each continuous or synthetic feature. A complete auxiliary baseline column is constructed using the hierarchical fallback country median, calendar-year median, and global median. These auxiliary columns provide stable and informed reference values that guide MICE away from implausible imputations in sparse country-year regions. The third group is the electoral-regime variable, which is first filled by a mode imputer and then entered as an ordinal numeric predictor.

Features with a missingness rate exceeding 5 % additionally receive binary indicators before MICE is fitted. These flags allow the imputer to be aware of systematic missing-data patterns [13]. Because iterative regression imputation is sensitive to feature scale, all columns in the MICE input matrix are standardised using training-set means and standard deviations before fitting. The imputer is run for a maximum of 20 iterations with convergence tolerance 10^{-3}. After convergence, the completed matrix is transformed back to the original units. Any values that remain missing

after the iterative procedure are filled by a global numeric median fallback as a final safety check, and the categorical regime variable is rounded to nearest integer within the accepted range.

Empirical comparison of imputation branches. The simple hierarchical median branch provides a stable and transparent fallback, but repeated assignment of country, year, or global medians can compress variance and create artificial probability mass. This is visible for renewable electricity share in Figure 5, where the simple branch concentrates imputed values around common fallback medians, whereas the MICE branch produces a smoother distribution closer to the observed density.

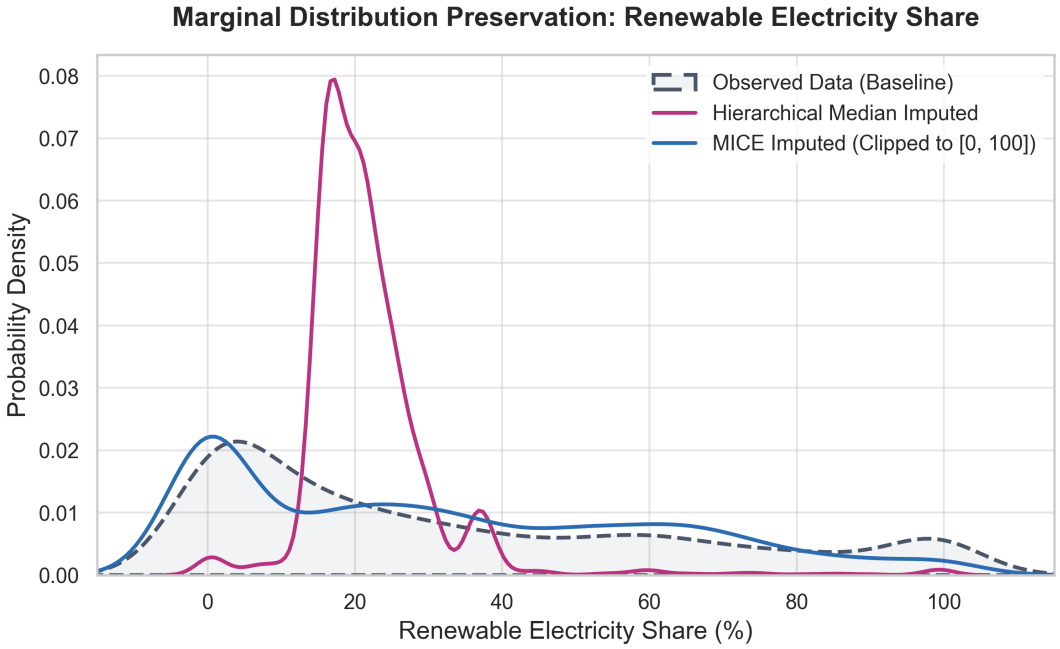

Figure 5: Marginal distribution comparison for renewable electricity share. The dashed curve denotes observed non-missing training values; the solid curves show imputed values from the hierarchical median and MICE branches.

A second diagnostic compares the reconstructed joint distribution of GDP per capita and energy intensity in Figure 6. The simple branch can create visible grid patterns because many missing values are mapped to identical fallback medians, weakening the structure between correlated indicators. the MICE branch places imputed observations closer to the observed data and better preserves the nonlinear association between economic scale and energy use.

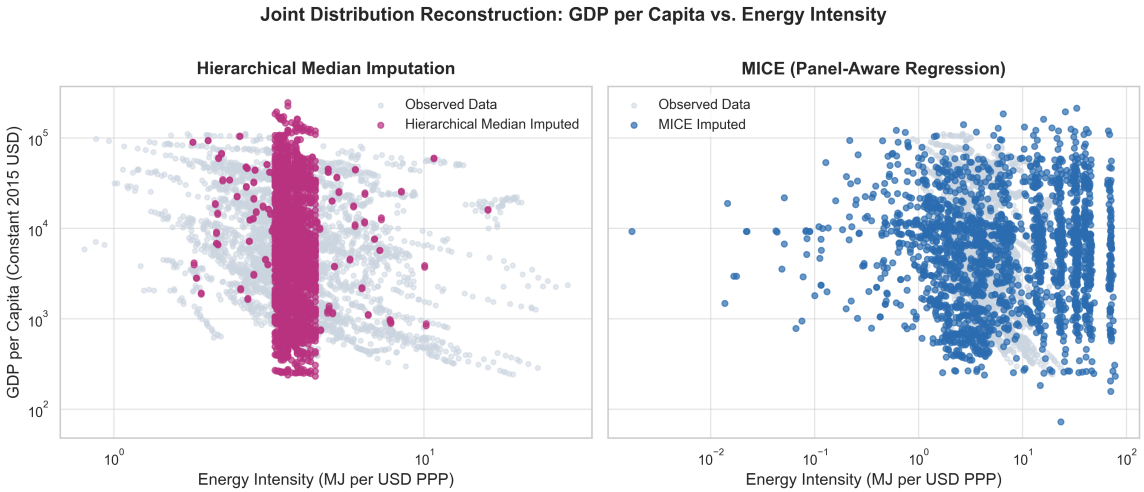

Figure 6: Joint distribution comparison for GDP per capita and energy intensity. The two panels compare hierarchical median and MICE imputations.

Andorra's Press Freedom Index is in Figure 7. For long country-specific gaps, hierarchical medians can introduce flat temporal plateaus that would are close to rolling-window features. In contrast, the MICE branch uses country embeddings, calendar-year terms, and correlated institutional features to reconstruct a smoother time path, reducing the risk that engineered temporal features are driven by imputation rather than plausible country dynamics.

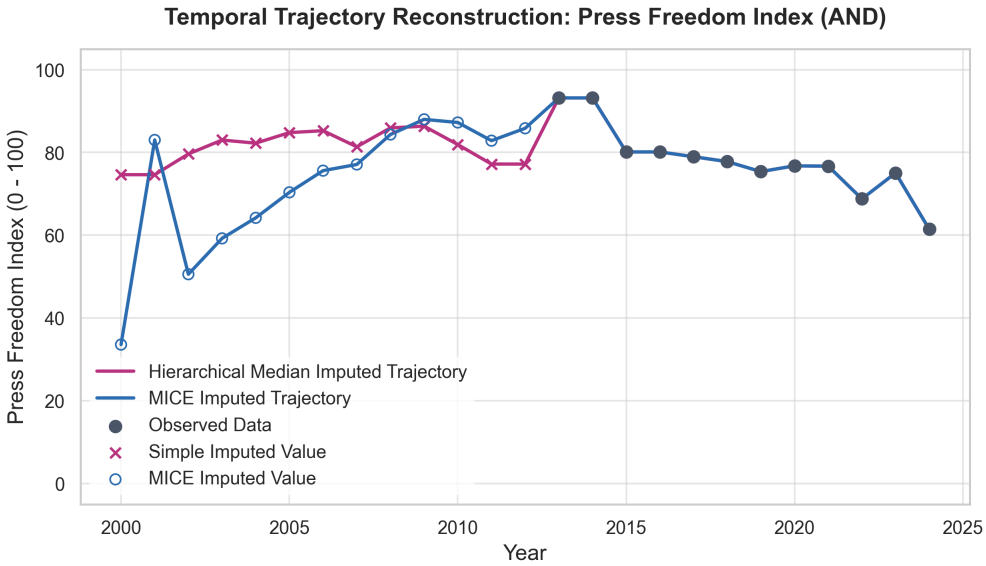

Figure 7: Temporal imputation comparison for Andorra's Press Freedom Index. Observed values are shown against the trajectories reconstructed by the hierarchical median and MICE branches.

4.1.3 Feature Engineering

Feature engineering transforms the imputed data into a richer representation by adding temporal and cross-variable features. All temporal features are computed from the imputed values, using the historical context of the training partition.

Seven-year rolling mean features. A seven-year backward rolling mean is added for each continuous feature. Rolling averages smooth year specific noise and provide a medium-term trend signal that is more stable than either the current level or the annual first difference [26]. A window of seven years is used to capture longer structural trends in emission-relevant indicators.

Rolling standard deviation features. Rolling standard deviations over three-year and five-year backward windows are added for each continuous feature. These features capture the recent variability of each indicator to complement the mean-based trend signals about the information about stability and volatility. For example, the rolling standard deviation of the GHG growth rate reflects whether a country's emissions trajectory has been erratic or consistent, providing additional signal beyond the level and trend alone.

Interaction terms. All pairwise product interactions are formed from all continuous features after imputation. These interaction terms capture multiplicative relationships between indicators that are not representable by additive terms alone. For instance, the product of total GHG emissions per capita and the INFORM Risk Index reflects the joint exposure of high emissions and high humanitarian risk. The interaction of energy intensity and GDP per capita distinguishes energy intensive high income economies from energy efficient ones. Expanding to all pairwise combinations avoids imposing prior assumptions about which cross-variable effects are theoretically relevant, leaving it to the feature selection step to identify informative interactions from the data.

Structural fill for temporal boundaries. Rolling features are structurally missing for the first observations of each country's time series. Such structural missingness has no information and is filled with zero. This prevents temporal boundary positions from propagating missing values into the model.

Encoding and scaling. The ordinal regime type variable is one-hot encoded using a fixed encoding fitted on the training data [11]. All continuous features are standardised to zero mean and unit variance using a Z-score (standard) scaler, fitted exclusively on the training partition and applied to validation and test partitions. Standardisation ensures that features with very different absolute magnitudes, such as total population ($\sim 10^8$) and the INFORM index (0–10), provides comparable influence on regularised and distance-based estimators.

4.1.4 Feature Selection

The complete feature engineering pipeline expands the feature space to several hundred columns, comprising raw features, seven-year rolling means, rolling standard deviations, interaction terms, country PCA embeddings, year terms, and encoded categories. Including all features increases risk of overfitting and inflates computational cost without improving generalisation. Thus, a feature selection step is added to retain the top $k = 24$ features with most predictive of the target variable.

Selection is performed using Recursive Feature Elimination (RFE) [27] with an Extra Trees regressor as the estimator. RFE iteratively fits the model on the current feature, ranks features by their contribution to the fitted model, and discards the least informative feature at each step, continuing until the desired number of features remains. Using Extra Trees as the base estimator provides feature importances that are sensitive to nonlinear interactions. It is robust to feature scaling, making it well suited to the heterogeneous feature space generated by the engineering step. Unlike filter-based selectors, RFE directly accounts for feature redundancy. A feature that

is informative by itself may be discarded if another feature in the set captures the same predictive signal more effectively. This encourages a diverse and non-collinear final subset. The detailed RFE configuration and Extra-Trees feature-selection settings are reported in Appendix B.

The selector is fitted on the common feature set, defined as the intersection of columns present across all inner fold training partitions, to ensure a coherent feature space across folds.

4.2 Modelling

4.2.1 Candidate Models

13 regression estimators are evaluated across all experimental configurations. The candidate set spans five algorithmic families that includes linear regularised models, kernel-based models, shallow ensembles, gradient-boosted ensembles, and a neural network. The aim of the selection procedure is to lessen the bias towards any single inductive class. All models are implemented in scikit-learn [28], except LightGBM [29] and CatBoost [30], which use their own Python packages. Table 8 summarises the thirteen candidates with their primary tuned hyperparameters.

Table 8: Candidate regression models and primary tuned hyperparameters. Complexity rank is used as a tie-breaker in winner selection. Lower rank indicates a preference when two models achieve equivalent MAE.

Model	**Primary tuned hyperparameters**	**Complexity rank**
Ridge regression	Regularisation α	1
Elastic net	Penalty α, ℓ_1 ratio	2
Kernel ridge regression	Regularisation α, kernel, γ	3
Support vector regression	C, ε, kernel, γ	4
AdaBoost	Number of estimators, learning rate, loss	5
Random forest	Depth, min. samples per leaf/split, features	6
Extra Trees	Depth, min. samples per leaf/split, features	7
Histogram gradient boosting	Iterations, depth, learning rate, leaves	8
XGBoost	Estimators, depth, learning rate, subsampling, ℓ_1/ℓ_2 regularisation	9
LightGBM	Estimators, depth, learning rate, leaves, ℓ_1/ℓ_2 regularisation	10
CatBoost	Iterations, depth, learning rate, ℓ_2 regularisation	11
Multi-layer perceptron	Hidden layers/sizes, activation, ℓ_2 penalty	12
Dummy mean	(none)	13

Ridge regression. Ridge regression minimises the ordinary least-squares objective with a ℓ_2 penalty on the coefficient vector. It provides a closed-form solution. Ridge introduces a single regularisation parameter α that controls the trade-off between fit and shrinkage. Ridge is the most interpretable and computationally efficient candidate. It serves as a strong linear baseline against more flexible models [11].

Elastic net. Elastic net combines a ℓ_1 (Lasso) and a ℓ_2 (Ridge) penalty, simultaneously inducing sparsity in the coefficient vector and controlling its magnitude [31]. The ℓ_1 component is useful when a subset of the engineered features is irrelevant, as it drives the corresponding coefficients exactly to zero. The mixing parameter ρ interpolates between pure Lasso ($\rho = 1$) and pure Ridge ($\rho = 0$).

Kernel ridge regression. Kernel ridge regression strengthens the ridge objective with a kernel-induced feature map, enabling the model to learn nonlinear decision surfaces while retaining the

closed-form ridge solution in the dual space [32]. The radial basis function and laplacin kernels are both evaluated, controlled by a length-scale parameter γ. Kernel ridge provides a nonlinear counterpart to the linear ridge baseline with a comparably simple regularisation.

Support vector regression. Support vector regression (SVR) finds the flattest function that fits the training data within an ε-insensitive tube around the predictions, mapping features into a kernel-induced feature space [32]. The radial basis function is chosen as kernel which captures nonlinear relationships . SVR is known to perform well on moderate-dimensional datasets after standardisation.

AdaBoost. AdaBoost builds an ensemble by iteratively reweighting training observations to concentrate subsequent learners on those observations with the highest loss [33]. A shallow decision tree serves as the base learner, providing a good bias-variance balance in the boosting context.

Random forest. The random forest regressor constructs an ensemble of deep decision trees, each trained on a bootstrap replicate of the training data and using a randomly selected feature subset at each split [34]. Aggregating across many de-correlated trees substantially reduces variance compared to any individual tree. Minimum samples per leaf and split are tuned to control overfitting.

Extremely randomised trees. Extra Trees augments the random forest strategy by additionally randomising the split threshold at each node, selecting the best threshold from a random draw rather than from an exhaustive search [35]. The increased randomisation further reduces variance and training time at the cost of a modest increase in bias. The hyperparameter search space mirrors that of random forests.

Histogram-based gradient boosting. The histogram-based gradient boosting regressor discretises continuous features into equal-frequency histograms before constructing each tree, substantially reducing training time on high-dimensional datasets [29]. The algorithm is otherwise equivalent to standard gradient boosting and shares its theoretical guarantees.

XGBoost. XGBoost extends gradient boosting with second-order Taylor approximations of the loss function, explicit ℓ_1 and ℓ_2 regularisation terms on the leaf weights, and hardware- optimised tree construction [9]. It provides a complementary gradient boosting implementation with a richer regularisation.

LightGBM. LightGBM is a gradient boosting framework that grows trees leaf-wise rather than level-wise, combined with histogram-based binning and gradient-based one-side sampling to reduce the number of data instances considered at each split [29]. This design yields substantially faster training on large, sparse feature matrices while maintaining accuracy.

CatBoost. CatBoost is a gradient boosting algorithm that employs ordered boosting, which computes residuals on a permutation of the data to prevent target leakage, together with symmetric decision trees that provide consistent predictions across different observation orderings [30]. Its built-in regularisation is well suited to the moderate-size, heterogeneous feature matrices.

Multi-layer perceptron. The MLP regressor is a feed-forward neural network trained by stochastic gradient descent with early stopping based on a hold-out validation loss. A range of compact architectures is considered together with tanh activations. Whereas all other candidates are invariant to the specific neural architecture, the MLP can in principle approximate arbitrary continuous functions [36], making it a useful bound on the complexity achievable by the overall model family.

Dummy mean. The dummy mean regressor always predicts the training-set mean of the target variable, regardless of the input features. It serves as an unconditional baseline, establishing the floor of performance that any informative model must exceed. Inclusion of this baseline ensures that reported metrics are interpretable relative to the simplest possible prediction strategy.

4.2.2 Nested Cross-Validation Structure

To get unbiased estimates of generalisation performance and to ensure that hyperparameter tuning does not introduce optimistic bias into the evaluation metrics, the training pipeline make use of a nested cross-validation design [37]. Nested cross-validation separates hyperparameter optimization (inner loop) from performance estimation (outer loop), preventing the information leakage that would arise if the same data were used for hyperparameter optimization and performance reporting.

Outer loop. The outer loop iterates over four hold-out test years (2018, 2019, 2020, and 2021). It constructs a distinct outer fold for each year. In each outer fold, all observations with year $< t_{\text{test}}$ are used as the outer training set, and observations from year t_{test} constructs the outer test set. The year 2022 is strictly kept as the final test set and is not used in any outer fold. This temporal split structure avoids temporal information leaking into the training process. This is an important constraint for time-indexed data [38]. The outer temporal holdout structure is summarised in Figure 8.

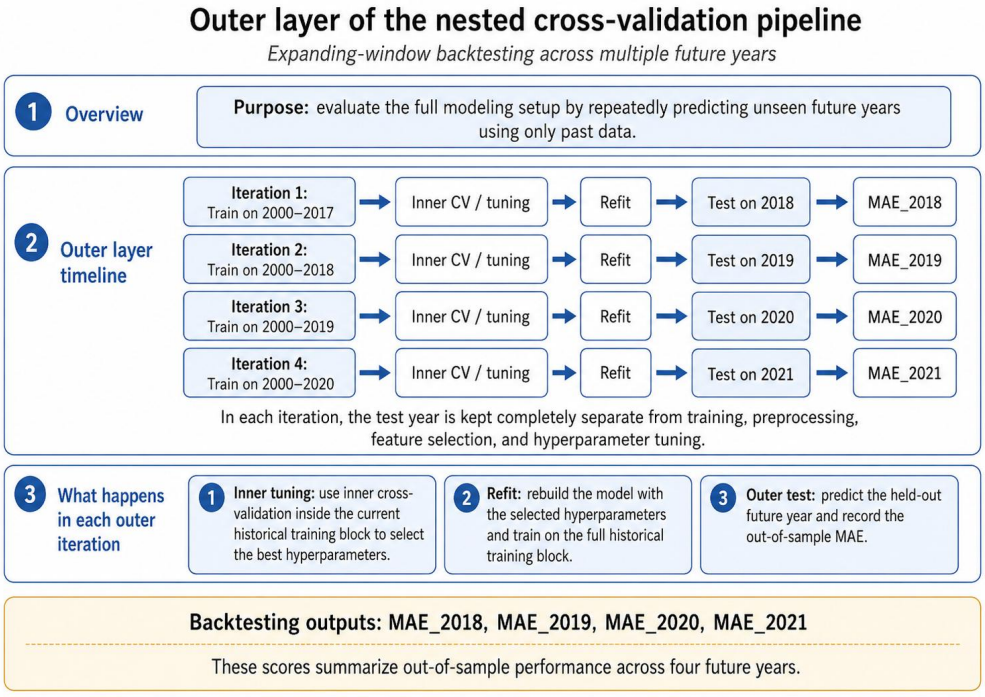

Figure 8: Outer-layer temporal holdout structure for nested cross-validation.

Dual cross-validation strategy. Two inner cross-validation strategies are considered, namely, temporal and random strategies. Each one is producing an independent set of outer results. Later, for each strategy separate winner models are subsequently selected.

The temporal strategy make use of an expanding-window scheme. The inner validation fold for year v uses all observations from year v, while the inner training fold uses all observations from years before year v. Five inner temporal folds are constructed which are corresponding to the five most recent years before the outer test year. Validation always takes place on data unseen to the models during inner training. Expanding-window cross-validation is the methodologically sound approach for time-ordered data, as it replicates the real-world scenario in which models are only trained on historical data and predicts the future observations [38]. The resulting expanding-window inner loop is shown in Figure 9.

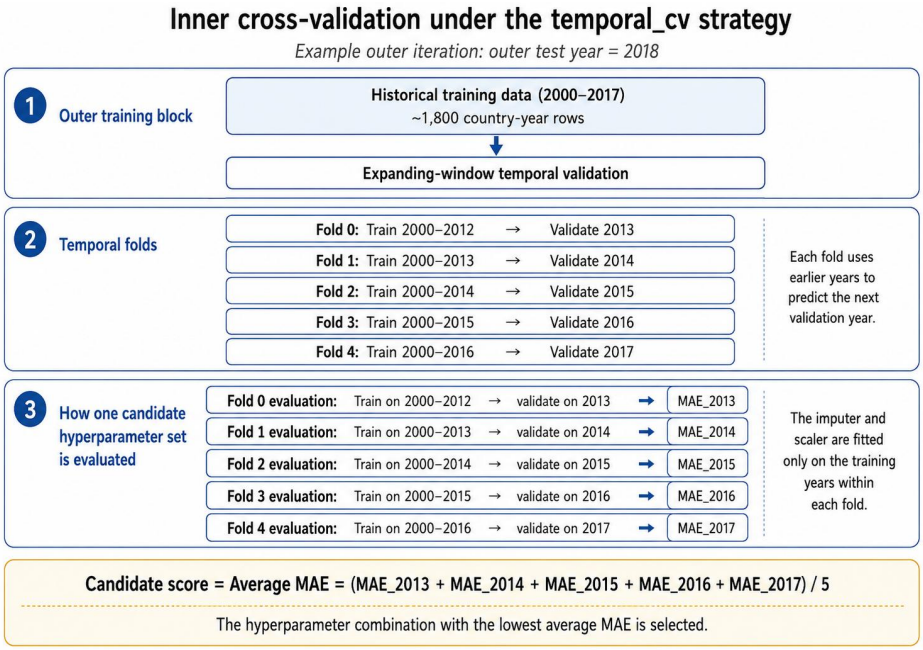

Figure 9: Temporal inner-loop expanding-window cross-validation structure.

The random strategy applies a standard five-fold split and it ignores the temporal ordering of observations. This provides an additional evaluation under the weaker assumption that all observations are approximately exchangeable. The random inner-loop structure is shown in Figure 10.

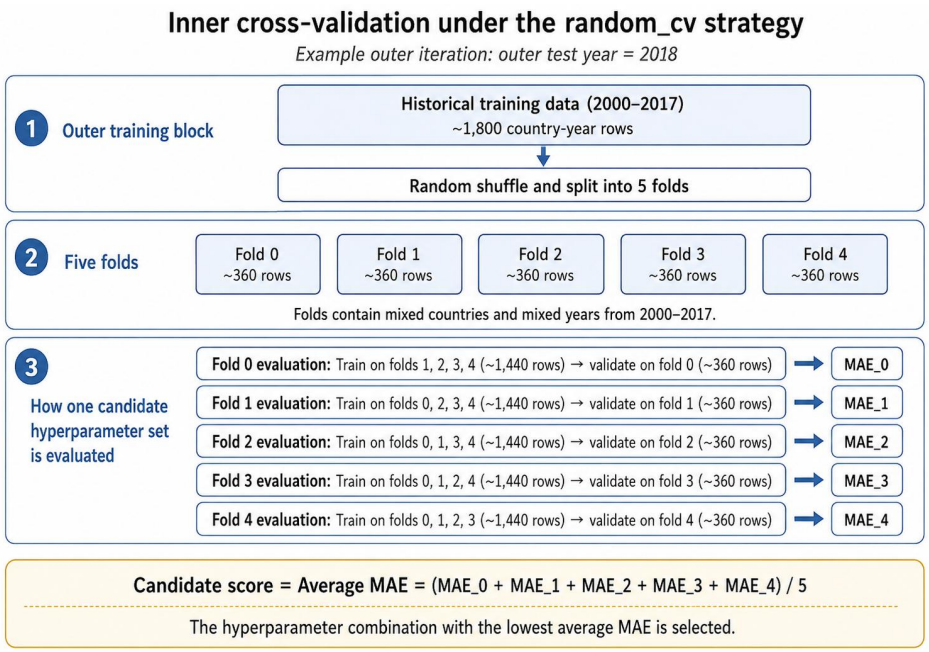

Figure 10: Random inner-loop cross-validation structure used as a complementary strategy.

4.2.3 Hyperparameter Optimisation

Within each inner cross-validation configuration, the hyperparameters of each candidate model are optimised using the Tree-structured Parzen Estimator (TPE) algorithm [12]. TPE is a sequential, model-based optimisation strategy that builds separate nonparametric density models for parameter configurations that produced good and poor objective values and proposes new candidates by maximising the ratio of these densities. Compared to grid search or random search, TPE concentrates evaluations in promising regions of the hyperparameter space and has been shown empirically to outperform grid search [39] and random search [12].

Each model is given a wall-clock budget of 240 seconds per combination. Every combination is subject to a minimum of three trials. The maximum number of trials is also capped per model according to the relative size of each model's hyperparameter space. The objective function is minimised during tuning is the penalised cross-validation MAE which is defined in Section 4.2.4. The full TPE search spaces, trial caps, and fixed model settings are reported in Appendix C.

4.2.4 Overfitting Diagnosis and Control

Overfitting is a significant risk in any machine learning application. For this purpose, the pipeline incorporates two additional mechanisms.

Penalised cross-validation objective. The tuning objective penalises the inner cross validation mean absolute error based on the magnitude of the train and validation gap:

$$\mathcal{L}_{\text{pen}} = \overline{\text{MAE}}_{\text{val}} \left[1 + \min(\max(\delta,\, 0)\,,\, 3.0) \times 0.5\right], \tag{7}$$

where

$$\delta = \frac{\overline{\text{MAE}}_{\text{val}} - \overline{\text{MAE}}_{\text{train}}}{\max\left(\overline{\text{MAE}}_{\text{train}},\, \epsilon\right)} \tag{8}$$

is the relative train-validation gap ratio, averaged across inner folds, with ϵ denoting the small positive floor used to avoid instability when the training MAE is near zero. Parameter combinations with large gaps are ranked below compared to similar validation MAE but tighter train-validation agreement.

Overfitting audit flags. Two separate overfitting diagnostics are computed in each training run. The first is an inner gap flag which is showed in Section 4.2.4. It mainly incurs an extra penalty during hyperparameter optimization and also flags the fold level result if the inner train-to-validation gap ratio satisfies $\delta \geq 0.50$. However, this does not influence winner selection.

The second is an outer gap flag, which is the primary overfitting signal. It is triggered when the outer generalisation gap (defined as the difference between outer test MAE and inner validation MAE) exceeds both an absolute threshold (0.1) and a relative threshold (0.5) simultaneously. A candidate model is classified as outer-gap flagged when its outer gap flag is triggered across all outer folds which indicates a systematic overfitting regardless of the test year.

4.2.5 Winner Selection

Winner selection determines the best (imputation branch, model) pair (in total 26 pairs) for each (strategy, target) combination (in total 6 combinations), based on the outer cross-validation results. Each remaining (branch, model) candidate is ranked by its mean outer test MAE averaged across the four outer test years 2018–2021. The three strongest candidates are kept as the primary shortlist for each (strategy, target) combination. For any primary candidate whose outer gap flag is triggered across all outer folds, the next best non-flagged candidate is added as a backup candidate, ensuring a non-overfitting alternative is available. The result is a shortlist of three to six candidates per (strategy, target) combination.

4.2.6 Final Evaluation

Model performance on the year-2022 holdout is reported using three standard regression metrics, namely, mean absolute error (MAE), root mean squared error (RMSE), and the coefficient of determination (R^2). MAE is the primary criterion for its interpretability and unit-consistent measure of prediction accuracy that is robust to occasional large-magnitude errors [40]. RMSE provides an additional information by placing greater weight on large deviations. Lastly, R^2 summarises the ratio of target variance that can be explained by the model. The final 2022 holdout workflow is summarised in Figure 11.

The test set is processed in two scored modes plus an all-entity prediction mode. First, in strict scored mode, observations are excluded if they have not target values or if their feature missingness is above the 30 % threshold. This corresponds to 109 observations. Second, in lenient scored mode, observations are excluded when they lack the target variables which provides 123 observations and relies on the imputation layers to handle higher feature missingness. In all-country mode, all 282 entities regardless of data completeness. All metrics are reported as averages across the 200 bootstrap iterations.

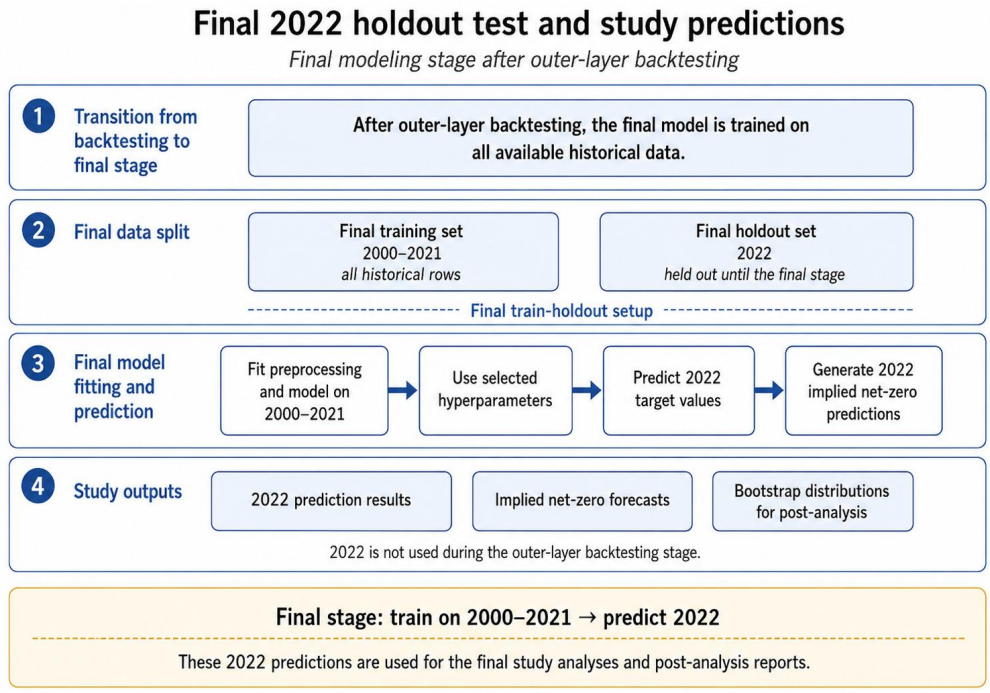

Figure 11: Final 2022 holdout evaluation after model selection and refitting.

4.2.7 Bootstrap Uncertainty Quantification

To quantify the uncertainty in the model predictions and the derived net-zero year estimates, the winner for each (strategy, target) combination is subjected to a bootstrap procedure [41]. First, the winner model is kept on the full pre-2022 data using hyperparameters re-tuned in Section 4.2.6. Resampling of pre-2022 training dataset is carried out at the country level rather than at the individual row level. Countries are sampled with replacement, and when a country is selected its pre-2022 training observations included in that bootstrap iteration. If not selected, then they are removed from training dataset for that bootstrap iteration. On average, about 30 %–40 % of the original training countries are absent from any single bootstrap iteration. The model is re-fitted on each bootstrap iteration using the selected hyperparameters, and predictions are generated for all countries observed in the year 2022, including those not used for scoring due to missing target values. Then 200 bootstrap iterations are performed.

This design makes the bootstrap uncertainty estimates sensitive to the composition of the country sample. Wide bootstrap intervals therefore indicate that a country's predicted budget or implied net-zero year depends strongly on which countries are present in the training set. Accordingly, narrow intervals indicate that the prediction is stable over alternative country samples. The bootstrap scores should be interpreted as a robustness check against training sample variability.

The collection of 200 per country predictions constructs the empirical predictive distribution. the mean, standard deviation, 5th percentile, and 95th percentile of the bootstrap predictions are computed for each country. Each bootstrap prediction is individually calculated back through the inverse target construction process in Section 2.4 to get a distribution over implied net-zero years. The resulting per country calculations reflect both model uncertainty due to sampling variability in the training countries and scenario uncertainty due to the three emission trajectory assumptions.

5 Results

This section reports the findings obtained from the modelling pipeline described in Section 4. Section 5.1 addresses model comparison and winner selection. Section 5.2 presents the final holdout evaluation. Section 5.3 shows the sensitivity to methodological choices. Section 5.4 examines feature importance via SHAP values. Section 5.5 reports bootstrap uncertainty. Section 5.6 presents the implied net-zero year projections. Section 5.7 benchmarks those projections against nationally pledged targets.

5.1 Model Comparison and Winner Selection

The nested cross validation procedure evaluated 13 regression estimators over four outer test years (2018–2021), two cross-validation strategies (temporal and random), two imputation branches (simple and MICE), and three emission scenarios (linear, optimistic, pessimistic). Table 9 summarises the training configuration.

Table 9: Overview of the training configuration.

Quantity	Value
Candidate models evaluated	13
Cross-validation strategies	2
Imputation branches	2
Target scenarios	3
Outer test folds (years)	4 (2018–2021)
Max trials per model	20-60 (model dependent)
Bootstrap iterations	200 (for 19 candidate models)
Final test countries (strict)	109
Final test countries (lenient)	123

Support vector regression (SVR) emerges as the dominant estimator of all experimental configurations. In the winner selection, SVR claims the rank-1 position in all six configurations (two strategies $\times$ three scenarios). In four of the six configurations, the second slot is also filled by SVR on the alternative imputation branch. The remaining second slots are occupied by HistGradientBoosting (linear scenario, random CV) and MLP (pessimistic scenario, random CV). The third slot is occupied by MLP in two configurations, by HistGradientBoosting in two configurations, by SVR in one, and by CatBoost in one. For the linear scenario under random CV, the outer-gap audit triggers backup candidate promotion, adding SVR (MICE branch) as a rank-4 backup.

Figure 12 displays the distribution of outer-test MAE across the four years for each candidate. The SVR distributions are showing low-median across all scenarios. SVR achieves the lowest mean error in the outer cross validation over all other models, followed by the histogram gradient boosting regressor and the multi-layer perceptron. Boosted ensembles (LightGBM, XGBoost, CatBoost) occupy the intermediate range, while AdaBoost, tree-based forests Linear models (ridge, elastic net) and display the highest error among non-trivial estimators. The mean-prediction dummy baseline yields an outer-test MAE of approximately 0.92, providing a naive reference against which all candidate models demonstrate substantial gains.

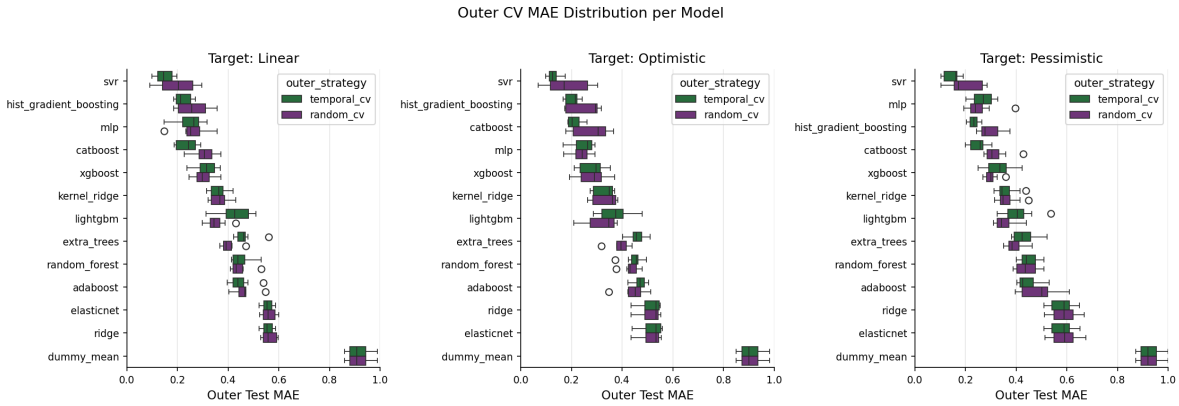

Figure 12: Outer cross-validation MAE distributions grouped by target scenario and CV strategy. Each box spans the interquartile range across four outer test folds (2018–2021). SVR candidates consistently yield the lowest median.

Figure 13 presents the selected models' mean outer validation MAE scores aggregated across all outer folds, strategies, and scenarios.

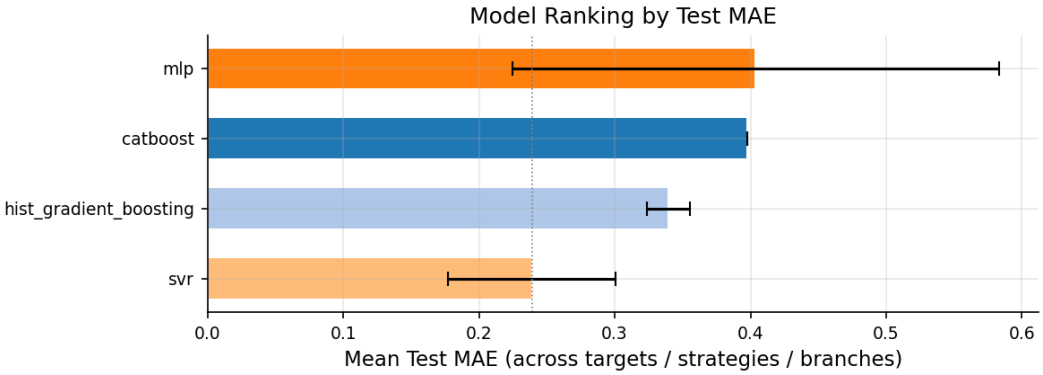

Figure 13: Mean outer cross-validation MAE for the winner candidate models, aggregated across all outer folds, two cross-validation strategies (temporal and random), two imputation branches (simple and MICE), and three emission scenarios (linear, optimistic, pessimistic). Error bars denote one standard deviation across folds. SVR achieves the lowest mean error in every (strategy, target) combination.

Overfitting audit. Despite SVR's extremely low inner training MAE (0.015–0.107 across winner configurations), the inner validation MAE is substantially higher (0.053–0.171). This yields inner relative gap ratio that exceed the 0.50 flag threshold in every fold and triggers the inner gap flag for all folds in every configuration. However, the inner gap flag is kept as a diagnostic per explanation in Section 4.2.4. Thus, model selection is governed by the outer generalisation gap. Since the outer gap (outer-test MAE minus inner-validation MAE) does not violate the combined absolute-and-relative threshold in any SVR winner fold, the outer-gap flag is not triggered for the selected SVR candidates (see Figure 14).

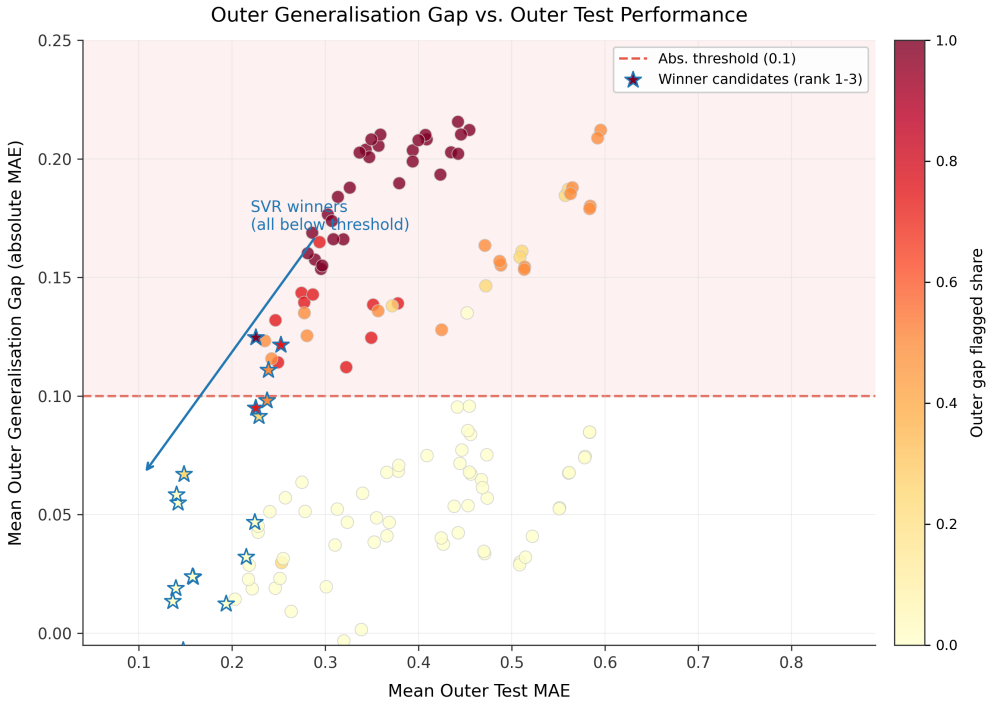

Figure 14: Mean outer generalisation gap (outer-test MAE minus inner-validation MAE, y-axis) plotted against mean outer-test MAE (x-axis) for all candidate models. The colour bar shows the outer-gap-flagged share in which a model's outer-test MAE exceeds its inner-validation MAE by more than both the absolute threshold (0.10) and the relative threshold (0.50), indicating possible overfitting. The dashed horizontal line marks the absolute outer-gap threshold (0.10). The shaded region above it indicates candidates that violates this criteria. Non-winner candidates are coloured by their outer-gap-flagged share.

5.2 Final Holdout Evaluation

All 19 winner models were refitted once on the pre-2022 training data and evaluated on the 2022 holdout partition. These point estimates report the performance of the model trained with maximum data availability, while the bootstrap columns quantify resampling uncertainty. The primary (strict) evaluation excludes observations with excessive feature missingness. The secondary (lenient) evaluation uses target-valid holdout observations regardless of feature completeness. Bootstrap resampling is performed at country-block level. When a country is sampled, its pre-2022 data is included together in that particular bootstrap iteration. Table 10 reports MAE and R^2 scores of the winner models.

Table 10: Final holdout performance for all 19 winner candidates. Bootstrap columns report the 5th and 95th percentiles and the mean of per-country MAE across 200 country-block resampling iterations evaluated on the strict set of countries.

Scenario	Strategy	Branch	Model	Rank	Test MAE (strict)	Test R^2 (strict)	Test MAE (lenient)	Test R^2 (lenient)	Boot. (strict) P05	Boot. (strict) P95	Boot. (strict) Mean
linear	random_cv	simple	SVR	1	0.213	0.917	0.295	0.837	0.342	0.466	0.397
linear	random_cv	simple	HistGradBoost	2	0.342	0.841	0.402	0.796	0.392	0.607	0.483
linear	random_cv	simple	MLP	3	0.282	0.903	0.332	0.866	0.372	0.504	0.435
linear	random_cv	mice	SVR	4[b]	0.264	0.869	0.344	0.785	0.378	0.493	0.433
linear	temporal_cv	simple	SVR	1	0.181	0.942	0.255	0.867	0.333	0.465	0.392
linear	temporal_cv	mice	SVR	2	0.240	0.894	0.304	0.819	0.375	0.499	0.431
linear	temporal_cv	simple	HistGradBoost	3	0.354	0.832	0.400	0.810	0.421	0.665	0.523
optimistic	random_cv	simple	SVR	1	0.153	0.956	0.223	0.883	0.294	0.414	0.350
optimistic	random_cv	mice	SVR	2	0.242	0.893	0.312	0.823	0.348	0.449	0.396
optimistic	random_cv	mice	MLP	3	0.320	0.863	0.371	0.811	0.363	0.501	0.429
optimistic	temporal_cv	simple	SVR	1	0.167	0.949	0.234	0.876	0.312	0.441	0.371
optimistic	temporal_cv	mice	SVR	2	0.226	0.917	0.291	0.845	0.353	0.467	0.405
optimistic	temporal_cv	simple	HistGradBoost	3	0.322	0.859	0.372	0.825	0.344	0.508	0.419
pessimistic	random_cv	simple	SVR	1	0.331	0.856	0.439	0.735	0.450	0.601	0.517
pessimistic	random_cv	simple	MLP	2	0.610	0.478	0.675	0.464	0.612	0.907	0.755
pessimistic	random_cv	mice	SVR	3	0.279	0.897	0.335	0.847	0.393	0.503	0.443
pessimistic	temporal_cv	simple	SVR	1	0.357	0.840	0.449	0.737	0.471	0.620	0.538
pessimistic	temporal_cv	mice	SVR	2	0.215	0.924	0.299	0.836	0.351	0.486	0.407
pessimistic	temporal_cv	simple	CatBoost	3	0.397	0.788	0.470	0.725	0.485	0.640	0.554

[b] Backup candidate appended after outer-gap flag triggered on the rank-2 primary.

On the strict partition, the random-CV SVR is the best-performing configuration under the optimistic scenario with simple imputation (Test MAE = 0.153, $R^2 = 0.956$). The coefficient of determination for SVR winner configurations on the strict holdout ranges from 0.856 to 0.956 which is an indication that the models explain a substantial proportion of the variance. Bootstrap mean MAE is higher than the point estimate because each country-block bootstrap training data omits roughly one-third of unique countries. This gap can be interpreted as a conservative resampling tax, rather than a contradiction of the point-estimate results.

Figure 15 presents the strict holdout MAE as a heatmap across all configurations. The optimistic scenario simple SVR (random CV) achieves the lowest strict MAE (0.153) and the pessimistic random CV MLP gets the highest MAE (0.610). This reflects MLP's sensitivity to this particular scenario under random strategy.

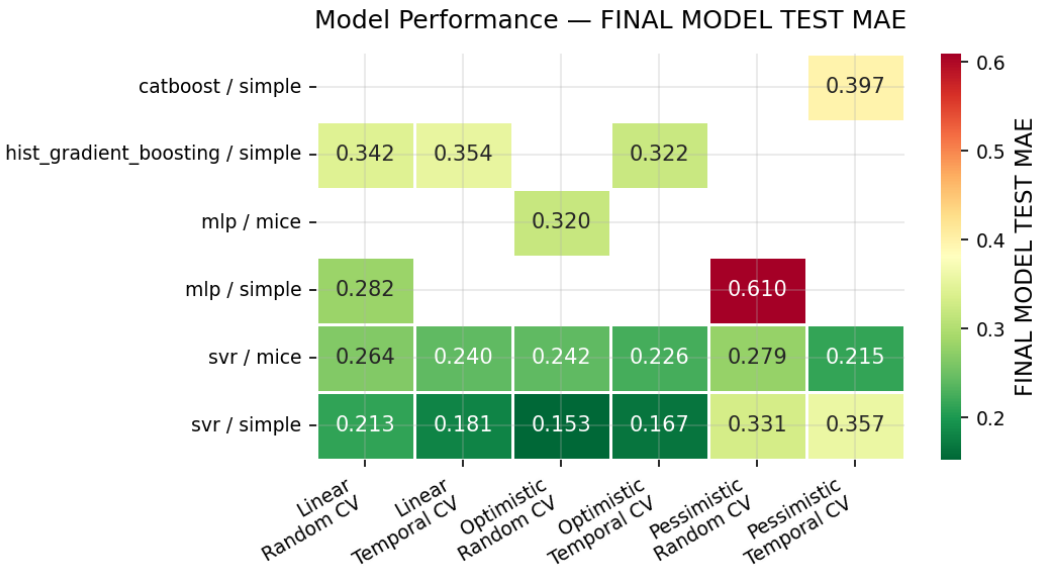

Figure 15: Final holdout MAE (strict partition, 109 observations, year 2022) for all 19 winner candidates, shown as a heatmap over model/imputation (rows) and scenario/strategy (columns).

Table 10 shows the systematic discrepancy between the strict and lenient MAE estimates, and between strict measures and the bootstrap mean. The lenient MAE is substantially higher (0.223–0.449 for SVR winners) than the strict MAE (0.153–0.357). This gap shows the sensitivity of the models when applied to data-sparse countries. In such cases, imputed features drive prediction rather than observed values. Similarly, the bootstrap mean MAE (0.350–0.538 for SVR) is substantially higher than the strict point estimate (0.153–0.357). Even though both are evaluated on the same test, the discrepancy arises from the bootstrap resampling procedure itself. In each of the 200 iterations, the training data is resampled with replacement at the country-block level, meaning that some countries are absent from any given bootstrap sample. Training on these less representative subsets degrades model performance relative to training on the full dataset and returns a mean test MAE values that is well above the single full-data point estimate. The width of this distribution, captured by the 5th and 95th percentiles, reflects the sensitivity of SVR to the composition of the training set. Figure 16 illustrates the relationship between final training MAE and final holdout MAE for all winner candidates.

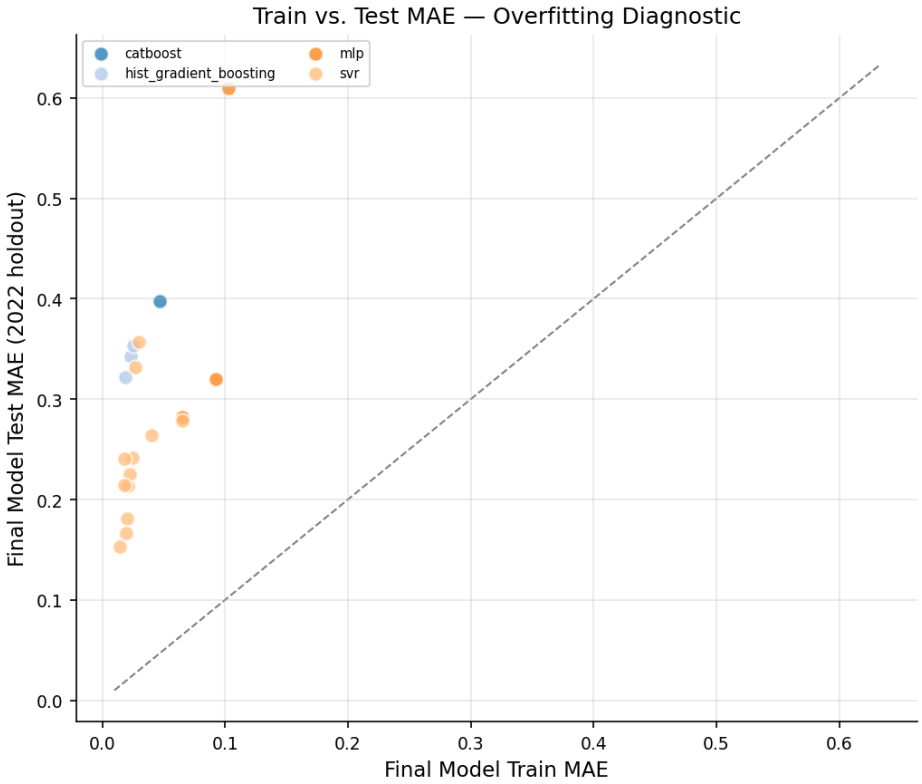

Figure 16: Final training MAE versus final holdout MAE (strict) for all 19 winner candidates. The diagonal line marks perfect agreement. Points above it indicate positive generalisation gap (higher test than training error). All SVR candidates are in the lower region compared to other models.

5.3 Sensitivity Analysis

Three dimensions of sensitivity are examined, namely, imputation branch, CV strategy, and emission scenario. Figure 17 provides a heatmap of pairwise differences over all three dimensions. The entries in the heatmaps are averaged pairwise differences in outer-CV MAE after aggregating over other dimensions. Thus, a value on the figure should be read as the average change in error by switching one methodological choice while holding the model family and scenario or strategy fixed.

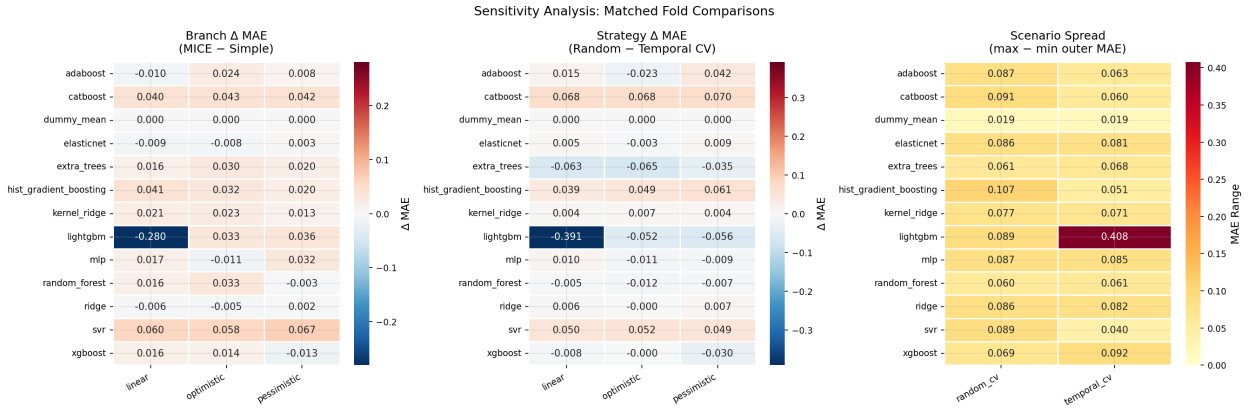

Figure 17: Sensitivity of outer-CV MAE to methodological choices. Warmer colours indicate cases where the first-listed option outperforms the second.

Imputation branch. The left panel reports the effect the branch as MICE MAE minus simple imputation MAE. These are averaged across the two CV strategies. For example, the (SVR, linear) cell is calculated from the random-CV difference ($0.2493 - 0.1485 = 0.1008$) and the temporal-CV difference ($0.1585 - 0.1398 = 0.0187$). This is giving an average branch penalty of $(0.1008 + 0.0187)/2 = 0.060$. Thus, the heatmap cell is their averaged difference. For most of model families, the direction of the branch effect is inconsistent over scenarios and strategies. This indicates that imputation related variation in model performance is secondary to model class and CV strategy effects. This shows that simple hierarchical median imputation is at least as competitive as MICE. This is consistent with the panel-aware auxiliary predictor design of the MICE branch. This, partially, compensates for the simpler branch's lack of multivariate dependency modelling [13]. However, the design of MICE may also introduce auxiliary predictor collinearity that degrades MICE branch performance.

Cross-validation strategy. The middle panel reports random-CV MAE minus temporal-CV MAE, averaged across imputation branches. In the linear SVR case, the simple branch contributes ($0.1485 - 0.1398 = 0.0087$), while the MICE branch contributes ($0.2493 - 0.1585 = 0.0908$). The average of strategy effect is approximately 0.050. That indicates a consistent advantage for temporal CV. For more complex boosted ensemble models (LightGBM, XGBoost), the two strategies are more comparable. The advantage of temporal CV for SVR is consistent for expanding-window evaluation on time-ordered panels. The alignment of inner-fold validation with future years provides a calibration that random variant does not, when the imputed feature matrix specifically preserves the strong temporal structure [38].

Emission scenario. Over all candidate models and configurations, the optimistic scenario ($\alpha = 2.0$) returns the lowest outer-test MAE (approximately 10–20% lower than linear) due to its smaller target budget. Since the target is the remaining cumulative per-capita budget, $C(t) = \int_t^{T_{NZ}} e(\tau)\, d\tau$ and pre-2030 part of the budget is a linear interpolation between current emissions and the 2030 target and is independent of α, the post-2030 part of power-law-decline becomes dominant. Approximating the post-2030 integral gives

$$\int_{2030}^{T_{NZ}} e_{2030} \left(1 - \frac{\tau - 2030}{T_{NZ} - 2030}\right)^{\alpha} d\tau \approx \frac{e_{2030}(T_{NZ} - 2030)}{\alpha + 1}.$$

The pessimistic pathway gives more weight to the smooth post-2030 component while the optimistic pathway gives more weight to pre-2030 component. The pessimistic pathway's scaling factor is $1/(0.5+1) = 0.67$, compared with 0.50 for the linear pathway and 0.33 for the optimistic pathway. Consequently, the pessimistic target is more protectes against noisy base-year emissions and reporting fluctuations, whereas the optimistic target is dominated by the pre-2030 transition. Thus, scenario sensitivity does not affect model rankings and SVR dominates across all three scenarios, confirming that its advantage is not pathway-specific.

Prediction agreement across strategies. Figure 18 shows the country level agreement between the random-CV and temporal-CV predictions for the linear scenario under SVR-simple configuration. The predictions of two strategies are highly correlated ($r > 0.98$), with minor divergence for countries at the extreme ends of the log-budget distribution. The agreement supports the robustness of the point estimates reported in the following sections.

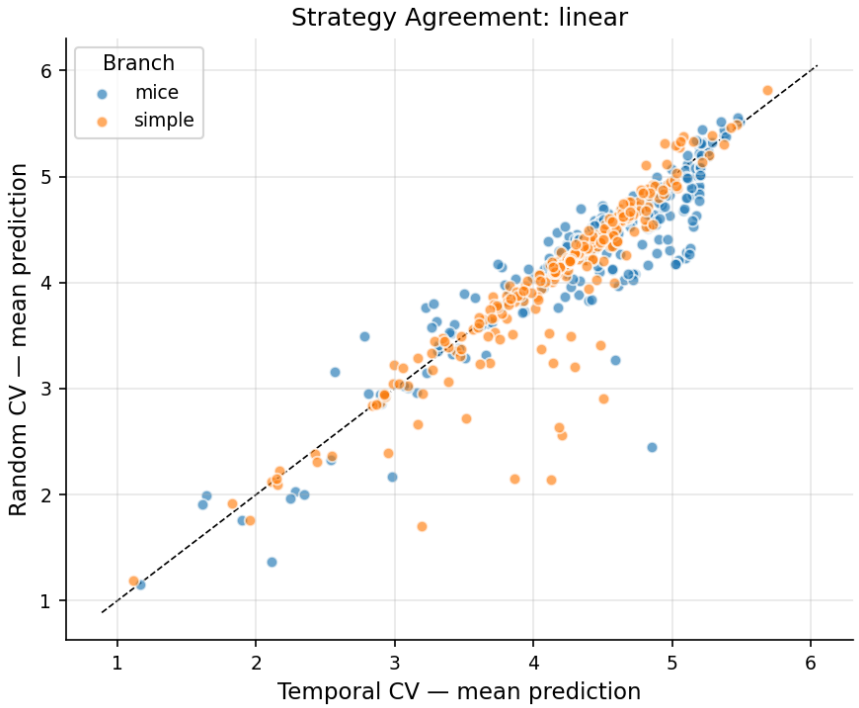

Figure 18: Country-level prediction agreement between the random-CV and temporal-CV SVR models for the linear scenario. Each point represents one country's mean predicted log-budget in year 2022. The diagonal line marks perfect agreement. High agreement indicates that CV strategy choice does not materially alter country level rankings.

5.4 Feature Importance and Interpretability

SHAP values [10] were computed for SVR models using a kernel based approximation to the SHAP value, which is applicable to any differentiable or callable predictor [10]. The analysis focuses on the primary winner for the linear scenario under random CV with simple imputation (SVR, rank 1). This configuration has the lowest lenient bootstrap MAE for the linear pathway and serves as the main model in Sections 5.6 and 5.7.

Figure 19 displays the mean absolute SHAP value for each feature, representing the average magnitude of each feature's contribution to the model's prediction across 2022 prediction set of 282 entities. This is different from the 123 target-valid countries used in the lenient final holdout evaluation and the 109 observations used in the strict evaluation.

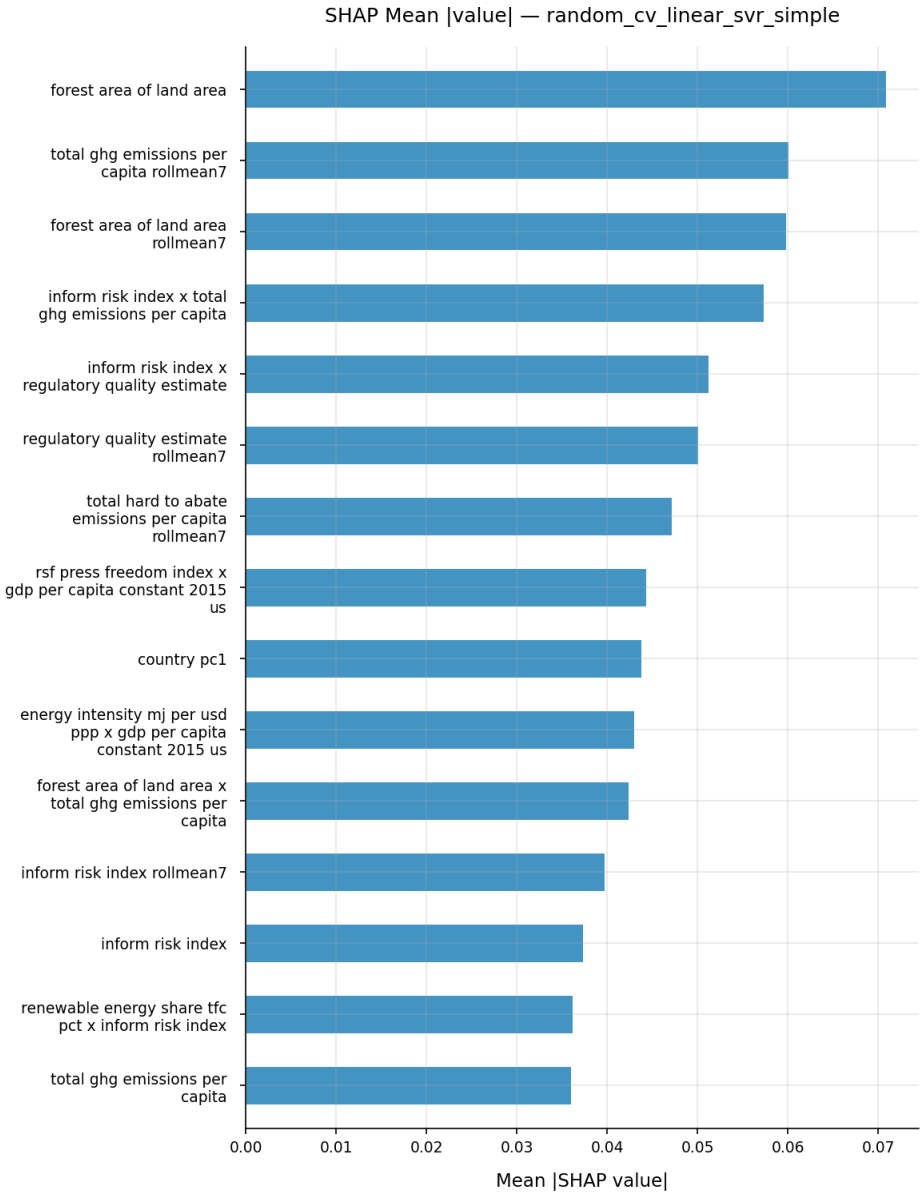

Figure 19: Mean absolute SHAP values for the linear-scenario SVR model (random CV, simple branch, rank 1 winner). The horizontal bars represent the average magnitude of each feature's contribution to the predicted log-budget across 2022 prediction set of 282 entities.

The most influential features cluster into four thematic groups. First, land use and emission structure which consists of forest area (current level and seven-year rolling mean) and total GHG emissions per capita (seven-year rolling mean) constructs the largest share of total model contribution. Countries with larger forest area relative to their land area are associated with lower cumulative emission budgets. At the same time, the rolling-mean GHG emissions level shows a

structural emission intensity beyond year specific fluctuations. Second group is interaction terms. The product of INFORM Risk Index and total GHG emissions per capita, and the product of INFORM Risk Index and regulatory quality, each contribute significantly. These interactions capture the amplifying effect of high humanitarian risk with high emissions or weak governance. This is a configuration linked with distinct decarbonisation profiles that additive terms cannot represent. Third group consists of governance and sectoral structure linked to total-hard-to-abate-emissions. The seven-year rolling mean of regulatory quality and total hard-to-abate emissions per capita are important continuous predictors to the model. The former is linked to institutional capacity to implement climate policy, while the latter reflects the dependence on sectors where decarbonisation is most technically difficult. Fourth group is the distinctive country structure. Country PCA component 1, derived from country-wise median profiles and missingness rates during training period, accounts for considerable average SHAP magnitude. Statistical inspection of first country PCA component indicates that it functions as a development and institutional-governance and explains 19.8% of the country-profile variance. Positive scores are driven by WGI index (high regulatory quality, control of corruption, political stability, voice and accountability), press freedom, animal-protein supply, and GDP per capita. The negative scores of the component are linked to INFORM risk, renewable energy share of total final consumption, GHG growth, energy intensity, and agricultural land share. The component makes a distinction between high-income, stable, high-governance countries and more vulnerable or data-sparse developing contexts. This dimensions are not fully represented by any single observed feature. Its importance in the SVR SHAP ranking indicates that the model uses macrodevelopment, governance capacity, and structural stability as key feature.

Figure 20 presents the SHAP beeswarm plot for 282-entity 2022 prediction set, showing both the magnitude and the direction of each feature's effect. For forest area, high values are associated with negative SHAP contributions. This displays that countries with larger forest shares receive lower predicted log-budgets. In contrast, for the INFORM risk index interaction terms, high values are associated with positive SHAP contributions. This reflects that higher risk countries with high emissions or weak governance have larger remaining per-capita emission.

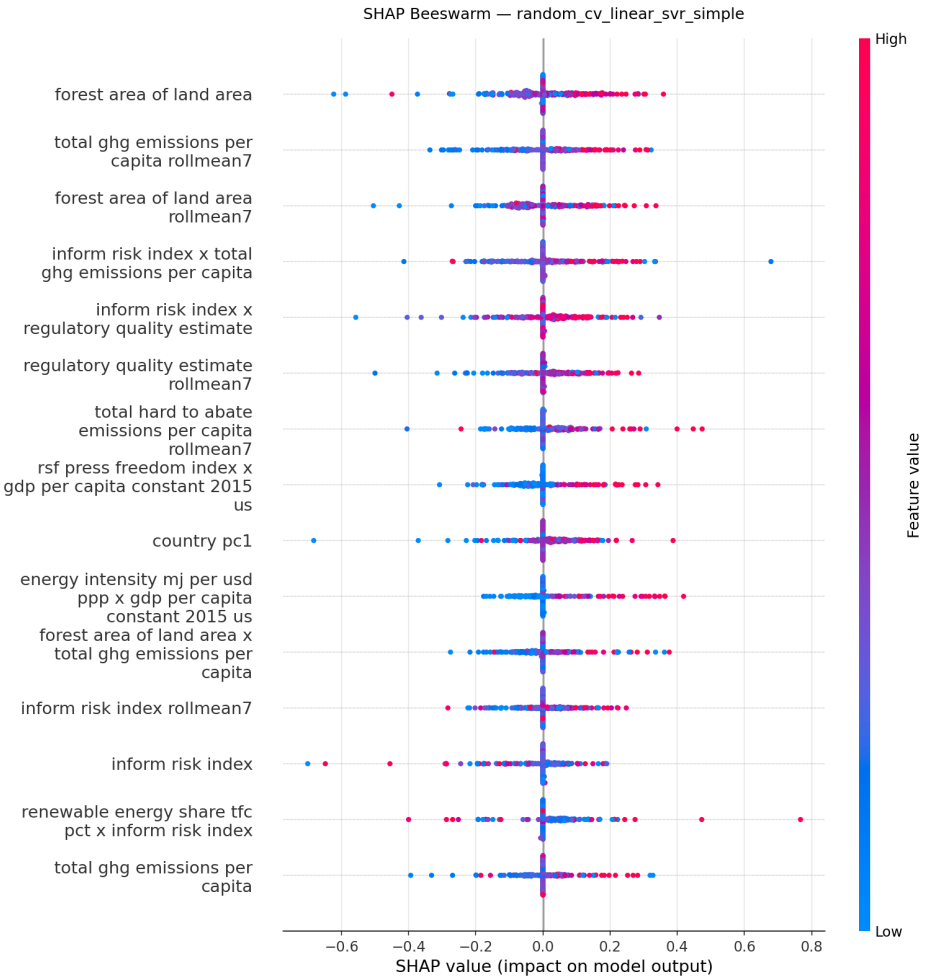

Figure 20: SHAP beeswarm plot for the linear-scenario SVR model. Each point represents one entity. Colour encodes the feature value (blue = low, red = high). Horizontal position shows the SHAP value.

5.5 Bootstrap Uncertainty Quantification

The 200-iteration bootstrap procedure described in Section 4.2.7 provides empirical distributions for every country. Uncertainty is summarised here as the width of the 90% predictive interval (5th to 95th percentile) on the log scale.

Figure 21 presents the distribution of confidence interval (CI) widths separately for each scenario. The y-axis counts intervals, not countries. In each scenario, one observation is one 90% bootstrap interval width for one entity under one winning pipeline configuration. Optimistic and Pessimistic panel has 1,692 interval widths which is equal to multiplication of 282 entities and six winning configurations, while Linear panel has 1,974 since Linear configuration has seven winning models. The figure uses 282 entities instead of smaller supervised-training. All three distributions are right-skewed, with scenario medians of 0.939 (linear), 0.869 (optimistic), and 1.002 (pessimistic) log-units. The heavy right tail is due to data sparse entities since each bootstrap samples constructs different imputed feature values and causes wide prediction intervals.

Figure 22 displays the geographic pattern of CI width for the linear scenario. High-uncertainty countries are concentrated in sub-Saharan Africa and Central Asia. Western Europe, East Asia, and

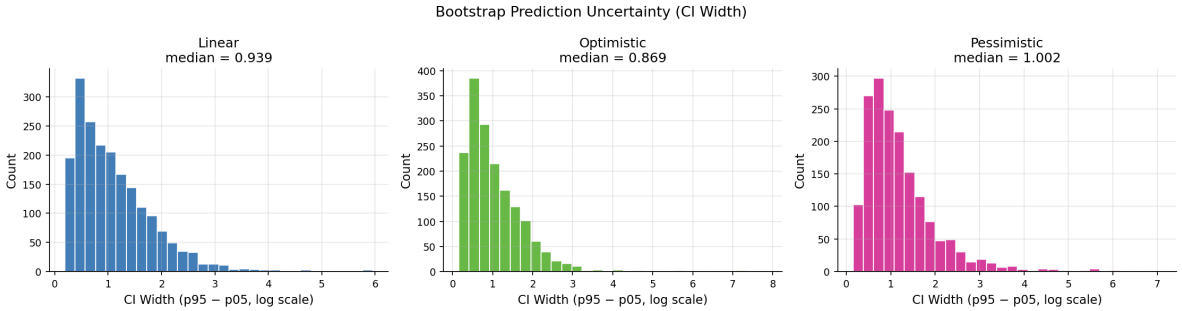

Figure 21: Distribution of 90% bootstrap predictive interval widths. The y-axis counts interval widths. Subplot titles report the median CI width per scenario.

major economies generally exhibit narrower intervals. The ten countries with the widest mean CI width include Niger (NER), Mali (MLI), Djibouti (DJI), Kiribati (KIR), and Kyrgyzstan (KGZ), which covers diverse economies and climates but having the characteristic of sparse data coverage. For these sub-Saharan, Central Asian, and island countries, high feature missingness combined with limited historical GHG reporting, the imputed values are becoming dominant factor.

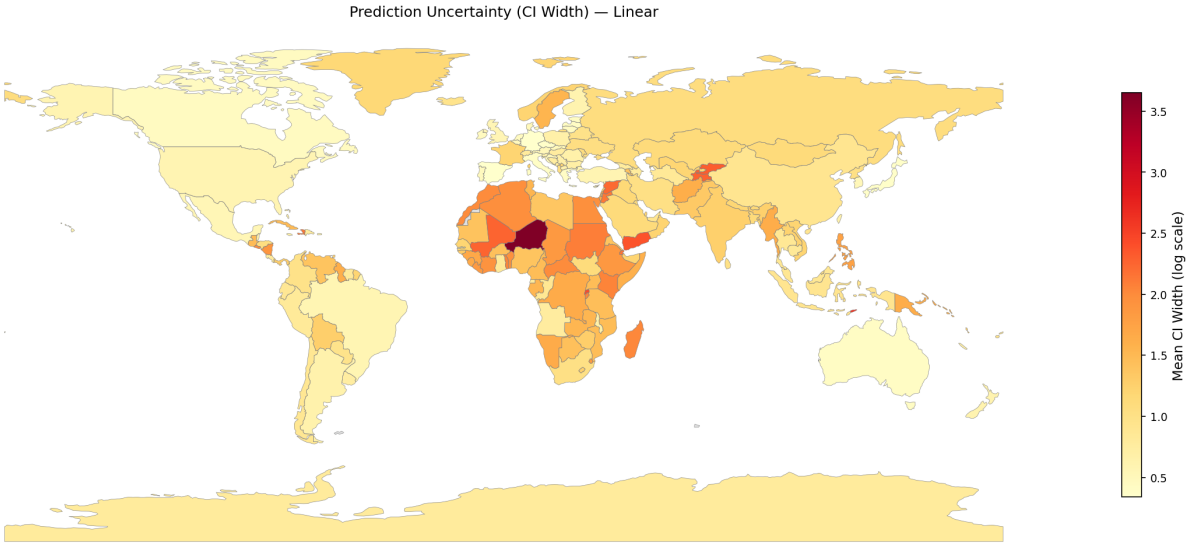

Figure 22: Geographic distribution of 90% bootstrap CI width for the linear scenario. Countries are coloured by mean CI width across all winner configurations for this scenario. Grey indicates countries with no valid prediction

Figure 23 shows kernel densities of bootstrap log-budget predictions for six countries under the linear scenario. The top row contains Germany (DEU), China (CHN), and India (IND), which are data-rich major emitters. Their pooled bootstrap distributions are narrow and approximately normally distributed. The bottom row contains Djibouti (DJI), Venezuela (VEN), and Somalia (SOM), which illustrates the data-sparse countries. These countries have much wider predictive distributions with visible multi-modality. The difference indicates that uncertainty is not uniform over the prediction set. This shows that prediction intervals narrowly bounded for data-rich emitters, whereas, it expands sharply when data becomes sparse and dominated by imputation step.

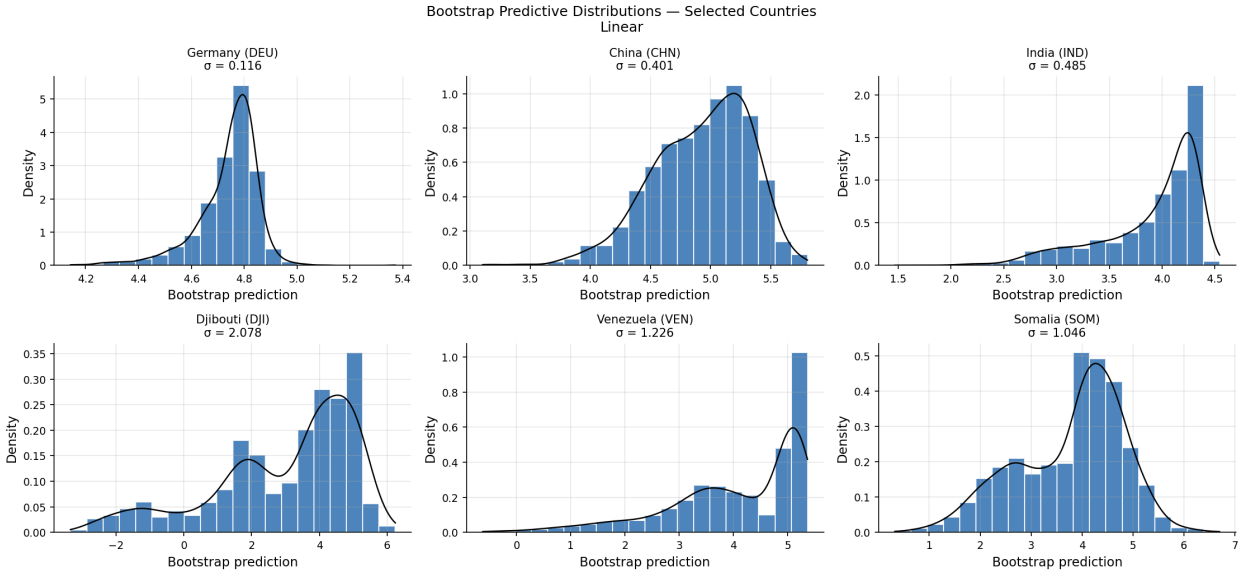

Figure 23: Bootstrap predictive distributions for selected countries under the linear scenario, comparing data-rich countries with data-sparse ones.

5.6 Implied Net-Zero Year Projections

Table 11 summarises the relevant sets used in predictions. Since model training, final evaluation, all-country prediction, physical T_{NZ} inversion, and pledge benchmarking requires different inclusion criterias, the analysis also makes use of different subset of data.

Table 11: Overview of entities used at different stages of the modelling and post-analysis.

Cohort	Size (N)	Inclusion criterion or use
Total compiled universe	282	All entities in the panel: 227 sovereign nations and dependent territories plus 55 regional, income, or global aggregates.
Target-valid / lenient set	123	Countries with a usable target pathway, used for lenient final holdout evaluation.
Strict evaluation set	109	Target-valid 2022 observations satisfying the $< 30\%$ feature-missingness level.
Forecast universe	282	All entities for which final 2022 budget predictions are generated in all-country mode.
Sovereign and territory forecast set	227	Sovereign nations and dependent territories are kept for physical net-zero year (T_{NZ}) inversion.
Successful T_{NZ} inversions	173	Sovereign nations and territories for which root-finding succeeds within the inversion rules.
Benchmark-valid set	119	Countries with both a successful implied T_{NZ} inversion and an announced national net-zero pledge.

The raw panel contains 282 entities, including 227 sovereign nations and dependent territories and 55 regional, income, or global aggregates. Supervised model evaluation has been made with 123 target-valid countries because observed target pathways are needed to make scoring possible. The strict holdout set is the subset of these countries after applying the $< 30\%$ feature missingness level. Following the training of the models, predictions are generated for 282-entity forecast universe.

However, conversion of predicted budgets into net-zero years is only attempted for the 227 sovereign nations and dependent territories. 173 of these 227 entities produce successful T_{NZ} inversions. The remaining entities fail because missing inputs and too early or too late predictions that are outside the range of 2024-2200. Finally, the benchmark against pledges make use of inversion and an announced national net-zero target, which translates to 119 benchmark valid countries.

The log-budget predictions obtained by bootstrap procedure are inverted to net-zero years (T_{NZ}) using the root-finding procedure described in Section 2.4. For each bootstrap draw, T_{NZ} is identified as the first year in the search window (2024–2200) for which the cumulative per-capita emission budget falls below zero.

T_{NZ} inversion succeeds for 173 of 227 entities. Inversion fails for 30 entities due to missing net-zero data or PRIMAP emissions, for 15 entities where the predicted budget corresponds to a negative emission starting value, and for 6 entities where T_{NZ} exceeds the upper search bound of 2200. One country's T_{NZ} falls below 2024, consistent with a current-policy emission trajectory that has already effectively reached net zero.

Among the 173 countries, the mean T_{NZ} is 2047 with standard deviation 17 years, median 2046, and interquartile range 2035–2053). Figure 24 presents the distribution of bootstrap-median T_{NZ} predictions for linear scenarios.

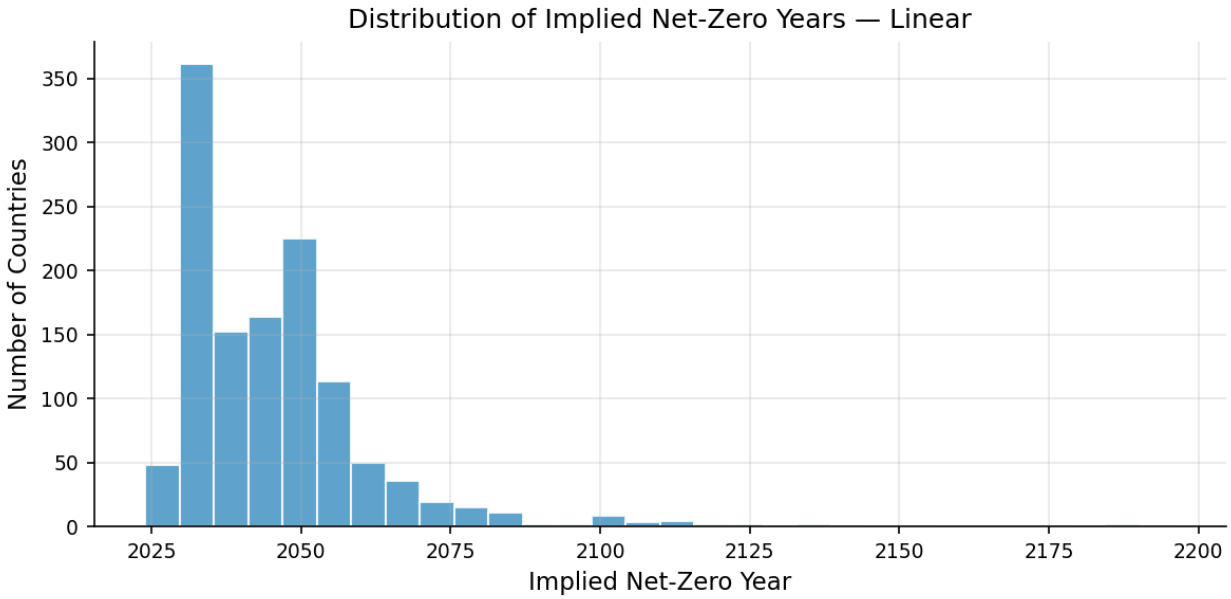

Figure 24: Distribution of implied net-zero years for the linear scenario. Each bar counts the number of (country, iteration). The distribution is right-skewed, with the bulk of predictions concentrated between 2030 and 2060 and a long tail extending beyond 2100.

The distribution of implied T_{NZ} years is concentrated between 2030 and 2060 window, with a long right tail extending to 2100 and beyond. The countries with T_{NZ} values in the 2025–2035 range are mainly small island states and developing nations with very low emissions and near-term pledges. Countries at the far right of the distribution are mostly large developing economies or countries with later pledge years.

5.7 Benchmark Against Pledged Targets

The implied net-zero years are benchmarked against nationally announced net zero target years to evaluate whether countries are going to realize their pledges. Among the 282 entities in the all-country prediction set, 119 are benchmark valid entities. For the primary reference configuration (linear scenario, random CV, simple branch, SVR), the gap between implied and announced T_{NZ} is:

- Ahead: 45 countries (37.8%) will achieve net zero more than five years before its pledged year.
- On track (broad): 48 countries (40.3%) are within a broad ±5-year window around the announced target. This group includes:
  - On track (tight): 26 countries (21.8%) with stricter ±2-year window.
  - On track (moderate): 22 countries (18.5%) are outside the tight window but within ±5-year band.
- Behind: 26 countries (21.8%) exceeds the announced target by more than the five-year window.

The mean gap across the 119 valid countries is −7.4 years (median −3.0 years). This indicates a negative gap that the model implies net zero will be achieved before the announced target. This negative mean is driven by a subset of countries which are major emitters with relatively late pledges. Figure 25 illustrates the country-level gap against the announced T_{NZ} year.

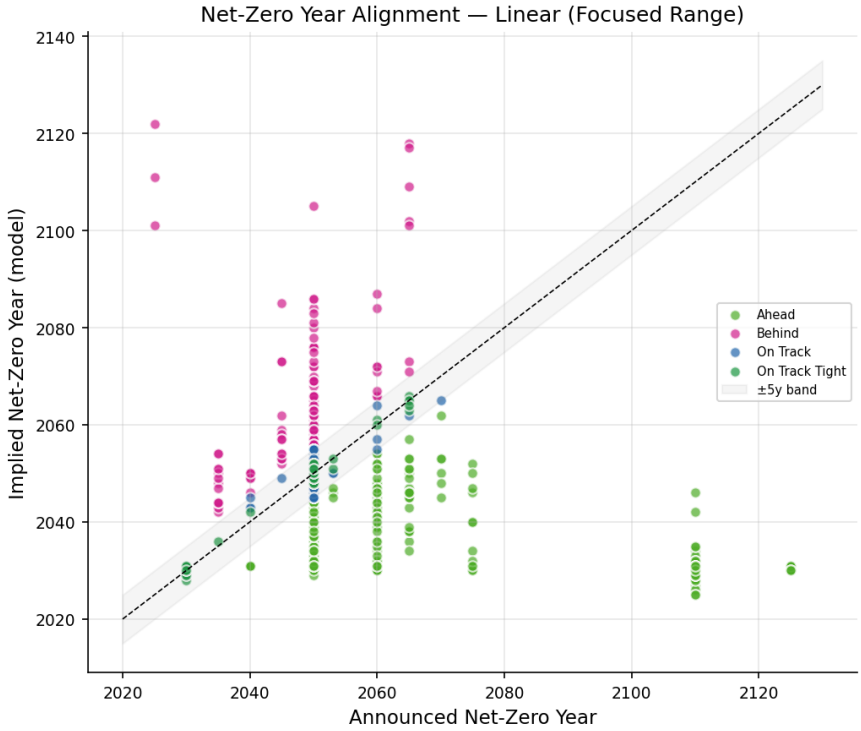

Figure 25: Gap between implied and announced net-zero year for 119 benchmark-valid countries. Colour indicates the acceptability band assigned to each country.

Figure 26 presents the geographic distribution of the T_{NZ} gap. Countries in Western Europe, North America, and parts of East Asia fall under ahead or on-track categories. In contrast, countries in South and Southeast Asia, Latin America, and Africa show a more heterogeneous distribution.

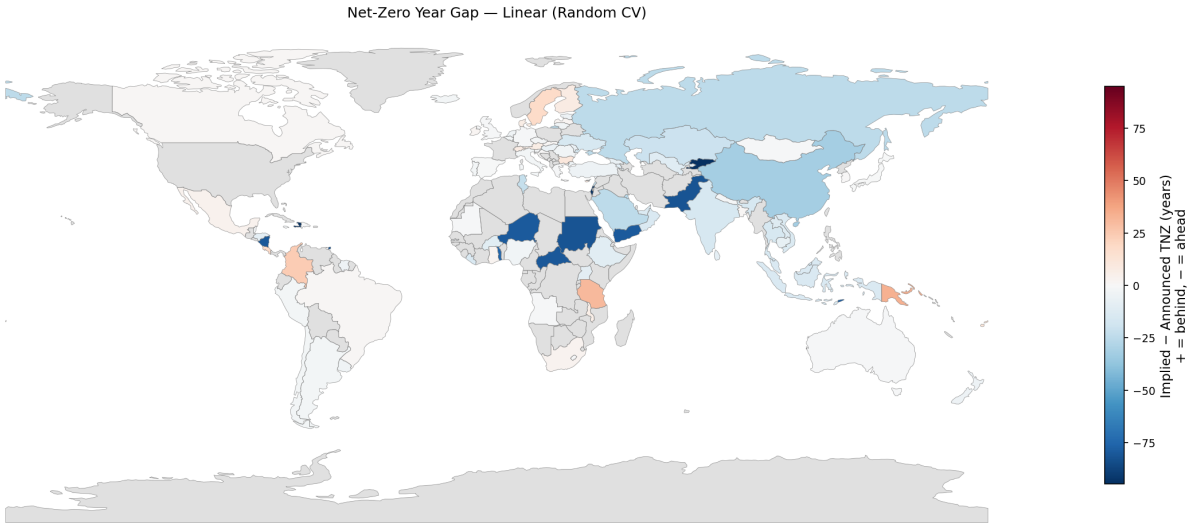

Figure 26: Choropleth map of the gap between implied and announced net-zero year for the linear scenario (random CV, SVR). Grey indicates countries excluded from the benchmark.

Figure 27 shows the distribution of gap values. The distribution is left-skewed, with mode at approximately -2 to $+2$ years and a negative tail. This confirms that the modal country is close to or slightly ahead of its pledged target. The far left tail is the same set of countries rather. It is driven mainly by countries with very late placeholder years because their net-zero commitments are only proposed / in discussion. Countries such as Mauritius (MUS), St. Kitts and Nevis (KNA), Haiti (HTI), Israel (ISR), Kyrgyzstan (KGZ), and Kiribati (KIR) are coded near 2125, while Pakistan (PAK), Trinidad and Tobago (TTO), Sudan (SDN), Togo (TGO), Nicaragua (NIC), Niger (NER), Yemen (YEM), Central African Republic (CAF), and Palau (PLW) are coded near 2110. Their T_{NZ} years are around the late 2020s or early 2030s since they have low predicted cumulative per-capita budgets. The left tail should be interpreted as an informal or incomplete pledges.

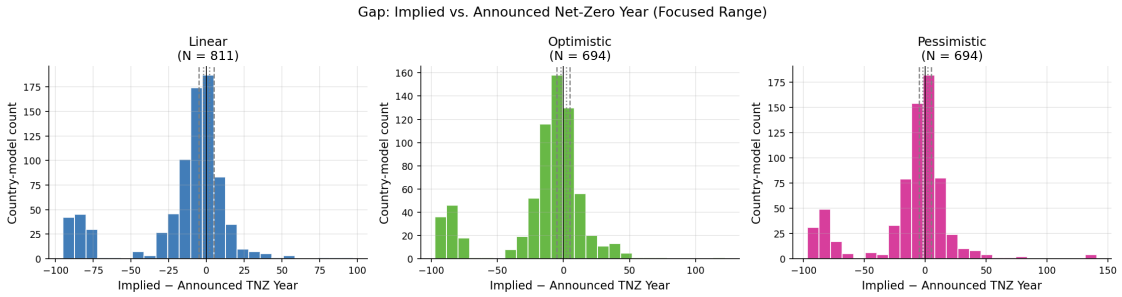

Figure 27: Distribution of T_{NZ} gaps.

6 Discussion and Conclusion

This section frames the empirical findings reported in Section 5 within the broader methodological and policy literature. Section 6.1 interprets the predictive performance of the selected models. Section 6.2 discusses the structural factors found by the SHAP analysis. Section 6.3 addresses methodological robustness and scenario consistency. Section 6.4 examines the implications of the

benchmark analysis for climate policy. Section 6.5 acknowledges the study's principal limitations. Section 6.6 concludes with a summary and directions for future research.

6.1 Model Performance in Context

The consistent dominance of SVR over all configurations is a key finding rather than a product of hyperparameter tuning. SVR with a radial basis function kernel is known to generalise well to high dimensional data with heterogeneous features [32]. The margin and the ε-insensitive tube penalises the model for fitting noisy or outlier observations in the training set. This is a desirable property given the heterogeneous data quality of the assembled panel. Several candidates from the gradient-boosted ensemble family (LightGBM, XGBoost) demonstrated low outer-CV MAE values but did not end up in the winner selection. In contrast, Histogram-based Gradient Boosting and CatBoost secured rank-2 and rank-3 winner positions in select configurations. This mixed pattern is consistent with the observation that boosted ensembles tend to overfit on short time series [42].

The inner-fold overfitting signature observed for SVR, near-zero inner training MAE contrasted with substantially higher inner validation MAE, necessitates interpretation. It does not imply that the inner validation folds are drawn from a different distribution than the outer test folds. This mainly shows the difference between in-sample fit and out-of-sample temporal prediction. SVR with a small ε and large capacity parameter C can fit the inner training subset almost exactly. This stems from the fact that country-level indicators and budgets are highly structured and autocorrelated over time. This returns a very low inner training MAE. However, inner validation MAE is measured on held-out country–year observations that were not used to fit the model, so it represents one-step-ahead validation task and is necessarily higher.

This is why the generalisation comparison is not made on inner training MAE versus outer-test MAE, but inner validation MAE versus outer-test MAE. Both are out-of-sample quantities. The inner validation folds approximate forward prediction during hyperparameter tuning. Similarly, the outer folds evaluate the selected configuration on a fully unseen test year.These two errors are close enough that the outer generalisation gap (outer-test MAE minus inner-validation MAE) does not trigger any all overfitting flag for any winner configuration. Thus, the large train–validation gap should be interpret as evidence of high in-sample capacity and not as distribution shift. This is the behaviour the nested cross-validation framework is designed to detect and control [37]. Thus, the inner folds tune hyperparameters and the outer folds provide an unbiased estimate of the generalisation error on data that is not seen by tuning or training.

The top SVR configurations achieved MAE of 0.153–0.215 log-units. This represents the prediction error of the cumulative per-capita GHG budget to net zero from the year 2022 under the natural lagorithm. It is important to note that the target variable covers 5 log-units. The mean error is 3–4% of the total target range. Exponentiating MAE of 0.213 log-units corresponds to a median multiplicative prediction factor of $e^{0.213} \approx 1.24$, a 24% error in the original budget scale. This is sufficient to rank countries by predicted net-zero readiness and to identify systematic over- or under-shooters. This also requires the acknowledgment of individual country predictions carry significant uncertainty, as quantified by the bootstrap intervals.

The difference between strict (109 observations) and lenient (123 observations) evaluations underscores the difficulty of applying machine-learning models trained on data-rich observations to data-sparse countries. This trade-off between coverage and precision is not unique to this study; it is a general property of regression with heterogeneous data quality [25].

6.2 Structural Drivers of Net-Zero Budget Depletion

The leading predictors of the log-cumulative per-capita GHG budget defined by SHAP is forest area and total GHG emissions per capita. This finding connects the model's behaviour to empirical regularities in the land-use change and emissions literature. Forest area is a strong proxy for a country's past and ongoing deforestation trajectory. The countries with large shares of forest land likely to have lower net emissions from the land use, land-use change, and forestry (LULUCF) sector and smaller remaining cumulative budgets relative to their pledges [3]. Seven year rolling average of total GHG emissions per capita is among the most important features. This shows how a country's sustained emission intensity determines the remaining budget. The predictive importance of land-use and emission features over energy-sector features is due to the data coverage structure. While PRIMAP total GHG emissions are available for all countries in the panel, IEA energy features are missing for almost half of observations. However, it is also important to acknowledge that for many developing countries, land use and emission intensity account for a disproportionately large share of emissions.

The country PCA components emerge as important predictors. This reflects the utility of latent structural embeddings in cross-country regression. First principal component represents a development and institutional-governance axis. Its positive side is driven by WGI index and income features, while its negative side is associated especially with INFORM risk, developmental features. It therefore distinguishes high-income, stable, high-governance countries from more vulnerable or data-sparse developing contexts. These, in return, approximates the macrodevelopment and institutional capacity connected to decarbonisation. The reliance of the model on principal component suggests that composite structural representations add more information than individual features.

The GDP per capita rolling mean and energy intensity are among the most economically interpretable predictors. The Environmental Kuznets Curve hypothesis assumes that emissions first rise with income and then fall as economies transition from manufacturing to services. This provides a foundation for the idea that as wealthier economies are more likely to complete the energy transition or to have the institutional and financial resources to accelerate it [43]. Energy intensity (MJ per USD PPP) measures the efficiency of economic production. Lower energy intensity means that to get the same economic output, less energy input is required. This is primarily associated with advanced economies and the entities invested in energy efficiency programmes.

INFORM risk index is a measure of humanitarian fragility. As it is a high-importance predictor with its rolling mean and interaction terms, it deserves an attention. Countries with high humanitarian risk scores are likely to have lower institutional capacity and higher vulnerability. These are barriers to rapid decarbonisation [3]. The negative SHAP direction of high INFORM values suggests that model has learnt fragile states with low per-capita absolute emissions face larger remaining challenges due to slow near-term reductions. Similarly, regulatory quality reflects the power of enforcement a state can implement environmental regulations.

6.3 Scenario Consistency and Methodological Robustness

The stability among model rankings and qualitative conclusions over three emission scenarios dictated by α and two imputation branches is a key feature of methodological robustness. The scenario exponent α is used in the construction of target variables. α is expected to affect both the scale of the target and the ordering of countries. Model performance by MAE scores tracks the scenario due to the variation in distributional properties of target values. The pessimistic pathway put weight on the smooth post-2030 power-law component of the cumulative budget and making the target more protected against noisy base-year emissions and reporting fluctuations. However, the optimistic

pathway assigns less weight to the post-2030 component. This helps in explanation of optimistic pathway's lower MAE without changing the overall model ranking.

The prediction agreement analysis in Figure 18 shows that the random-CV and temporal-CV SVR models yields the same country level rankings ($r > 0.98$). This is important in the benchmark analysis. The conclusion is that the majority of countries are predicted to reach net zero before or broadly near their pledged date does not depend on CV strategy choice. Simultaneously, the countries that are identified as behind schedule is same across strategies and scenarios, with the same set of countries that are typically lower-income nations with late pledges and high data missingness. This happens irrespective of pathway assumption.

The sensitivity analysis of the simple imputation branch and MICE for SVR is another finding. Theoretically, MICE should produce better imputed values, since it models the joint feature distribution [13]. In practice, the MICE branch is a deliberately hybrid rather than a pure chained-equations imputer. The imputation matrix is aware of country and year median features. This should stabilise MICE under the high missingness where entire feature histories may be absent for some countries. The distinction between the two branches is partially collapsed due to awareness of median values. The large number of auxiliary predictors can also increase the risk of overfitting the imputation model in inner folds with less observations.

The MICE branch could have been constructed in a different way. One alternative would be a pure MICE that uses only the observed continuous predictors. However, it would likely be less stable for rows with extensive missingness. A second alternative would be tree-based chained imputation, such as Random Forest or Extra Trees imputation that can capture nonlinear feature interactions. However, this approach is computationally more expensive inside nested cross-validation. A third option could be a group-wise MICE by region or income group. This would allow for different correlation structures for developed and developing economies, but at the cost of smaller and noisier imputation samples within each group. The present design can be considered as a pragmatic trade-off between stability and flexibility. The finding that simple imputation is comparable to MICE in this context can not be generalised to panels with more complete data.

6.4 Net-Zero Feasibility: Pledges Versus Predicted Trajectories

The benchmark analysis reveals that significant portion of the 119 benchmark-valid countries (approximately 78%) are predicted to reach net zero before or within an acceptable window relative to their formally pledged target years. Approximately 22% are predicted to be later than their announced commitment. The mean implied-minus-announced gap is -7.4 years. This is mainly result of countries with late-dated pledges years (2070–2110). For these countries, the model projects rapid near-term budget depletion. The median gap, which is less sensitive to outliers, is -3 years. This suggests that the typical country is predicted to achieve net zero only slightly ahead of their own pledge.

These results must be interpreted with caution. This does not indicate that current policies are sufficient to achieve pledged net-zero targets. The results indicate that if the patterns observed between country features and emission budgets until now persists into the future, then many countries' pledged timelines are consistent with the trajectory of their emission features. Three interpretive caveats are particularly important.

First of all, the model predicts the cumulative per-capita emission budget to net zero by relying on country characteristics in 2022 and it does not predict the actual emission trajectory. The implied T_{NZ} year is calculated by reversing the budget prediction calculation steps in Section 2.3 and in Section 2.4. These involves through the assumed power-law pathway which uses a specific decarbonisation shape parameter (α) and also relies on the 2030 current-policy projection from

Climate Action Tracker and related sources. If any assumption on actual policies diverge from this current-policy baseline, the implied T_{NZ} will shift accordingly in ways that the model cannot anticipate.

Secondly, the analysis evaluates countries against their announced commitments and not against a science-based global carbon budget that is consistent with the 1.5 °C or 2 °C temperature targets of the Paris Agreement. A country that is predicted to realize its own national pledge may still be increasing the aggregate global warming beyond these limits if the pledge is insufficient [3]. Thus, the alignment of national pledges and global temperature targets is a separate question, and it is outside the scope of this study.

Lastly, the negative skew in the gap distribution may partly reflect an optimistic bias in the model. Countries with large remaining emission budgets is likely to be those with delayed near-term reductions. For this reason, the structural profile the model associates them with a slow depletion rate and a later-than-pledged T_{NZ}. However, the model is trained on historical patterns, and historical decarbonisation rates may systematically underestimate the requirement of accelaration to realize the pledges. This would particularly apply to countries with commitment to rapid transitions from current high-emission levels.

6.5 Limitations

Several limitations of the study requires acknowledgement.

Target variable construction. The target variable is calculated through a combination of data sources. Each of these sources carries their own uncertainties. PRIMAP-hist emissions are adjusted and standardised from UNFCCC national inventory reports. UNFCCC national inventory reports vary in completeness and methodology over countries, as it is acknowledged by Gütschow et al. [18] The power-law pathway parameterisation ($\alpha \in \{0.5, 1.0, 2.0\}$) provides three decarbonisation shapes. However, it does not cover all plausible trajectories. Also, the shape parameters by themselves do not represent trajectories of all countries in the dataset.

Data coverage and generalisability. The strict evaluation set of 109 observations represents a biased sample. It mainly takes into account the higher-income countries with comprehensive statistical systems. The models are trained on observations from data-rich. This limits the capacity of the models and also the confidence of their predictions for data-sparse countries.

Causal inference and policy application. By their design, the models are predictive tools and not a causal one. High SHAP importance for a feature does not imply that policies target that feature feature will reduce the country's cumulative remaining emissions. It rather reflects that feature correlates with some unobserved combination of factors. The model is useful for ranking countries by predicted net-zero readiness, but not for precisely assessing individual policy actions.

Temporal scope and out-of-sample extrapolation. The model is trained and tested on data from 2000 to 2022. Using it to estimate T_{NZ} years over a horizon from 2024 to 2200 requires prediction beyond the training data. The further the estimated T_{NZ} year, the more uncertainty introduced to the prediction since the model assumes the observed patterns will persist into the future. In reality, it is highly probable that major changes by energy transitions, geopolitical shocks and so forth will happen and the models do not fully captures those effects.

Cross-sectional homogeneity. The model combines all countries and train under one regression model. The differences between countries are represented through the use of country embeddings, rolling statistics and governance indices. While this approach is computationally efficient and statistically reasonable, it still assumes that the relationship between the inputs and target variable is same for every country. Country-specific models or hierarchical specifications [25] might capture these patterns but they would require more data for individual countries and more coverage.

6.6 Conclusion

This study introduces and evaluates a machine-learning framework for predicting the log-transformed cumulative per-capita greenhouse gas emission budget from 2022 to the net-zero year for countries with formal net-zero commitments. The framework builds three target variables from a nine source panel dataset that covers 282 entities from 2000 to 2024. It trains 13 regression estimators with a nested cross-validation and bootstrap uncertainty quantification.

Support vector regression achieves the lowest MAE scores between 0.153 and 0.215 across all configurations against a dummy-baseline MAE of approximately 0.92. The model explains more than 90% of the variance in the log-budget on the 2022 holdout partition and the bootstrap procedure with 200 iterations provides predictive distributions for all countries.

SHAP analysis identifies land-use structure (forest area), total GHG emissions per capita, latent country level structural profiles (PCA embeddings), economic capacity (GDP per capita and energy intensity), hard-to-abate sector emissions, and regulatory quality as the highest contributing features. These factors collectively reflect both the emission-generating and emission-abating structural characteristics of national economies. Their importance is consistent with established empirical climate-economy literature.

The benchmark analysis suggests that the majority of countries are predicted to achieve net zero before or within an acceptable window compared to their announced target years. The mean implied-versus-pledged gap is -7.4 years (median -3 years). This indicates that most countries' pledged years are consistent with the emission reduction trajectories. However, 26 countries are predicted to be behind schedule. The predicted years for several large developing economies carry wide confidence intervals which can be attributed to their data-coverage limitations.

These findings should be interpreted with caution. The target variables are not a directly observed. They are based on measurements and/or calculations combination of data sources with their own assumptions and uncertainties. Thus, models do not predict when each country will reach net-zero, but rather estimates implied net-zero year based on each countries' emission related features. The model is best used as a tool to compare pledges. It would show whether a given country reach net-zero level early, later or on time. It should not be treated as a forecasting model. The lenient scores show that data-sparsity of countries has a significant effect on the models capacity to predict.

Despite the limitations, the study shows that machine learning applied to harmonised country–year panel data can identify useful and interpretable patterns in countries' net-zero trajectories. As the ambition of net-zero pledges are facing a growing policy attention [3], this type of quantitative assessments can be useful complement to qualitative policy analysis.

References

[1] Michel G. J. den Elzen, S. C. J. Weenk, Leonardo Nascimento, Said Çetinkaya, S. P. Troost, B. van Os, and Ioannis Dafnomilis. Climate ambition explained: key indicators and net-zero target projections using machine learning. *Mitigation and Adaptation Strategies for Global Change*, 31:60, 2026. doi: 10.1007/s11027-026-10330-4.

[2] United Nations Framework Convention on Climate Change. Paris agreement. Treaties and agreements, 2015. URL https://unfccc.int/documents/37107. Publication date: 12 December 2015.

[3] IPCC. *Climate Change 2022: Mitigation of Climate Change. Contribution of Working Group III to the Sixth Assessment Report of the Intergovernmental Panel on Climate Change*. Cambridge University Press, Cambridge, UK, 2022. doi: 10.1017/9781009157926.

[4] Heleen L. van Soest, Michel G. J. den Elzen, and Detlef P. van Vuuren. Net-zero emission targets for major emitting countries consistent with the Paris Agreement. *Nature Communications*, 12:2140, 2021. doi: 10.1038/s41467-021-22294-x.

[5] William F. Lamb and Jan C. Minx. The political economy of national climate policy: Architectures of constraint and a typology of countries. *Energy Research & Social Science*, 64:101429, 2020. doi: 10.1016/j.erss.2020.101429.

[6] Leonardo Nascimento, Michel den Elzen, Takeshi Kuramochi, Sophie Woollands, Ioannis Dafnomilis, Milla Moisio, Mark Roelfsema, Nicklas Forsell, and Zoila Araujo Gutierrez. Comparing the sequence of climate change mitigation targets and policies in major emitting economies. *Journal of Comparative Policy Analysis: Research and Practice*, 26(3–4):233–250, 2024. doi: 10.1080/13876988.2023.2255151.

[7] Yann Robiou du Pont, Louise Jeffery, Johannes Gütschow, Joeri Rogelj, Peter Christoff, and Malte Meinshausen. Equitable mitigation to achieve the Paris Agreement goals. *Nature Climate Change*, 7(1):38–43, 2017. doi: 10.1038/nclimate3186.

[8] Andreas von Dulong and Achim Hagen. Institutions make a difference: assessing the predictors of climate policy stringency using machine learning. *Environmental Research Letters*, 20(1): 014056, 2024. doi: 10.1088/1748-9326/ada0cb.

[9] Tianqi Chen and Carlos Guestrin. XGBoost: A scalable tree boosting system. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 785–794, 2016. doi: 10.1145/2939672.2939785.

[10] Scott M. Lundberg and Su-In Lee. A unified approach to interpreting model predictions. In *Advances in Neural Information Processing Systems*, volume 30, 2017. doi: 10.48550/arXiv.1705.07874. URL https://proceedings.neurips.cc/paper_files/paper/2017/hash/8a20a8621978632d76c43dfd28b67767-Abstract.html.

[11] Trevor Hastie, Robert Tibshirani, and Jerome Friedman. *The Elements of Statistical Learning*. Springer, 2nd edition, 2009. doi: 10.1007/978-0-387-84858-7.

[12] James Bergstra, Rémi Bardenet, Yoshua Bengio, and Balázs Kégl. Algorithms for hyper-parameter optimization. In *Advances in Neural Information Processing Systems*, volume 24, pages 2546–2554, 2011. URL https://papers.nips.cc/paper/4443.

[13] Stef van Buuren and Karin Groothuis-Oudshoorn. mice: Multivariate imputation by chained equations in R. *Journal of Statistical Software*, 45(3):1–67, 2011. doi: 10.18637/jss.v045.i03.

[14] John Lang, Camilla Hyslop, Diego Manya, Sybrig Smit, Peter Chalkley, John Bervin Galang, Frances Green, Thomas Hale, Frederic Hans, Nick Hay, Angel Hsu, Denise Hsu, Takeshi Kuramochi, Steve Smith, and Helen Tatlow. Net zero tracker, 2025. URL https://zerotracker.net/. Accessed 19 May 2026.

[15] PBL Netherlands Environmental Assessment Agency. The pbl climate pledge ndc tool, 2025. URL https://themasites.pbl.nl/o/climate-ndc-policies-tool/. Updated 23 April 2025; accessed 19 May 2026.

[16] Gokul Iyer, Yang Ou, Jae Edmonds, Allen A. Fawcett, Nate Hultman, James McFarland, Jay Fuhrman, Stephanie Waldhoff, and Haewon McJeon. Ratcheting of climate pledges needed to limit peak global warming. *Nature Climate Change*, 12(12):1129–1135, 2022. doi: 10.1038/s41558-022-01508-0.

[17] Ioannis Dafnomilis, Michel den Elzen, and Detlef van Vuuren. Paris targets within reach by aligning, broadening and strengthening net-zero pledges. *Communications Earth & Environment*, 5(1):48, 2024. doi: 10.1038/s43247-023-01184-8.

[18] Johannes Gütschow, Daniel Busch, and Mika Pflüger. The primap-hist national historical emissions time series v2.6 (1750–2023), 2024. URL https://doi.org/10.5281/zenodo.13752654. Accessed 2026-04-01.

[19] Joint Research Centre (European Commission). Inform report 2024: 10 years of inform – shared evidence for managing crises and disasters. Technical report, Publications Office of the European Union, 2024. URL https://op.europa.eu/en/publication-detail/-/publication/b0ac7f32-e4f1-11ee-8b2b-01aa75ed71a1/. Catalogue number: KJ-NA-31-820-EN-N.

[20] Michael Coppedge, John Gerring, Carl Henrik Knutsen, Staffan I. Lindberg, et al. V-dem dataset v14, 2024. URL https://v-dem.net/dsarchive.html. Country-year dataset, version 14 (published March 2024).

[21] United Nations, Department of Economic and Social Affairs, Population Division. World population prospects 2024: Summary of results, 2024. URL https://desapublications.un.org/publications/world-population-prospects-2024-summary-results. Official citation format provided by UN DESA.

[22] Michel G. J. den Elzen, Ioannis Dafnomilis, Leonardo Nascimento, Arthur Beusen, Nicklas Forsell, Joost Gubbels, Mathijs Harmsen, Elena Hooijschuur, Zuelclady Araujo Gutierrez, and Takeshi Kuramochi. Uncertainties around net-zero climate targets have major impact on greenhouse gas emissions projections. *Annals of the New York Academy of Sciences*, 1544:209–222, 2025. doi: 10.1111/nyas.15285.

[23] John W. Tukey. *Exploratory Data Analysis*. Addison-Wesley, 1977.

[24] Ian T. Jolliffe. *Principal Component Analysis*. Springer, 2nd edition, 2002. doi: 10.1007/b98835.

[25] Andrew Gelman and Jennifer Hill. *Data Analysis Using Regression and Multilevel/Hierarchical Models*. Cambridge University Press, 2007. doi: 10.1017/CBO9780511790942.

[26] Rob J. Hyndman and George Athanasopoulos. *Forecasting: Principles and Practice*. OTexts, Melbourne, Australia, 2nd edition, 2018. URL https://otexts.com/fpp2/.

[27] Isabelle Guyon, Jason Weston, Stephen Barnhill, and Vladimir Vapnik. Gene selection for cancer classification using support vector machines. *Machine Learning*, 46(1–3):389–422, 2002. doi: 10.1023/A:1012487302797.

[28] Fabian Pedregosa et al. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830, 2011. URL https://jmlr.org/papers/v12/pedregosa11a.html.

[29] Guolin Ke, Qi Meng, Thomas Finley, Taifeng Wang, Wei Chen, Weidong Ma, Qiwei Ye, and Tie-Yan Liu. LightGBM: A highly efficient gradient boosting decision tree. In *Advances in Neural Information Processing Systems*, volume 30, 2017. URL https://papers.nips.cc/paper/6907.

[30] Liudmila Prokhorenkova, Gleb Gusev, Aleksandr Vorobev, Anna Veronika Dorogush, and Andrey Gulin. CatBoost: Unbiased boosting with categorical features. In *Advances in Neural Information Processing Systems*, volume 31, 2018. URL https://papers.nips.cc/paper/7898.

[31] Hui Zou and Trevor Hastie. Regularization and variable selection via the elastic net. *Journal of the Royal Statistical Society: Series B*, 67(2):301–320, 2005. doi: 10.1111/j.1467-9868.2005.00503.x.

[32] Alex J. Smola and Bernhard Sch"olkopf. A tutorial on support vector regression. *Statistics and Computing*, 14(3):199–222, 2004. doi: 10.1023/B:STCO.0000035301.49549.88.

[33] Yoav Freund and Robert E. Schapire. A decision-theoretic generalization of on-line learning and an application to boosting. *Journal of Computer and System Sciences*, 55(1):119–139, 1997. doi: 10.1006/jcss.1997.1504.

[34] Leo Breiman. Random forests. *Machine Learning*, 45(1):5–32, 2001. doi: 10.1023/A:1010933404324.

[35] Pierre Geurts, Damien Ernst, and Louis Wehenkel. Extremely randomized trees. *Machine Learning*, 63(1):3–42, 2006. doi: 10.1007/s10994-006-6226-1.

[36] Kurt Hornik, Maxwell Stinchcombe, and Halbert White. Multilayer feedforward networks are universal approximators. *Neural Networks*, 2(5):359–366, 1989. doi: 10.1016/0893-6080(89)90020-8.

[37] Sudhir Varma and Richard Simon. Bias in error estimation when using cross-validation for model selection. *BMC Bioinformatics*, 7:91, 2006. doi: 10.1186/1471-2105-7-91.

[38] Christoph Bergmeir and José M. Benítez. On the use of cross-validation for time series predictor evaluation. *Information Sciences*, 191:192–213, 2012. doi: 10.1016/j.ins.2011.12.028.

[39] James Bergstra and Yoshua Bengio. Random search for hyper-parameter optimization. *Journal of Machine Learning Research*, 13:281–305, 2012. URL https://jmlr.org/papers/v13/bergstra12a.html.

[40] Cort J. Willmott and Kenji Matsuura. Advantages of the mean absolute error (MAE) over the root mean square error (RMSE) in assessing average model performance. *Climate Research*, 30(1):79–82, 2005. doi: 10.3354/cr030079.

[41] Bradley Efron. Bootstrap methods: Another look at the jackknife. *The Annals of Statistics*, 7 (1):1–26, 1979. doi: 10.1214/aos/1176344552.

[42] Jerome H. Friedman. Greedy function approximation: A gradient boosting machine. *The Annals of Statistics*, 29(5):1189–1232, October 2001. doi: 10.1214/aos/1013203451. URL https://doi.org/10.1214/aos/1013203451.

[43] Gene M. Grossman and Alan B. Krueger. Environmental impacts of a north american free trade agreement. NBER Working Paper 3914, National Bureau of Economic Research, Cambridge, MA, November 1991. URL https://www.nber.org/papers/w3914.

A Iceland Pathway and Target Construction

This appendix provides a complete worked example of the target-variable construction procedure for Iceland (ISL). The purpose is to make the calculation in Section 2.3 transparent by showing how historical emissions, the 2024 prediction cut-off, the 2030 target point, the pledged net-zero year, population normalisation, backwards accumulation, and logarithmic transformation are combined for one country.

Figure 28 visualises the emission pathway inputs used for the example. The black line shows Iceland's historical PRIMAP greenhouse gas emissions from 2000 to 2024. The blue marker identifies the 2022 baseline used for final model prediction, the green marker shows the 2023 observed emissions point, and the orange marker and vertical line indicate the 2024 prediction cut-off value from which the forward pathway is started. Emissions then decline linearly from the 2024 cut-off to the 2030 NDC target of 9.14 Mt CO_2e, shown by the red marker and vertical line. From 2030 to the 2040 pledged net-zero year, the three scenario pathways diverge according to the power-law exponent α: the optimistic pathway ($\alpha = 2.0$) front-loads reductions, the linear pathway ($\alpha = 1.0$) declines at a constant rate, and the pessimistic pathway ($\alpha = 0.5$) delays more of the reduction until later in the transition period. All three pathways converge to zero emissions at the 2040 net-zero target year.

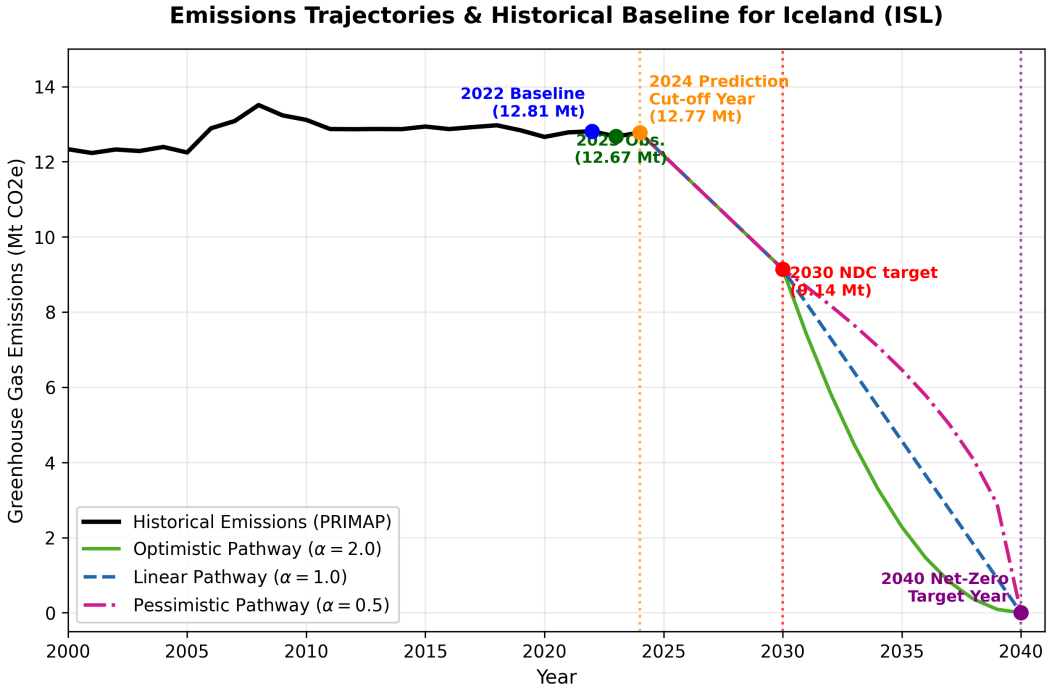

Figure 28: Iceland emission pathways used in target variable construction.

Table 12 gives the corresponding spreadsheet-style calculation for the optimistic pathway. The rows are ordered backwards from the 2040 net-zero year to 2000 because the target at each year is a remaining cumulative budget from that year through net zero. Column B records observed historical emissions where available, while Column C gives the population series used to express annual emissions on a per-capita basis. Column D is the linear interpolation weight for the 2024–2030 segment, and Column E applies that weight to move from the 2024 cut-off emissions value to the 2030 NDC target. Columns F–H construct the post-2030 power-law decay: Column F is the normalised transition time τ_t, Column G is the remaining fraction $1-\tau_t$, and Column H squares this fraction for

the optimistic scenario ($\alpha = 2.0$). Column I converts the squared fraction into the net-zero pathway emission value between 2030 and 2040. Column J selects the annual emissions value used in the final pathway, taking observed emissions for historical years, interpolated emissions for 2024–2030, and the power-law pathway for 2030–2040. Column K divides annual emissions by population to obtain tonnes CO_2e per person. Column L then accumulates Column K backwards, so that each row gives the remaining cumulative per-capita budget from that year to 2040. Finally, Column M applies the natural logarithm to Column L, producing the log-budget target value $y_t = \ln(B_t)$ used for model training whenever $B_t > 0$.

Table 12: Granular Excel-Style Backwards Target Variable Construction for Iceland (ISL)

Year	**Col B**	**Col C**	**Col D**	**Col E**	**Col F**	**Col G**	**Col H**	**Col I**	**Col J**	**Col K**	**Col L**	**Col M**
(t)	E_{hist} [Mt]	P_t [k]	w_t [Ratio]	E_{NDC} [Mt]	τ_t [Ratio]	$1-\tau_t$ [Ratio]	$(1-\tau_t)^2$ [Ratio]	E_{NZ} [Mt]	E_t [Mt]	e_t^{pc} [t/cap]	B_t [t/cap]	$y_t = \ln(B_t)$ [Log]
2040	–	429.5	–	–	1.00	0.00	0.0000	0.000	0.000	0.000	0.000	–
2039	–	428.5	–	–	0.90	0.10	0.0100	0.091	0.091	0.213	0.213	-1.5447
2038	–	427.5	–	–	0.80	0.20	0.0400	0.366	0.366	0.856	1.069	0.0667
2037	–	426.3	–	–	0.70	0.30	0.0900	0.823	0.823	1.930	2.999	1.0984
2036	–	425.0	–	–	0.60	0.40	0.1600	1.463	1.463	3.442	6.441	1.8627
2035	–	423.6	–	–	0.50	0.50	0.2500	2.286	2.286	5.397	11.838	2.4713
2034	–	422.1	–	–	0.40	0.60	0.3600	3.292	3.292	7.799	19.637	2.9774
2033	–	420.4	–	–	0.30	0.70	0.4900	4.480	4.480	10.657	30.294	3.4110
2032	–	418.6	–	–	0.20	0.80	0.6400	5.852	5.852	13.979	44.274	3.7904
2031	–	416.6	–	–	0.10	0.90	0.8100	7.406	7.406	17.778	62.051	4.1280
2030	–	414.3	1.000	9.144	0.00	1.00	1.0000	9.144	9.144	22.068	84.119	4.4322
2029	–	411.9	0.833	9.749	–	–	–	–	9.749	23.670	107.789	4.6802
2028	–	409.0	0.667	10.354	–	–	–	–	10.354	25.313	133.102	4.8911
2027	–	405.8	0.500	10.959	–	–	–	–	10.959	27.003	160.105	5.0758
2026	–	402.3	0.333	11.564	–	–	–	–	11.564	28.742	188.848	5.2409
2025	–	398.3	0.167	12.169	–	–	–	–	12.169	30.555	219.403	5.3909
2024	12.774	386.5	0.000	12.774	–	–	–	–	12.774	33.050	251.874	5.5289
2023	12.671	385.7	–	–	–	–	–	–	12.671	32.856	284.729	5.6515
2022	12.810	382.0	–	–	–	–	–	–	12.810	33.534	318.264	5.7629
2021	12.781	372.5	–	–	–	–	–	–	12.781	34.310	352.574	5.8653
2020	12.660	366.5	–	–	–	–	–	–	12.660	34.547	387.121	5.9587
2019	12.835	360.6	–	–	–	–	–	–	12.835	35.598	422.719	6.0467
2018	12.970	352.7	–	–	–	–	–	–	12.970	36.771	459.489	6.1301
2017	12.922	343.4	–	–	–	–	–	–	12.922	37.629	497.118	6.2088
2016	12.869	335.4	–	–	–	–	–	–	12.869	38.364	535.482	6.2832
2015	12.936	330.8	–	–	–	–	–	–	12.936	39.104	574.585	6.3536
2014	12.866	327.4	–	–	–	–	–	–	12.866	39.300	613.886	6.4198
2013	12.871	323.8	–	–	–	–	–	–	12.871	39.756	653.641	6.4826
2012	12.865	320.7	–	–	–	–	–	–	12.865	40.114	693.756	6.5421
2011	12.870	319.0	–	–	–	–	–	–	12.870	40.343	734.099	6.5986
2010	13.114	318.0	–	–	–	–	–	–	13.114	41.234	775.332	6.6533
2009	13.235	318.5	–	–	–	–	–	–	13.235	41.553	816.885	6.7055
2008	13.508	317.4	–	–	–	–	–	–	13.508	42.558	859.443	6.7563
2007	13.084	311.6	–	–	–	–	–	–	13.084	41.995	901.438	6.8040
2006	12.889	303.8	–	–	–	–	–	–	12.889	42.429	943.867	6.8500
2005	12.246	296.7	–	–	–	–	–	–	12.246	41.270	985.137	6.8928
2004	12.394	292.1	–	–	–	–	–	–	12.394	42.433	1027.570	6.9350
2003	12.286	289.5	–	–	–	–	–	–	12.286	42.434	1070.004	6.9754
2002	12.326	287.5	–	–	–	–	–	–	12.326	42.869	1112.873	7.0147
2001	12.232	285.0	–	–	–	–	–	–	12.232	42.925	1155.798	7.0525
2000	12.333	281.2	–	–	–	–	–	–	12.333	43.859	1199.657	7.0898

B Feature Selection via Recursive Feature Elimination

To mitigate the curse of dimensionality and eliminate redundant or uninformative predictors, a Recursive Feature Elimination (RFE) procedure is employed. The feature selection workflow is governed by a set of hyperparameters that control the pruning process and the behavior of the underlying ensemble estimator:

1. **Target Feature Subset Size (N):** The feature elimination process terminates when the active feature space is reduced to exactly $N = 24$ predictors. This threshold balances model parsimony with the preservation of predictive capacity.

2. **Pruning Step Fraction (γ):** In each step of the recursive pruning cycle, $\gamma = 50\%$ of the remaining features—specifically those exhibiting the lowest importance scores—are pruned simultaneously. This geometric step size accelerates computation relative to single-feature elimination while maintaining stable, robust feature rankings.

3. **Base Estimator:** An Extremely Randomized Trees (Extra-Trees) regressor is utilized to evaluate feature importances at each elimination step. The regressor itself is configured with the following hyperparameters:

   - **Ensemble Size (M):** The forest comprises $M = 50$ independent decision trees.
   - **Splitting Feature Subset Size ($m_{\mathbf{try}}$):** At each node split, the algorithm randomly samples $m_{\text{try}} = \sqrt{p}$ features from the p active features in that iteration, selecting the optimal split from randomly generated thresholds.
   - **Minimum Leaf Size ($n_{\mathbf{min}}$):** A minimum of $n_{\text{min}} = 2$ training observations is required to form a terminal node, which dampens model variance and regularizes individual trees.
   - **Randomization Seed:** A fixed seed of 42 is specified to ensure deterministic split generation and exact reproducibility of the feature selection rankings.

Table 13: Hyperparameter Configuration for the Recursive Feature Elimination (RFE) and Extremely Randomized Trees (Extra-Trees) Pipeline

Component / Hyperparameter	**Symbol / Value**	**Description**
Recursive Feature Elimination (RFE)		
Target Feature Subset Size	$N = 24$	Final number of selected features to retain
Pruning Step Fraction	$\gamma = 0.50$	Proportion of low-ranking features eliminated per step
Base Estimator	Extra-Trees	Ensemble estimator used to compute feature importances
Extremely Randomized Trees (Extra-Trees)		
Ensemble Size	$M = 50$	Number of decision trees in the forest ensemble
Node Splitting Feature Pool	$m_{\text{try}} = \sqrt{p}$	Size of the random feature subset evaluated at each node split
Minimum Leaf Size	$n_{\text{min}} = 2$	Minimum number of training samples required in a terminal node
Deterministic Seed	42	Seed for the pseudo-random number generator for reproducibility

C TPE Hyperparameter Search Spaces

Table 14: Comprehensive TPE Hyperparameter Search Spaces, Optimization Budgets, and Fixed Configurations for All Models

Model (Max Trials)	Hyperparameter	Notation	Search Space / Range	Fixed Parameter(s)
XGBoost (50)	Number of Trees	M	$\mathcal{U}_{\mathbb{Z}}(50, 500, 10)$	Objective: Squared Error
	Maximum Depth	d	$\mathcal{U}_{\mathbb{Z}}(2, 5, 1)$	Tree Method: Histogram
	Learning Rate	η	$\log\text{-}\mathcal{U}(0.005, 0.3)$	Random Seed: 42
	Row Subsample Ratio	s	$\mathcal{U}(0.5, 1.0)$	
	Feature Subsample Ratio	c	$\mathcal{U}(0.5, 0.9)$	
	L2 Regularization	λ	$\log\text{-}\mathcal{U}(0.1, 100.0)$	
	L1 Regularization	α	$\log\text{-}\mathcal{U}(10^{-4}, 50.0)$	
	Minimum Child Weight	w	$\mathcal{U}_{\mathbb{Z}}(5.0, 50.0, 1.0)$	
	Minimum Loss Reduction	γ	$\mathcal{U}(0.0, 1.0)$	
LightGBM (50)	Boosting Type	–	Categorical({gbdt, dart})	Objective: Absolute Loss (L_1)
	Number of Trees	M	$\mathcal{U}_{\mathbb{Z}}(50, 400, 10)$	Random Seed: 42
	Maximum Depth	d	$\mathcal{U}_{\mathbb{Z}}(2, 6, 1)$	
	Maximum Leaf Nodes	L	$\mathcal{U}_{\mathbb{Z}}(8, 32, 4)$	
	Learning Rate	η	$\log\text{-}\mathcal{U}(0.005, 0.2)$	
	Minimum Samples per Leaf	n_{leaf}	$\mathcal{U}_{\mathbb{Z}}(20, 80, 5)$	
	Row Subsample Ratio	s	$\mathcal{U}(0.5, 1.0)$	
	Feature Subsample Ratio	c	$\mathcal{U}(0.4, 1.0)$	
	L1 Regularization	α	$\log\text{-}\mathcal{U}(10^{-3}, 50.0)$	
	L2 Regularization	λ	$\log\text{-}\mathcal{U}(10^{-3}, 100.0)$	
CatBoost (45)	Number of Iterations	M	$\mathcal{U}_{\mathbb{Z}}(150, 400, 25)$	Loss Function: Absolute Error
	Maximum Depth	d	$\mathcal{U}_{\mathbb{Z}}(5, 7, 1)$	Random Seed: 42
	Learning Rate	η	$\log\text{-}\mathcal{U}(0.005, 0.3)$	
	L2 Leaf Regularization	λ	$\log\text{-}\mathcal{U}(1.0, 100.0)$	
	Random Strength	–	$\log\text{-}\mathcal{U}(0.01, 50.0)$	
	Bagging Temperature	T	$\mathcal{U}(0.0, 1.0)$	
Random Forest (40)	Number of Trees	M	$\mathcal{U}_{\mathbb{Z}}(100, 500, 25)$	Random Seed: 42
	Maximum Depth	d	Categorical({5, 10, 12, 15})	
	Minimum Split Samples	n_{split}	$\mathcal{U}_{\mathbb{Z}}(10, 100, 2)$	
	Minimum Leaf Samples	n_{leaf}	$\mathcal{U}_{\mathbb{Z}}(15, 60, 1)$	
	Feature Pool Ratio	c	Categorical({0.5, 0.7})	
Extra Trees (40)	Number of Trees	M	$\mathcal{U}_{\mathbb{Z}}(100, 500, 25)$	Random Seed: 42
	Maximum Depth	d	Categorical({5, 10, 12, 15})	
	Minimum Split Samples	n_{split}	$\mathcal{U}_{\mathbb{Z}}(10, 100, 2)$	
	Minimum Leaf Samples	n_{leaf}	$\mathcal{U}_{\mathbb{Z}}(10, 60, 1)$	
	Feature Pool Ratio	c	Categorical({0.5, 0.7})	
HistGradBoost (45)	Maximum Iterations	M	$\mathcal{U}_{\mathbb{Z}}(100, 400, 10)$	Early Stopping: False
	Maximum Depth	d	Categorical({None, 4, 6})	Random Seed: 42
	Learning Rate	η	$\log\text{-}\mathcal{U}(0.005, 0.15)$	
	Minimum Leaf Samples	n_{leaf}	$\mathcal{U}_{\mathbb{Z}}(20, 80, 5)$	
	Maximum Leaf Nodes	L	$\mathcal{U}_{\mathbb{Z}}(15, 50, 1)$	
	L2 Regularization	λ	$\log\text{-}\mathcal{U}(0.1, 100.0)$	
AdaBoost (35)	Number of Trees	M	$\mathcal{U}_{\mathbb{Z}}(25, 400, 5)$	Base Estimator: Decision Tree
	Learning Rate	η	$\log\text{-}\mathcal{U}(0.005, 2.0)$	Random Seed: 42
	Loss Metric Choice	–	Categorical({square, exponential})	
	Base Tree Max Depth	d_{base}	$\mathcal{U}_{\mathbb{Z}}(1, 3, 1)$	
Ridge (25)	L2 Regularization Strength	α	$\log\text{-}\mathcal{U}(10^{-5}, 10^{3})$	None
ElasticNet (40)	Penalty Intensity	α	$\log\text{-}\mathcal{U}(10^{-4}, 10^{2})$	Max Iterations: 50,000
	L1 mixing Ratio	ρ	$\mathcal{U}(0.0, 1.0)$	Random Seed: 42
SVR (60)	Kernel Type	–	Categorical({rbf})	Max Iterations: 10,000,000
	Regularization Parameter	C	$\log\text{-}\mathcal{U}(0.01, 50.0)$	
	Loss Insensitivity Zone	ϵ	$\log\text{-}\mathcal{U}(0.005, 1.0)$	
	RBF Kernel Coefficient	γ	$\log\text{-}\mathcal{U}(10^{-4}, 0.25)$	
MLP (40)	Hidden Layer Architecture	–	Categorical($\mathcal{S}_{\text{layer}}$)[a]	Activation: Hyperbolic Tangent
	L2 Regularization	α	$\log\text{-}\mathcal{U}(10^{-3}, 10^{-1})$	Early Stopping: True
	Initial Learning Rate	η_0	$\log\text{-}\mathcal{U}(10^{-4}, 10^{-1})$	Tolerance Patience: 20 epochs
	Mini-Batch Size	B	Categorical({128, 256})	Max Epochs: 2,000
				Random Seed: 42
Kernel Ridge (35)	Regularization Penalty	α	$\log\text{-}\mathcal{U}(1.0, 10^{3})$	None
	Kernel Function	–	Categorical({rbf, laplacian})	
	Kernel Bandwidth	γ	$\log\text{-}\mathcal{U}(0.001, 0.1)$	
Dummy Mean (1)	None	–	–	Strategy: Sample Mean

[a] Note: $\mathcal{S}_{\text{layer}} = \{(32,), (64,), (128,), (8, 16), (16, 8), (32, 16), (64, 32)\}$.