IMPERIAL

Imperial College London

Department of Mathematics

Proximal Policy Optimisation for Stock Carbon-Aware Portfolio Construction

Author: Jérôme ALLIER (CID: 06012483)

A thesis submitted for the degree of

MSc in Mathematics and Finance, 2024-2025

Declaration

I declare that this thesis is the result of my own work, unless otherwise indicated. The thesis was written during my research predoctoral contract at the Institut Louis Bachelier. Part of the research conducted under the supervision of Professor Peter Tankov and Professor Olivier-David Zerbib at ENSAE Paris, specifically the greenhouse gas forecasting presented in Chapter 2, is included in this work. The data used in this thesis were kindly provided by the Institut Louis Bachelier and Kepler Chevreux.

Acknowledgements

The authors express their sincere gratitude to Professor Peter Tankov and Professor Olivier-David Zerbib, who supervised my predoctoral research at the Institut Louis Bachelier and provided invaluable guidance throughout this work. I am especially thankful for their confidence in me and for the time they dedicated, which made it possible for me to write this thesis.

I am also grateful to Thibault and Mohamed from the PLADIFES Group and the Data-Lab at the Institut Louis Bachelier for granting me access to data. I am deeply grateful to my colleague Lita for her kind support, well-timed jokes, and research cartography in sustainable finance. I would like to thank Ines from Kepler Cheuvreux for her help with data and for her valuable suggestions.

I wish to thank Professor Johannes Muhle-Karbe, my master's tutor and thesis supervisor on the university side, for his guidance during this year.

On a personal note, I am deeply thankful to Mathilde for her constant support, encouragement throughout this work.

Abstract

This thesis investigates how carbon-awareness can be integrated into portfolio construction. It combines predictive modelling of corporate greenhouse-gas emissions with reinforcement learning techniques for asset allocation. The motivation comes from the pressing role of the financial sector in supporting the low-carbon transition. The work develops a unified framework that connects environmental forecasting with advanced portfolio optimisation. The first contribution lies in forecasting firm-level carbon intensities across the three Scopes using three large-scale proprietary databases (COGEM, MSCI, Refinitiv). Results show that advanced machine learning models consistently outperforms simpler models. The second contribution integrates these forecasts into a portfolio optimisation setting. We formalise the joint minimisation of financial risk and carbon footprint as a multi-objective problem. Analytical results show that modest reductions in portfolio carbon exposure can be achieved at negligible variance cost. We further implement a reinforcement learning approach using Proximal Policy Optimisation. The algorithm successfully approximates the quadratic programming frontier while providing greater flexibility to incorporate additional ecological and operational objectives. Empirical tests on U.S. equity portfolios demonstrate three main findings: (i) investors can reduce carbon exposure substantially without deteriorating risk-adjusted performance; (ii) replacing naive historical predictors with machine-learning forecasts improves the risk–carbon frontier; and (iii) the reinforcement learning framework delivers robust and adaptive solutions, outperforming quadratic programming in out-of-sample evaluation.

Overall, this thesis contributes to the literature on green finance by showing how predictive modelling and reinforcement learning can be combined to design transparent, flexible, and scalable tools for carbon-aware investment strategies.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

Global climate change poses one of the most pressing challenges of the twenty-first century, and the financial sector is uniquely positioned to influence corporate behavior through capital allocation. In particular, the quantification and forecasting of corporate greenhouse-gas (GHG) emissions have become cornerstones of "green finance" research. Despite rapid advances in both data availability and machine learning methodologies, practitioners and scholars still lack a unified, interpretable, and scalable framework for predicting Scope 1, 2, and 3 emissions, integrating carbon intensity into asset selection, and embedding environmental constraints into portfolio theory. This thesis aims to provide a quantitative approach to integrate the GHG emissions into portfolio optimisation.

First, in Chapter 2 we review related literature in carbon accounting, predictive modelling, and environmental portfolio theory. We then develop a mathematical finance toolkit for low carbon portfolio construction. Building on classical mean–variance theory, we introduce carbon-aware constraints and reformulate the problem as a multi-objective optimisation. We derive conditions under which a small sacrifice in financial metrics, typically a negligible increase in risk, can yield large reductions in the portfolio carbon footprint, and we propose a multi-objective reinforcement-learning approach to navigate the risk–carbon trade-off dynamically. Compared with Quadratic Programming (QP), the principal advantage of the Reinforcement Learning (RL) framework is flexibility: it readily accommodates additional environmental objectives and operational constraints (e.g., scope or sector specific targets, financed-emissions caps, transition-path or budget constraints, turnover and transaction costs, and non-convex or path-dependent rules) for which closed-form or convex QP formulations become impractical.

Second, Chapter 3 develops and benchmarks predictive models for corporate GHG intensities. We compile three proprietary datasets (COGEM, MSCI, Refinitiv) covering Scope 1–3 emissions for a panel of publicly traded and small firms, and engineer a suite of temporal and categorial features reflecting both raw and smoothed historical emissions. We benchmark with simple baselines used in practice: company-wise historical mean, company-wise exponentially weighted moving average, company-wise temporal OLS. We evaluate regularised panel linear regression (Ridge) and pandel gradient-boosted decision trees (CatBoost) under rolling-window scheme with a one-year forecast horizon on log-transformed intensity targets. Both models are trained with the same features, exploiting temporal and cross-sectional informations. Out-of-sample, CatBoost consistently outperforms Ridge and all baselines across providers. SHAP value analyses indicate that smoothed lagged emissions dominate predictive power. We also compare alternative denominators for intensity: assets, revenue, COGS, market capitalisation. We document context-dependent performance: assets and revenue behave similarly overall, while market capitalisation and COGS are noisier in COGEM and MSCI/Refinitiv respectively.

Scope-level difficulty persists across datasets: Scope 2 < Scope 1 < Scope 3 in forecast error, supporting the robustness and generalisability of the framework.

Third, Chapter 4 operationalises the framework on a balanced universe of 50 U.S. firms with complete carbon and financial histories across the three providers. Portfolios are long-only and fully invested, rebalanced weekly. Risk is estimated with an exponentially smoothed (EWMA) covariance. The carbon input is the firm-level intensity vector (Scopes 1–3), held piecewise-constant within each year. We evaluate three carbon information sets for carbon: ex-post realised, naïve firm-wise mean, CatBoost forecast with two optimisation frameworks a weekly scalarised QP and a weekly scalarised actor–critic policy trained with PPO. Empirically, (i) a minimum-variance investor can reduce portfolio emissions at negligible variance cost; (ii) replacing naïve predictors with CatBoost forecasts shifts the risk–carbon frontier inward; and (iii) the PPO policy not only closely tracks but also outperforms the penalised QP frontier while offering greater flexibility for additional constraints. Results are robust across emissions providers, despite differences in corporate GHG emissions estimates.

The remainder of the thesis is organised as follows. Chapter 5 concludes with avenues for future research. Key technical details and proofs are deferred to Appendix A.

Chapter 2

Portfolio Optimisation Theory

2.1 Literature Review

GHG Metrics Forecasts

A growing body of research has developed predictive models for GHG emissions and intensities. While our approach shares commonalities with these studies, it also introduces several key innovations. We briefly review relevant contributions below.

Machine learning for firm-level emissions projections: Nguyen et al. [1], Kosgahakumbura et al. [21], and Nguemkam et al. [26] employ meta-learners that blend multiple algorithms to forecast emissions, achieving high predictive accuracy at the expense of interpretability, reproducibility, and operational simplicity. More recently, Nguyen et al. [27] focus on Scope 3 emissions using linear models, XGBoost, and random forests, confirming both the difficulty of Scope 3 emissions forecasting and the effectiveness of tree-based approaches.

Regularised linear approaches: Griffin et al. [16] and Goldhammer et al. [15] compare traditional regression frameworks for projecting emission intensities, while Xia et al. [40] apply Lasso regression to identify the main drivers of firm-level emissions. These approaches are highly interpretable but often lack the flexibility to capture non-linear dependencies in large cross-sectional panels.

Alternative scales and architectures: At higher aggregation levels, Dong et al. [13] use gradient-boosted decision trees to predict provincial emissions in China, while Xia et al. [41] apply a CNN–LSTM hybrid at the industrial-park level. At the city level, Ayasrah et al. employ a VAR–LSTM model [3].

Our contribution: We contribute to the literature on carbon forecasting by introducing several novel elements. First, we provide the first systematic comparison of alternative denominators for scaling firm-level GHG emissions in a predictive framework. Second, we are, to our knowledge, the first to forecast with a large-scale corporate emissions database such as COGEM, enabling analysis at unprecedented coverage and granularity. Third, our modelling strategy emphasises transparency and interpretability: we rely on a single, scope-specific model with a concise set of engineered temporal and firm-level features, and we directly benchmark a tree-based model against regularised linear regression using the exact same feature set. This design allows a clean assessment of model complexity versus predictive gains. Finally, we establish robustness through a cross-database comparison

across three leading providers and through a global study covering all three emission scopes.

Taken together, these contributions show that it is possible to achieve accurate, generalisable, and practically relevant forecasts with relatively simple models, while providing the first large-sample evidence on how methodological and definitional choices affect carbon-intensity prediction for portfolio optimisation.

Portfolio Optimisation

The use of carbon metrics in portfolio optimisation, as well as reinforcement learning techniques for portfolio construction or combining both approaches has been widely studied in the literature. Our work introduces several novel aspects. We summarise related contributions below.

Green portfolio construction: Barahhou et al. [5], Benslimane et al. [6], and Bolton et al. [8] propose aligning portfolios with net-zero targets by incorporating carbon metrics, while minimising tracking error relative to a benchmark. Garcia et al. [18] adopt a multi-objective framework that combines two risk metrics with a Bloomberg ESG score. Fitzgibbons et al. [28] derive an ESG-efficient frontier, showing the highest attainable Sharpe ratio for each ESG level.

Reinforcement learning portfolios: Yang et al. [42] propose an ensemble strategy that leverages deep reinforcement learning to optimise stock trading strategies by maximising returns, outperforming other RL-based approaches. Almahdi & Yang [2] employ recurrent reinforcement learning to minimise expected maximum drawdown.

Reinforcement learning with ESG objectives: He et al. [38] use an LSTM to predict stock returns and apply reinforcement learning to optimise a triple objective: maximising returns, minimising volatility, and maximising ESG scores. Maree & Omlin [24] show that multi-agent deep deterministic policy gradients fail to converge in a bi-objective risk–ESG portfolio optimisation setting, and propose an alternative algorithm. Choudhary et al. [11] extend this line of work by designing a multivariate reward function to jointly optimise portfolio risk and ESG performance.

Our contribution: This is, to our knowledge, the first study to integrate a real forecasting framework for firm-level carbon intensities directly into portfolio optimisation, and the first to do so within a reinforcement learning setting. We demonstrate for the first time how predictive accuracy in carbon intensities translates into measurable reductions in portfolio emissions. Methodologically, we benchmark classical quadratic programming against reinforcement learning, showing that PPO not only reproduces the efficient risk–carbon frontier but can in some regions strictly dominate the static approach. We further establish robustness by evaluating portfolio outcomes across three distinct emissions providers: while absolute levels vary, relative results are consistent. By focusing on a single, transparent, and auditable environmental metric—GHG emissions—we avoid the opacity and divergence of composite ESG scores, highlighting that credible decarbonisation can be achieved with reliable data and advanced optimisation techniques. As Berg et al. [7] highlighted, ESG scores diverge substantially depending on the provider.

2.2 Notations

Let $\Omega := \{1, \ldots, n\}$ denote a universe of n assets, and let $r_t \in \mathbb{R}^n$, $t = 1, 2, \ldots, T$, be their return vector at time t, where $r_{t,i}$ is the return of asset i from time $t-1$ to t. Returns are modelled as random variables observable through their realisations.

Definition 2.2.1 (Asset return, Covariance and Volatility)**.** The return of an asset i at the time t is defined as

$$r_{t,i} := \frac{P_{t,i} - P_{t-1,i}}{P_{t-1,i}}, \quad \text{with } P_{t,i} \text{ the price of } i \text{ at time } t.$$

The covariance matrix $\Sigma_t \in \mathbb{S}_+^n$, where $\mathbb{S}_+^n$ denotes the cone of positive semidefinite matrices, is defined as:

$$\Sigma_t := \text{cov}[r_t] = \mathbb{E}[r_t r_t'] - \mathbb{E}[r_t]\mathbb{E}[r_t]'.$$

The volatility vector is given by $\sigma_t \in \mathbb{R}^n$ with:

$$\sigma_{t,i} := \sqrt{\mathbb{E}[(r_{t,i} - \mathbb{E}[r_{t,i}])^2]}.$$

Definition 2.2.2 (GHG Emissions)**.** The GHG emissions of an asset i at time t is defined as $\xi_{t,i}$ the sum of the three scopes emissions (see definition 3.1.1) of the company between time $t-1$ and t. As we only have annual emissions data, we consider a constant emission rate to calculate weekly emissions.

Definition 2.2.3 (GHG Intensities)**.** The GHG intensity of an asset i at time t is defined as

$$I_{t,i} := \frac{\xi_{t,i}}{R_{t,i}}$$

the GHG emissions during the corresponding year divided by the revenue of the company during the same year (see definition 3.1.2).

Definition 2.2.4 (Portfolio)**.** Let $w \in \mathbb{R}^n$ be the portfolio weight vector, with w_i representing the proportion of wealth invested in asset i. We impose the standard budget constraint

$$\mathbf{1}'w = 1, \quad w \geq 0,$$

where short-selling is excluded for simplicity. The expected return of the portfolio at time t is

$$\mu_t(w) := \mathbb{E}[r_t]'w,$$

where $\mathbb{E}[r_t] \in \mathbb{R}^n$ is the vector of expected asset returns. The portfolio variance is

$$\sigma_t^2(w) := w'\Sigma_t w,$$

where Σ_t is the covariance matrix of returns at time t.

2.3 Theorical Framework

2.3.1 Portfolio Optimisation Theory

The classical mean–variance problem stated by Markowitz [25] is formulated as:

$$\min_{w \in \mathbb{R}^n} \; w'\Sigma_t w \quad \text{s.t. } \mathbf{1}'w = 1, \;\; \mu_t(w) \geq \mu^*,$$

where μ^* is a target expected return. Alternatively, one can solve the unconstrained mean–variance trade-off:

$$\max_{w \in \mathbb{R}^n} \ \mu_t(w) - \frac{\lambda}{2} w' \Sigma_t w,$$

where $\lambda > 0$ is the investor's risk-aversion parameter. This quadratic program yields the efficient frontier of portfolios.

Despite its foundational role, the classical mean–variance framework exhibits important limitations in practice. First, the formulation relies on estimates of expected returns $\mathbb{E}[r_t]$, which are notoriously difficult to predict with sufficient accuracy. Chopra & Ziemba [10] show that return forecasts are often dominated by estimation error, leading to unstable and unreliable optimal weights. For this reason, much of the empirical portfolio optimisation literature, and this thesis in particular, focuses on variance minimisation as a more robust objective.
A second critical challenge lies in the estimation of the covariance matrix Σ_t. The sample covariance estimator

$$\hat{\Sigma}_t := \frac{1}{T-1} \sum_{s=1}^{T} (r_s - \bar{r})(r_s - \bar{r})',$$

is known to be highly unstable when the number of assets n is of the same order of magnitude as the sample size T. In such cases, $\hat{\Sigma}_t$ becomes ill-conditioned, and the resulting portfolio weights are extremely sensitive to estimation noise. Introducing more data to estimation would mean, in case of large asset universe, introducing outdated data with the same weight as recent data. Ledoit and Wolf [23] proposed a shrinkage estimator, combining the sample covariance with a structured target (e.g. the identity matrix), which significantly improves conditioning and out-of-sample portfolio performance.
Furthermore, random matrix theory (RMT) has shown that most eigenvalues of $\hat{\Sigma}_t$ lie within the Marčenko–Pastur distribution and thus correspond largely to noise rather than information (Bouchaud et al.[22, 9]). Filtering the covariance matrix by retaining only the informative part of the spectrum can lead to more stable and robust portfolio allocations.
Finally, alternative estimators such as the exponentially weighted moving average (EWMA) covariance,

$$\hat{\Sigma}_t^{\mathrm{EWMA}} := (1-\alpha) \sum_{s=1}^{t} \alpha^{t-s} (r_s - \bar{r})(r_s - \bar{r})', \quad 0 < \alpha < 1,$$

are widely used in both academia and industry. The EWMA estimator is simple to implement, adapts quickly to changing market conditions. It enables using more historical data while putting more weight on recent data. It has been shown to perform well empirically in volatility and risk forecasting by RiskMetrics [19], using generally $\alpha = 0.94$ for an half-life of approximately 11 days. For these reasons, it is often employed as a practical and robust alternative to the sample covariance, and we will use it in the empirical part of this thesis.

2.3.2 Extending the Problem with Ecological Objectives

The quadratic program admits a natural extension: because the efficient frontier is typically flat around the optimum, one can often introduce additional constraints or secondary objectives at negligible cost in terms of variance as shown by Scherer [33] (Chap 2, Sec 2.4). In particular, we show that environmental criteria, such as minimising portfolio-level carbon intensity, can be embedded alongside the traditional risk objective without significantly sacrificing financial performance. This motivates our integration of carbon-aware constraints into the optimisation problem.

To incorporate environmental considerations, we introduce portfolio-level carbon intensity:

$$I_t(w) := I_t'\omega = \sum_{i=1}^{n} w_i I_{t,i},$$

where $I_{t,i}$ denotes the carbon intensity of asset i at time t. This allows us to penalise carbon exposure within the optimisation. Two formulations are possible:

1. **Constrained optimisation:**

$$\min_{w\in\mathbb{R}^n} \; w'\Sigma_t w \quad \text{s.t. } \mathbf{1}'w = 1, \;\; I_t(w) \le I^*,$$

   where I^* is a maximum admissible portfolio carbon intensity.

2. **Scalarised ulti-objective optimisation:** we embed carbon directly into the objective via a trade-off parameter $\gamma \ge 0$:

$$\min_{w\in\mathbb{R}^n} \; w'\Sigma_t w + \gamma I_t(w).$$

The two formulations above are closely related.

Proposition 2.3.1. *Let $\Sigma_t \succ 0$ and suppose $I_t(w) = I'w$ is linear in w. Assume the constrained problem is feasible. Then, for every admissible intensity threshold I^* in*

$$\min_{w\in\mathbb{R}^n} \; w'\Sigma_t w \quad \textit{s.t. } \mathbf{1}'w = 1, \;\; I'w \le I^*, \tag{C}$$

there exists a penalty parameter $\gamma^\star \ge 0$ such that the solution coincides with that of the penalised problem

$$\min_{w\in\mathbb{R}^n} \; w'\Sigma_t w + \gamma^\star I'w \quad \textit{s.t. } \mathbf{1}'w = 1. \tag{P_γ}$$

Conversely, for any $\gamma \ge 0$, the solution w_γ of (P_γ) solves (C) with threshold $I^*_\gamma := I'w_\gamma$.

Proof. Since $\Sigma_t \succ 0$, both problems admit unique minimisers on the affine set $\{w : \mathbf{1}'w = 1\}$.

$(C \Rightarrow P_{\gamma^\star})$. The Lagrangian of (C) is

$$\mathcal{L}(w,\lambda,\gamma) = w'\Sigma_t w + \lambda(\mathbf{1}'w - 1) + \gamma(I'w - I^*), \qquad \lambda \in \mathbb{R}, \;\; \gamma \ge 0.$$

By Slater's condition, the KKT conditions are necessary and sufficient. Thus, at the optimum $(w^\star, \lambda^\star, \gamma^\star)$,

$$\begin{aligned}
&\text{(stationarity)} && 2\Sigma_t w^\star + \lambda^\star \mathbf{1} + \gamma^\star I = 0,\\
&\text{(primal feasibility)} && \mathbf{1}'w^\star = 1, \quad I'w^\star \le I^*,\\
&\text{(dual feasibility)} && \gamma^\star \ge 0,\\
&\text{(complementary slackness)} && \gamma^\star(I'w^\star - I^*) = 0.
\end{aligned}$$

The term $-\gamma^\star I^*$ is constant in w, hence minimisation of $\mathcal{L}$ over w is equivalent to minimisation of

$$w'\Sigma_t w + \gamma^\star I'w \quad \text{s.t. } \mathbf{1}'w = 1.$$

Therefore $w^\star$ also solves $(P_{\gamma^\star})$. If the inequality is inactive, then $\gamma^\star = 0$.

$(P_\gamma \Rightarrow C)$. Let w_γ solve (P_γ) and set $I^*_\gamma := I'w_\gamma$. For any feasible w of (C) with threshold I^*_γ we have

$$w'\Sigma_t w + \gamma I'w \;\ge\; w_\gamma'\Sigma_t w_\gamma + \gamma I'w_\gamma.$$

Since $I'w \le I^*_\gamma = I'w_\gamma$, it follows that

$$w'\Sigma_t w \;\ge\; w'_\gamma \Sigma_t w_\gamma + \gamma\big(I'w_\gamma - I'w\big) \;\ge\; w'_\gamma \Sigma_t w_\gamma.$$

Thus w_γ is the minimiser of (C) with threshold I^*_γ. This establishes the equivalence between the constrained and penalised formulations. □

This equivalence is the exact analogue of the classical duality between the mean–variance formulation with a target return

$$\min_w \; w'\Sigma_t w \quad \text{s.t. } \mu_t(w) \ge \mu^*$$

and the unconstrained risk–return trade-off

$$\max_w \; \mu_t(w) - \tfrac{\rho}{2} w'\Sigma_t w.$$

In both cases, the constrained and the penalised problems generate the same efficient frontier, with a one-to-one correspondence between constraint levels (I^* or μ^*) and penalty parameters (γ or ρ).
From a practical standpoint, the constrained approach is often more interpretable for investors, who may impose a maximum admissible carbon footprint, whereas the penalised approach is computationally simpler, as it avoids introducing inequality constraints and reduces to a single quadratic minimisation problem. Both perspectives are useful and complementary: one emphasises explicit environmental targets, while the other highlights the trade-off between financial risk and ecological impact.
The scalarisation approach allows one to recover a Pareto-efficient frontier in the risk–carbon space by varying λ. It highlights the trade-off between financial and ecological objectives.

A formal "small-cost" result. We focus on variance minimisation as a robust baseline, and we ask: what is the theoretical variance cost of imposing a carbon constraint?

Consider the minimum-variance problem with a carbon target I^*:

$$\min_w \; \tfrac{1}{2} w'\Sigma w \quad \text{s.t.} \quad \mathbf{1}'w = 1, \;\; I'w = I^*,$$

Define the scalars

$$A := \mathbf{1}'\Sigma^{-1}\mathbf{1}, \quad B := \mathbf{1}'\Sigma^{-1}I \quad C := I'\Sigma^{-1}I, \quad \Delta := AC - B^2 \;(>0).$$

Let $I_0 := B/A$ denote the carbon intensity of the unconstrained global minimum-variance (GMV) portfolio, $w_{gmv} = \Sigma^{-1}\mathbf{1}/A$.

Proposition 2.3.2 (Closed-form variance–carbon trade-off)**.** *Let $\Sigma \succ 0$. Consider*

$$\min_w \; w'\Sigma w \quad \textit{s.t.} \quad \mathbf{1}'w = 1, \;\; I'w = I^*.$$

Then the optimiser has the form

$$w^*(I^*) = \Sigma^{-1}\big(\lambda\,\mathbf{1} + \theta\, I\big),$$

where (λ, θ) solves

$$\begin{bmatrix} A & B \\ B & C \end{bmatrix} \begin{bmatrix} \lambda \\ \theta \end{bmatrix} = \begin{bmatrix} 1 \\ I^* \end{bmatrix}.$$

Moreover, the minimal variance $\sigma^2(I^*) = w^{*\top}\Sigma w^*$ is

$$\sigma^2(I^*) \;=\; \frac{1}{A} \;+\; \frac{A}{\Delta}\,\big(I^* - I_0\big)^2.$$

Proof sketch. Form the Lagrangian $\mathcal{L}(w, \lambda, \theta) = w^\top \Sigma w - \lambda(\mathbf{1}^\top w - 1) - \theta(I^\top w - I^*)$. Stationarity gives $2\Sigma w - \lambda \mathbf{1} - \theta I = 0$, hence $w = \frac{\lambda}{2}\Sigma^{-1}\mathbf{1} + \frac{\theta}{2}\Sigma^{-1} I$; renaming multipliers absorbs the factor $1/2$ and yields $w = \Sigma^{-1}(\lambda \mathbf{1} + \theta I)$. Imposing the two linear constraints produces the 2×2 system shown, which has a unique solution since $\Delta > 0$. Substituting w^* back gives $\sigma^2(I^*) = w^{*\top} \Sigma w^* = \lambda + \theta I^* = \frac{1}{A} + \frac{A}{\Delta}(I^* - I_0)^2$. (See Appendix A.1 for a more detailed proof) □

.

Implications. (i) If the inequality constraint $I'w \leq I^*$ is used, it is slack whenever $I^* \geq I_0$, and the optimum remains w_{gmv} with $\sigma^2 = 1/A$ (zero cost).
(ii) When tightening around I_0, the variance penalty is *second order*:

$$\sigma^2(I_0 + \delta) \;=\; \frac{1}{A} \;+\; \frac{A}{\Delta}\,\delta^2 \;=\; \sigma^2_{gmv} \;+\; \mathcal{O}(\delta^2),$$

so small carbon improvements (small $|\delta|$) incur negligible variance cost.
(iii) The marginal variance cost is

$$\frac{\mathrm{d}\sigma^2}{\mathrm{d}k} \;=\; \frac{2A}{\Delta}\,(I^* - I_0),$$

which is zero at $I^* = I_0$ and grows linearly away from it; the curvature $2A/\Delta > 0$ quantifies how flat the efficient set is around the GMV point in the carbon direction.

Equivalent penalised (scalarised) form. Consider instead the convex penalised problem

$$\min_{w} \;\tfrac{1}{2}\, w'\Sigma w \;+\; \gamma\, I'w \quad \text{s.t.} \quad \mathbf{1}'w = 1, \qquad \gamma \geq 0.$$

The solution is $w(\gamma) = \Sigma^{-1}\big(\lambda(\gamma)\,\mathbf{1} - \gamma\, I\big)$ with $\lambda(\gamma) = (1 + \gamma B)/A$. It implies

$$I'w(\gamma) \;=\; I_0 \;-\; \gamma\Big(C - \frac{B^2}{A}\Big) \;=\; I_0 \;-\; \gamma\,\frac{\Delta}{A},$$

and the corresponding variance satisfies

$$\sigma^2(\gamma) \;=\; \frac{1}{A} \;+\; \gamma^2\,\frac{\Delta}{A}.$$

Thus, for small γ, variance increases only quadratically in the penalty strength, again formalising the "small-cost" claim. Moreover, for any target $k \leq I_0$, choosing $\gamma = \frac{A}{\Delta}(I_0 - k)$ reproduces the constrained solution (Lagrangian duality).

2.3.3 Reinforcement Learning Approach

RL provides a general framework for sequential decision-making under uncertainty, and is particularly well-suited to portfolio optimisation problems where investors rebalance dynamically in response to evolving market conditions. Formally, RL models the environment as a Markov Decision Process (MDP), and seeks to learn a policy that maps observable states to portfolio allocations in order to optimise a cumulative objective.

On the following section and corresponding Appendix, the contraction and fixed-point results Lemmas A.2.1–A.2.2) follow standard arguments (see Sutton and Barto [35]; Puterman [31]). The policy gradient theorem (Theorem 2.3.10) was introduced in Sutton et al. [36]. The advantage form and actor–critic variants trace back to Konda Tsitsiklis [20].

Definition 2.3.3 (Markov decision process)**.** A (discounted, infinite-horizon) MDP is a tuple $(\mathcal{S}, \mathcal{A}, P, R, \beta)$ where:

- $\mathcal{S}$ is the state space; $s_t \in \mathcal{S}$.
- $\mathcal{A}$ is the action space (possibly state-dependent $\mathcal{A}(s)$).
- $P(\cdot \mid s, a)$ is a Markov transition kernel on $\mathcal{S}$.
- $R : \mathcal{S} \times \mathcal{A} \to \mathbb{R}$ is a (bounded) reward function.
- $\beta \in (0, 1)$ is the discount factor.

A (Markov) policy π maps each $s \in \mathcal{S}$ to a distribution on $\mathcal{A}$. Its value is

$$V^\pi(s) := \mathbb{E}\left[\sum_{t=0}^{\infty} \beta^t R(s_t, a_t) \,\middle|\, s_0 = s,\ a_t \sim \pi(\cdot \mid s_t),\ s_{t+1} \sim P(\cdot \mid s_t, a_t)\right].$$

Portfolio-control instantiation.

- **State s_t:** information available to the investor, including predicted risk and sustainability signals. For example, s_t contains a covariance forecast $\Sigma_t = \Sigma(s_t) \succeq 0$ (e.g. EWMA) and a vector of asset-level carbon intensities $I_t = I(s_t) \in \mathbb{R}^n_+$.
- **Action a_t:** the portfolio weights $w_t \in \Delta^{n-1}$, where

$$\Delta^{n-1} \ := \ \{w \in \mathbb{R}^n_+ : \mathbf{1}'w = 1\}.$$

- **Transition kernel $P(\cdot \mid s_t, a_t)$:** describes the stochastic evolution of market and signals. Under price-taking and no impact, one typically has $P(\cdot \mid s_t, a_t) = P(\cdot \mid s_t)$, with Σ_{t+1} and I_{t+1} updated using new returns and carbon informations.
- **Reward R:** encodes the trade-off between financial risk and carbon considerations. Formally, it is a measurable bounded function of the state-dependent risk estimate Σ_t and the portfolio's carbon intensity $I_t(w_t)$,

$$R(s_t, w_t) \ := \ f\big(w_t'\Sigma_t w_t,\ I_t(w_t)\big),$$

  for some chosen functional form $f : \mathbb{S}^n_+ \times \mathbb{R}_+ \to \mathbb{R}$.
- **Discount factor $\beta \in (0, 1)$:** weights future rewards.

Definition 2.3.4 (Policies and value functions)**.** A policy specifies the investor's portfolio allocation rule.

- A *deterministic policy* is a mapping

$$\pi : \mathcal{S} \to \mathcal{A}, \qquad w_t = \pi(s_t),$$

  assigning a single portfolio weight vector to each state.
- More generally, a *stochastic policy* is represented by a conditional distribution

$$\pi(w|s) \in \Delta(\mathcal{A}),$$

  where $\Delta(\mathcal{A})$ denotes the set of probability distributions on $\mathcal{A}$, giving the probability of choosing portfolio w in state s.
- In applications, stochastic policies are often chosen from a parametric family $\{\pi_\theta : \theta \in \mathbb{R}^d\}$, where $\pi_\theta(w|s)$ is differentiable in its parameters θ (e.g. the weights of a neural network).

The performance of a policy π is measured by its value function $V^\pi : \mathcal{S} \to \mathbb{R}$:

$$V^\pi(s) := \mathbb{E}_\pi\left[\sum_{t=0}^{\infty} \beta^t R(s_t, w_t) \;\middle|\; s_0 = s\right],$$

assigning to each state s the expected discounted cumulative reward.

The objective of reinforcement learning is to find an optimal policy

$$\pi^* = \arg\max_\pi V^\pi(s), \quad \forall s \in \mathcal{S},$$

corresponding to the best dynamic portfolio allocation strategy.

Definition 2.3.5 (Action-value function)**.** For a policy π, the action-value function $Q^\pi : \mathcal{S} \times \mathcal{A} \to \mathbb{R}$ is defined as

$$Q^\pi(s,a) \;:=\; \mathbb{E}_\pi\Big[\sum_{t=0}^{\infty} \beta^t R(s_t, w_t) \,\Big|\, s_0 = s,\, w_0 = a\Big],$$

Definition 2.3.6 (Bellman operators)**.** For any policy π, define the policy Bellman operator T^π acting on $Q : \mathcal{S} \times \mathcal{A} \to \mathbb{R}$ by

$$(T^\pi Q)(s,a) \;=\; R(s,a) \;+\; \beta\, \mathbb{E}_{s'|s,a}\Big[\mathbb{E}_{a' \sim \pi(\cdot|s')} Q(s', a')\Big].$$

Define the optimality operator T by

$$(T^* Q)(s,a) \;=\; R(s,a) \;+\; \beta\, \mathbb{E}_{s'|s,a}\Big[\max_{a' \in \mathcal{A}} Q(s', a')\Big].$$

For continuous action spaces, replace max by sup and the supremum is attained under some hypothesis (e.g., compact $\mathcal{A}$ and continuity).

Theorem 2.3.7 (Bellman optimality for Q^*)**.** *Define $Q^*(s,a) := \sup_\pi Q^\pi(s,a)$. Then Q^* is the unique fixed point of T^*, i.e.,*

$$Q^*(s,a) \;=\; R(s,a) \;+\; \beta\, \mathbb{E}_{s'|s,a}\Big[\max_{a'} Q^*(s', a')\Big],$$

and the greedy policy $\pi^*(s) \in \arg\max_a Q^*(s,a)$ is optimal.

Proof. See Appendix A.2. □

Corollary 2.3.8 (Optimal value function)**.** *The optimal state value satisfies $V^*(s) = \max_a Q^*(s,a)$, and V^* is the unique fixed point of the operator*

$$(TV)(s) \;=\; \max_a \Big\{R(s,a) + \beta\, \mathbb{E}_{s'|s,a}[V(s')]\Big\}.$$

Proof. Define $\widetilde{V}(s) := \max_a Q^*(s,a)$. Using the Bellman optimality equation for Q^*,

$$Q^*(s,a) \;=\; R(s,a) + \beta\, \mathbb{E}_{s'|s,a}\Big[\max_{a'} Q^*(s', a')\Big] \;=\; R(s,a) + \beta\, \mathbb{E}_{s'|s,a}\big[\widetilde{V}(s')\big].$$

Taking $\max_a$ on both sides yields

$$\widetilde{V}(s) \;=\; \max_a \Big\{R(s,a) + \beta\, \mathbb{E}_{s'|s,a}[\widetilde{V}(s')]\Big\} \;=\; (T\widetilde{V})(s),$$

so $\widetilde{V}$ is a fixed point of T.

Next, T is a β-contraction on the Banach space of bounded functions $V : \mathcal{S} \to \mathbb{R}$ under $\|\cdot\|_\infty$: for any bounded V_1, V_2,

$$|(TV_1 - TV_2)(s)| = \Big| \max_a \{R(s,a) + \beta \mathbb{E}_{s'|s,a}[V_1(s')]\} - \max_a \{R(s,a) + \beta \mathbb{E}_{s'|s,a}[V_2(s')]\} \Big|$$
$$\leq \beta \sup_a \mathbb{E}_{s'|s,a}\big[\, |V_1(s') - V_2(s')| \,\big] \;\leq\; \beta \|V_1 - V_2\|_\infty.$$

Hence, by Banach's fixed-point theorem, T has a unique fixed point. Since $\widetilde{V}$ is a fixed point, it must be the unique one; denote it by V^*. Therefore $V^*(s) = \max_a Q^*(s,a)$ and $V^* = TV^*$.

Finally, standard dominance arguments show $V^*(s) = \sup_\pi V^\pi(s)$ (because Q^* dominates all Q^π and the greedy policy w.r.t. Q^* attains Q^*), so $\widetilde{V}$ is indeed the optimal state value function. □

Definition 2.3.9 (Performance objective)**.** We define the infinite-horizon, discounted performance objective $J : \{\pi_\theta : \theta \in \mathbb{R}^d\} \to \mathbb{R}$ as

$$J(\pi_\theta) \;=\; \mathbb{E}_\pi\left[\sum_{t=0}^{\infty} \beta^t \, R(s_t, a_t)\right],$$

where the expectation is taken over trajectories generated by the policy π_θ and transition kernel P, starting from some initial distribution ρ over s_0

Theorem 2.3.10 (Policy Gradient Theorem)**.** *Suppose that*

- *the discount factor satisfies $\beta \in (0,1)$, the reward function is bounded,*
- *the policy class $\{\pi_\theta : \theta \in \mathbb{R}^d\}$ consists of differentiable probability distributions $\pi_\theta(a|s)$ with well-defined log-gradients $\nabla_\theta \log \pi_\theta(a|s)$,*
- *and expectations and gradients can be interchanged (dominated convergence assumptions).*

Then the gradient of the performance objective is

$$\nabla_\theta J(\theta) \;=\; \frac{1}{1-\beta} \mathbb{E}_{\substack{s\sim d_\beta^{\pi_\theta} \\ a\sim\pi_\theta(\cdot|s)}} \Big[\nabla_\theta \log \pi_\theta(a|s) \, Q^{\pi_\theta}(s,a) \Big].$$

Equivalently,

$$\nabla_\theta J(\theta) \;=\; \mathbb{E}_{\pi_\theta}\left[\sum_{t=0}^{\infty} \beta^t \, \nabla_\theta \log \pi_\theta(a_t|s_t) \, Q^{\pi_\theta}(s_t, a_t)\right].$$

Proof. See Appendix A.3. □

Solution approaches in RL. The challenge is then to approximate the optimal (or Pareto-optimal) policy. The main families of methods are:

1. **Value-based methods.** The central idea developed by is to evaluate policies indirectly through their *action–value function* (Q-function). For a given policy π, this is defined as

$$Q^\pi(s,a) = \mathbb{E}_\pi\left[\sum_{k=0}^{\infty} \beta^k R(s_{t+k}, a_{t+k}) \;\Big|\; s_t = s,\; a_t = a\right],$$

i.e. the expected discounted reward obtained by taking action a in state s and following π thereafter. The associated *value function* is the expectation of Q^π under the policy:

$$V^\pi(s) = \mathbb{E}_{a\sim\pi(\cdot|s)}[Q^\pi(s,a)]\,.$$

The optimal Q-function Q^* satisfies the Bellman optimality equation (Theorem 2.3.7)

$$Q^*(s,a) = R(s,a) + \beta\,\mathbb{E}_{s'\sim P(\cdot|s,a)}\Big[\max_{a'} Q^*(s',a')\Big],$$

which expresses the recursive structure of optimal decisions. Once Q^* is known, the optimal policy is simply

$$\pi^*(s) = \arg\max_{a\in\mathcal{A}} Q^*(s,a).$$

Q-learning is a model-free algorithm that approximates Q^* directly from data. At each step $(s_t, a_t, R(s_t,a_t), s_{t+1})$, the estimate is updated according to

$$Q(s_t,a_t) \;\leftarrow\; Q(s_t,a_t) + \alpha\Big(R(s_t,a_t) + \beta\max_{a'} Q(s_{t+1},a') - Q(s_t,a_t)\Big),$$

with learning rate α. Over time, Q converges toward Q^*, and the greedy policy $\pi(s) = \arg\max_a Q(s,a)$ approaches optimality (see Watkins & Dayans [39]).

Limitation for portfolio optimisation. In portfolio problems, the action a corresponds to a weight vector in the n-dimensional simplex. This action space is continuous and high-dimensional. Computing $\max_{a'} Q(s,a')$ is intractable unless one discretises the weights, which quickly becomes impossible (the curse of dimensionality). Thus, pure value-based methods like tabular Q-learning are rarely used directly for portfolio optimisation. Instead, continuous-action variants (e.g. Deep Q-Networks with function approximation, or extensions like DDPG and SAC) are employed, which combine value estimation with policy parametrisation.

2. **Policy-based methods.** These methods dispense with explicit value-function estimation and instead directly parametrise the policy. The policy is written as a distribution $\pi_\theta(a|s)$ over actions, with parameters θ, and the goal is to maximise the expected discounted return

$$J(\theta) = \mathbb{E}_{\pi_\theta}\left[\sum_{t=0}^{\infty}\beta^t R(s_t,a_t)\right].$$

A key result is the policy gradient theorem 2.3.10, which provides an expression for the gradient of $J(\theta)$:

$$\nabla_\theta J(\theta) = \mathbb{E}_{\pi_\theta}[\nabla_\theta \log\pi_\theta(a_t|s_t)\,Q^{\pi_\theta}(s_t,a_t)]\,.$$

In practice, the action–value Q^{π_θ} is not known, so one substitutes the Monte Carlo return

$$G_t = \sum_{k=0}^{\infty}\beta^k R(s_{t+k},a_{t+k}),$$

which yields the update rule

$$\nabla_\theta J(\theta) \approx \mathbb{E}_{\pi_\theta}[\nabla_\theta \log\pi_\theta(a_t|s_t)\,G_t]\,.$$

Advantages: Naturally suited to continuous action spaces. By choosing π_θ appropriately (e.g. a softmax or Dirichlet distribution for weights on the simplex), portfolio

weights can be parametrised without discretisation. Directly optimises the objective $J(\theta)$, avoiding the need for maximisation over actions (as in Q-learning).

Drawbacks: Gradient estimates have high variance, since G_t is a noisy Monte Carlo return. Convergence can be slow without variance-reduction techniques, such as introducing a baseline $b(s_t)$ to replace G_t by $G_t - b(s_t)$, reducing variance while keeping the estimator unbiased.

3. **Actor–critic methods.** Actor–critic algorithms combine the strengths of value-based and policy-based approaches. The actor is the policy $\pi_\theta(a|s)$, parametrised by θ, which is updated in the direction suggested by the policy gradient theorem (Theorem 2.3.10). The critic is a function approximator (with parameters ϕ) that estimates the value function V^π or the action–value function Q^π, thereby reducing the variance of the gradient estimate.

   Policy gradient revisited. From Theorem 2.3.10, the gradient of the performance objective can be expressed as

$$\nabla_\theta J(\theta) = \mathbb{E}_{\pi_\theta}[\nabla_\theta \log \pi_\theta(a_t|s_t)\, Q^{\pi_\theta}(s_t, a_t)]\,.$$

   However, the true action–value Q^{π_θ} is unknown. A naive substitution with Monte Carlo returns G_t produces an unbiased estimator (REINFORCE), but with very high variance.

   Variance reduction via baselines. Lemma A.3.1 shows that we can subtract any baseline $b(s_t)$ without changing the expectation. Choosing $b(s_t) = V^{\pi_\theta}(s_t)$ yields the advantage function

$$A^{\pi_\theta}(s_t, a_t) = Q^{\pi_\theta}(s_t, a_t) - V^{\pi_\theta}(s_t),$$

   leading to the lower-variance update rule (Corollary A.3.2)

$$\nabla_\theta J(\theta) = \mathbb{E}_{\pi_\theta}[\nabla_\theta \log \pi_\theta(a_t|s_t)\, A^{\pi_\theta}(s_t, a_t)]\,.$$

   Role of the critic. Since both Q^{π_θ} and V^{π_θ} are unknown, the critic is introduced to estimate them. In practice, a parametrised function $V_\phi(s)$ or $Q_\phi(s, a)$ is learned by regression to the observed returns, typically minimising a squared-error loss:

$$\min_\phi \ \mathbb{E}\Big[\big(V_\phi(s_t) - G_t\big)^2\Big].$$

   The critic's estimate is then substituted into the advantage, yielding the actor update

$$\nabla_\theta J(\theta) \ \approx \ \mathbb{E}\big[\nabla_\theta \log \pi_\theta(a_t|s_t)\, \big(Q_\phi(s_t, a_t) - V_\phi(s_t)\big)\big]\,.$$

   Advantages and drawbacks. Actor–critic methods therefore strike a balance:

   - *Advantages:* lower variance than pure policy gradients, and more sample efficient. They can handle continuous action spaces naturally through the actor's parametric distribution.
   - *Drawbacks:* both actor and critic must be trained simultaneously, which can be unstable. Training the critic well is essential, as a poor critic leads to poor policy updates.

   Variants such as Advantage Actor–Critic (A2C), Asynchronous A3C, and Proximal Policy Optimisation (PPO) are widely used in practice.

Multi-objective reinforcement learning (MORL). RL can treat risk and carbon separately with a vector-valued reward

$$\mathbf{R}(s_t, a_t) = \Big(-\tfrac{1}{2}\, a_t' \Sigma_t a_t,\ -I_t(a_t) \Big) \in \mathbb{R}^2.$$

The corresponding value function is vector-valued,

$$\mathbf{V}^\pi(s) = \mathbb{E}\left[\sum_{t=0}^{\infty} \beta^t \mathbf{R}(s_t, a_t) \;\Big|\; s_0 = s,\ a_t = \pi(s_t) \right],$$

measuring the long-term trade-off between variance reduction and carbon reduction. The set of all attainable $\mathbf{V}^\pi(s)$ defines the *Pareto frontier* of efficient risk–carbon policies. Two standard approaches in the MORL literature are:

- **Scalarisation:** introduce a preference vector $\alpha = (\alpha_1, \alpha_2) \in \Delta_2$ and maximise the weighted sum
$$V_\alpha^\pi(s) = \alpha_1 V_{\text{var}}^\pi(s) + \alpha_2 V_{\text{carbon}}^\pi(s).$$
Varying α traces the efficient frontier. This recovers the penalised quadratic program, with $\gamma = \alpha_2/\alpha_1$.

- **Pareto-based methods:** directly approximate the set of non-dominated policies. The RL agent learns a family of policies spanning the efficient frontier, leaving the trade-off choice to the investor ex post.

Remark. The main motivation for RL is also extensibility: once we move beyond the static QP— to include transaction costs/turnover limits, more ESG constraints and objectives, an exact QP is no longer available, while actor–critic methods (e.g., PPO) remain applicable and may deliver a big advantage.

Chapter 3

GHG Emissions and Intensities Forecasts

3.1 Definitions

Regulators and capital allocators, in the context of the escalating climate emergency, must consider a firm's carbon footprint as a first-order metric of environmental impact.

Definition 3.1.1 (Corporate GHG Emissions)**.** GHG emissions are accounted, according to the standard corporate-footprint taxonomy, following the Greenhouse Gas Protocol ([32]):

- Scope 1: Direct on-site emissions resulting from sources that are owned or controlled by the company.
- Scope 2: Indirect emissions associated with the generation of purchased electricity, heat, steam, or cooling consumed by the company.
- Scope 3: All other upstream and downstream supply-chain emissions beyond Scopes 1 and 2, including financed emissions via investments.

GHG emissions are expressed in terms of carbon-dioxide equivalents (CO_2e) to account for the varying radiative forcing of CO_2, CH_4, N_2O and other greenhouse gases. It offers a transparent, comparable and quantitative measure.

Estimating Scope 3 emissions presents unique challenges, as it requires tracing the full upstream and downstream supply chain. Providers and purchasers often do not disclose their own emissions, leading to gaps in the estimations. By contrast, Scope 2 emissions are more straightforward to estimate, even when undeclared, because energy usage is controlled and energy carbon intensities are standardised geographically. Scope 1 emissions, which encompass a company's direct on-site outputs, are also simpler to quantify than Scope 3, yet they can exhibit significant year-to-year volatility if a firm alters its production processes, upgrades equipment, or inaugurates new facilities.

Definition 3.1.2 (Corporate GHG Intensities:)**.** The GHG intensity of a company is the ratio of its scope carbon emissions to an economic or physical output. In 1990, the World Business Council for Sustainable Development (WBCSD) represented by Verfaillie & Bidwell [37], was already pointing out the needs of ecological-economic ratio to measure and compare the impact of companies. Porter & Klaas [29] state that a higher carbon intensity suggests that a firm uses its resources inefficiently, emitting more GHG at the same production level and is then a poor performer.

In order to compare the carbon emissions efficiency across companies and to incorporate changes in a company's business activities over time, it is important to consider a ratio of the absolute different scopes carbon emissions to a business metric. The use of revenues as denominator is natural and consistent with prior environmental studies. However, Clarkson et al. [12] pointed out that when firms are not homogenous in terms of production processes, the ratio is not directly comparable across firms. Inspired from Bajic & Kiesel [4] and Verfaillie & Bidwell [37], we will compare different denominators to scale total carbon emission to provide a clear review of their meaning and performance in a projection framework:

- Total Revenues. Company's revenue-generating capacity, aligning emissions with economic output; however, it may be distorted by pricing or outsourcing.
- Total Assets. Overall resource base and enables assessment of emissions efficiency relative to asset utilisation; yet, it may not correspond directly to emissions-generating activities.
- COGS (Cost of Goods Sold). Direct production costs associated with the goods sold, linking emissions to core manufacturing or procurement activities; however, it excludes indirect upstream and downstream emissions.
- Market Capitalisation. Total equity market value of outstanding shares, aligning emissions with investor valuation and perceived company scale; however, it is driven by share-price volatility and market sentiment rather than operational performance.

Other economic output of the companies could have been used as a denominator, but for data availability and comparability across a large number of company, we will focus these 4 business metrics.

3.2 GHG Emissions Databases

COGEM Emission Database The raw data originate from the COGEM database constructed by the PLADIFES group at the Institut Louis Bachelier [14], which provides complete estimates for Scope 1, Scope 2, Scope 3 emissions. The COGEM methodology employs machine-learning model (CatBoost) to infer corporate GHG emissions. They use publicly available inputs from CDP and Refinitiv, supplemented by various open-source datasets. The model is calibrated to perform robustly across firms of all sizes, sector and region. This approach yields a consistent, transparent, and comprehensive emissions dataset for an exceptionally large cross-section of companies, mitigating the inconsistencies and opacity often encountered with other GHG estimation providers. The PLADIFES group has fully documented their methodology and sources, ensuring reproducibility and precision in corporate emissions estimation. Using the COGEM database is then justified by the lack of reliable raw data on corporate carbon emissions. Firms often fail to transparently disclose and existing data providers rely on proprietary, non-reproducible models. In contrast, the PLADIFES group has openly published the COGEM methodology and sources, combining multiple data inputs to produce consistent and transparent estimates. To ensure robust temporal modelling, we retained only companies with more than 10 consecutive years of complete Scope 1, 2, 3 emissions history, a maximum span of 19 years is available. We get a panel dataset of **31,677** companies spanning 125 countries, 163 sub-industries, each observed annually for between 11 and 19 years.

MSCI Emission Database The MSCI database initially comprises annual Scope 1 + 2 and Scope 3 emissions for publicly listed firms over a 14-years window (2009–2022). To construct our analysis sample, we applied the data filters, removing any firm exhibiting year-over-year changes exceeding 10 000%, with less than 9 years of historical data, reporting less than 100 kg (CO_2e). After filtering, our balanced panel contains **431** firms, each observed annually from 2009 through 2022.

Refinitiv Emission Database The Refinitiv database initially comprises annual Scope 1, Scope 2 and Scope 3 emissions for publicly listed firms over a 19-years window (2005–2023). To construct our analysis sample, we applied data filters, removing any firm exhibiting year-over-year changes exceeding 10 000%, with less than 9 years of historical data, reporting less than 100 kg CO_2e. After filtering, our balanced panel contains **341** firms, each observed annually from 2005 through 2023 with between 11 and 19 years of continuous emissions history.

Filtering Methodology for Scaling In order to compute intensities from emission databases and financial data, we have to ensure the completeness and validity of the denominator. To ensure robustness and comparability across firms, we apply the following four-step filtering procedure to each denominator series:

1. Temporal Completeness. Require a fully observed history of the denominator for every fiscal year from 2010 onward. Any company with one or more missing values in this period is excluded in its entirety.

2. Year-to-Year Consistency. Compute the annual relative change

$$\Delta_t = \left|v_t - v_{t-1}\right| / v_{t-1} \, .$$

   If $\Delta_t \geq 100$ for any t, the corresponding firm is removed, thereby eliminating implausible jumps or data-entry errors.

3. Lower-Tail Truncation (1% Cut-off). Calculate the 1% quantile $q_{0.01}$ of the pooled series. Firms whose minimum annual value falls below $q_{0.01}$ are dropped to mitigate the influence of extreme low-end outliers.

4. Intersection of Valid Companies. We kept the intersection of valid companies for the four denominators.

After this data cleaning steps, for COGEM each intensity serie has 13616 companies with a total of 242,727 observations, 311 companies and 4,249 observations for MSCI and 249 companies and 3,956 observations for Refinitiv.

Table 3.1: Descriptive Statistics of Emissions and Financial Variables (COGEM database)

Statistic	**Scope 1(kg)**	**Scope 2(kg)**	**Scope 3(kg)**	**Revenue($)**	**Market Cap.($)**
Count	2.43e+05	2.43e+05	2.43e+05	2.43e+05	2.43e+05
Mean	4.50e+05	8.01e+04	6.13e+05	2.25e+09	3.92e+09
Std	4.09e+06	3.68e+05	1.07e+07	1.21e+10	8.57e+10
Median	8.04e+03	1.23e+04	3.70e+03	1.91e+08	1.86e+08

Descriptive statistics The MSCI and Refinitiv database are smaller, with some missing value in their scope 3 data. They regroup only large cap companies, with higher mean

Table 3.2: Descriptive Statistics of Emissions and Financial Variables (MSCI Database)

Statistic	Scope 1+2 (kg)	Scope 3 (kg)	Revenue ($)	Market Cap. ($)
Count	4.24e+03	3.68e+03	4.24e+03	4.24e+03
Mean	6.94e+06	2.23e+07	2.56e+10	4.15e+10
Std	2.00e+07	6.82e+07	4.11e+10	9.25e+10
Median	6.30e+05	1.82e+06	1.28e+10	1.73e+10

Table 3.3: Descriptive Statistics of Emissions and Financial Variables (Refinitiv Database)

Statistic	Scope 1(kg)	Scope 2(kg)	Scope 3(kg)	Revenue($)	Market Cap.($)
Count	3.89e+03	3.82e+03	3.59e+03	3.96e+03	3.95e+03
Mean	5.30e+06	9.37e+05	1.83e+07	2.30e+10	3.79e+10
Std	1.97e+07	2.41e+06	6.81e+07	3.74e+10	1.20e+11
Median	1.10e+05	1.94e+05	5.11e+05	1.04e+10	1.25e+10

revenues than the COGEM database. Another detail worth to highlight is that the market cap has much higher standard deviation that the revenue for all databases with comparable mean. It could be a drawback for using it as a denominator for GHG intensity.
The COGEM database cover 161 sub-industries and 75 different countries of headquarters while the MSCI database covers 23 countries and 100 sub-industries and the Refinitiv database covers 31 countries and 90 industries. Their regional distribution varies a lot.

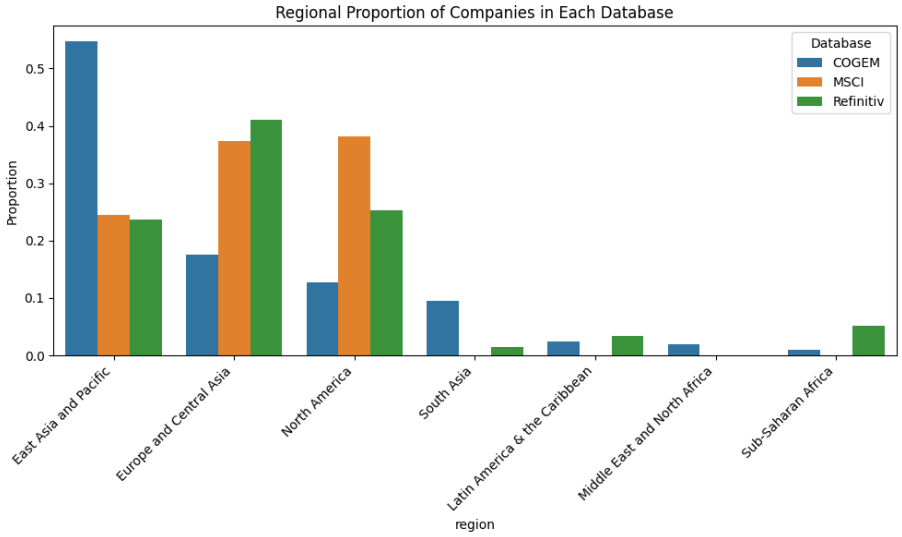

Figure 3.1: Regional distribution of companies across the three databases (COGEM, MSCI, Refinitiv).

3.3 Forecasting Framework

The goal is to forecast corporate GHG intensities across Scope 1, Scope 2, and Scope 3. We evaluated few models ranging from simple baselines to regularised linear regression and gradient-boosted trees. All targets are transformed to log-scale to stabilise variance and mitigate skewness. The logarithm transformation was employed due to the highly skewed

distribution of intensity variables. Corporate GHG intensities are highly right-skewed, with a small number of firms responsible for extremely large emissions and the majority emitting far less.

To ensure robust out-of-sample evaluation, we adopt a rolling window forecasting framework with a fixed one-year prediction horizon. At each iteration, models are trained on all available historical data up to year t, and evaluated by predicting emissions in year t+1. Model hyper-parameters for both CatBoost and Ridge regression are selected via grid search over a fixed validation period spanning depending on the dataset used (2014-2016 for COGEM Data, 2017-2018 for MSCI and Refinitiv Data). Once optimal hyperparameters are identified, we perform rolling forecasts for the test years depending on the dataset (2017-2023 for COGEM Data, 2019-2022 for MSCI Data and 2019-2023 for Refinitiv Data).

3.3.1 Models

We compare five prediction models for each scope category. The first three models are baseline naive models and serve as a comparison, the following two models are machine learning models elaborated and optimised in order to minimise mean square error:

1. Company-wise Mean : Predicts each firm's log-intensity as the historical average of its past log-intensities.

2. Company-wise EWMA: Uses the exponentially weighted moving average of past log-intensities per company.

3. Company-wise Temporal OLS: Calculates and uses the historical trend of each company log-intensity by an OLS estimation per company.

4. Panel Ridge Regression: Unique penalised regression model per scope, on the entire panel dataset, trained to minimise squared error on the log-transformed target.

5. Panel CatBoost Regression: A unique gradient-boosted decision tree model per scope, on the entire panel dataset, handling categorical features natively, trained to minimise squared error on the log-transformed target.

Comparing both linear and non-linear machine learning methods is essential to determine whether the added complexity of a more advanced model delivers significant benefits. Based on a literature review, we selected Ridge regression and CatBoost regression: tree-based methods like CatBoost are not only faster and better suited to this type of data and prediction task than Neural Network, but also often more interpretable. Although other Machine Learning approaches might perform well, our primary goal here is to demonstrate that tree-based approaches can increase the forecasting accuracy and more precisely that CatBoost outperforms both linear and baseline models.

Ridge Regression on Log-Transformed Target

We regress the log-transformed scope intensities target $y_{s,i,t}$ (company i, scope s, year t) on a set of lagged and trend features plus covariates (detailed in a following subsection), regularising the coefficients to mitigate over-fitting and multicollinearity. It profits from the cross-section information in addition to the temporal firm-wise information.

We apply a ColumnTransformer that handles numerical and categorical blocks separately:

- Numerical pipeline: No numerical metric is missing in the COGEM database, an Imputer should be used otherwise. A standard scaler to center and scale each feature to zero mean and unit variance.

- Categorical pipeline: A simple imputer fixes a single value for all missing entries. A one hot encoder is used to encode each categorical feature into a binary variable, dropping the first category to avoid perfect collinearity and safely ignoring any unseen categories at prediction time.

After preprocessing, we fit a ridge-penalised least-squares model:

$$\hat{\boldsymbol{\beta}} = \arg\min_{\boldsymbol{\beta}} \sum_{i=1}^{n} \left(y_i - \mathbf{x}_i^\top \boldsymbol{\beta}\right)^2 \; + \; \lambda \, \|\boldsymbol{\beta}\|_2^2,$$

where $\lambda > 0$ controls the shrinkage of coefficients toward zero. We select λ via a grid search on the validation set, optimising the mean squared error.
This ridge approach combines the interpretability and stability of linear models with feature standardisation and regularisation, yielding robust forecasts of corporate GHG intensities on a log-transformed scale. We fit one global model per scope using all historical data available, maximising the cross-sectional and temporal informations.

CatBoost Regression on Log-Transformed Target

CatBoost is a gradient boosting library developed by Prokhorenkova et al. [30] that is designed for handling categorical features natively by computing category-specific target statistics via an ordered permutation scheme, capturing interactions and preventing target leakage. An important property here as we use Geographic and Sectorial features. It offers fast training, strong performance on tabular data, and robust over-fitting control, making it well-suited for this prediction problem. As Grinsztajn et al. [17] outlined " *This superiority is explained by specific features of tabular data: irregular patterns in the target function, uninformative features, and non rotationally-invariant data where linear combinations of features misrepresent the information.*"
The CatBoost model extends the Ridge Regression framework by using gradient-boosted decision trees to capture non-linear relationships and interactions among predictors. The log transformed target y_i is regressed on the same features as for the Ridge Regression, which are detailed in the following subsection.

We configure the model hyper-parameters with a grid search on validation for each scope and database: iterations, learning rate, tree depth, L2 leaf regularisation. All features are converted to strings for categorical columns. The model iteratively fits new trees to the negative gradient of the RMSE loss.
This CatBoost approach typically yields superior performance by leveraging non-linear partitioning of the feature space and efficient, unbiased handling of categorical variables. We fit one global model per scope using all historical data available, maximising the cross-sectional and temporal informations.

3.3.2 Feature Selection and Engineering

To construct a parsimonious yet expressive predictor set, we employed a two-fold strategy: (i) domain-driven retention of economically interpretable variables, (ii) statistical screening, (iii) temporal smoothing and persistence features.

i) Economic Interpretability Filter We first retained only the metrics with clear links to future emissions:

- `past intensities`: enable to capture temporal trends for each company,
- `Financial metrics`: proxies for firm scale and production intensity,
- `Geographic Information`: The carbon footprint of the energy used by the company depends on the geographic zone,
- `Sectorial Information`: Capture information on specific activities of the company.

ii) Statistical Screening We kept only the features with clear importance on the training set using both the regularised linear model easily interpretable and the Shap values of the CatBoost model.

iii) Temporal Feature Engineering To capture both short-term inertia and long-term trends, we applied two smoothing constructs:

- **Lag features** for each scope s and lag $k \in \{1, 2, 3\}$ $\mathrm{Lag}^{(k)}_{s,i,t} = \log_{10}\big(I_{s,i,t-k} + \varepsilon\big)$ This captures year-to-year persistence in emissions, reflecting operational inertia and gradual changes in production processes.
- **log-EWMA smoothing** ($\alpha = 0.9$) on revenues and each scope's emissions $\mathrm{EWMA}_{s,i,t} = \alpha\, E_{s,i,t} + (1-\alpha)\,\mathrm{EWMA}_{s,i,t-1}$, then take the log of the $\mathrm{EWMA}_{s,i,t}$ This highlights longer-term trends while damping idiosyncratic fluctuations in annual reports.
- **Rolling volatility** (5-year window) on the log-emissions series:

$$\mathrm{RollStd}^{(5)}_{s,i,t} = \mathrm{std}\Big\{\log_{10}(E_{s,i,t-k} + \varepsilon) \;:\; k = 0, ..., 5\Big\}.$$

  This measures variability in emissions over a multi-year horizon, capturing the magnitude of fluctuations and informing the model about the stability or volatility of a company's emission profile.

These smoothing constructs mitigate year-to-year volatility while preserving underlying trends and temporal dependencies, with lag features explicitly encoding persistence and rolling standard deviation quantifying emission variability, thus improving model robustness and predictive accuracy.

iv) Final Feature Vector Combining all elements, each firm–year observation is represented by the features Table 3.4.

This rigorously and transparently selected and engineered feature space supports our machine learning models, balancing interpretability, statistical validity, and predictive performance.

Table 3.4: Final Feature Vector for Firm–Year Observations

Feature	**Type**	**Description**
Intensity $\text{Lag}_{1\text{–}3}$	Dynamic	Log_{10} intensities at $t-1$, $t-2$, $t-3$; captures year-to-year persistence.
Intensity EWMA	Dynamic	Log_{10} of EWMA of intensities; highlights long-term trend.
RollStd_5	Dynamic	5-year rolling standard deviation of $\log_{10}$ intensities; measures volatility.
Country	Categorical	geographic indicator.
GICS Sub Industries	Categorical	Industry classification at finest level.

3.4 Model Evaluation

The models were tested on a rolling window with a one year horizon on the three databases to ensure a robust framework for different providers of emissions data.

3.4.1 Results on COGEM Database

Model	Scope 1	Scope 2	Scope 3
CatBoost	0.0094	0.0078	0.0145
Ridge	0.0222	0.0139	0.0230
EWMA	0.1497	0.1119	0.2290
OLS	0.2123	0.1657	0.3841
Mean	0.2752	0.2620	0.5779

Table 3.5: MAE for **Asset** Scaling (COGEM)

Model	Scope 1	Scope 2	Scope 3
CatBoost	0.0095	0.0076	0.0143
Ridge	0.0247	0.0158	0.0231
EWMA	0.1625	0.1241	0.2280
OLS	0.2246	0.1744	0.3856
Mean	0.2682	0.2423	0.6193

Table 3.6: MAE for **Revenue** Scaling (COGEM)

Model	Scope 1	Scope 2	Scope 3
CatBoost	0.0115	0.0094	0.0154
Ridge	0.0297	0.0223	0.0275
EWMA	0.2130	0.1749	0.2745
OLS	0.3032	0.2524	0.4250
Mean	0.3563	0.3291	0.6063

Table 3.7: MAE for **Market Cap.** Scaling (COGEM)

Model	Scope 1	Scope 2	Scope 3
CatBoost	0.0097	0.0072	0.0145
Ridge	0.0233	0.0157	0.0230
EWMA	0.1573	0.1246	0.2274
OLS	0.2208	0.1795	0.3898
Mean	0.2650	0.2490	0.6251

Table 3.8: MAE for **COGS** Scaling (COGEM)

Analysis: The results in Tables 3.5–3.8 show that **CatBoost** is by far the best-performing model across all scopes. Its MAE values are approximately a factor of two lower than those obtained with Ridge regression, the second-best approach. Both CatBoost and Ridge, as more advanced models, clearly outperform the baseline models (EWMA, Mean, and OLS). Importantly, relying on a simple temporal trend through OLS is not a suitable strategy for modelling carbon intensities: the errors are both large and unstable, making it a poor choice in practice. Among the baseline methods, EWMA offers a better compromise: it is fast, simple to implement, and consistently delivers lower errors than the Mean or OLS. Therefore, while advanced methods such as CatBoost should be preferred whenever possible, EWMA can serve as a reasonable lightweight baseline.

The results reported in Tables 3.5–3.8 allow us to compare the effect of the chosen scaling variable on the accuracy of carbon intensity forecasts. We observe that using assets, revenue, or COGS as scalars leads to very similar levels of predictive accuracy across all scopes and models. In contrast, the forecasts obtained when scaling by market capitalisation display consistently higher errors. This result may be explained by the following: market capitalisation is subject to speculation, volatility, and noise from financial markets, making it a less reliable denominator for stable intensity measures. We see on the other datasets that the market capitalisation works as well as the others scalers.

Another important pattern concerns the relative predictability of the different scopes. Across all models and scaling choices, Scope 2 emissions are consistently forecasted with the lowest error, followed by Scope 1, while Scope 3 forecasts are markedly less accurate. This hierarchy can be rationalised by the underlying nature of the data. Scope 2 emissions are derived from purchased electricity and heat, which are typically well-documented, regularly reported, and strongly correlated with observable firm characteristics such as production activity and energy intensity. By contrast, Scope 1 emissions reflect direct emissions from a company's operations, which can vary with production processes, technology, introducing additional variability. The greatest challenge, however, lies in Scope 3: these emissions arise along the entire value chain (both upstream and downstream), are often estimated rather than directly measured, and depend on data from suppliers and customers that may be incomplete or inconsistent. Consequently, Scope 3 emissions inherently exhibit higher noise levels and weaker links with the predictors available, which explains why all models show substantially lower predictive accuracy in this scope.

Interpretability: For tree based model, the SHAP Values are commonly used to understand better the model behaviour. The SHAP values provide insight into how each predictor contributes to the model's output for a particular instance. Positive SHAP values indicate a predictor that pushes the prediction higher, while negative values indicate a predictor that pushes it lower.

Figure 3.2 visualises the SHAP value distributions for our CatBoost models. The EWMA of the historical intensity dominates model output: higher EWMA values (red points) uniformly increase the predicted log intensities, reflecting the strong persistence in historical intensity trends. The lag-1 feature is the next most influential, with high lag-values tending to push forecasts down. Rolling-standard-deviation features exhibit smaller but discernible impacts, indicating that recent volatility modestly adjusts the baseline trend. In contrast, categorical variables (GICS sub-industry, country) contribute little to the instantaneous prediction, as evidenced by their narrow SHAP distributions centred near zero. Overall, these plots confirm that temporal intensities features encapsulate the bulk of predictive power in our emissions forecasting models.

Figures 3.3–3.6 reveal an almost perfectly linear relationship between the EWMA of past Scope 3 emissions and its SHAP contribution, confirming that this smoothed trend feature drives the forecast nearly linearly. It explains the good performances of the EWMA predictor and the Ridge predictor. In contrast, it shows a non-linear, U-shaped effect of the one-period and two period lags: SHAP values decrease (more negative impact) as the lag increases from low to moderate levels—reflecting regression toward the long-term trend—but then flatten and even rise slightly at very high lag values, indicating the model partially corrects for extreme past emissions. These patterns underscore the CatBoost model's ability to combine a dominant linear trend component with nuanced, non-linear adjustments based on recent deviations. It explains why the Catboost has superior result on this prediction problem.

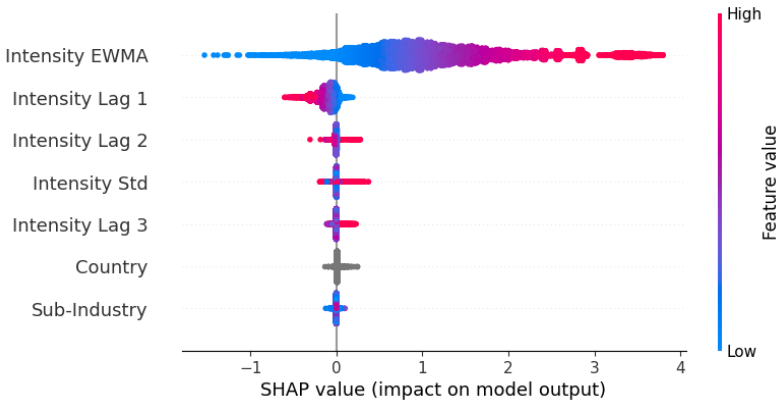

Figure 3.2: SHAP Summary

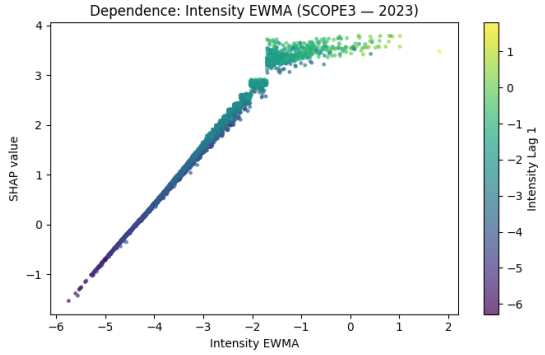

Figure 3.3: Dependance Plot EWMA

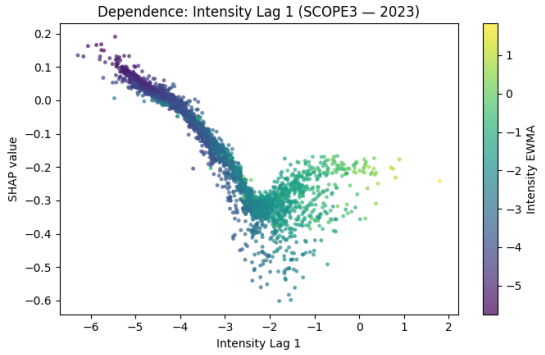

Figure 3.4: Dependance Plot Lag-1

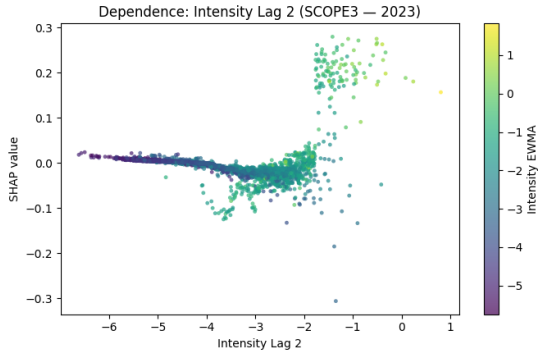

Figure 3.5: Dependance Plot Lag-2

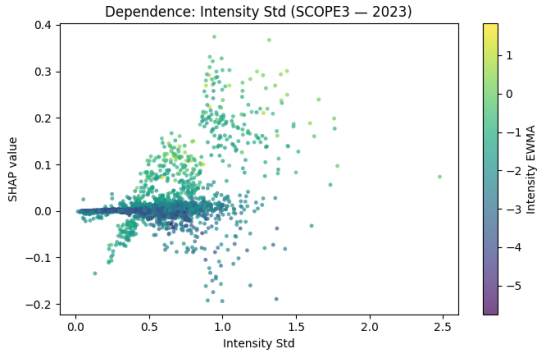

Figure 3.6: Dependance Plot Std

3.4.2 Results on MSCI and Refinitiv Databases

The goal is to apply here exactly the same methodology on two other independent databases, as there is some divergence in the estimation methodology between different providers. The MSCI and Refinitiv databases are smaller and less smooth than the COGEM database, it could be harder to calibrate ML models.

The COGEM database covers a very broad universe of around 13,616 companies worldwide, including smaller firms and emerging markets, and relies heavily on machine-learning estimation techniques to fill data gaps, resulting in smoother but sometimes less granular

values. By contrast, MSCI and Refinitiv restrict their coverage to large-cap firms, mainly from developed economies, with more direct and consistent reporting.

Model	Scope 1+2	Scope 3
CatBoost	0.0079	0.0466
Ridge	0.0173	0.0483
EWMA	0.0704	0.1903
OLS	0.1220	0.5518
Mean	0.1939	0.6671

Table 3.9: MAE for **Asset** Scaling (MSCI)

Model	Scope 1+2	Scope 3
CatBoost	0.0097	0.0503
Ridge	0.0187	0.0552
EWMA	0.0758	0.2008
OLS	0.1230	0.5878
Mean	0.1705	0.7478

Table 3.10: MAE for **Revenue** Scaling (MSCI)

Model	Scope 1+2	Scope 3
CatBoost	0.0098	0.0556
Ridge	0.0225	0.0599
EWMA	0.1070	0.2176
OLS	0.1522	0.5634
Mean	0.2543	0.6518

Table 3.11: MAE for **Market Cap.** Scaling (MSCI)

Model	Scope 1+2	Scope 3
CatBoost	0.0195	0.0550
Ridge	0.0198	0.0581
EWMA	0.084	0.2101
OLS	0.1386	0.6402
Mean	0.1845	0.8029

Table 3.12: MAE for **COGS** Scaling (MSCI)

Model	Scope 1	Scope 2	Scope 3
CatBoost	0.0120	0.0116	0.0321
Ridge	0.0204	0.0170	0.0477
EWMA	0.0675	0.0820	0.1522
OLS	0.1289	0.1855	0.4422
Mean	0.1800	0.2064	0.5401

Table 3.13: MAE for **Asset** Scaling (Refinitiv)

Model	Scope 1	Scope 2	Scope 3
CatBoost	0.0139	0.1333	0.0359
Ridge	0.0215	0.0188	0.0502
EWMA	0.0782	0.0916	0.1565
OLS	0.1313	0.2011	0.4648
Mean	0.1721	0.2154	0.6112

Table 3.14: MAE for **Revenue** Scaling (Refinitiv)

Model	Scope 1	Scope 2	Scope 3
CatBoost	0.0136	0.0123	0.0360
Ridge	0.0227	0.0208	0.0552
EWMA	0.1056	0.1240	0.1931
OLS	0.1738	0.2241	0.4689
Mean	0.2356	0.2800	0.5646

Table 3.15: MAE for **Market Cap.** Scaling (Refinitiv)

Model	Scope 1	Scope 2	Scope 3
CatBoost	0.0172	0.0149	0.0369
Ridge	0.0239	0.0220	0.0546
EWMA	0.0846	0.1015	0.1693
OLS	0.1522	0.2256	0.5073
Mean	0.1988	0.2406	0.6608

Table 3.16: MAE for **COGS** Scaling (Refinitiv)

When extending the analysis to the MSCI and Refinitiv databases, Tables 3.9–3.16, we observe a consistent hierarchy of model performances. In both datasets, CatBoost remains the most accurate model, followed by Ridge regression, while the baseline methods perform substantially worse. The error magnitudes are of the same order as those obtained with

the COGEM database, which underlines the robustness of these conclusions across distinct data providers.
Some differences nevertheless emerge. While in the COGEM dataset market capitalisation proved to be the least reliable scaling variable, in the MSCI and Refinitiv cases it is COGS that leads to the largest errors. We use later the revenue as scaler to compute the intensities, as it is economically sense-full and statistically reliable.
Finally, the relative predictability of emission scopes is preserved across all three databases: Scope 2 emissions are consistently forecasted with lower errors than Scope 1, while Scope 3 remains the most challenging to model.

Chapter 4

Practical Approaches to Low-Carbon Portfolio Optimisation

4.1 Portfolio Optimisation Strategy

This chapter operationalises the theoretical framework of Chapter 2 and the forecasting models of Chapter 3. We construct weekly long-only, fully-invested portfolios that jointly target financial risk reduction using variance and environmental impact reduction using portfolio-level carbon intensity. We compare two approaches: (i) a quadratic program that scalarises risk and carbon, and (ii) a scalarised reinforcement learning policy trained with Proximal Policy Optimisation (PPO), an actor-critic algorithm.

4.1.1 Risk Estimation: EWMA Covariance

Let $r_t \in \mathbb{R}^n$ denote asset returns at day t. We estimate risk via an exponentially weighted moving-average (EWMA) covariance:

$$\widehat{\Sigma}_t^{\mathrm{EWMA}} \;=\; (1-\alpha)\sum_{s=1}^{t} \alpha^{\,t-s}\,(r_s-\bar{r})(r_s-\bar{r})^{\top}, \qquad \alpha = 0.94,$$

using all data available up to t. This choice balances reactivity to new information with stability.

4.1.2 Carbon Signal: Annual Predicted Intensities

Portfolio carbon exposure is measured by the portfolio intensity $I^{(y)\prime}w$. Firm-level intensities are observed annually; hence, as in Chapter 3, we form one-year-ahead predictions with CatBoost using only information up to year $y-1$. At the start of calendar year y, we freeze the predicted intensity vector $I^{(y)} \in \mathbb{R}^n$ for all trading weeks t in that year. To isolate the role of forecasting accuracy, we run each strategy under three intensity inputs:

1. Ex-post realised intensities.
2. Naïve mean (firm-wise historical average up to $y-1$).
3. CatBoost forecast.

4.1.3 Weekly Portfolio Construction: Two Frameworks

Let $\Delta^{n-1} := \{w \in \mathbb{R}^n : \mathbf{1}'w = 1,\ w \geq 0\}$ denote the long-only simplex, and let $y(t)$ be the calendar year of week t.

(i) Penalised QP. Given $\widehat{\Sigma}_t^{\mathrm{EWMA}}$ and $I^{(y(t))}$, we solve

$$\min_{w \in \Delta^{n-1}} \quad w' \widehat{\Sigma}_t^{\mathrm{EWMA}} w \ + \ \gamma\, I^{(y(t))\prime} w, \tag{4.1.1}$$

with trade-off $\gamma \geq 0$ (cf. 2.3.1). Varying γ traces a risk–carbon efficient set.

(ii) Scalarised RL. We cast the sequential problem as an MDP (Definitions in Chapter 2): the state s_t includes $\widehat{\Sigma}_t^{\mathrm{EWMA}}$ and $I^{(y(t))}$; the action is $w_t \in \Delta^{n-1}$; the (scalarised) reward is

$$R(s_t, w_t) \ = \ -\tfrac{1}{2}\, w_t' \widehat{\Sigma}_t^{\mathrm{EWMA}} w_t \ - \ \gamma\, I^{(y(t))\prime} w_t. \tag{4.1.2}$$

We train a stochastic policy $\pi_\theta(w|s)$ with a proximal policy optimisation algorithm (PPO) from Dhariwal et al. [34], leveraging the policy gradient theorem (Theorem 2.3.10). To improve stability and feasibility, we warm-start the actor from the QP solution of (4.1.1) at the same γ.

Remark. The convex-combination scalarisation with $\lambda \in (0,1)$ is equivalent to the penalised form $\min_w\, w'\widehat{\Sigma}_t w + \gamma\, I_t' w$ via $\gamma \ = \ \frac{1-\lambda}{\lambda}$, with corners $\lambda = 1 \Leftrightarrow \gamma = 0$ (pure variance) and $\lambda = 0 \Leftrightarrow \gamma = \infty$ (pure carbon). The λ formulation is used in our implementation.

4.2 Proximal Policy Optimisation Implementation

We detail the PPO implementation tailored to portfolio weights on the simplex.

4.2.1 Policy Class and Constraints

To enforce $w_t \in \Delta^{n-1}$, we parametrise actions via a softmax:

$$w_t \ = \ \mathrm{softmax}\big(z_\theta(s_t)\big), \qquad \text{so that } w_{t,i} \geq 0,\ \sum_i w_{t,i} = 1.$$

(Equivalently, a Dirichlet policy is possible; we found softmax logits numerically stable.) The critic $V_\phi(s)$ estimates the value function.

4.2.2 Learning Objective (PPO)

Let $\kappa_t(\theta) = \frac{\pi_\theta(w_t|s_t)}{\pi_{\theta_{\mathrm{old}}}(w_t|s_t)}$ and $\widehat{A}_t$ be the advantage. Advantages use GAE with parameter $\lambda_{\mathrm{GAE}} \in [0,1]$. The clipped actor loss, critic loss, and entropy bonus are

$$\mathcal{L}^{\mathrm{actor}}(\theta) = \mathbb{E}\Big[\min\big(\kappa_t(\theta)\widehat{A}_t,\ \mathrm{clip}(\kappa_t(\theta), 1-\epsilon, 1+\epsilon)\,\widehat{A}_t\big)\Big],$$

$$\mathcal{L}^{\mathrm{critic}}(\phi) = \mathbb{E}\big[(V_\phi(s_t) - \widehat{G}_t)^2\big],$$

$$\mathcal{H} = \mathbb{E}[\,\mathrm{Ent}(\pi_\theta(\cdot|s_t))\,],$$

we optimise the composite objective

$$\mathcal{L}^{\mathrm{actor}} - \beta_V \mathcal{L}^{\mathrm{critic}} + \beta_H \mathcal{H}.$$

Advantage estimation. The advantage quantifies how much better the sampled action w_t is compared to the critic's baseline for the same state:

$$A^{\pi}(s_t, w_t) \;=\; Q^{\pi}(s_t, w_t) \;-\; V^{\pi}(s_t).$$

In practice we form an on–policy estimator $\widehat{A}_t$ using the value critic V_ϕ and generalised advantage estimation (GAE). Let the one–step TD residual be

$$\delta_t \;=\; R_t \;+\; \beta\, V_\phi(s_{t+1}) \;-\; V_\phi(s_t),$$

(with $V_\phi(s_{t+1}){:=}0$ if s_{t+1} is terminal). Then for a trajectory segment ending at T,

$$\widehat{A}_t \;=\; \sum_{l=0}^{T-t-1} (\beta\,\lambda_{\mathrm{GAE}})^l\, \delta_{t+l} \quad\Longleftrightarrow\quad \widehat{A}_t \;=\; \delta_t \;+\; \beta\,\lambda_{\mathrm{GAE}}\,\widehat{A}_{t+1},\;\; \widehat{A}_T{:=}0,$$

where $\lambda_{\mathrm{GAE}} \in [0,1]$ controls the bias–variance trade–off ($\lambda_{\mathrm{GAE}}{=}1$: MC–like, higher variance; $\lambda_{\mathrm{GAE}}{=}0$: one–step TD, lower variance). The critic target used in $\mathcal{L}^{\mathrm{critic}}$ is

$$\widehat{G}_t \;=\; \widehat{A}_t \;+\; V_\phi(s_t).$$

In PPO, the sign and magnitude of $\widehat{A}_t$ modulate the policy update via the clipped ratio $\kappa_t(\theta)$: positive (resp. negative) advantages increase (resp. decrease) $\pi_\theta(w_t|s_t)$. For numerical stability we normalise advantages within each batch,

$$\widehat{A}_t \leftarrow \frac{\widehat{A}_t - \overline{\widehat{A}}}{\mathrm{std}(\widehat{A}) + \varepsilon}.$$

4.2.3 Algorithm

In our experiments we use a fixed risk-carbon trade-off: a scalarisation weight $\lambda \in [0,1]$ combines variance and carbon into a single reward. At week t, given an EWMA covariance $\widehat{\Sigma}_t^{\mathrm{EWMA}}$ (decay $\alpha = 0.94$) and a weekly carbon vector I_t, the reward is

$$R_t \;=\; \lambda\big(-\, w_t' \widehat{\Sigma}_t^{\mathrm{EWMA}}\, w_t\big) \;+\; (1-\lambda)\big(-\, I_t' w_t\big).$$

Actions $w_t \in \Delta^{n-1}$ (long-only, fully invested) are produced by a softmax policy over logits $z_\theta(s_t)$ to enforce simplex constraints. We train one PPO agent per value of λ and compare the resulting policies. We use the standard PPO algorithm [34] via the off-the-shelf implementation in Stable-Baselines3; we make no modifications to the core algorithm.

Algorithm 1 PPO for Carbon-Aware Portfolio (Fixed Scalarisation λ, Weekly Rebalancing)

Inputs: discount $\beta \in (0,1)$; learning rates $\alpha_\theta, \alpha_\phi$; scalarisation $\lambda \in [0,1]$; EWMA decay α=0.94; daily window W
Initialise: actor parameters θ, critic parameters ϕ
Choose the first rebalancing week t_0 such that at least W daily returns are available
for $t = t_0, t_0{+}1, \ldots, T$ **do**
    Weekly loop
Build the W-day return window up to week t; compute EWMA covariance $\widehat{\Sigma}_t$
Obtain weekly carbon vector I_t (piecewise-constant within the current year)
Form state s_t (e.g., $\widehat{\Sigma}_t$, I_t)
Sample action $w_t \sim \pi_\theta(\cdot \mid s_t)$ and project onto $\Delta^{n-1} = \{w \geq 0,\ \mathbf{1}'w = 1\}$
Compute reward: $R_t \leftarrow \lambda\big(-\, w_t^\top \widehat{\Sigma}_t\, w_t\big) \;+\; (1-\lambda)\big(-\, I_t^\top w_t\big)$
Observe next state s_{t+1}
Critic (TD): $\delta_t \leftarrow R_t + \beta\, V_\phi(s_{t+1}) - V_\phi(s_t)$; $\phi \leftarrow \phi + \alpha_\phi\, \nabla_\phi V_\phi(s_t)\, \delta_t$
Estimate advantage $\widehat{A}_t$ (e.g., GAE)
Actor (PPO): update θ with the clipped surrogate objective using $\widehat{A}_t$

Explanation and design choices.

- **Stationarity.** Within each year, the carbon signal I_t is fixed by the annual forecast (Chapter 3). Keeping λ fixed ensures the reward is stationary during training, aiding PPO stability.
- **Connection to theory.** The scalarised reward implements the penalised quadratic program of Chapter 2 in a dynamic setting. Varying λ traces a risk–carbon efficient set, directly comparable to the static frontier obtained by sweeping the penalty parameter.
- **Action constraints.** The softmax parametrisation yields $w_t \in \Delta^{n-1}$ by construction; an explicit projection step can be kept as a safeguard.
- **State representation.** We supply either the last W daily returns (used to reconstruct $\widehat{\Sigma}_t$), concatenated with I_t.

In the next section we report out-of-sample results across a grid of λ values, comparing variance, carbon intensity, and turnover against the static quadratic program and baseline strategies.

4.3 Empirical Results

4.3.1 Data and Experimental Setup

We evaluate all strategies out of sample over **2019–2022** on a balanced universe of **50 U.S. listed companies**. The sample consists of the U.S. firms with complete multi-year coverage of GHG data across the three emissions providers considered in this thesis (see Chapter 3). Portfolios are rebalanced weekly. Risk is estimated with the EWMA covariance (decay $\alpha = 0.94$) using all daily returns available up to each rebalancing date. The carbon signal is updated annually: at the start of year y, we freeze the firm-level carbon vector $I^{(y)}$ (either realised or predicted) for the entire year, as discussed in Chapter 2.
We compare two optimisation frameworks:

- **Quadratic program (QP):** $\min_w \ \lambda w'\Sigma_t w + (1-\lambda)\, I^{(y)\prime} w$ with $w_t \in \Delta^{n-1}$.
- **Reinforcement learning (PPO):** the reward is $R_t = \lambda(-w_t'\Sigma_t w_t)+(1-\lambda)(-I^{(y)\prime} w_t)$ with $w_t \in \Delta^{n-1}$, trained as in §4.2.

For robustness, each framework is run under three information sets for $I^{(y)}$: (i) ex-post realised intensities ("Actual"), (ii) naïve firmwise mean ("Mean"), and (iii) CatBoost forecasts ("CatBoost"). Predictions are constructed as in Chapter 3.

Evaluation Metrics We report two complementary metrics.

- **Annual portfolio total carbon (absolute emissions):** Let $E_{y,i}$ denote firm i's annual total GHG (Scopes 1+2+3) in year y. With weekly portfolio weights $w_{t,i}$ and trading weeks $\mathcal{T}_y$, define the average annual weight

$$\bar{w}_{y,i} \ = \ \frac{1}{|\mathcal{T}_y|} \sum_{t \in \mathcal{T}_y} w_{t,i},$$

  and the portfolio's annual *total* carbon

$$C_y \ = \ \sum_{i=1}^{n} \bar{w}_{y,i}\, E_{y,i}.$$

  We report C_y and the sample average $\bar{C} = |\mathcal{Y}|^{-1} \sum_{y \in \mathcal{Y}} C_y$, with $\mathcal{Y} = \{2019, 2020, 2021, 2022\}$.

- **Realised portfolio variance:** Let $R_{t+1} = w_t' r_{t+1}$. For each year y, compute the weekly variance over $t \in \mathcal{T}_y$ and annualise:

$$\sigma_y^2 \ = \ 52\, \mathrm{Var}_y[R], \qquad \bar{\sigma}^2 \ = \ \frac{1}{|\mathcal{Y}|} \sum_{y \in \mathcal{Y}} \sigma_y^2.$$

Risk-Carbon Frontier. For QP we sweep $\lambda \in [0,1]$ and backtest to obtain pairs $\big(\bar{\sigma}^2(\lambda), \bar{C}(\lambda)\big)$. For PPO, we train one agent per $\lambda \in [0,1]$ (Algorithm 1) and backtest to obtain $\big(\bar{\sigma}^2(\lambda), \bar{C}(\lambda)\big)$.

4.3.2 QP Results: Risk–Carbon Frontiers

Figure 4.1 displays quadratic-programming frontiers, plotting average total portfolio emissions (x-axis) against average annualised variance (y-axis). We observe that using CatBoost forecasts shifts the frontier inward relative to the naïve Mean baseline: for any given level of variance, predicted intensities achieve materially lower emissions—confirming that forecast quality translates into tangible economic value. The equal-weight portfolio is shown for reference. Notably, portfolios with trade-off parameters $\lambda \in (0.5, 1)$ attain *lower variance* than the pure variance-optimisation case ($\lambda = 1$). In this region, portfolio emissions can be reduced by a factor of about three while maintaining the same optimal variance, illustrating that substantial decarbonisation can be achieved at negligible, or even negative, cost in terms of variance.
Finally, we remark that the emission–variance frontier retains a close shape to the classical mean–variance efficient frontier: portfolios lie on a smooth, convex trade-off curve. This shows that in practice there is indeed a frontier capturing the fundamental trade-off between variance and carbon emissions.

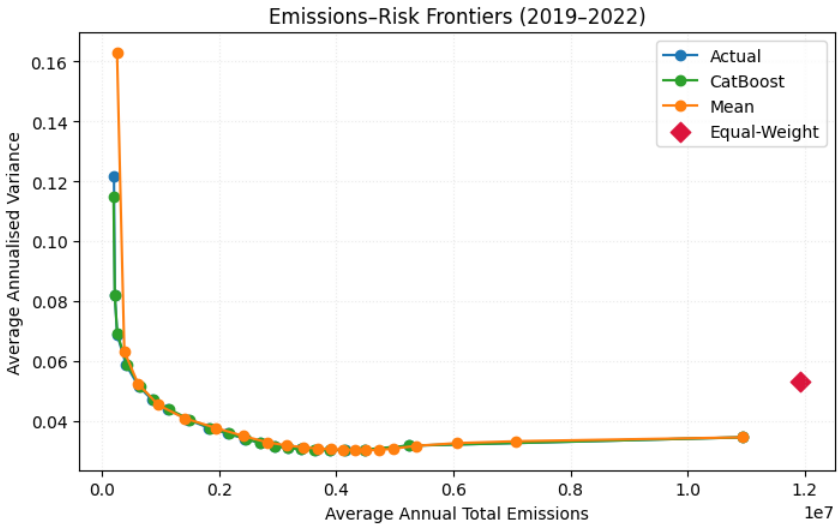

Figure 4.1: **QP frontiers (2019–2022).** Average annual total emissions vs. annualised variance, comparing Actual, CatBoost, and Mean intensity information. Better forecasts move the frontier inward (lower carbon for the same risk).

Table 4.1 quantifies the effect of forecasting firm-level carbon intensities on portfolio emissions. For each value of the penalty parameter λ, we compute the difference in average portfolio total emissions between a forecast-based optimisation and the benchmark case with actual realised intensities:

$$\Delta_{\text{cat}} = \bar{C}_{\text{CatBoost}} - \bar{C}_{\text{Actual}}, \qquad \Delta_{\text{mean}} = \bar{C}_{\text{Mean}} - \bar{C}_{\text{Actual}}.$$

By construction, positive values indicate that forecasting leads to higher total portfolio emissions than if actual intensities were known at the optimisation stage.

The results highlight two main points. First, the CatBoost forecasts track realised intensities closely: across all values of λ, Δ_{cat} is small (on the order of 10^3–10^4 tCO_2e), which is negligible compared to the scale of total portfolio emissions (10^6 tCO_2e). Second, naïve mean forecasts lead to very large gaps: Δ_{mean} ranges from $3.5 \cdot 10^5$ to $1.6 \cdot 10^6$ tCO_2e, i.e. between 30% and more than 100% higher emissions relative to the benchmark. This shows that poor forecasts of carbon intensities can severely degrade portfolio decarbonisation, whereas a machine-learning approach maintains emissions very close to the actual-information optimum.

Table 4.1: Average total emissions gap vs. Actual (QP, 2019–2022).

λ	$\bar{C}_{\text{Actual}}(TCO_2e)$	Δ_{cat}	Δ_{mean}
0.90	$4.49 * 10^6$	$1.19 * 10^3$	$1.58 * 10^6$
0.60	$2.94 * 10^6$	$2.76 * 10^3$	$1.18 * 10^6$
0.40	$1.82 * 10^6$	$2.43 * 10^4$	$1.32 * 10^6$
0.20	$0.62 * 10^6$	$2.00 * 10^4$	$7.89 * 10^5$
0.10	$0.26 * 10^6$	$1.64 * 10^3$	$3.54 * 10^5$

Cross-provider evaluation. We further assess robustness by re-evaluating the same QP portfolio weights against an alternative emissions database (e.g., MSCI or Refinitiv) for $E_{y,i}$, while keeping the optimisation inputs fixed to COGEM. Figure 4.2 overlays the resulting frontiers when portfolio emissions are measured with COGEM versus the alternative provider. The absolute levels of total emissions shift substantially across databases,

with differences reaching up to 80% higher emissions when evaluated with MSCI or Refinitiv instead of COGEM for the global minimum variance portfolio. This reflects heterogeneity in raw emissions data rather than instability of the optimisation procedure. The key implication is that portfolio-level decarbonisation results are consistent within each provider, but absolute levels should not be compared across providers without adjusting for data-quality differences.

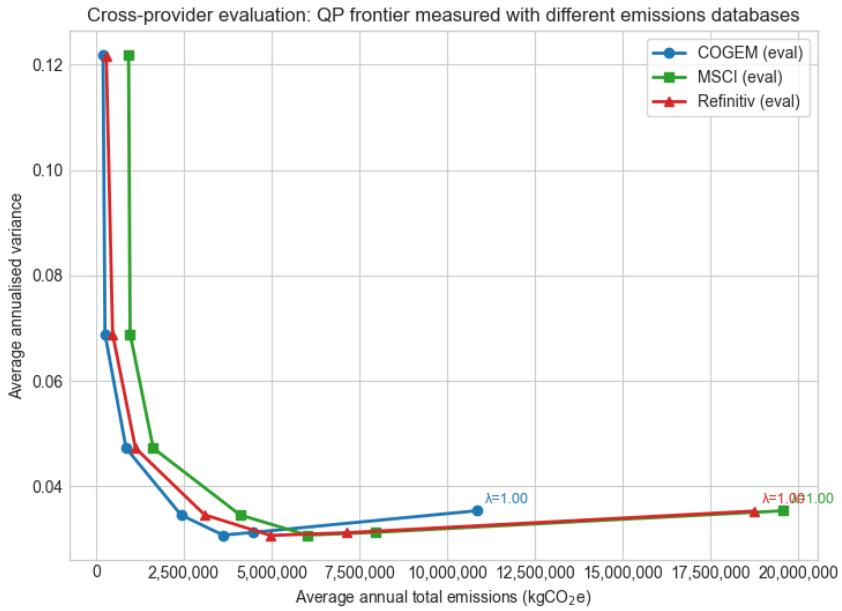

Figure 4.2: **Cross-database evaluation (QP).** Emissions measured with COGEM vs. an alternative provider. The relative ordering across information sets is stable.

4.3.3 Reinforcement Learning Results

We implement PPO as in Section 4.2 (Algorithm 1), restricting the testing years to 2021–2022 in order to retain a larger training set. The resulting RL frontier in the risk–carbon plane closely tracks the QP frontier, as expected from the scalarised objective. For direct comparability, Figure 4.3 overlays the QP and PPO frontiers (same λ grid, same testing years). Overall differences are minimal, indicating that PPO effectively learns the penalised optimisation structure (see Appendix A.4 for hyperparameters).

A number of key patterns emerge. First, for all values of λ, both QP and PPO portfolios dominate the equal-weight benchmark in terms of emissions. Second, for $\lambda \geq 0.45$, both methods also outperform the equal-weight portfolio in terms of variance. Third, PPO consistently achieves lower emissions than QP for the same λ, showing that reinforcement learning can exploit the trade-off more efficiently.

The strongest joint improvement occurs at $\lambda = 0.85$: the PPO portfolio delivers the lowest variance along the frontier while simultaneously reducing emissions by about 50% relative to the equal-weight portfolio. At $\lambda = 0.85$, PPO reduces variance by around 20% compared to the global minimum-variance case, while dividing total emissions by a factor of three. These results show that PPO not only reproduces the efficient risk–carbon frontier but, in certain regions, dominates both QP and the equal-weight benchmark by delivering lower risk and substantially lower emissions.

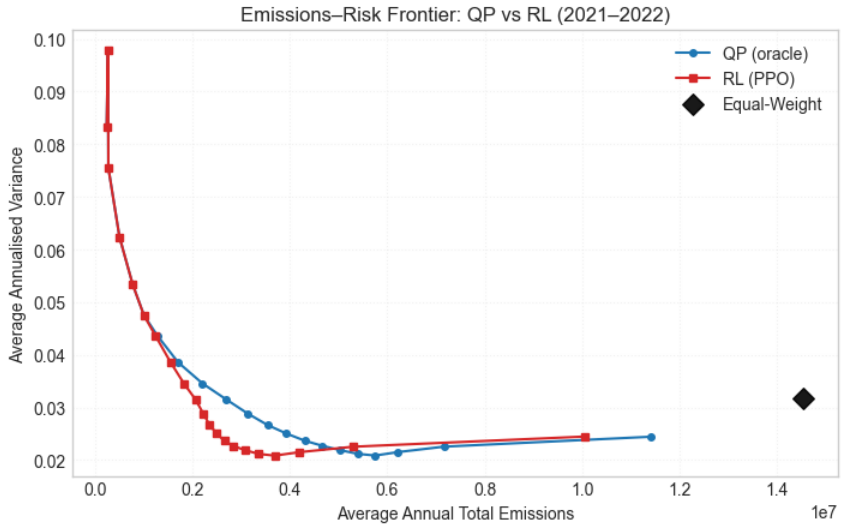

Figure 4.3: **QP vs. PPO frontiers.** Overlaid frontiers confirm that the actor–critic policy outperforms the penalised quadratic solution in practice.

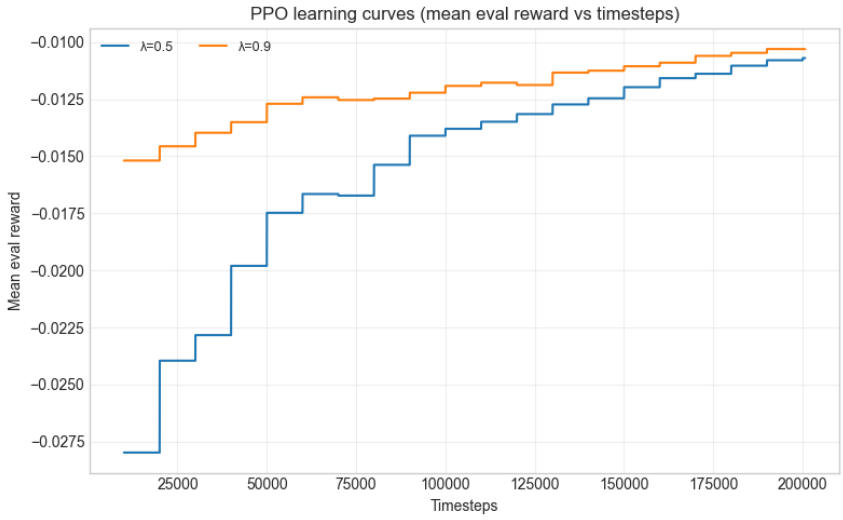

Figure 4.4: **PPO Learning curve.**

Figure 4.4 reports the PPO learning dynamics for two trade-off parameters, $\lambda = 0.5$ and $\lambda = 0.9$. Both policies converge steadily as timesteps increase, with rewards stabilising around -0.01 to -0.0105. The $\lambda = 0.9$ policy exhibits faster initial learning and systematically higher rewards throughout training, suggesting that portfolios with stronger emphasis on variance reduction learn more efficiently. By contrast, $\lambda = 0.5$ converges more slowly but eventually approaches similar performance. Overall, these results confirm that PPO achieves stable convergence and that the choice of λ affects both learning speed and final policy quality.

Discussion and Key Takeaways.

- *Risk–carbon efficiency:* Both QP and PPO achieve substantial carbon reductions without increasing portfolio risk. In particular, it is possible to construct portfolios with equal or lower variance than the global minimum-variance benchmark while dividing total emissions by a factor of three.

- *Forecasting accuracy:* The quality of carbon-intensity forecasts is critical. Machine-learning forecasts (CatBoost) consistently shift the entire risk–carbon frontier downward, whereas naïve mean forecasts lead to portfolios with substantially higher emissions.
- *Methodological choice:* Reinforcement learning (PPO) reproduces the efficient frontier and, in several regions, strictly dominates the penalised QP solution. This makes PPO particularly attractive in dynamic contexts where adaptability and sequential decision-making matter.
- *Data-provider heterogeneity:* Absolute levels of portfolio emissions vary widely across databases (e.g., COGEM vs. MSCI or Refinitiv). However, within each provider, results are internally consistent. Cross-provider comparisons should therefore be interpreted cautiously.

Chapter 5

Conclusion

This thesis has demonstrated that integrating predictive modelling of corporate greenhouse gas emissions with advanced portfolio optimisation frameworks enables investors to decarbonise portfolios without compromising financial performance. By combining CatBoost forecasts with both quadratic programming (QP) and reinforcement learning (PPO), we showed that methodological advances in forecasting and optimisation can shift the risk–carbon frontier inward, delivering portfolios that are simultaneously financially robust and environmentally more sustainable. In particular, we documented that portfolio total emissions can be reduced by a factor of three while maintaining, or even lowering, variance compared to classical global minimum variance.

Beyond the quantitative gains, several broader lessons emerge. First, accurate forecasts of firm-level carbon intensities are indispensable: naïve predictors generate portfolios with up to 100% higher emissions, whereas machine-learning forecasts remain close to the actual-information optimum. Second, the methodological choice matters: reinforcement learning not only reproduces the efficient frontier but can, in some regions, strictly dominate static QP solutions by delivering lower emissions and lower risk simultaneously. Third, results are consistent within each emissions database, but cross-provider heterogeneity remains substantial: absolute portfolio carbon levels can differ by as much as 80% depending on the chosen provider. This underscores the need for careful consideration of data quality and comparability in empirical studies and in practice.

Looking forward, the reinforcement learning framework offers particular promise. Unlike static QP, PPO naturally accommodates sequential decision-making and can incorporate additional ecological or operational constraints, such as sector caps, transaction costs, or alignment with net-zero pathways. Extending the approach to include broader sustainability objectives, such as water use, biodiversity, or social criteria, is a natural direction for future work. Another avenue is to move beyond the scalarised reward formulation and explore genuinely multi-objective reinforcement learning, which would allow investors to trace out full Pareto frontiers between financial and ecological objectives without imposing fixed trade-off weights.

Overall, this thesis contributes to the growing literature at the intersection of finance, machine learning, and climate economics by providing evidence that better data and better methods can materially shift portfolio outcomes. For investors and policymakers alike, the results highlight a key message: credible decarbonisation of financial portfolios is feasible at low or even negative cost in terms of financial risk, provided that one leverages accurate emissions forecasts and advanced optimisation techniques.

While these findings are encouraging, several avenues for future research remain open:

- **Closer company- and sector-level analysis:** simple divestment strategies can be misleading. More nuanced sectorial allocation, such as maintaining equal weights across sectors while favouring lower-emission firms within each sector, could lead to more effective outcomes.
- **ESG data limitations and Scope 3 double counting:** Scope 3 emissions remain highly uncertain and subject to methodological inconsistencies, reflecting the broader challenges of ESG data quality. Better disclosure standards and harmonisation are necessary to improve reliability.
- **Inclusion of impact investing and broader ecological constraints:** extending the framework beyond carbon footprints to incorporate biodiversity, water use, or transition alignment would bring portfolio optimisation closer to holistic sustainability.
- **Ambiguous effects of divestment:** empirical evidence suggests that excluding "brown" companies does not necessarily reduce real-world emissions, highlighting the need for complementary engagement strategies.

Ultimately, these directions underscore a deeper and unresolved question: can financial engineering alone deliver the ecological transition we need? Or, put differently, is it possible to decouple economic growth from environmental degradation, and is green finance truly an efficient way to foster a sustainable world?

Appendix A

Technical Proofs

A.1 Proof of Proposition 2.3.2: Closed-form variance–carbon trade-off

Carbon-constrained GMV: KKT conditions, solution, and value. Let $I \in \mathbb{R}^n$ denote asset carbon intensities and consider the minimum-variance problem with a carbon target k:

$$\min_{w \in \mathbb{R}^n} \ \tfrac{1}{2}\, w'\Sigma w \quad \text{s.t.} \quad \mathbf{1}'w = 1, \ \ I'w = k. \tag{A.1.1}$$

(We omit no-short-selling constraints to obtain closed forms; with $w \geq 0$ the active-set must be identified and the solution is piecewise.)

KKT conditions. The Lagrangian is

$$\mathcal{L}(w, \lambda, \theta) = \tfrac{1}{2}\, w'\Sigma w - \lambda(\mathbf{1}'w - 1) - \theta(I'w - k).$$

The KKT conditions (primal feasibility + stationarity; there are only equalities) are

$$\mathbf{1}'w = 1, \quad I'w = k, \qquad \nabla_w \mathcal{L} = \Sigma w - \lambda \mathbf{1} - \theta I = 0.$$

Hence

$$w^*(k) = \Sigma^{-1}(\lambda \mathbf{1} + \theta I). \tag{A.1.2}$$

Substituting $w^*(k) = \Sigma^{-1}(\lambda \mathbf{1} + \theta I)$ into the budget constraint $\mathbf{1}'w = 1$ yields

$$\lambda\, \mathbf{1}'\Sigma^{-1}\mathbf{1} \ + \ \theta\, \mathbf{1}'\Sigma^{-1} I \ = \ 1,$$

while substituting into the carbon constraint $I'w = k$ gives

$$\lambda\, I'\Sigma^{-1}\mathbf{1} \ + \ \theta\, I'\Sigma^{-1} I \ = \ k.$$

Defining

$$A := \mathbf{1}'\Sigma^{-1}\mathbf{1}, \qquad B := \mathbf{1}'\Sigma^{-1} I, \qquad C := I'\Sigma^{-1} I,$$

these two equations can be written compactly as the linear system

$$\begin{bmatrix} A & B \\ B & C \end{bmatrix} \begin{bmatrix} \lambda \\ \theta \end{bmatrix} = \begin{bmatrix} 1 \\ k \end{bmatrix}. \tag{A.1.3}$$

Let $\Delta := AC - B^2$. Since $\Sigma \succ 0$, we have $\Sigma^{-1} \succ 0$ and $\Delta > 0$ unless I is collinear with $\mathbf{1}$.

Solving (A.1.3) gives

$$\lambda(k) = \frac{C - Bk}{\Delta}, \qquad \theta(k) = \frac{Ak - B}{\Delta}, \tag{A.1.4}$$

and, by (A.1.2),

$$w^*(k) = \Sigma^{-1}\Big(\tfrac{C-Bk}{\Delta}\,\mathbf{1} + \tfrac{Ak-B}{\Delta}\,I\Big) = . \tag{A.1.5}$$

Closed-form value $\sigma^2(k)$. The minimal (unscaled) variance at target k is

$$\sigma^2(k) := w^*(k)'\Sigma\, w^*(k) = (\lambda\mathbf{1} + \theta c)'\Sigma^{-1}(\lambda\mathbf{1} + \theta c) = \lambda^2 A + 2\lambda\theta B + \theta^2 C.$$

A compact derivation uses the general quadratic form identity:

$$\sigma^2(k) = \min_w\{w'\Sigma w :\ F'w = g\} = g'\,(F'\Sigma^{-1}F)^{-1}g, \quad F := [\mathbf{1}\ I],\ \ g := [1\ k]',$$

so with $M := F'\Sigma^{-1}F = \begin{bmatrix} A & B \\ B & C \end{bmatrix}$ and $M^{-1} = \frac{1}{\Delta}\begin{bmatrix} C & -B \\ -B & A \end{bmatrix}$,

$$\sigma^2(k) = \frac{1}{\Delta}\left(C - 2Bk + Ak^2\right). \tag{A.1.6}$$

Let $I_0 := B/A$ denote the carbon intensity of the unconstrained GMV portfolio $w_{gmv} = \Sigma^{-1}\mathbf{1}/A$. Completing the square in (A.1.6) yields

$$\sigma^2(k) = \frac{1}{A} \ + \ \frac{A}{\Delta}\,(k - I_0)^2. \tag{A.1.7}$$

Proof: $Ak^2 - 2Bk + C = A[(k - I_0)^2 - I_0^2] + C = A(k - I_0)^2 + (C - B^2/A)$, and since $C - B^2/A = \Delta/A$, dividing by Δ gives (A.1.7). □

Inequality constraint $I'w \le k$. If the carbon constraint is $\le$, complementary slackness implies it is inactive when $k \ge I_0$; hence $w^* = w_{\mathrm{GMV}}$ and $\sigma^2 = 1/A$ (zero cost). For $k < I_0$ the constraint binds and the solution coincides with (A.1.5); the variance is given by (A.1.7), which increases only *quadratically* in the tightening $(I_0 - k)$.

A.2 Proof of Theorem 2.3.7: Bellman Optimality for the Action–Value Function

We consider a discounted MDP $(\mathcal{S}, \mathcal{A}, P, R, \beta)$ with discount $\beta \in (0, 1)$ and bounded rewards $|R(s, a)| \le R_{\max} < \infty$. For simplicity we write expectations under $P(\cdot \,|\, s, a)$ as $\mathbb{E}_{s'|s,a}[\cdot]$.

Lemma A.2.1 (Contraction and fixed points)**.** *Both T^π and T are β-contractions on the Banach space of bounded Q-functions with the sup norm $\|\cdot\|_\infty$. Hence each has a unique fixed point. Moreover, Q^π is the unique fixed point of T^π.*

Proof. For any bounded Q_1, Q_2 and any (s, a),

$$\begin{aligned} |(T^\pi Q_1 - T^\pi Q_2)(s, a)| &= \beta\left|\mathbb{E}_{s'|s,a}\big[\mathbb{E}_{a'\sim\pi}[Q_1(s', a') - Q_2(s', a')]\big]\right| \\ &\le \beta\,\|Q_1 - Q_2\|_\infty. \end{aligned}$$

Taking sup over (s, a) yields $\|T^\pi Q_1 - T^\pi Q_2\|_\infty \le \beta\|Q_1 - Q_2\|_\infty$. Similarly,

$$\begin{aligned} |(T^* Q_1 - T^* Q_2)(s, a)| &= \beta\left|\mathbb{E}_{s'|s,a}\big[\max_{a'} Q_1(s', a') - \max_{a'} Q_2(s', a')\big]\right| \\ &\le \beta\,\mathbb{E}_{s'|s,a}\Big[\sup_{a'} |Q_1(s', a') - Q_2(s', a')|\Big] \\ &\le \beta\,\|Q_1 - Q_2\|_\infty. \end{aligned}$$

Thus T^π and T are β-contractions. By the Banach fixed-point theorem, each has a unique fixed point. Standard dynamic programming arguments show Q^π satisfies $Q^\pi = T^\pi Q^\pi$ and, by uniqueness, is the fixed point of T^π. □

Lemma A.2.2 (Monotonicity)**.** *If $Q_1 \le Q_2$ pointwise, then $T^\pi Q_1 \le T^\pi Q_2$ and $T^* Q_1 \le T^* Q_2$.*

Proof. Immediate from linearity of expectation and the fact that $x \mapsto \max_{a'} x(a')$ is monotone. □

Proof. (Theorem) *(i) Q^* dominates all Q^π.* For any π, $Q^\pi = T^\pi Q^\pi \le T^* Q^\pi$ by Lemma A.2.2. Iterating and using monotonicity,

$$Q^\pi \;\le\; (T^*)^k Q^\pi \quad \text{for all } k \ge 1.$$

By Lemma A.2.1, $(T^*)^k Q^\pi \to \tilde{Q}$ where $\tilde{Q}$ is the unique fixed point of T^*. Hence $Q^\pi \le \tilde{Q}$ for all π, so $Q^* \le \tilde{Q}$.

(ii) $\tilde{Q}$ is achievable and thus equals Q^*. Let $\pi_{\tilde{Q}}$ be any greedy policy w.r.t. $\tilde{Q}$: $\pi_{\tilde{Q}}(s) \in \arg\max_a \tilde{Q}(s,a)$. Then

$$T^{\pi_{\tilde{Q}}} \tilde{Q} \;=\; R + \beta\, \mathbb{E}_{s'|s,a}\Big[\mathbb{E}_{a' \sim \pi_{\tilde{Q}}(\cdot|s')} \tilde{Q}(s',a')\Big] \;=\; R + \beta\, \mathbb{E}_{s'|s,a}\Big[\max_{a'} \tilde{Q}(s',a')\Big] \;=\; T^* \tilde{Q} \;=\; \tilde{Q}.$$

So $\tilde{Q}$ is a fixed point of $T^{\pi_{\tilde{Q}}}$. By Lemma A.2.1, this fixed point equals $Q^{\pi_{\tilde{Q}}}$. Therefore $\tilde{Q} = Q^{\pi_{\tilde{Q}}} \le Q^*$, since Q^* is the supremum over all policies. Combined with (i), we get $\tilde{Q} = Q^*$.

Finally, since $Q^* = Q^{\pi_{\tilde{Q}}}$, the greedy policy $\pi_{\tilde{Q}}$ attains the optimal action–value function and is thus optimal; denote it by π^*. □

Remark (continuous actions). If $\mathcal{A}$ is continuous, replace max by sup in the definitions. Under mild regularity (e.g., compact $\mathcal{A}$ and continuity of R and Q in a), the supremum is attained.

A.3 Proof of Theorem 2.3.10: Policy Gradient Theorem

We consider a discounted MDP $(\mathcal{S}, \mathcal{A}, P, R, \beta)$ with $\beta \in (0,1)$ and bounded rewards $|R(s,a)| \le R_{\max} < \infty$. Let $\pi_\theta(a|s)$ be a differentiable stochastic policy with parameters θ.

Assumptions. We assume:

1. Rewards are bounded, $|R(s,a)| \le R_{\max} < \infty$.
2. For each θ, $\pi_\theta(\cdot|s)$ is a valid probability distribution, differentiable in θ, with well-defined log-gradient $\nabla_\theta \log \pi_\theta(a|s)$.
3. Support condition: if $\pi_\theta(a|s) = 0$, then $\nabla_\theta \pi_\theta(a|s) = 0$.
4. The transition kernel P and initial distribution ρ do not depend on θ.
5. There exists an integrable bound so that expectations and derivatives may be interchanged (e.g. by dominated convergence).

Proof. Let a trajectory be $\tau = (s_0, a_0, s_1, a_1, \ldots)$ with density

$$p_\theta(\tau) \;=\; \rho(s_0) \prod_{t=0}^{\infty} \pi_\theta(a_t|s_t)\, P(s_{t+1}|s_t, a_t).$$

Write the (discounted) return as $G(\tau) = \sum_{t=0}^{\infty} \beta^t R(s_t, a_t)$. Then

$$J(\theta) \;=\; \int p_\theta(\tau)\, G(\tau)\, d\tau.$$

Differentiate under the integral (dominated convergence holds by bounded rewards and differentiability of π_θ):

$$\nabla_\theta J(\theta) = \int \nabla_\theta p_\theta(\tau)\, G(\tau)\, d\tau = \int p_\theta(\tau)\, \nabla_\theta \log p_\theta(\tau)\, G(\tau)\, d\tau.$$

Since P and ρ do not depend on θ,

$$\log p_\theta(\tau) = \log \rho(s_0) + \sum_{t=0}^{\infty} \big(\log \pi_\theta(a_t|s_t) + \log P(s_{t+1}|s_t, a_t)\big) \;\Rightarrow\; \nabla_\theta \log p_\theta(\tau) = \sum_{t=0}^{\infty} \nabla_\theta \log \pi_\theta(a_t|s_t).$$

Hence

$$\nabla_\theta J(\theta) = \mathbb{E}_{\pi_\theta}\left[\sum_{t=0}^{\infty} \nabla_\theta \log \pi_\theta(a_t|s_t)\, G(\tau)\right].$$

Now condition on (s_t, a_t) and decompose $G(\tau) = \sum_{k=0}^{t-1} \beta^k R(s_k, a_k) + \beta^t \sum_{j=0}^{\infty} \beta^j R(s_{t+j}, a_{t+j})$. The first (pre-$t$) part is independent of a_t given s_t and has zero contribution in expectation because $\mathbb{E}[\nabla_\theta \log \pi_\theta(a_t|s_t) \,|\, s_t] = \nabla_\theta \sum_a \pi_\theta(a|s_t) = 0$. Therefore,

$$\mathbb{E}\Big[\nabla_\theta \log \pi_\theta(a_t|s_t)\, G(\tau) \,\Big|\, s_t, a_t\Big] = \mathbb{E}\left[\nabla_\theta \log \pi_\theta(a_t|s_t)\, \beta^t \sum_{j=0}^{\infty} \beta^j R(s_{t+j}, a_{t+j}) \,\Big|\, s_t, a_t\right].$$

Taking total expectation and using the tower property,

$$\nabla_\theta J(\theta) = \mathbb{E}_{\pi_\theta}\left[\sum_{t=0}^{\infty} \beta^t\, \nabla_\theta \log \pi_\theta(a_t|s_t)\, Q^{\pi_\theta}(s_t, a_t)\right],$$

which is the second expression in the theorem. Grouping by states and using $d_\beta^{\pi_\theta}$ yields the first expression with the factor $(1-\beta)^{-1}$. □

Lemma A.3.1 (Baseline invariance)**.** *For any function $b : \mathcal{S} \to \mathbb{R}$,*

$$\mathbb{E}_{\pi_\theta}\left[\sum_{t=0}^{\infty} \beta^t\, \nabla_\theta \log \pi_\theta(a_t|s_t)\, b(s_t)\right] \;=\; 0.$$

Proof. Condition on s_t: $\mathbb{E}\big[\nabla_\theta \log \pi_\theta(a_t|s_t)\, b(s_t) \,\big|\, s_t\big] = b(s_t) \sum_a \pi_\theta(a|s_t) \nabla_\theta \log \pi_\theta(a|s_t) = b(s_t) \sum_a \nabla_\theta \pi_\theta(a|s_t) = \nabla_\theta \sum_a \pi_\theta(a|s_t) = \nabla_\theta 1 = 0$. For continuous $\mathcal{A}$ replace the sum with an integral; the argument is identical. □

Corollary A.3.2 (Advantage form)**.** *Let $A^{\pi_\theta}(s,a) := Q^{\pi_\theta}(s,a) - b(s)$ with any baseline $b(s)$. Using Lemma A.3.1 in Theorem 2.3.10,*

$$\nabla_\theta J(\theta) = \mathbb{E}_{\pi_\theta}\left[\sum_{t=0}^{\infty} \beta^t\, \nabla_\theta \log \pi_\theta(a_t|s_t)\, A^{\pi_\theta}(s_t, a_t)\right].$$

Choosing $b(s) = V^{\pi_\theta}(s)$ yields $A^{\pi_\theta}(s,a) = Q^{\pi_\theta}(s,a) - V^{\pi_\theta}(s)$, which reduces the variance of the gradient estimator without introducing bias.

A.4 PPO hyperparameters

Hyperparameter	Value
algo	PPO
policy	ActorCriticPolicy
net_arch	{'pi': [128, 128], 'vf': [128, 128]}
learning_rate	0.0003
n_steps	2048
batch_size	64
n_epochs	10
gamma	0.99
gae_lambda	0.95
clip_range	callable
ent_coef	0.0
vf_coef	0.5
max_grad_norm	0.5
target_kl	–
seed	–
cov_window_days	120
EWMA decay	0.94
training years	(2016, 2017, 2018, 2019, 2020)
min_obs	30
total_timesteps	200000

Table A.1: PPO hyperparameters and environment settings (identical across all λ).

Bibliography

[1] *Predicting corporate carbon footprints for climate finance risk analyses: A machine learning approach*, Energy Economics, 95 (2021), p. 105129.

[2] S. Almahdi and S. Y. Yang, *An adaptive portfolio trading system: A risk-return portfolio optimization using recurrent reinforcement learning with expected maximum drawdown*, Expert Systems with Applications, 87 (2017), pp. 267–279.

[3] F. T. Ayasrah, N. S. Alsharafa, S. Sivaprakash, B. Srinivasa Rao, S. Sengan, and N. Kumaran, *Strategizing low carbon urban planning through environmental impact assessment by artificial intelligence driven carbon footprint forecasting*, Journal of Machine and Computing, 4 (2024), pp. 1140–1151.

[4] A. Bajic and R. Kiesel, *Climate transition matrix: Assessing carbon performance of companies*, Working paper, (2024).

[5] I. Barahhou, M. Ben Slimane, T. Roncalli, and N. Oulid Azouz, *Net zero investment portfolios – part 1: The comprehensive integrated approach*, working paper, Amundi Institute, December 2022. Posted: 13 Dec 2022; Date written: 13 Oct 2022.

[6] M. Ben Slimane, D. Lucius, T. Roncalli, and J. Xu, *Net zero investment portfolios – part 2: The core-satellite approach*, working paper, Amundi Institute, October 2023. Posted: 25 Oct 2023; Last revised: 3 Nov 2023; Date written: 24 Oct 2023.

[7] F. Berg, J. F. Kölbel, and R. Rigobon, *Aggregate confusion: The divergence of esg ratings**, Review of Finance, 26 (2022), pp. 1315–1344.

[8] P. Bolton, M. Kacperczyk, and F. Samama, *Net-zero carbon portfolio alignment*, (2022). This draft: January 13, 2022.

[9] J. Bun, M. Potters, and J.-P. Bouchaud, *Cleaning large correlation matrices: tools from random matrix theory*, Physics Reports, 666 (2017), pp. 1–109.

[10] V. K. Chopra and W. T. Ziemba, *The effect of errors in means, variances, and covariances on optimal portfolio choice*, The Journal of Portfolio Management, 19 (1993), pp. 6–11.

[11] H. Choudhary, A. Orra, K. Sahoo, and M. Thakur, *Risk-adjusted deep reinforcement learning for portfolio optimization: A multi-reward approach*, International Journal of Computational Intelligence Systems, 18 (2025), p. 126.

[12] P. M. Clarkson, Y. Li, G. D. Richardson, and F. P. Vasvari, *Revisiting the relation between environmental performance and environmental disclosure: An empirical analysis*, Accounting, Organizations and Society, 33 (2008), pp. 303–327.

[13] Z. Dong, J. Shi, and S. Pan, *The interactions of carbon emission driving forces: Analysis based on interpretable machine learning*, Urban Climate, 59 (2025), p. 102323.

[14] M. Fahmaoui, T. Barreau, and P. Tankov, *Estimating corporate greenhouse gas emissions*, working paper, Institut Louis Bachlier, CREST, 2024.

[15] B. Goldhammer, C. Busse, and T. Busch, *Estimating corporate carbon footprints with externally available data*, Journal of Industrial Ecology, 21 (2017), pp. 1165–1179.

[16] P. A. Griffin, D. H. Lont, and E. Y. Sun, *The relevance to investors of greenhouse gas emission disclosures*, Contemporary Accounting Research, 34 (2017), pp. 1265–1297.

[17] L. Grinsztajn, E. Oyallon, and G. Varoquaux, *Why do tree-based models still outperform deep learning on tabular data?*, 2022.

[18] Y. Hanine, Y. Lamrani Alaoui, M. Tkiouat, and Y. Lahrichi, *Socially responsible portfolio selection: An interactive intuitionistic fuzzy approach*, Mathematics, 9 (2021).

[19] J.P. Morgan/Reuters, *Riskmetrics–technical document*, tech. rep., J.P. Morgan, 1996.

[20] V. R. Konda and J. N. Tsitsiklis, *Actor-critic algorithms*, in Advances in Neural Information Processing Systems (NIPS 12), 2000, pp. 1008–1014.

[21] K. Kosgahakumbura, P. Kumara, and S. Lakmali, *A novel stacked machine learning framework for enhanced prediction of corporate co2 emissions in multisectoral analysis*, in 2024 8th SLAAI International Conference on Artificial Intelligence (SLAAI-ICAI), 2024, pp. 1–6.

[22] L. Laloux, P. Cizeau, J.-P. Bouchaud, and M. Potters, *Noise dressing of financial correlation matrices*, Phys. Rev. Lett., 83 (1999), pp. 1467–1470.

[23] O. Ledoit and M. Wolf, *A well-conditioned estimator for large-dimensional covariance matrices*, Journal of Multivariate Analysis, 88 (2004), pp. 365–411.

[24] C. Maree and C. W. Omlin, *Balancing profit, risk, and sustainability for portfolio management*, (2022), pp. 1–8.

[25] H. Markowitz, *Portfolio selection*, The Journal of Finance, 7 (1952), pp. 77–91.

[26] I. P. Nguemkam Tebou, N. Tsopze, and D. Tchuente, *Explaining meta-learner's predictions: Case of corporate co2 emissions*, in Research in Computer Science, P. Melatagia Yonta, K. Barkaoui, R. Ndoundam, and O.-B. Yenke, eds., Cham, 2024, Springer Nature Switzerland, pp. 92–105.

[27] Q. Nguyen, I. Diaz-Rainey, A. Kitto, B. I. McNeil, N. A. Pittman, and R. Zhang, *Scope 3 emissions: Data quality and machine learning prediction accuracy*, PLOS Climate, (2023).

[28] L. H. Pedersen, S. Fitzgibbons, and L. Pomorski, *Responsible investing: The esg-efficient frontier*, Journal of Financial Economics, 142 (2021), pp. 572–597.

[29] M. E. Porter and C. van der Linde, *Green and competitive: Ending the stalemate*, Harvard Business Review, 73 (1995). Harvard Business School Publishing.

[30] L. Prokhorenkova, G. Gusev, A. Vorobev, A. V. Dorogush, and A. Gulin, *Catboost: unbiased boosting with categorical features*, Advances in neural information processing systems, 31 (2018).

[31] M. L. Puterman, *Markov Decision Processes: Discrete Stochastic Dynamic Programming*, Wiley, New York, 1994.

[32] J. Ranganathan and P. Bhatia, *The Greenhouse Gas Protocol: a Corporate Accounting and Reporting Standard, Revised Edition*, 03 2004.

[33] B. Scherer, *Portfolio Construction and Risk Budgeting*, Springer, London, 2nd ed., 2002.

[34] J. Schulman, F. Wolski, P. Dhariwal, A. Radford, and O. Klimov, *Proximal policy optimization algorithms*, arXiv preprint arXiv:1707.06347, (2017).

[35] R. S. Sutton and A. G. Barto, *Reinforcement Learning: An Introduction*, MIT Press, Cambridge, MA, 2nd ed., 2018.

[36] R. S. Sutton, D. McAllester, S. Singh, and Y. Mansour, *Policy gradient methods for reinforcement learning with function approximation*, in Proceedings of the 13th International Conference on Neural Information Processing Systems, NIPS'99, Cambridge, MA, USA, 1999, MIT Press, p. 1057–1063.

[37] H. A. Verfaillie and R. Bidwell, *Measuring eco-efficiency: A guide to reporting company performance*, 2000. Produced by Monsanto Company and Environmental Resources Management plc.

[38] N. N. Vo, X. He, S. Liu, and G. Xu, *Deep learning for decision making and the optimization of socially responsible investments and portfolio*, Decision Support Systems, 124 (2019), p. 113097.

[39] C. J. C. H. Watkins and P. Dayan, *Q-learning*, Machine Learning, 8 (1992), pp. 279–292.

[40] M. Xia and H. Cai, H.}, *The driving factors of corporate carbon emissions: an application of the lasso model with survey data*, Environmental Science and Pollution Research, 30 (2023), pp. 56484–56512.

[41] X. Xia, D. Zhu, J. Sha, R. Ma, and W. Kang, *Research on industrial carbon emission prediction method based on cnn–lstm under dual carbon goals*, International Journal of Low-Carbon Technologies, 20 (2025), pp. 580–589.

[42] H. Yang, X.-Y. Liu, S. Zhong, and A. Walid, *Deep reinforcement learning for automated stock trading: an ensemble strategy*, (2021).