Trading Frictions in Cap-and-Trade Markets*

Nicola Borri, Yukun Liu, Aleh Tsyvinski, and Xi Wu.

April 2026

Abstract

We develop a model of cap-and-trade markets with standard finance frictions: slow participation, limited intermediation, and heterogeneous information. The model predicts infrequent trading, surrender-month price pressure, and return predictability from informative operator flow. We test these predictions in the EU ETS using the universe of registry transactions and compliance data over the full available history, 2005–2021. The evidence supports the model. First, about 40% of operators do not trade annually, falling to 20% by early Phase IV. Second, operators concentrate purchases in surrender months, when returns are predictably high. Third, some operators speculate using private information.

*Nicola Borri, LUISS University, email: nborri@luiss.it. Yukun Liu, University of Rochester, email: yliu229@ur.rochester.edu. Aleh Tsyvinski, Corresponding author, Department of Economics at Yale University, email: a.tsyvinski@yale.edu. Xi Wu, University of California Berkeley, email: xiwu@berkeley.edu

1 Introduction

Carbon emissions are a classic example of externalities, where economic agents do not fully internalize their environmental impact, leading to overproduction of emissions that drive global warming (Matthews et al., 2009; Allen et al., 2009). Reducing carbon emissions is therefore a central component of virtually all scenarios aiming to limit the increase in global mean temperature (Rogelj et al., 2018). A cap-and-trade system, where agents can trade allowances to emit carbon, offers a market-based mechanism for firms to internalize the environmental impact.

A recent meta-analysis of 70 carbon-pricing schemes (Döbbeling-Hildebrandt et al., 2024) concludes that carbon pricing substantially reduces emissions. Yet for a market-based mechanism, emissions reductions are not the only criterion. It also matters whether allowances are reallocated in a cost-effective way across firms and over time, and whether prices incorporate and aggregate relevant scarcity information. Recent evidence on the European Union Emissions Trading System (EU ETS) points to different economic margins. Känzig (2023) finds that restrictive carbon price shocks reduce emissions but also lower aggregate economic activity, while Colmer et al. (2025) finds that the EU ETS reduced regulated firms' emissions with little effect on output or employment. This contrast highlights the importance of understanding how carbon markets allocate compliance costs in practice and how cap-and-trade systems operate when allowances are traded as financial assets. We therefore study cap-and-trade markets with realistic finance frictions—slow participation, limited intermediation, and heterogeneous information—using tools from finance and asset pricing to analyze trading, intermediation, and price formation, dimensions that have received relatively limited attention in this setting.

To organize the analysis, consider a cap-and-trade market with bankable allowances and no trading frictions. Firms then trade until marginal abatement costs are equalized across firms and over time. This benchmark delivers three sharp implications: non-trading should arise only when a firm's initial allowance holdings already coincide with its cost-minimizing compliance and banking needs; predictable compliance-related demand should be reflected in prices before the surrender deadline; and operator order flow based only on public compliance needs should not forecast future returns (Dales, 1968; Montgomery, 1972; Rubin, 1996).

These benchmark implications are useful precisely because several standard frictions may be important in allowance markets. Regulated firms may face internal attention, treasury, or approval costs that slow

participation, analogous to slow-moving capital and search frictions in financial markets (Duffie, 2010; Duffie et al., 2007); intermediaries may have limited capacity to absorb predictable order imbalances, as in models of constrained arbitrage and demand pressure (Gromb and Vayanos, 2002; Gârleanu et al., 2009); and some operators may have information about future scarcity that is not yet fully reflected in prices, as in canonical models of heterogeneous information and informed trading (Glosten and Milgrom, 1985; Admati and Pfleiderer, 1988). We therefore introduce three classic finance frictions in a cap-and-trade market—slow participation, limited intermediation, and heterogeneous information—and ask what observable predictions they generate.

We first present a two-period cap-and-trade model with frictions to isolate the basic mechanism, and then extend the framework to continuous time—with Poisson market access, affine price impact, and informational advantages—to characterize the dynamic implications. The model generates three empirical predictions: persistent non-trading, surrender-month demand pressure, and informative operator flow.

We test these predictions using granular transaction and compliance data from the European Union Transaction Log (EUTL). Our sample includes the universe of transactions and compliance information currently available from the EUTL, spanning February 2005 to September 2021. It covers the first three phases of the EU ETS, as well as part of the ongoing fourth phase, which is scheduled to conclude in 2030. There are three main groups of participants in the market: regulated firms, intermediaries, and regulators. They are largely represented in the EU ETS market by operator holding accounts, person holding accounts, and administrative accounts, respectively. Our analyses focus on regulated firms, or operators, and study their transactions with other regulated firms and intermediaries.

The data reveal three main facts, corresponding to the three predictions generated by the model with frictions. First, a large fraction of operators—that is, regulated firms subject to compliance obligations—do not trade in a given year. Approximately 40% of operators do not trade in the overall sample, and this share remains about 20% by early Phase IV. Related patterns of low participation were documented in early policy-oriented work focusing on the initial years of Phase I of the EU ETS (Betz and Schmidt, 2016; Jaraitė et al., 2010), a period marked by design and implementation challenges that subsequent reforms were expected to address. We show that non-trading is persistent throughout the history of the EU ETS. Yet a cap-and-trade system can reallocate allowances smoothly across firms and over time only if firms participate in the market and actively trade.

One possible explanation for the large fraction of non-trading firms is that they receive an allocation of

free allowances closely aligned with their surrender obligations under the EU ETS. However, the data do not support this interpretation. Many of these firms exhibit imbalances in their allowance holdings even in the years they do not trade. An alternative explanation—put forward by, among others, Betz and Schmidt (2016) and Cantillon and Slechten (2023)—is that non-trading operators bank surplus allowances, possibly in anticipation of higher future prices. Yet, the data do not support this hypothesis either. We find that a lack of trading in a given year significantly and positively predicts continued non-trading in subsequent years. In the continuous-time model, such persistent non-trading corresponds to operators with low market-access intensity, for whom expected compliance gaps are large and adjustment is slow.

Second, we demonstrate that many regulated firms delay purchases until the surrender month, when compliance demand is concentrated and allowance prices are systematically higher, leading to significant additional private costs. The EU ETS market operates on a fixed verifying and surrendering schedule each year. By the end of April, firms must surrender a sufficient number of emission allowances to the administrator account to match their verified emissions. This fixed schedule leads to a predictable pattern in emission allowance trading, with April consistently showing the largest net purchases of allowances by regulated firms throughout the sample period. The magnitude of these purchases is striking, with net purchases in April being an order of magnitude larger than in typical non-April months. This concentration of compliance demand is associated with systematically high April returns, averaging more than 10%. In addition to a large fraction of non-trading firms, many firms that do trade concentrate their activity in the surrender month, when compliance is immediate and the price of emission allowances is predictably high. A back-of-the-envelope calculation puts the implied private cost from this surrender pattern at about €5 billion, representing additional compliance costs relative to purchasing allowances earlier at lower prices. This number should be interpreted as a private-cost measure for delayed buyers rather than as a direct measure of aggregate resource costs, because part of the higher April price is paid to counterparties. Importantly, this delayed and frictional trading pattern persists into the most recent phase of the EU ETS, indicating limited improvement over time.

Third, we find that about 10% of regulated firms trade an unusually large amount of emission allowances each year, and that their trading patterns are consistent with informational advantages and economically meaningful return predictability. Using a conservative definition, these firms have trading volumes more than four times the amount they surrender in a given year. We document trading patterns consistent with these firms having informational advantages about future scarcity and prices. In particular, they buy emission

allowances ahead of price increases and sell them ahead of price declines. Back-of-the-envelope calculations that compare their net purchases and sales to subsequent price movements imply economically meaningful ex post trading performance. To capture the net purchases of regulated firms in a month, we construct a variable called *Operator Flow Imbalance*, defined as the net flow from person holding accounts to operator holding accounts divided by total bilateral transaction volume between the two account types in that month. We establish that Operator Flow Imbalance is a strong predictor of future cumulative emission allowance returns. Using Operator Flow Imbalance to predict cumulative allowance returns from one month to twelve months ahead, we find that the coefficient estimates are all positive, becoming significant at the 5% level at the two-month horizon. The results are consistent with the notion that these operators have informational advantages and tend to accumulate emission allowances ahead of future price increases. Consistent with the model, the predictive content is concentrated among frequent traders—operators with higher effective market-access intensity—whose order flow reflects informational advantages not yet fully impounded in prices. A back-of-the-envelope calculation puts the implied ex post trading performance at roughly €8 billion over the sample period.

The paper contributes to two strands of the literature. The first studies how trading frictions shape asset-market participation, intermediation, and price formation. We build on models of slow-moving capital and search frictions (Duffie, 2010; Duffie et al., 2007), constrained intermediation and demand pressure (Gromb and Vayanos, 2002; Acharya et al., 2013; Gârleanu et al., 2009; Hendershott and Menkveld, 2014), and heterogeneous information and informative order flow (Subrahmanyam, 1991). Our contribution is to bring these core finance frictions into cap-and-trade markets, where allowances are bankable assets and predictable compliance demand must be absorbed through the secondary market.

The second strand studies cap-and-trade and environmental markets. Classic work establishes the frictionless benchmark in which trading equalizes marginal abatement costs across firms and over time (Dales, 1968; Montgomery, 1972; Rubin, 1996). Recent work by Biais et al. (2026) studies imperfect hedging and financial constraints under transition risk, and Biais and Landier (2026) study emission caps, green investment, and government commitment. We study a core feature of cap-and-trade markets—the secondary market through which permits are traded and compliance demand is reallocated—and show how slow participation, limited intermediation, and heterogeneous information shape trading timing, seasonal demand pressure, and return predictability.

2 A Two-Period Model of Trading Frictions in a Cap-and-Trade Market

This section develops a simple two-period model of a cap-and-trade market with frictions.[1] The framework is deliberately stylized: its role is to provide intuition by isolating the basic mechanism that we then embed in a continuous-time dynamic model in Section 3. The building blocks are a permit-market model in the spirit of Harstad and Eskeland (2010), combined with ingredients motivated by finance models of constrained intermediation, order imbalances, and informed trading (Gromb and Vayanos, 2002; Acharya et al., 2013; Gârleanu et al., 2009; Hendershott and Menkveld, 2014; Subrahmanyam, 1991). The model introduces the smallest set of frictions needed to rationalize the empirical facts that we test below. Operators face the EU ETS's fixed annual verification and surrender schedule, and intermediaries absorb net order flow subject to a quadratic immediacy-provision cost. Full derivations are provided in the Supplemental Appendix.

Consider a competitive market for bankable emission allowances in two periods, $t = 0, 1$. Firm i is a permit-compliant operator which chooses emissions x_{it}, net purchases of allowances m_{it}, and the quantity of allowances banked from period 0 to period 1, denoted by b_{i1}. The firm's flow profit in each period is

$$\pi_{it} = \kappa_i - \frac{1}{2}(\theta_{it} - x_{it})^2 - p_t m_{it}, \tag{1}$$

where a firm's type is θ_{it}, which denotes business-as-usual emissions. Allowance holdings satisfy

$$b_{i1} = b_{i0} + q_{i0} + m_{i0} - x_{i0}, \qquad 0 = b_{i1} + q_{i1} + m_{i1} - x_{i1}, \qquad b_{i1} \geq 0, \tag{2}$$

where q_{it} denotes the quantity of allowances allocated to the firm. Let $M_t = \sum_i m_{it}$ denote aggregate operator net demand. Throughout, $M_t > 0$ denotes net operator buying pressure, equivalently net allowance flow from intermediaries or person holding accounts to operator holding accounts; $M_t < 0$ denotes net operator selling. A representative competitive intermediary sector supplies immediacy by absorbing this order flow, with a quadratic immediacy-provision (or balance-sheet) cost $\frac{\phi}{2}M_t^2$, where $\phi > 0$ indexes limited intermediation capacity. Because M_t is a flow rate rather than an inventory stock, $\frac{\phi}{2}M_t^2$ should be read as the marginal cost of absorbing contemporaneous order flow, in the spirit of Hendershott and Menkveld (2014), rather than as

[1] Goulder (2013); Schmalensee and Stavins (2013); Cantillon and Slechten (2023) provide excellent reviews of the classic cap-and-trade models.

a literal inventory-holding cost.

In the frictionless benchmark, $\phi = 0$, participation costs are zero, and no trader possesses private information[2]. Let λ_{i0} and λ_{i1} denote the shadow values of one additional allowance in periods 0 and 1, and let p_t^* denote the corresponding frictionless price. For an interior solution, the firm's optimality conditions imply

$$\theta_{i0} - x_{i0} = \lambda_{i0}, \qquad \theta_{i1} - x_{i1} = \lambda_{i1}, \tag{3}$$

$$p_0^* = \lambda_{i0}, \qquad p_1^* = \lambda_{i1}, \tag{4}$$

$$p_0^* = \delta E_0[p_1^*]. \tag{5}$$

Equations (3)–(5) imply that, absent frictions, marginal abatement costs are equalized across firms and frictionless allowance prices satisfy the standard no-arbitrage condition for a bankable asset. The banking constraint $b_{i1} \geq 0$ is asymmetric: firms can store surplus allowances across compliance years but cannot run a negative bank, so borrowing future allocations to cover current obligations is ruled out. The first-order conditions (3)–(5) hold for an interior banking solution; when the no-borrowing corner $b_{i1} = 0$ binds, the Euler equation (5) holds with inequality and the firm has no spare allowances to draw down to meet a current shortfall. Empirically, this is the relevant corner: the firms who delay purchases until April are precisely those with $d_i > 0$ for whom $b_{i1} = 0$ binds, so that the asymmetric banking constraint reinforces the surrender-month demand pressure formalized below. This benchmark yields three immediate implications. First, non-trading is a knife-edge outcome: a firm does not trade only if its initial allowance position already coincides with its cost-minimizing compliance and banking plan under the frictionless benchmark. Second, predictable seasonal compliance demand, such as the April surrender deadline, should already be reflected in prices before the deadline arrives. Third, order flow based only on public compliance needs should not predict future returns. We interpret the empirical patterns in the paper as departures from this benchmark.

When intermediation is limited, $\phi > 0$, actual prices satisfy

$$p_t = p_t^* + \phi M_t. \tag{6}$$

Equation (6) says that prices can depart from frictionless benchmark prices when intermediaries must absorb

[2]Throughout the paper, the frictionless benchmark is used to organize the analysis of secondary-market functioning: cost-effective reallocation of allowances, the timing with which predictable compliance demand is incorporated into prices, and the informational content of order flow. We do not use this benchmark to claim that observed market outcomes are Pareto-inefficient once trading frictions and institutional constraints are taken as given.

net order imbalances and charge a premium for immediacy provision. For example, if many operators delay purchases until the surrender month, intermediaries must absorb a large wave of compliance demand, which can push April prices above their frictionless benchmark level.

The frictionless benchmark above describes the cost-minimizing allocation with bankable allowances. To study the timing of compliance trades within a yearly compliance cycle, we now use a simplified within-cycle representation with an early trading date e and a surrender-month date A. This within-cycle block is not meant to replace the banking benchmark; rather, it isolates the participation and timing mechanism that generates delayed trading and surrender-month price pressure.

2.1 Participation Costs and Surrender-Month Trading

To interpret the large fraction of firms that do not trade and the concentration of purchases in April, suppose that firms face a fixed participation cost $K_i > 0$ whenever they trade actively before the surrender deadline. This cost can be interpreted broadly as the internal cost of attention, treasury management, internal approval, compliance monitoring, or establishing and maintaining a trading relationship.

Consider one compliance cycle with two trading dates: an early date e and the surrender month A. Let $d_i \equiv \max\{\text{expected allowance shortfall}_i, 0\}$ denote firm i's expected positive shortfall at the surrender date. To keep the mechanism parsimonious, we abstract from verification uncertainty and treat firms' compliance needs as known in expectation before the surrender date. A firm that becomes active at date e can purchase allowances early at price p_e rather than wait and buy in April at price p_A. Ignoring within-year discounting, the firm trades early if and only if

$$d_i \left(E_0[p_A] - p_e\right) \geq K_i. \tag{7}$$

Condition (7) governs the timing of purchases among firms with positive expected shortfalls: firms for which the condition fails delay purchases until the surrender month, while firms for which the condition holds buy early. Equivalently, one can write the threshold as $d_i(E_0[p_A] - p_e) \geq K_i^e - K_i^A$, where K_i^e is the cost of early market participation and K_i^A is the lower incremental cost of deadline-month compliance trading; we normalize $K_i^A = 0$ so that $K_i \equiv K_i^e$ is the incremental cost of early relative to deadline-month compliance trading. Complete non-trading arises in a different regime: it applies to firms whose expected shortfall is zero or can be covered with initial or banked allowances, or to firms whose effective market-access costs

prevent discretionary rebalancing despite nonzero imbalances.[3] Because intermediaries absorb net order flow subject to immediacy-provision costs, prices at each within-cycle date $\tau \in \{e, A\}$ satisfy

$$p_\tau = p^*_\tau + \phi M_\tau. \tag{8}$$

When many firms delay participation, compliance demand is concentrated at the surrender date, so M_A becomes large and the April price can rise above the early price even though the compliance wave is predictable. This mechanism is closely related to models in which limited arbitrage and intermediary balance-sheet constraints affect asset prices (Gârleanu et al., 2009; Hendershott and Menkveld, 2014; Gabaix and Maggiori, 2015). It generates two empirical predictions: a large share of firms that do not trade at all and substantial compliance buying in April. The latter can be mapped into avoidable private compliance costs, which we estimate below.

2.2 Private Information and Speculative Trading

Participation costs alone cannot explain why operator order flow predicts future allowance returns. To capture this fact, suppose that a subset of active firms receives a private signal about future permit scarcity or future permit values. Let $\mathcal{I}_{it}$ denote firm i's information set at time t, which contains both public information $\mathcal{F}_t$ and any private information available to the firm. Define μ_{it} as the private component of expected benchmark returns,

$$\mu_{it} = E_t\left[p^*_{t+1} - p^*_t \mid \mathcal{I}_{it}\right] - E_t\left[p^*_{t+1} - p^*_t \mid \mathcal{F}_t\right], \tag{9}$$

where p^*_t is the frictionless benchmark price. This convention isolates information not yet reflected in the public state and avoids treating predictable public April price pressure as private information. Taking prices as given, an active firm can, in addition to its compliance-motivated position, take a speculative inventory position n_{it} and solves

$$\max_{n_{it}} \left\{\mu_{it} n_{it} - \frac{\gamma_i}{2} n^2_{it}\right\}, \tag{10}$$

[3] Condition (7) should be interpreted as a reduced-form participation condition rather than as the solution to a fully specified equilibrium problem. The expected April premium is taken as given by the individual firm, even though in equilibrium it is shaped by delayed compliance demand and intermediary immediacy-provision costs.

where $\gamma_i > 0$ captures inventory, balance-sheet, or risk management costs. The optimal speculative position is therefore

$$n_{it} = \frac{\mu_{it}}{\gamma_i}. \tag{11}$$

Equation (11) implies that aggregate operator order flow can predict future returns when a subset of firms trades on private information and intermediaries absorb order imbalances only gradually. The model therefore suggests that return predictability should be strongest for frequent traders, since these are the firms most likely to have low participation costs and to trade on private information. This is the pattern we test below in the return-predictability tests, where forecasting power is concentrated among frequent traders and the implied ex post trading performance is economically large.

2.3 Private Compliance Costs

The two-period model clarifies the private cost borne by regulated firms that delay compliance purchases. If firm i expects to be short by d_i allowances at the surrender date, its expected private cost from delaying purchases until April is

$$E_0[\text{Cost}_i^{\text{delay}}] = d_i(E_0[p_A] - p_e).$$

Aggregating across delayed buyers yields the back-of-the-envelope private compliance-cost measure used below. This object is not a direct measure of aggregate resource costs, because allowance payments are partly transfers across market participants. Instead, it should be interpreted as the private cost borne by delayed buyers and as an indication of the size of the participation and adjustment frictions needed to rationalize delayed trading.

3 A Continuous-Time Model with Slow Participation and Informative Order Flow

The two-period model is useful for intuition, but it compresses the recurring compliance cycle into a single early-versus-late trading decision. In practice, regulated firms manage allowance positions over repeated annual surrender cycles, access the market gradually rather than continuously, and trade in a market where prices adjust as intermediaries absorb order imbalances. This section extends the mechanism of the

two-period model to a continuous-time framework. The continuous-time model is a reduced-form dynamic extension of the two-period mechanism: its purpose is to discipline the timing and return-predictability implications, not to provide a fully structural equilibrium model of all EU ETS trading. It preserves the same three economic forces as the two-period model—participation frictions, limited intermediation, and informational advantages among active operators—and yields closed-form expressions in a leading seasonal case together with transparent comparative statics more generally.[4]

The mapping from the two-period model to continuous time is direct. The high participation cost of the two-period model becomes a low Poisson market-access intensity: in simple terms, firms do not rebalance continuously, but instead receive trading opportunities only intermittently, so a lower arrival rate means slower effective access to the market. Limited intermediary capacity becomes affine price impact: when intermediaries must absorb more net buying or selling pressure at a given moment, prices move more, and in the model this effect is summarized by a linear relation between contemporaneous net order flow and the price deviation from the frictionless benchmark. Private information becomes private signals received by a subset of operators when they gain market access, implying that information enters prices gradually rather than all at once.

3.1 Dynamic Environment and Frictionless Benchmark

Time is continuous, and compliance obligations recur on a yearly calendar. Let s_t denote the time remaining until the next surrender date. The public state at date t is

$$X_t = (B_t, z_t, s_t), \tag{12}$$

where B_t is the aggregate bank of allowances and z_t is a public scarcity state that captures tightness of the cap and aggregate compliance demand. We use an affine benchmark price,

$$P_t^* = A_0(s_t) + A_B(s_t)B_t + A_z(s_t)z_t, \tag{13}$$

which obtains under affine state dynamics and linear marginal compliance value and provides a tractable approximation more generally. The coefficient functions $A_0(s_t)$, $A_B(s_t)$, and $A_z(s_t)$ capture the smoothing

[4]Full derivations and proofs are in the Supplemental Appendix.

of predictable compliance demand across the compliance cycle. In the frictionless benchmark, predictable compliance demand is reflected in prices well before the surrender date, surrender-month returns are not systematically elevated, and public operator order flow does not forecast future returns.

3.2 Slow Participation as Poisson Market Access

The key departure from the frictionless benchmark is that operators adjust compliance positions only intermittently. Rather than modeling a continuous-time fixed-cost participation problem, we follow the continuous-time analog of fixed participation costs and assume Poisson market access. Let c_{it} denote operator i's *compliance holdings* and h^*_{it} the operator's desired compliance holdings, which reflect expected end-of-cycle obligations and any optimal banking motive. Compliance positions can be rebalanced only when a Poisson shock arrives; upon arrival, the operator resets compliance holdings to the desired level,

$$dc_{it} = \left(h^*_{it} - c_{it^-}\right) dN_{it}, \qquad N_{it} \text{ is a Poisson process with intensity } \lambda_i, \qquad E_t\left[dN_{it}\right] = \lambda_i\, dt. \tag{14}$$

Low arrival intensity λ_i is the continuous-time reduced-form analog of a large participation cost in the two-period model: it captures treasury frictions, internal approvals, inattention, compliance management frictions, or the cost of maintaining a trading relationship. The Poisson access process governs discretionary market access during the compliance cycle. Because surrender is mandatory, any remaining positive compliance shortfall must be closed at the surrender date through a terminal compliance trade, forced access event, or equivalent deadline-month access spike. Thus, if $g_{iA-} > 0$ denotes operator i's residual gap immediately before surrender, the terminal compliance trade is

$$\Delta c_{iA} = g_{iA-}, \tag{15}$$

which maps unresolved compliance gaps into concentrated surrender-month demand. The same reduced-form mechanism can be interpreted as an end-March verification shock to desired compliance holdings that must be resolved before the April surrender deadline.

The terminal compliance trade converts unresolved pre-surrender gaps into surrender-month order flow. Let $g_{iA-} = h^*_{iA} - c_{iA-}$ denote operator i's compliance gap immediately before the surrender date. The

terminal compliance order flow is

$$M_A^{\text{term}} = \sum_i g_{iA-}^{+}, \tag{16}$$

where $g_{iA-}^{+} = \max\{g_{iA-}, 0\}$, since only positive shortfalls generate mandatory surrender-month purchases. Under the same reduced-form immediacy-supply relation introduced below, the surrender-month price impact is

$$P_A - P_A^*(X_A) = \phi\left(M_A^{\text{term}} + M_A^{\text{info}}\right). \tag{17}$$

Thus, unresolved compliance gaps accumulated during the cycle translate directly into terminal operator buying pressure at the surrender date. Equivalently, one can view the surrender month as a sharp increase in effective access intensity, which converts accumulated gaps into observed order flow.

Define the operator-level compliance gap as

$$g_{it} = h_{it}^* - c_{it}. \tag{18}$$

Speculative or information-motivated positions, introduced in Section 3.4 below, are kept separate from c_{it} and therefore do not enter the compliance-gap law of motion. We treat h_{it}^* as a stochastic target: it has a deterministic drift component capturing the calendar build-up of compliance demand toward surrender, an idiosyncratic component reflecting firm-level innovations to expected obligations between verification cycles, and a public-state component that loads on the aggregate scarcity state z_t. Specifying h_{it}^* as stochastic ensures that operator-level compliance gaps are not fully predictable away from access shocks and gives private signals about future scarcity a non-trivial object to be informative about. We do not solve for the law of h_{it}^* from primitives; the analysis below is reduced form in this respect. In expectation, unresolved compliance demand at the operator level obeys the linear adjustment equation

$$E_t[dg_{it}] = E_t[dh_{it}^*] - \lambda_i g_{it}\, dt. \tag{19}$$

Aggregating across operators, aggregate unresolved compliance demand $G_t = \sum_i g_{it}$ satisfies

$$E_t[dG_t] = A(s_t, X_t)\,dt - \sum_i \lambda_i g_{it}\,dt = A(s_t, X_t)\,dt - \bar{\lambda}_t\,G_t\,dt, \qquad \bar{\lambda}_t \equiv \frac{\sum_i \lambda_i g_{it}}{G_t}, \tag{20}$$

where $A(s_t, X_t) = \sum_i a_i(s_t, X_t)$ captures how desired compliance demand rises as the surrender date approaches and $\bar{\lambda}_t$ is the gap-weighted average access intensity. When desired compliance demand rises toward surrender and operators adjust only gradually, unresolved demand accumulates and is largest near surrender.

3.3 Intermediation and Affine Price Impact

A representative competitive intermediary sector supplies immediacy by absorbing contemporaneous net operator order flow, with a quadratic immediacy-provision (or balance-sheet) cost indexed by a parameter $\phi > 0$. Taking $P_t - P_t^*$ as given, intermediary supply M_t^s satisfies the inverse-supply condition $P_t - P_t^* = \phi M_t^s$, and market clearing at operator demand $M_t^s = M_t$ delivers the affine pricing equation

$$P_t = P_t^*(X_t) + \phi M_t, \tag{21}$$

where M_t denotes the contemporaneous net operator order-flow *rate*—positive when operators are net buyers, equivalently when allowances flow from intermediaries or person holding accounts to operator holding accounts—and $\phi > 0$ is the reduced-form price impact of that contemporaneous net flow. It is useful to decompose M_t into a compliance-related component M_t^{comp}, driven by the aggregate unresolved gap G_t, and an informational component M_t^{info} coming from informed speculative orders submitted when operators gain access:

$$M_t = M_t^{\text{comp}} + M_t^{\text{info}}. \tag{22}$$

For the compliance block, expected contemporaneous compliance flow is the access-weighted sum across operators of the share of the gap each closes upon access. Letting $\alpha_i \in (0, 1]$ denote the fraction of operator

i's gap closed conditional on access (full reset corresponds to $\alpha_i = 1$), expected compliance flow is

$$E_t[M_t^{\text{comp}}] = \sum_i \alpha_i \lambda_i g_{it} \equiv \eta_t G_t, \qquad \eta_t \equiv \frac{\sum_i \alpha_i \lambda_i g_{it}}{G_t}, \tag{23}$$

so that contemporaneous compliance order flow is proportional to the aggregate unresolved compliance gap with an effective access-weighted closing rate η_t. We treat η_t as a reduced-form composite, not as a primitive independent of the access intensities λ_i or the closing fractions α_i. Equation (23) should be interpreted as a reduced-form mapping rather than as a fully micro-founded market-clearing condition. The sign of ϕ is the direct analog of the slope of the dealers' inverse supply curve for immediacy in the two-period model: $\phi > 0$ implies that price pressure is stronger when unresolved demand is larger, and the pressure becomes more pronounced as the surrender date approaches or when the market is tighter. Because M_t is a flow rate rather than an inventory stock, $\frac{\phi}{2} M_t^2$ should be interpreted as the marginal cost of absorbing contemporaneous order flow rather than as a literal inventory-holding cost; in the resource-cost decomposition reported in the appendix, this term is treated as an aggregate resource cost only to the extent that it represents a balance-sheet or risk-bearing cost rather than a transfer between operators and intermediaries.

3.4 Informational Advantages and Operator Flow

To preserve a return-predictability prediction, we introduce a parsimonious private-signal block. A subset $\mathcal{I}$ of operators receives a private signal about short-run innovations to the frictionless benchmark price. The private-information block is kept separate from the compliance stock c_{it}: informed operators submit speculative orders only conditional on gaining market access, and these orders do not enter the compliance-gap law of motion (19). Define operator i's informational advantage in return form as

$$\mu_{it} = E_t\left[P^*_{t+\Delta} - P^*_t \mid \mathcal{I}_{it}\right] - E_t\left[P^*_{t+\Delta} - P^*_t \mid X_t\right], \tag{24}$$

where $\mathcal{I}_{it}$ denotes operator i's information set, including any private signal, and X_t is the public state. Upon access, taking prices as given, operator i submits a speculative order q_{it}^{info} that solves

$$\max_{q_{it}^{\text{info}}} \left\{ \mu_{it}\, q_{it}^{\text{info}} - \frac{\gamma_i}{2} \left(q_{it}^{\text{info}}\right)^2 \right\}, \qquad q_{it}^{\text{info}} = \frac{\mu_{it}}{\gamma_i}, \tag{25}$$

where $\gamma_i > 0$ captures inventory, risk-management, or balance-sheet costs. Because access arrivals are Poisson with intensity λ_i, the contemporaneous informative order-flow rate contributed by operator i is $\lambda_i \mu_{it}/\gamma_i$ in expectation. Aggregating in expected-flow form across informed operators,

$$M_t^{\text{info}} = \sum_{i \in \mathcal{I}} \lambda_i \frac{\mu_{it}}{\gamma_i}, \tag{26}$$

which makes information incorporation an expected-rate object rather than a function of the ambiguous contemporaneous active set. Because access is intermittent, information enters prices gradually through operator order flow, and higher λ_i implies faster and more informative flow. Operator flow imbalance therefore predicts future returns when informed operators trade on signals about future scarcity that are not yet fully embedded in the price; moreover, predictive content should be stronger for operators with higher access rates λ_i and lower inventory costs γ_i, which maps directly to the frequent-versus-infrequent trader split we exploit in the data.

Because actual prices include both the benchmark price and contemporaneous price impact, $P_t = P_t^*(X_t) + \phi M_t$, informative order flow M_t^{info} predicts actual future returns only if contemporaneous price impact does not fully reveal the private signal. We assume that the signal content of M_t^{info} is incorporated into the benchmark price gradually, so that subsequent adjustment in P_t^* exceeds any temporary price-impact reversal. Concretely, if

$$E_t\left[P_{t+\Delta}^* - P_t^* \mid M_t^{\text{info}}\right] = \beta M_t^{\text{info}}, \tag{27}$$

then positive predictability of actual returns obtains when the signal-revelation slope β exceeds the expected reversal of contemporaneous price impact. The private-information block is therefore reduced form: it isolates the empirical content of informative order flow without solving an underlying filtering or Kyle-style problem.

3.5 Model Predictions

The continuous-time model yields three empirical predictions that guide the analysis below.

First, the model predicts slow participation. With finite Poisson access rates λ_i, operator compliance holdings c_{it} adjust gradually toward desired compliance levels h_{it}^*. Operators with lower access intensity

have larger expected compliance gaps g_{it} and adjust more slowly. As a result, they trade less often overall and concentrate a larger share of purchases near the surrender date. This prediction is the dynamic analog of the participation-cost mechanism in the two-period model: high participation costs correspond to low effective market-access intensity. Empirically, this prediction maps into persistent non-trading and the concentration of purchases near surrender.

Second, the model predicts seasonal price pressure. When desired compliance demand rises as surrender approaches, slow participation creates unresolved aggregate demand G_t. During the compliance cycle, this gap generates compliance-related order flow as active operators rebalance, with expected contemporaneous compliance flow $E_t[M_t^{\text{comp}}] = \eta_t G_t$, where η_t is the access-weighted closing rate defined in (23). At the surrender date, any remaining positive gap G_{A-}^{+} is converted into terminal compliance order flow M_A^{term} as in (16). With positive intermediary price impact $\phi > 0$, both within-cycle rebalancing demand and terminal surrender-month demand raise the allowance price above the frictionless benchmark. The surrender-month premium is therefore stronger when intermediary capacity is more limited, aggregate scarcity is higher, or the allowance bank is smaller. Importantly, this price pressure is an equilibrium object, not an expression of market irrationality: it arises because slow participation leaves compliance demand unresolved until intermediaries must absorb a predictable, calendar-driven wave of demand at the surrender deadline. Empirically, this prediction maps into concentrated April purchases by operators and predictably high April allowance returns.

Third, the model predicts informative operator flow. If a subset $\mathcal{I}$ of operators receives private signals about short-run innovations to the frictionless benchmark price, then the informational component of operator order flow, M_t^{info}, forecasts future allowance returns. Because informed operators can trade on their signals only when they gain market access, the expected contribution of operator i to M_t^{info} scales with λ_i. Higher access rates make informative order flow appear more quickly and more strongly in the market. The predictive content of operator flow should therefore be stronger among operators with higher market-access intensity, or lower effective participation frictions. Empirically, this prediction maps into Operator Flow Imbalance forecasting future allowance returns, especially for frequent traders.

Taken together, the model implies that gradual participation, limited intermediation, and informational advantages among a subset of active operators can generate delayed trading, seasonal price pressure, and informative operator flow in cap-and-trade markets. The next section evaluates these predictions using the universe of EU ETS transaction-level data.

4 Empirical Tests of the Model's Predictions

In this section, we test the three empirical predictions generated by the model with frictions. Section 4.1 describes the institutional setting and transaction data. Sections 4.2–4.5 then evaluate, in turn, the slow-participation prediction, the seasonal price-pressure prediction, the return-predictability prediction, and the cross-sectional prediction that return predictability is concentrated among more active operators.

4.1 Institutional Setting and Data

The EU ETS is the world's first major carbon market, covering over 13,000 facilities across multiple industrial sectors. It operates as a cap-and-trade system, allocating emission permits (EUAs) to regulated firms. The system is structured into "phases" and follows a fixed annual schedule: by the end of March, emissions for the previous year must be verified, and by the end of April, firms must surrender enough allowances to cover their emissions. Thus, April is the uniform surrender deadline for all regulated operators and corresponds to the surrender date $s_t = 0$ in the continuous-time model of Section 3.

Our sample includes the universe of transactions and compliance data from the EUTL between February 2005 and September 2021, covering the first three EU ETS phases and part of the fourth. The EUTL features three account types (Abrell, 2023): operator holding accounts (used by regulated installations to receive, transfer, and surrender allowances), person holding accounts (used by non-regulated participants, including financial intermediaries), and administrative accounts (used by regulators for allocation and surrender). Accounts are held by companies or individuals, and a single company may hold multiple accounts of different types. The Supplemental Appendix provides a detailed description of the market and the data.

Our transaction data come from registry transfers rather than exchange-level execution records. We therefore observe the direction and quantity of allowance movements across account types, which we aggregate to the company level, but not a structural decomposition of motives or execution prices for each trade. Accordingly, throughout the paper we interpret account-type flows as informative proxies for compliance demand, intermediation, and informational trading rather than as direct observations of intermediary inventories or private information.

4.2 Prediction 1: Slow Participation

The first prediction of the model is that finite market-access intensity generates persistent gaps between desired and actual allowance holdings, implying delayed trading and non-participation for a subset of operators. We evaluate this prediction in two steps. We first document the cross-sectional distribution of trading activity across operators, and then show that non-trading is persistent over time—a direct implication of low effective λ_i in the continuous-time model.

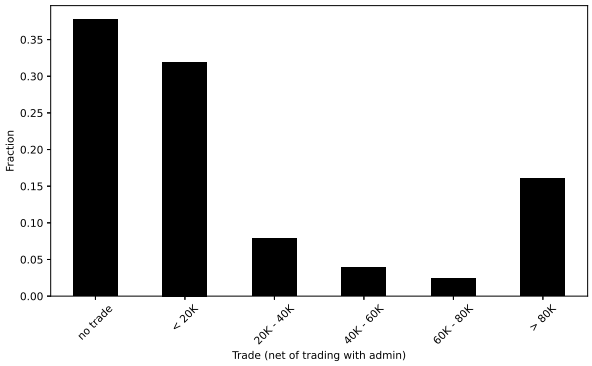
In any cap-and-trade system, cost-effective reallocation of allowances through the secondary market requires firms to actively participate and trade. We analyze the trading behavior of firms with operator holding accounts in the EU ETS. The top panels of Figure 1 show the yearly share of firms by trading volume—Panel A for the full sample, Panel B by EU ETS phases. A large fraction of firms do not trade in a given year: about 40% overall, nearly 60% in Phase I, and over 35% in Phases II and III. In the first nine months of Phase IV, the non-trading share falls to around 20%. The large fraction of non-trading firms is consistent with low effective market-access intensity λ_i in the continuous-time model, which implies that at any point in the compliance cycle a non-trivial mass of operators has not yet had an opportunity to rebalance toward desired compliance holdings.

The top panels group firms by trading volume, excluding transactions with administrative accounts: no trade; 0-20k; 20k-40k; 40k-60k; 60k-80k; and 80k+ contracts. The bottom panels show trading ratios—defined as the volume traded (excluding administrative transactions) divided by annual surrendering amounts. Panels C and D display results for the full sample and by phase, with categories: no trade; ratio < 1; $1-2$; $2-3$; $3-4$; and 4+.

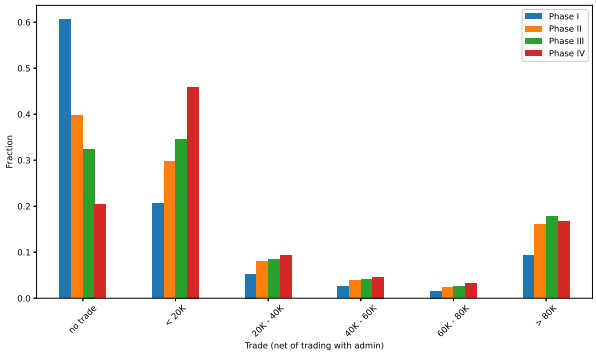
Figure 1: Distribution of Operator Trading Amounts

The top panels plot the fraction of regulated companies with different trading volumes in each year, excluding transactions involving administration accounts. Panels A, and B correspond to the overall sample and Phases I through IV, respectively. For each figure, we group the operators into the following categories: no trade; $0 \leq$ trades $<$ 20k; 20k $\leq$ trades $<$ 40k; 40k $\leq$ trades $<$ 60k; 60k $\leq$ trades $<$ 80k; and 80k $\leq$ trades. The bottom panels plot the fraction of regulated companies with different trading ratios in each year, where the trading ratio is defined as the amount traded, excluding transactions with administration accounts, divided by the company's surrendering amount. Panels C, and D correspond to the overall sample and Phases I through IV, respectively. For each figure, we group the operators into the following categories: no trade; trade ratios $<$ 1; 1 $\leq$ trade ratios $<$ 2; 2 $\leq$ trade ratios $<$ 3; 3 $\leq$ trade ratios $<$ 4; 4 $\leq$ trade ratios. Data is for the period from February 2005 to September 2021.

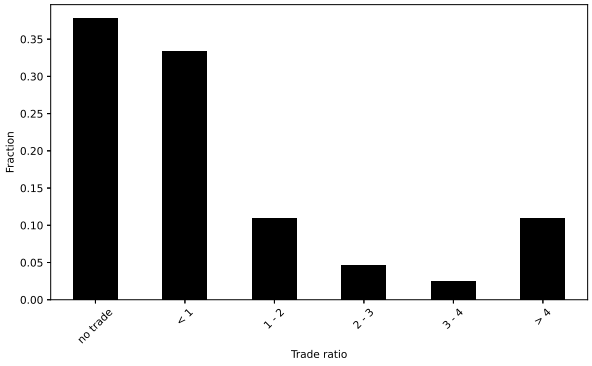
(A) Overall



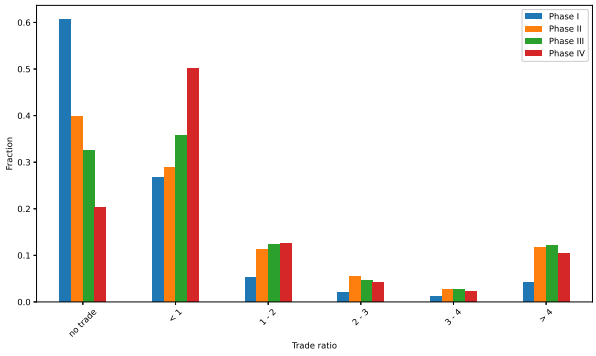
(B) EU-ETS Phases



(C) Overall



(D) EU-ETS Phases



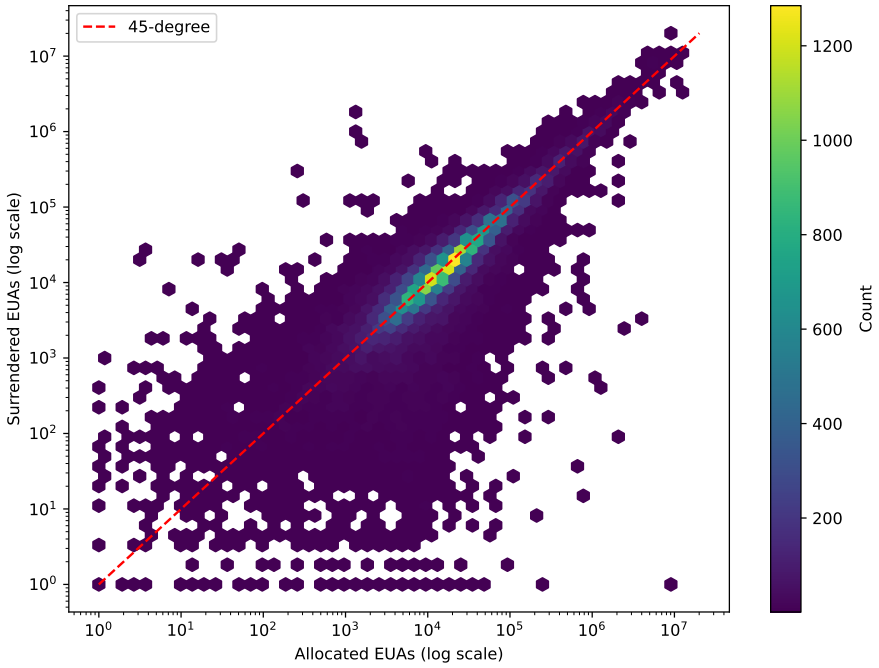
Among firms that do trade, average annual volume is approximately 95,000 contracts, but the distribution is highly skewed. The median is about 3,700; 10% of firms trade fewer than 100 contracts, while another 10% trade more than 86,000. To control for firm size, we normalize trading volume by the surrendering amount. The bottom panels of Figure 1 show that 35% of firms trade less than they surrender (ratio $<$ 1), while over 10% trade more than four times their surrender obligation. This pattern holds across EU ETS phases.

Two patterns emerge. First, 30–40% of regulated firms do not trade in a given year, despite differing emissions and abatement costs—a pattern that in the model is attributable to low effective λ_i. Second, over 10% of firms trade far more than needed for compliance, consistent with informational advantages and speculative positions taken by operators with higher access intensity; we return to this group in Sections 4.4–4.5.

In the frictionless benchmark, non-trading should arise only in the narrow case in which initial allocations already match firms' cost-minimizing compliance and banking needs; any heterogeneity in costs or allocations would otherwise induce trade. The patterns in the data are difficult to reconcile with this benchmark.

Figure 2: Allocated vs Surrendered EUAs

This figure plots the relationship between total allocated and surrendered allowances for operators with no trade in a given year. Each hexagon represents the number of operator-year observations falling within a bin of allocated and surrendered EUA values, as indicated by the color bar. Both axes are in log scale. The red dashed line denotes the 45-degree line. Data is from February 2005 to September 2021.



One possible justification for the large fraction of non-trading firms illustrated in Figure 1 is that they receive a number of free allowances roughly equal to the amount they are required to surrender. However, the data does not support this explanation. Among these firms, the 5th and 25th percentiles of the allocated-to-surrendered ratio are 0 and 0.75, respectively, while the 75th and 95th percentiles are 1.37 and 5.41. This

suggests that many non-trading firms face substantial imbalances in the years they do not trade. Figure 2 plots the relationship between total allocated and surrendered EUAs, for non-trading operators, and confirms this pattern, with a large number of operator-year observations falling above or below the 45-degree line.

A second possible justification for the large fraction of non-trading firms is that they save—or bank—excess allowances (Betz and Schmidt, 2016; Cantillon and Slechten, 2023), perhaps in anticipation of higher future prices and a reduction in the number of freely allocated allowances. One way to test this hypothesis is to examine whether the absence of trading in a given year predicts continued inactivity in subsequent years. If banking were the main reason for non-trading, we would expect that firms eventually return to the market, and firms that bank more are more likely to engage in trading in the future. Thus, no trade in one year should not systematically and positively predict no trade in future years.

Table 1 presents panel regression results where the dependent variable is an operator-level dummy indicating no trade in year $t+h$ (for $h = 1,\ldots,5$), and the main explanatory variable is a no-trade dummy in year t. Panel A reports estimates for the full sample (2005-2021), while Panel B focuses on the post-2013 period, covering Phase III and the beginning of Phase IV.[5]

We report both baseline specifications and those including year fixed effects. Across all horizons from $t+1$ to $t+5$, the estimated coefficients are positive and statistically significant at the 1% level, in both samples and regardless of specification. These estimates can be interpreted as changes in probability. For instance, in the full sample baseline model, operators that did not trade in year t are, on average, 40 percentage points more likely to also not trade in year $t+1$, compared to those that did trade.

Overall, the high prevalence of non-trading firms, together with its persistence, is consistent with the first model prediction: a non-trivial subset of operators has low effective access intensity λ_i, so that expected compliance gaps remain unresolved and trading is systematically delayed. The gradual decline in non-trading rates over time points to an increase in effective λ_i, plausibly driven by market development and policy changes across phases, but a substantial share of firms remains effectively inactive even in the current phase.

[5] It is worth noting that banking was not allowed between Phase I and Phase II, nor between Phase III and Phase IV; allowances not used at the end of these phases would be forfeited. Yet, in 2007 and 2020—the final years of Phases I and III—approximately 60% and 25% of operators, respectively, did not trade, figures that are very similar to other years within those phases (e.g., 53% in 2006 and 27% in 2019). By contrast, in 2012—the final year of Phase II, when banking to Phase III was permitted—the non-trading share was only slightly different, at 36%.

Table 1: No trade predictability

This table reports estimates from panel regressions testing whether the absence of trading by an operator in year t predicts continued non-trading behavior in subsequent years ($t+1$ to $t+5$). The dependent variable is a no-trade dummy in future years, and the main regressor is a no-trade dummy in year t. Panel A presents results for the full sample (2005-2021), while Panel B restricts the sample to Phase III and early Phase IV (2013-2021), during which banking across phases is not allowed. Both baseline specifications and models with year fixed effects are reported. Standard errors are Newey-West (Newey and West, 1987) adjusted with n lags where n is the number of overlapping periods. Standard errors are clustered at the operator level. t-statistics are reported in parentheses. *, **, and *** denote significance at the 10%, 5%, and 1% levels, respectively.

Panel A: Full sample					
	$t+1$	$t+2$	$t+3$	$t+4$	$t+5$
Baseline					
no trade	0.404***	0.323***	0.255***	0.215***	0.196***
	(127.120)	(92.780)	(66.460)	(51.820)	(43.590)
R2	0.169	0.109	0.070	0.050	0.042
Year Fixed Effects					
no trade	0.394***	0.314***	0.264***	0.225***	0.202***
	(11.040)	(10.710)	(12.100)	(12.590)	(16.720)
R2	0.158	0.101	0.073	0.053	0.043
Panel B: Phase III and IV					
	$t+1$	$t+2$	$t+3$	$t+4$	$t+5$
Baseline					
no trade	0.439***	0.361***	0.303***	0.251***	0.217***
	(96.720)	(71.200)	(52.740)	(39.840)	(31.180)
R2	0.195	0.134	0.096	0.068	0.053
Year Fixed Effects					
no trade	0.431***	0.362***	0.317***	0.272***	0.233***
	(9.940)	(11.950)	(20.340)	(23.320)	(50.600)
R2	0.189	0.136	0.107	0.081	0.061

4.3 Prediction 2: Surrender-Month Demand Pressure

The second prediction of the model is that, when desired compliance demand rises as surrender approaches and participation is gradual, unresolved aggregate demand G_t translates—through limited intermediary capacity $\phi > 0$—into predictable surrender-month price pressure. The EU ETS system has a fixed verifying and surrendering schedule each year: by the end of April, firms must surrender a sufficient number of emission allowances to the administrator account to match their emissions. April is therefore the designated month for surrendering allowances for all firms under the EU ETS system and corresponds to the smallest value of s_t within each compliance cycle.

We compute the monthly net purchase of emission allowances by firms, excluding transactions involving admin accounts. The results are presented in Panel A of Figure 3. A striking pattern emerges: the purchasing of emission allowances predominantly occurs in April each year, as highlighted in red in the figure. April consistently shows the largest net purchases of emission allowances by firms throughout the entire sample period. The difference in magnitudes between surrendering and non-surrender months is substantial. The second largest net purchase month is typically December, possibly due to tax reasons or the expiration of the most common futures contract on the EUA. While part of the April buying may reflect uncertainty resolved by the end-March verification process, verification timing alone is unlikely to account for the size and persistence of the April spike.

Next, we examine the average monthly emission allowance return, which we document in Panel B of Figure 3. The emission allowance return is computed as the average monthly log change in emission allowance prices. We observe that the average return is highest in April, exceeding 9.5%. This monthly return is significantly higher than returns in other months of the year. The concentrated demand in April drives a predictable price surge in emission allowances.

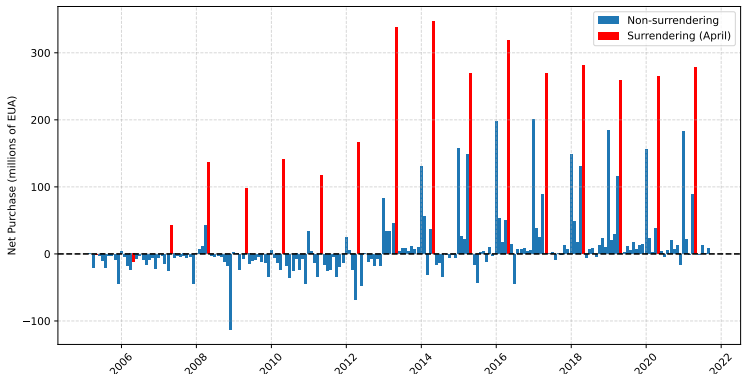
To quantify the implied private cost to regulated firms that purchase emission allowances when prices are predictably high during the surrender month, we use a back-of-the-envelope calculation. Using the total net amount purchased by the regulated firms during the surrender month, the emission allowance prices at the time, and assuming the average 10% predictable appreciation for the surrender month, we estimate that the total loss by the regulated firms amounts to about €5 billion, or 2% of the total traded volume of the regulated firms. As emphasized in Section 2.3, this is a private compliance-cost measure for delayed buyers rather than a direct measure of aggregate resource costs, because part of the higher April price is paid to

counterparties. The persistence of the April buying and the April return premium is consistent with the second model prediction.

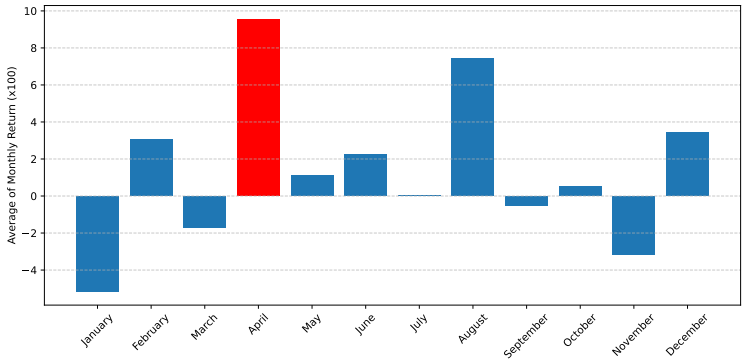
Figure 3: Trading and Prices in the Surrendering Month

This figure plots the monthly net purchase of emission allowances by operators and the average emission allowance returns for each month. Panel A depicts the average total amount of transactions over time, while Panel B shows the average emission contract returns by month since 2008. Transactions involving admin accounts are excluded. At least one party in a transaction must be an operator for the transaction to be included in the analysis.

(A) Net Transaction Amount



(B) Average EUA Return



4.4 Prediction 3: Operator Flow Imbalance and Return Predictability

The third prediction of the model is that operator flow imbalance forecasts future allowance returns when a subset of active operators trades on private signals about future scarcity and prices absorb order flow gradually. We now test this prediction. A small number of regulated firms account for a disproportionate share of trading: these are precisely the operators whose higher effective market-access intensity λ_i allows them, in the model, to act on private signals when they arrive. If so, allowance prices should rise following

net purchases by these firms and fall following net sales.

To test this prediction, we construct a monthly *Operator Flow Imbalance*—the net flow from person holding accounts (non-regulated participants, including financial intermediaries) to operator holding accounts (regulated firms), scaled by total bilateral flows between the two groups—and use it to predict future cumulative returns. In the spirit of the model, Operator Flow Imbalance is a monthly proxy for the informative-order-flow component M_t^{info}. Table 2 shows that the imbalance significantly forecasts emission allowance returns up to twelve months ahead, with increasing coefficients and R-squared values over longer horizons. These results are robust to Newey-West standard errors (Newey and West, 1987) and are consistent with active operators trading on information not yet fully impounded in prices.

Because trading activity clusters around compliance deadlines, we include month fixed effects to control for seasonality. This strengthens the predictive power of *Operator Flow Imbalance*, with all coefficients remaining positive and significant. The Supplemental Appendix further shows that this return-predictability result is robust to standard long-horizon predictability concerns. Appendix A.4 reports Valkanov's $t/\sqrt{T}$ inference and shows that Stambaugh-type bias is quantitatively small in this setting.

A back-of-the-envelope calculation based on the estimated return-predictability patterns implies roughly €8 billion in cumulative ex post trading performance over our sample—about 3.5% of total traded volume under a conservative three-month holding period. We view this figure as an illustrative magnitude consistent with informational advantages, not as a direct estimate of structural abnormal returns. We interpret these forecasting patterns as consistent with informational advantages among a subset of active operators; the empirical measure is not a pure structural measure of informed trading.

Table 2: Time-Series Return Predictability

This table reports the results of time-series EUA return predictability tests. *Operator Flow Imbalance* is defined as the net flow from person holding accounts to operator holding accounts divided by total bilateral transaction volume between the two groups in each month. Standard errors are Newey-West (Newey and West, 1987) adjusted with n lags where n is the number of overlapping periods. t-statistics are reported in parentheses. *, **, and *** denote significance at the 10%, 5%, and 1% levels, respectively.

	$r_{t\to t+1}$	$r_{t\to t+2}$	$r_{t\to t+3}$	$r_{t\to t+4}$	$r_{t\to t+6}$	$r_{t\to t+8}$	$r_{t\to t+12}$
Baseline							
Operator Flow Imbalance	0.039*	0.097***	0.137**	0.213***	0.322**	0.414***	0.565***
	(1.958)	(2.626)	(2.471)	(2.806)	(2.513)	(2.681)	(2.770)
R2	0.013	0.040	0.050	0.087	0.115	0.132	0.145
Month Fixed Effects							
Operator Flow Imbalance	0.070**	0.132**	0.187**	0.252**	0.369**	0.472**	0.776***
	(2.350)	(2.505)	(2.527)	(2.599)	(2.417)	(2.585)	(2.901)
R2	0.098	0.115	0.139	0.153	0.160	0.148	0.202

4.5 Frequent versus Infrequent Traders

A further implication of the third model prediction is that the predictive content of operator flow should be concentrated among operators with higher market-access intensity λ_i or lower inventory costs γ_i—that is, among frequent traders. The intuition is direct: an operator who rarely accesses the market has few opportunities to act on private signals, so informational advantages, even when present, generate little realized order flow.

To test this prediction, we calculate the number of trades companies engage in over the sample period. We then create two subsamples based on the median number of trades and compute *Operator Flow Imbalance* separately within each subsample. We denote *Operator Flow Imbalance*$^{\text{high}}$ and *Operator Flow Imbalance*$^{\text{low}}$ as the imbalance for firms above and below the sample median, respectively.

Table 3 presents the results from time-series return predictability regressions using both *Operator Flow Imbalance*$^{\text{high}}$ and *Operator Flow Imbalance*$^{\text{low}}$. We first report the baseline results without month fixed effects and then report the results with month fixed effects. When both variables are included in the regressions, the coefficient estimates of *Operator Flow Imbalance*$^{\text{high}}$ are always positive, with or without month

fixed effects. Without month fixed effects, the coefficient estimates of *Operator Flow Imbalance*$^{\text{high}}$ are significant from the two-month horizon onward; with month fixed effects, they are significant at the 5% level or beyond at every horizon. By contrast, the coefficient estimates of *Operator Flow Imbalance*$^{\text{low}}$ are never significant in any specification. Return predictability is therefore driven by trading by frequent operators, as implied by the third model prediction. In the model, this frequent-versus-infrequent split is a reduced-form way to capture heterogeneity in effective market access and balance-sheet capacity rather than a direct estimate of structural λ_i or γ_i.

Table 3: Time-Series Return Predictability: Frequent Traders

This table reports the results of time-series EUA return predictability tests for frequent and infrequent operators. *Operator Flow Imbalance*$^{\text{low}}$ and *Operator Flow Imbalance*$^{\text{high}}$ are defined as the net flow from person holding accounts to operator holding accounts divided by total bilateral transaction volume between the two groups, computed separately for operators below and above the sample median number of transactions. Standard errors are Newey-West (Newey and West, 1987) adjusted with n lags, where n is the number of overlapping periods. t-statistics are reported in parentheses. *, **, and *** denote significance at the 10%, 5%, and 1% levels, respectively.

	$r_{t\to t+1}$	$r_{t\to t+2}$	$r_{t\to t+3}$	$r_{t\to t+4}$	$r_{t\to t+6}$	$r_{t\to t+8}$	$r_{t\to t+12}$
Baseline							
Operator Flow Imbalance$^{\text{low}}$	-0.025	-0.003	0.014	0.002	0.065	0.079	-0.045
	(-0.627)	(-0.071)	(0.247)	(0.022)	(0.660)	(0.580)	(-0.283)
Operator Flow Imbalance$^{\text{high}}$	0.050	0.082*	0.105*	0.188***	0.253**	0.326**	0.556**
	(1.405)	(1.760)	(1.846)	(2.714)	(2.208)	(2.274)	(2.571)
R2	0.012	0.027	0.035	0.070	0.100	0.113	0.130
Month Fixed Effects							
Operator Flow Imbalance$^{\text{low}}$	-0.026	-0.013	0.008	-0.022	0.037	0.076	0.042
	(-0.652)	(-0.254)	(0.117)	(-0.275)	(0.345)	(0.551)	(0.231)
Operator Flow Imbalance$^{\text{high}}$	0.078**	0.125**	0.160**	0.237***	0.308**	0.375**	0.718***
	(2.105)	(2.381)	(2.385)	(2.794)	(2.280)	(2.271)	(2.878)
R2	0.094	0.104	0.127	0.142	0.145	0.130	0.183

5 Conclusions

Cap-and-trade has become a widely adopted mechanism for pricing emissions (Grubb, 2012), implemented in the European Union and in large economies such as California (Fowlie et al., 2012) and, more recently, China (Liu et al., 2022) (Newell et al., 2014). These systems rely on secondary markets to absorb predictable compliance demand and to incorporate information into prices, which makes their performance a question about a regulated asset market as much as a question about environmental policy.

This paper studies how a major regulated asset market—the EU ETS—absorbs predictable compliance demand and information. Using the universe of EU ETS registry transactions from 2005 to 2021, we document three facts that are difficult to reconcile with a frictionless allowance market. A substantial fraction of regulated firms do not trade in a given year. Many operators concentrate purchases in the April surrender month, when prices are systematically high. And operator flow imbalance predicts future allowance returns, with stronger predictive content among more active traders. We organize these facts around a model of trading frictions in cap-and-trade markets, first in a two-period version that isolates the basic mechanism and then in a continuous-time version with Poisson market access, affine price impact, and informational advantages. The framework implies that gradual participation and limited intermediary capacity can generate delayed trading, surrender-month price pressure, and informative operator flow.

These results speak to the financial implementation of cap-and-trade. Market-based climate policy relies on a secondary allowance market to transmit prices and information, and our findings suggest that trading frictions shape how that market functions. They have direct implications for the design of market microstructure in the EU ETS and in emerging systems in jurisdictions such as China, the United States, and beyond.

Data Availability

The EUTL registry transaction and compliance data used in the paper are publicly available from the European Commission and from the EUETS.info database. EUA price data are obtained from Datastream and are subject to Datastream's data-use restrictions. The replication package will include the code needed to construct the analysis sample from the public registry files and to reproduce all tables and figures, except for proprietary price data that must be obtained separately.

Code Availability

The replication code and documentation will be deposited in a public repository upon acceptance, subject to data-license restrictions for proprietary price data.

References

Abrell, J. (2023). Database for the european union transaction log. Available at Euets.info.

Acharya, V. V., Lochstoer, L. A., and Ramadorai, T. (2013). Limits to arbitrage and hedging: Evidence from commodity markets. *Journal of Financial Economics*, 109(2):441–465.

Admati, A. R. and Pfleiderer, P. (1988). A theory of intraday patterns: Volume and price variability. *Review of Financial Studies*, 1(1):3–40.

Allen, M. R., Frame, D. J., Huntingford, C., Jones, C. D., Lowe, J. A., Meinshausen, M., and Meinshausen, N. (2009). Warming caused by cumulative carbon emissions towards the trillionth tonne. *Nature*, 458(7242):1163–1166.

Betz, R. and Schmidt, T. S. (2016). Transfer patterns in phase i of the european union emissions trading system: a first reality check based on cluster analysis. *Climate Policy*, 16(4):474–495.

Biais, B., Hombert, J., Schmidt, D., and Weill, P.-O. (2026). Cap and trade with imperfect hedging. Working paper, February 26, 2026. HEC Research Papers Series No. 1574. Available at SSRN.

Biais, B. and Landier, A. (2026). Emission caps and investment in green technologies. *Review of Financial Studies*. Forthcoming.

Cantillon, E. and Slechten, A. (2023). Market design for the environment. In *Handbook of the Economics of Corporate Finance*.

Colmer, J., Martin, R., Muûls, M., and Wagner, U. J. (2025). Does pricing carbon mitigate climate change? firm-level evidence from the european union emissions trading system. *Review of Economic Studies*. Forthcoming.

Dales, J. H. (1968). *Pollution, property, and prices*. University of Toronto Press.

Döbbeling-Hildebrandt, N., Miersch, K., Khanna, T. M., Bachelet, M., Bruns, S. B., Callaghan, M. W., Edenhofer, O., Flachsland, C., Forster, P. M., Kalkuhl, M., et al. (2024). Systematic review and meta-analysis of ex-post evaluations on the effectiveness of carbon pricing. *Nature Communications*, 15(1):4147.

Duffie, D. (2010). Presidential address: Asset price dynamics with slow-moving capital. *Journal of Finance*, 65(4):1237–1267.

Duffie, D., Gârleanu, N., and Pedersen, L. H. (2007). Valuation in over-the-counter markets. *Review of Financial Studies*, 20(6):1865–1900.

Fowlie, M., Holland, S. P., and Mansur, E. T. (2012). What do emissions markets deliver and to whom? evidence from southern california's nox trading program. *American Economic Review*, 102(2):965–993.

Gabaix, X. and Maggiori, M. (2015). International liquidity and exchange rate dynamics. *Quarterly Journal of Economics*, 130(3):1369–1420.

Gârleanu, N., Pedersen, L. H., and Poteshman, A. M. (2009). Demand-based option pricing. *Review of Financial Studies*, 22(10):4259–4299.

Glosten, L. R. and Milgrom, P. R. (1985). Bid, ask and transaction prices in a specialist market with heterogeneously informed traders. *Journal of Financial Economics*, 14(1):71–100.

Goulder, L. H. (2013). Markets for pollution allowances: what are the (new) lessons? *Journal of Economic Perspectives*, 27(1):87–102.

Gromb, D. and Vayanos, D. (2002). Equilibrium and welfare in markets with financially constrained arbitrageurs. *Journal of Financial Economics*, 66(2-3):361–407.

Grubb, M. (2012). Cap and trade finds new energy. *Nature*, 491:666–667.

Harstad, B. and Eskeland, G. S. (2010). Trading for the future: Signaling in permit markets. *Journal of Public Economics*, 94(9-10):749–760.

Hendershott, T. and Menkveld, A. J. (2014). Price pressures. *Journal of Financial Economics*, 114(3):405–423.

Jaraitė, J., Convery, F., and Di Maria, C. (2010). Transaction costs for firms in the eu ets: lessons from ireland. *Climate Policy*, 10(2):190–215.

Känzig, D. R. (2023). The unequal economic consequences of carbon pricing. Technical report, National Bureau of Economic Research.

Liu, L., Chen, C., Zhao, Y., and Zhao, E. (2022). China's carbon-emissions trading: Overview, challenges and future. *Renewable and Sustainable Energy Reviews*.

Matthews, H. D., Gillett, N. P., Stott, P. A., and Zickfeld, K. (2009). The proportionality of global warming to cumulative carbon emissions. *Nature*, 459(7248):829–832.

Montgomery, W. D. (1972). Markets in licenses and efficient pollution control programs. *Journal of Economic Theory*, 5(3):395–418.

Newell, R. G., Pizer, W. A., and Raimi, D. (2014). Carbon markets: Past, present, and future. *Annual Review of Resource Economics*, 6:191–215.

Newey, W. and West, K. (1987). A simple, positive semi-definite, heteroskedasticity and autocorrelation consistent covariance matrix. *Econometrica*, 55(3):703–708.

Rogelj, J., Popp, A., Calvin, K. V., Luderer, G., Emmerling, J., Gernaat, D., Fujimori, S., Strefler, J., Hasegawa, T., Marangoni, G., et al. (2018). Scenarios towards limiting global mean temperature increase below 1.5°c. *Nature Climate Change*, 8(4):325–332.

Rubin, J. D. (1996). A model of intertemporal emission trading, banking, and borrowing. *Journal of Environmental Economics and Management*, 31(3):269–286.

Schmalensee, R. and Stavins, R. N. (2013). The so2 allowance trading system: The ironic history of a grand policy experiment. *Journal of Economic Perspectives*, 27(1):103–122.

Stambaugh, R. F. (1999). Predictive regressions. *Journal of Financial Economics*, 54(3):375–421.

Subrahmanyam, A. (1991). Risk aversion, market liquidity, and price efficiency. *Review of Financial Studies*, 4(3):417–441.

Valkanov, R. (2003). Long-horizon regressions: theoretical results and applications. *Journal of Financial Economics*, 68(2):201–232.

A Supplemental Appendix

A.1 Market Overview

The EU ETS is the first large carbon market in the world. The system covers more than 13,000 facilities spanning diverse industrial sectors. Functioning as a cap-and-trade system, the EU ETS allocates emission permits, commonly known as European Union Allowances (EUAs), to regulated companies. Regulated companies are required to report their annual carbon emission amount to the regulatory authorities and also must surrender a corresponding amount of permits.

The EU ETS was launched in 2005 as a cornerstone of the EU policy to reduce greenhouse gas emissions. The evolution of the EU ETS is organized in various "phases". The first trading phase (Phase I) goes from 2005 to 2007. This is a pilot phase that lays the groundwork for subsequent phases by establishing trading rules and procedures. The second trading phase (Phase II) goes from 2008 to 2012. Building upon Phase I, the EU ETS expanded its scope and refined its mechanisms, including efforts to standardize the allocation rules and improve the pricing and functionality of the trading system. The third phase (Phase III) goes from 2013 to 2020, while the current phase (Phase IV) started in 2021 and will end in 2030.

During our sample period, the EU ETS has a fixed verifying and surrendering schedule each year. In particular, each year operators must submit an emission report. The emissions data for a given year must be verified by an accredited verifier by the end of March of each year. Then, operators must surrender the equivalent number of allowances to match the installation's emissions by the end of April. In other words, April is the month of surrendering for all regulated firms under the EU ETS. For each ton of emissions where no allowance is surrendered on time, there is a penalty of €100, in addition to the cost of surrendering allowances. The names of penalized operators are also made public.

A.2 Data

The EU ETS individual trading and compliance data are publicly available and are from the European Union Transaction Log (EUTL) (Abrell, 2023).[6] The EUTL serves as the primary platform for reporting and monitoring within the EU ETS. This tool enables the European Commission to transparently disclose crucial data regarding compliance by regulated entities, the participation of stakeholders within the system,

[6]The data used in the paper is publicly available through the EUTL website (https://ec.europa.eu/clima/ets/). We downloaded the data using the tools available at https://www.euets.info.

and the transactions conducted among these participants. Our sample contains the universe of transactions and compliance information from the EUTL from February 2005 to September 2021, the first three phases of the EU ETS, and part of the current fourth phase. The EUTL data that we use in this paper contain a number of components, including compliance, account, and transaction information, which we describe separately below.

Compliance Information Compliance within the EU ETS occurs at the level of installations. The installations have the obligation to surrender emission allowances for the verified amounts. The EUTL data contain detailed information about the identities of the installations as well as the compliance information. The compliance information contains the allocated amount and the surrendered amount for each installation in each year. In the sample, the compliance rate of installations is close to 100%, and therefore, non-compliance is unlikely to be an issue in our analysis.

Account Information There are three basic account types in the EUTL (Abrell, 2023): operator holding accounts, person holding accounts, and administrative accounts. An operator holding account represents an installation—the regulated participant. The operator holding account allows the installation to receive, transfer, and surrender emission allowances. Non-regulated participants (or financial intermediaries) can participate in the EU ETS by using personal holding accounts to trade emission allowances. Regulators must use administrative accounts to allocate and receive surrendered emission allowances. Accounts are held by account holders or companies. Sometimes, a company may have multiple accounts, including operator holding accounts and/or person holding accounts.

Transaction Information The EUTL records all emission allowance transactions between accounts. Transactions can happen between any two accounts, including operator, person, and administrative accounts. Reallocations of emission allowances within the same firm are recorded as transactions in the EUTL. Receiving and surrendering emission allowances from and to administrative accounts are also recorded as transactions. In this paper, we exclude all EUTL recorded transactions involving administrative accounts, as our focus is on actual market activity—particularly transactions involving regulated entities—rather than allowance allocation or surrender. However, we retain information on surrender activity and the identities of firms that only interact with administrative accounts. Furthermore, since we aggregate transaction information at the company level, transactions recorded within the same company in the EUTL always net out to zero.

The composition and behavior of different account types in the EU ETS highlight important market dynamics. While person holding accounts are fewer in number than operator holding accounts, they account for the majority of trading value in the market. Notably, their activity has increased over time, particularly from Phase I to Phase III. In contrast, operators typically engage in relatively few transactions, most often with intermediaries rather than other operators. Further details on account type evolution, transaction volumes, and operator-level trading statistics are provided in the Supplemental Appendix.

A.3 EU ETS Transaction information

The number of person holding accounts was at its lowest (257 accounts) at the onset of EU ETS. This number increased, peaking at over 1,421 at the beginning of Phase II, and then steadily declined almost every year, reaching 652 in 2021. In contrast, the number of operator holding accounts has been increasing, starting at 3,755 at the onset of the EU ETS, reaching a peak of 7,758 in 2013, and stabilizing around 6,198 in 2021. While the number of person holding accounts is a fraction of the number of operator holding accounts, person holding accounts represent the largest share of the total value of EUAs sold and purchased, as illustrated in Figure A.1. Furthermore, the figure shows that the total value of EUAs sold and purchased by person holding accounts has been increasing from Phase I to Phase III of the EU ETS, reaching approximately €800 billion (50 billion EUAs) EUAs sold and €700 billion (45 billion EUAs) EUAs purchased in Phase III. Operator holding accounts, instead, account for approximately €20 billion in EUAs sold (1.7 billion EUAs) and €63 billion in EUAs purchased (5.3 billion EUAs) in Phase III.[7]

Table A.1 provides descriptive statistics on operator trading throughout the entire sample period, as well as across the three distinct phases of the EU ETS market and the first months of the current Phase IV. The statistics reflect the number of transactions—both sales and purchases—rather than the volume of EUAs traded. It is important to note that individual transactions may correspond to very different quantities of EUAs, with some involving small transfers and others representing large blocks of allowances.

Panel A shows that, on average, an operator engaged in approximately 8 sell transactions and 13 buy transactions over the entire sample. The average share of transactions involving sales to other operators was 19%, while purchases from other operators accounted for 16%. Overall, the average operator conducted around 20 transactions in total, with 19% involving another operator as the counterparty. These propor-

[7]We use the daily EUA spot price, obtained from Datastream, to estimate the euro value of the EUAs purchased and sold. Values for Phase IV correspond to the first 9 months of 2021.

tions are relatively stable across the first three market phases. Panel B highlights that the distribution of transaction counts is right-skewed, indicating that median operators trade significantly less than the average. Specifically, the median operator engaged in just 3 sell and 5 buy transactions, typically interacting with intermediaries rather than with other operators—a pattern that holds across all phases of the market.

Figure A.1: EUA Buy and Sell Transactions by Holding Type

This figure presents the evolution in the total euro amount (top panels) and transaction counts (bottom panel) of EUA sold and purchased by holding type. Data is from February 2005 to September 2021.

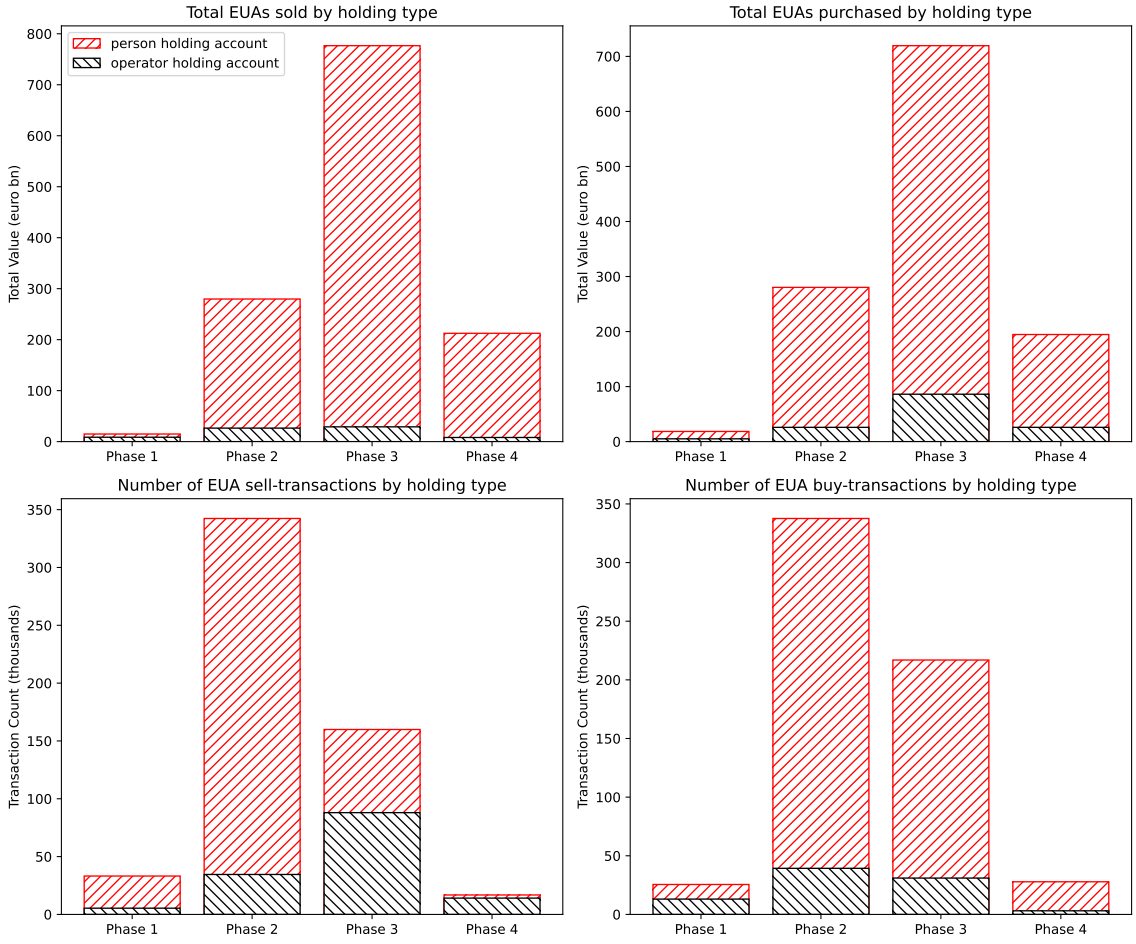


Table A.1: Operator Trading in the EU ETS Market

This table shows the fraction of trades operators engage with other operators. Panel A reports the average numbers of sales, buys, and total trades as well as the fractions of sales, buys, and total trades with other operators. Panel B reports the median numbers of sales, buys, and total trades as well as the fractions of sales, buys, and total trades with other operators.

	Sales	To Operators (%)	Buys	From Operators (%)	Total Trades	With Operators (%)
Panel A: Averages						
Overall	7.86	0.19	12.91	0.16	20.78	0.19
Phase 1	3.49	0.13	1.44	0.16	4.93	0.22
Phase 2	5.87	0.17	5.16	0.19	11.03	0.22
Phase 3	3.46	0.20	9.81	0.14	13.27	0.20
Phase 4	0.65	0.15	2.96	0.15	3.61	0.22
	Sales	To Operators (%)	Buys	From Operators (%)	Total Trades	With Operators (%)
Panel B: Medians						
Overall	3	0	5	0	10	0
Phase 1	1	0	0	0	2	0
Phase 2	3	0	2	0	5	0
Phase 3	1	0	4	0	6	0
Phase 4	0	0	2	0	2	0

A.4 Further results in terms of emission allowance return predictability

We provide several robustness checks and confirm the baseline return predictability results. In the return predictability literature, several econometric issues have been documented in long-horizon predictive regressions with overlapping cumulative returns and persistent regressors. First, such regressions can lead to inference problems, particularly with the t-statistics, which may diverge over long horizons (Valkanov, 2003). Second, an inference problem can arise from using persistent regressors when the shocks to the dependent and independent variables are correlated (Stambaugh, 1999).

To address the first critique, we employ the $t/\sqrt{T}$ test (Valkanov, 2003). We compute the $t/\sqrt{T}$ test and simulate the critical values with parameters calibrated to our data (Valkanov, 2003). We show that the coefficient estimates are significant at the 5% level starting from the two-month-ahead horizon with the $t/\sqrt{T}$ test. To address the second critique, we examine the extent of this bias (Stambaugh, 1999) and determine that its economic impact is minor; the bias amounts to less than 1% of the coefficient estimate. This is due to the relatively low autocorrelation of the independent variable.

A.5 Continuous-Time Model: Derivations and Proofs

This subsection provides the derivations and proofs underlying the continuous-time model in Section 3. We follow the notation of the main text: calendar time is continuous with a recurring annual compliance cycle, s_t denotes time remaining until the next surrender date, and $X_t = (B_t, z_t, s_t)$ is the public state. The aggregate bank B_t evolves deterministically with aggregate net compliance use, and the public scarcity state z_t follows a mean-reverting diffusion

$$dz_t = -\kappa_z(z_t - \bar{z})\,dt + \sigma_z\,dW_t, \tag{A.1}$$

where $\kappa_z \geq 0$ and $\bar{z}$ is the long-run mean. The specification is stylized: any affine state dynamics delivers the same conclusions.

Affine Benchmark Pricing

In the frictionless benchmark, intermediary capacity is unlimited, participation is continuous, and information is public. Consider a pricing functional $P_t^* = P^*(X_t)$ and conjecture the affine form

$$P^*(X_t) = A_0(s_t) + A_B(s_t)\,B_t + A_z(s_t)\,z_t. \tag{A.2}$$

Absent frictions, allowance prices satisfy a no-arbitrage condition relating the instantaneous discount rate to the drift of P^* plus any expected compliance-driven dividend flow. Substituting (A.2) and matching coefficients on B_t, z_t, and the constant yields a system of ordinary differential equations in the within-cycle calendar variable s_t for $A_0(s_t)$, $A_B(s_t)$, and $A_z(s_t)$, with periodic boundary conditions imposed by the recurrence of the compliance cycle. Closed-form solutions obtain under constant aggregate compliance demand and mean-reverting z_t; in general, the coefficients are smooth functions of s_t. Further details are standard in affine asset pricing and omitted.

Aggregation of Compliance Demand

Let c_{it} denote operator i's compliance holdings and h_{it}^* desired compliance holdings. Under the Poisson access specification (14), the compliance gap $g_{it} = h_{it}^* - c_{it}$ satisfies the reduced-form adjustment equation

in the main text,

$$E_t[dg_{it}] = E_t[dh^*_{it}] - \lambda_i g_{it}\, dt. \tag{A.3}$$

Speculative positions are tracked separately from c_{it} in Section 3.4 of the main text and do not enter (A.3). Write $E_t[dh^*_{it}] = a_i(s_t, X_t)\, dt$, where $a_i(s_t, X_t)$ captures how desired compliance demand rises as the surrender date approaches ($s_t \to 0$). Consistent with the discussion in Section 3, we treat h^*_{it} as a stochastic target with deterministic seasonal drift, idiosyncratic firm-level innovations, and a public-state component that loads on z_t; this stochastic structure ensures that operator-level gaps are not fully predictable away from access shocks and gives the private-signal block in Section 3.4 a non-trivial object to be informative about. The analysis is reduced form in this respect: we work with the conditional drift $a_i(s_t, X_t)$ rather than solving for h^*_{it} from primitives. Under mild integrability conditions, the solution of (A.3) is

$$E_t[g_{it}] = e^{-\lambda_i t}\, g_{i0} + \int_0^t e^{-\lambda_i (t-u)}\, a_i(s_u, X_u)\, du. \tag{A.4}$$

Aggregating across operators yields

$$E_t[dG_t] = A(s_t, X_t)\, dt - \sum_i \lambda_i g_{it}\, dt = A(s_t, X_t)\, dt - \bar{\lambda}_t\, G_t\, dt, \qquad G_t = \sum_i g_{it}, \tag{A.5}$$

with $A(s_t, X_t) = \sum_i a_i(s_t, X_t)$ and $\bar{\lambda}_t \equiv \sum_i \lambda_i g_{it} / G_t$ the gap-weighted access intensity. Equation (A.5) is the aggregate analog of (A.3) and is the law of motion for unresolved aggregate compliance demand. A closed-form expression for $E[G_t]$ under a simple seasonal profile and constant $\bar{\lambda}$ is derived below.

Intermediary Price Impact

A representative competitive intermediary sector supplies immediacy by absorbing the contemporaneous net operator order-flow rate M_t, with an instantaneous immediacy-provision (or balance-sheet) cost $\frac{\phi}{2} M_t^2$, where $\phi > 0$ indexes limited intermediary capacity. Taking $P_t - P^*_t$ as given, intermediary supply M^s_t satisfies the inverse-supply condition $P_t - P^*_t = \phi\, M^s_t$. Market clearing at operator demand $M^s_t = M_t$ then yields the equilibrium pricing relation

$$P_t = P^*_t(X_t) + \phi\, M_t, \tag{A.6}$$

which is equation (21) of the main text. The intermediary does not choose operator order flow; rather, ϕ is the slope of the dealers' inverse supply curve for immediacy, and (A.6) obtains by clearing. The decomposition $M_t = M_t^{\text{comp}} + M_t^{\text{info}}$ follows by splitting operator flow into a compliance-related component and an informational component. For the compliance block, we use the access-weighted aggregation $E_t[M_t^{\text{comp}}] = \sum_i \alpha_i \lambda_i g_{it} \equiv \eta_t G_t$ stated in (23) of the main text; η_t is the effective access-weighted closing rate at which active operators close unresolved gaps and is a reduced-form composite of the access intensities λ_i and the closing fractions $\alpha_i \in (0, 1]$.

Informed Demand

Let $\mathcal{I}_{it} \supseteq \sigma(X_t)$ denote operator i's information set, including any private signal, and define the informational advantage μ_{it} in return form as in (24). Upon access, operator $i \in \mathcal{I}$ submits a speculative order q_{it}^{info} that solves $\max_q \{\mu_{it} q - \frac{\gamma_i}{2} q^2\}$, which yields the closed-form rule $q_{it}^{\text{info}} = \mu_{it}/\gamma_i$. Because access arrivals are Poisson with intensity λ_i, the expected contribution of operator i to the informative order-flow rate is $\lambda_i \mu_{it}/\gamma_i$. Aggregating in expected-rate form across informed operators,

$$M_t^{\text{info}} = \sum_{i \in \mathcal{I}} \lambda_i \frac{\mu_{it}}{\gamma_i}, \tag{A.7}$$

which avoids conditioning on a time-varying active set of operators and makes information incorporation a conditional expected-flow object. Because access is intermittent, information is impounded into prices gradually through the arrival of market-access opportunities, and higher λ_i implies faster and more informative flow.

Derivation of Prediction 1 (Slow Participation)

From (A.4), $E_t[g_{it}]$ is decreasing in λ_i pointwise in t when $a_i(\cdot) \geq 0$; in particular, lower λ_i implies a strictly larger expected compliance gap at every point in the compliance cycle. Because a realized trade at i requires a Poisson access arrival, the probability that operator i trades in any interval of length Δ equals $1 - e^{-\lambda_i \Delta}$, which is monotonically increasing in λ_i. Consequently, operators with lower λ_i trade less often and have larger unresolved compliance gaps. Because $a_i(s_t, X_t)$ is larger for smaller s_t (the surrender date is approaching), conditional on eventual trading the expected share of purchases realized near surrender is also larger for operators with lower λ_i. This establishes Prediction 1. □

Derivation of Prediction 2 (Seasonal Price Pressure)

Combining (A.6) with the decomposition $M_t = M_t^{\text{comp}} + M_t^{\text{info}}$, and using the access-weighted aggregation $E_t[M_t^{\text{comp}}] = \eta_t G_t$, where $\eta_t \equiv \sum_i \alpha_i \lambda_i g_{it} / G_t$ is the effective access-weighted closing rate,

$$E_t[P_t - P_t^*(X_t)] = \phi \eta_t G_t + \phi E_t[M_t^{\text{info}}]. \tag{A.8}$$

Equation (A.5) implies that $E_t[G_t]$ is strictly larger in calendar subintervals where $A(s_t, X_t)$ is larger, in particular as $s_t \to 0$ from above. Hence $E_t[P_t - P_t^*(X_t)]$ is strictly positive near surrender whenever $\phi > 0$ and $\eta_t > 0$. Comparative statics in ϕ are immediate from (A.8). Comparative statics in scarcity and the allowance bank follow from the dependence of $A(s_t, X_t)$ on z_t and B_t: specifically, tighter cap (z_t high) or a smaller bank (B_t low) raises desired compliance demand $\sum_i a_i(s_t, X_t)$, and hence raises G_t near surrender. This establishes Prediction 2. □

Derivation of Prediction 3 (Informative Operator Flow)

Under the joint law of $(\mu_{it}, P_{t+\Delta}^* - P_t^*)$ implied by the signal structure, conditional expectations satisfy $E[P_{t+\Delta}^* - P_t^* \mid \mu_{it}] = c\,\mu_{it}$ with $c > 0$ whenever $\mathcal{I}_{it}$ contains a non-trivial signal about the future benchmark price. Because M_t^{info} in (A.7) is a linear combination of $\lambda_i \mu_{it} / \gamma_i$ across informed operators, $E[P_{t+\Delta}^* - P_t^* \mid M_t^{\text{info}}]$ is strictly increasing in M_t^{info}, and hence operator flow imbalance forecasts future returns. The expected contribution of operator i to M_t^{info} is $\lambda_i \mu_{it} / \gamma_i$, so that informative order flow generated by operator i scales linearly with its access intensity. Therefore the predictive content of M_t^{info} is larger for the frequent-trader subsample. Cross-sectional comparative statics in γ_i follow from $q_{it}^{\text{info}} = \mu_{it} / \gamma_i$. This establishes Prediction 3. □

The empirical tests use actual allowance returns, not benchmark-price innovations. Actual prices satisfy

$$P_t = P_t^*(X_t) + \phi M_t, \tag{A.9}$$

so that

$$P_{t+\Delta} - P_t = \left(P_{t+\Delta}^* - P_t^*\right) + \phi\left(M_{t+\Delta} - M_t\right). \tag{A.10}$$

The reduced-form return-predictability implication requires that the signal content of M_t^{info} is not fully incorporated into the contemporaneous price. In particular, if

$$E_t\left[P_{t+\Delta}^* - P_t^* \mid M_t^{\text{info}}\right] = \beta M_t^{\text{info}}, \tag{A.11}$$

and the expected reversal of contemporaneous price impact is smaller than the signal-revelation component, then $E_t[P_{t+\Delta} - P_t \mid M_t^{\text{info}}]$ is increasing in M_t^{info}. In the special case in which $E_t[M_{t+\Delta} \mid M_t^{\text{info}}]$ does not load on M_t^{info}, a sufficient condition for positive predictability of actual returns is $\beta > \phi$; more generally, the required condition is that contemporaneous price impact under-reveals the private signal relative to the subsequent adjustment in the benchmark price. This maintained reduced-form condition maps the model's informative-flow prediction into the empirical return-predictability tests. We do not derive this condition from a market-maker filtering or Kyle-style equilibrium; we treat it as a maintained reduced-form assumption.

Private Compliance Costs and Real Frictions

This subsection is not intended as a full welfare analysis. Its purpose is to clarify how the private compliance-cost measure used in the empirical section differs from aggregate resource costs. The empirical calculation measures the private cost borne by delayed buyers; aggregate resource costs depend on the underlying participation, adjustment, and immediacy-provision frictions.

A simple quadratic benchmark illustrates the resource-cost components associated with deviations from the frictionless allocation:

$$W = E_0 \int_0^\infty e^{-\rho t}\left[\sum_i \left(\kappa_i - \frac{1}{2}(\theta_{it} - x_{it})^2\right) - \sum_i a_{it} K_i - \frac{\phi}{2} M_t^2\right] dt, \tag{A.12}$$

where a_{it} indicates whether operator i participates at date t and $\rho > 0$ is a continuous-time discount rate. Relative to the frictionless allocation, the resource-cost components are (i) misallocation of abatement across firms and over time, (ii) real participation or attention costs, and (iii) intermediary immediacy-provision (or order-flow absorption) costs. The third component enters as an aggregate resource cost only to the extent that $\frac{\phi}{2} M_t^2$ represents a balance-sheet or risk-bearing cost borne by intermediaries; if instead ϕM_t is interpreted as pure transitory price impact, then $(P_t - P_t^*) M_t = \phi M_t^2$ is a transfer to intermediaries and the corresponding

aggregate resource cost is $\frac{\phi}{2}M_t^2$ rather than ϕM_t^2. The private costs borne by delayed buyers in April are measurable, but they are not direct measures of aggregate resource costs because allowance payments are partly transfers across market participants.

Closed-Form Seasonal Compliance Gap and Price Premium

This section displays one explicit closed-form object that underlies Prediction 2. Let $u \in [0,T]$ denote time within the compliance cycle, with surrender at $u = T$, and consider the piecewise-constant aggregate desired-compliance profile

$$A(u) = \begin{cases} a_0, & 0 \le u < T - \tau, \\ a_0 + a_1, & T - \tau \le u \le T, \end{cases} \tag{A.13}$$

with $a_1 > 0$ capturing the run-up in desired compliance adjustment near surrender and $\tau \in (0,T)$ the length of the surrender run-up. Set the aggregate gap dynamics to

$$\frac{dG(u)}{du} = A(u) - \bar{\lambda} G(u), \tag{A.14}$$

and, for expositional simplicity, take $G(0) = 0$. Equation (A.14) is a linear first-order ODE with piecewise-constant forcing, and its solution is

$$G(u) = \frac{a_0}{\bar{\lambda}} \left(1 - e^{-\bar{\lambda} u}\right), \qquad 0 \le u < T - \tau, \tag{A.15}$$

and, for $T - \tau \le u \le T$,

$$G(u) = G(T-\tau)\, e^{-\bar{\lambda}(u-(T-\tau))} + \frac{a_0 + a_1}{\bar{\lambda}} \left(1 - e^{-\bar{\lambda}(u-(T-\tau))}\right), \tag{A.16}$$

with $G(T-\tau) = \frac{a_0}{\bar{\lambda}}(1 - e^{-\bar{\lambda}(T-\tau)})$. Combining (A.6) with $M_t^{\text{comp}} = \eta G_t$ and abstracting from M_t^{info} in expectation delivers the seasonal-premium representation

$$P(u) - P^*(u) = \phi \eta G(u), \tag{A.17}$$

so that the price premium inherits the piecewise-closed-form structure of $G(u)$.

Equations (A.15)–(A.17) describe the pre-surrender accumulation of unresolved compliance demand within the cycle. Because surrender is mandatory, the remaining positive gap at $u = T^-$ is converted into terminal compliance order flow:

$$M_T^{\text{term}} = \sum_i g_{iT-}^{+}. \tag{A.18}$$

Under the same reduced-form immediacy-supply relation (A.6),

$$P_T - P_T^*(X_T) = \phi \left(M_T^{\text{term}} + M_T^{\text{info}} \right). \tag{A.19}$$

Thus, the closed-form expression for $G(T^-)$ provides the model counterpart of the surrender-month demand pressure documented in the data; the relevant object for terminal price pressure is the pre-compliance gap $G(T^-)$, not the post-compliance value $G(T)$. Equations (A.15)–(A.17) illustrate Prediction 2: the expected equilibrium premium is largest near surrender because aggregate unresolved compliance demand accumulates most rapidly when $A(u) = a_0 + a_1$, and the premium scales linearly with the product $\phi\eta$, which aggregates intermediary price impact and the access-weighted closing rate at which active operators close unresolved gaps. In this displayed example we treat η as constant for transparency; in the general formulation, the access-weighted closing rate η_t depends on the cross-sectional distribution of g_{it} and λ_i.

We emphasize the scope of this closed-form object. The affine benchmark price P_t^* and the aggregate compliance-gap law (A.5) are standard. The continuous-time model is a tractable reduced-form dynamic extension of the two-period mechanism in Section 2: $E_t[M_t^{\text{comp}}] = \eta_t G_t$ is an access-weighted aggregation step, not a fully micro-founded market-clearing condition, and the informed-flow block (A.7) is represented in conditional expected-flow form for tractability rather than as a solution to a filtering or equilibrium Kyle-style problem. The full periodic affine system with time-varying coefficients in B_t and z_t admits closed-form solutions in special limits (for instance, constant aggregate compliance demand together with mean-reverting z_t), but we do not attempt a general exact solution in the main appendix text: the displayed closed-form object above is the core tractable example we use to illustrate Prediction 2. An analogous smoother alternative is the linear ramp $A(u) = \alpha + \beta(1 - s_u/s_0)$ with $\beta > 0$ as surrender approaches, which delivers a qualitatively identical pattern.

A.6 Two-Period Model Derivations

This subsection provides the derivations underlying the simple model introduced in Section 2.

Frictionless Benchmark and Intermediation

Consider the two-period problem of firm i:

$$\begin{aligned}\max_{x_{i0},x_{i1},m_{i0},m_{i1},b_{i1}} &\left[\kappa_i - \frac{1}{2}(\theta_{i0} - x_{i0})^2 - p_0^* m_{i0}\right] \\ &+\delta E_0\left[\kappa_i - \frac{1}{2}(\theta_{i1} - x_{i1})^2 - p_1^* m_{i1}\right]\end{aligned} \tag{A.20}$$

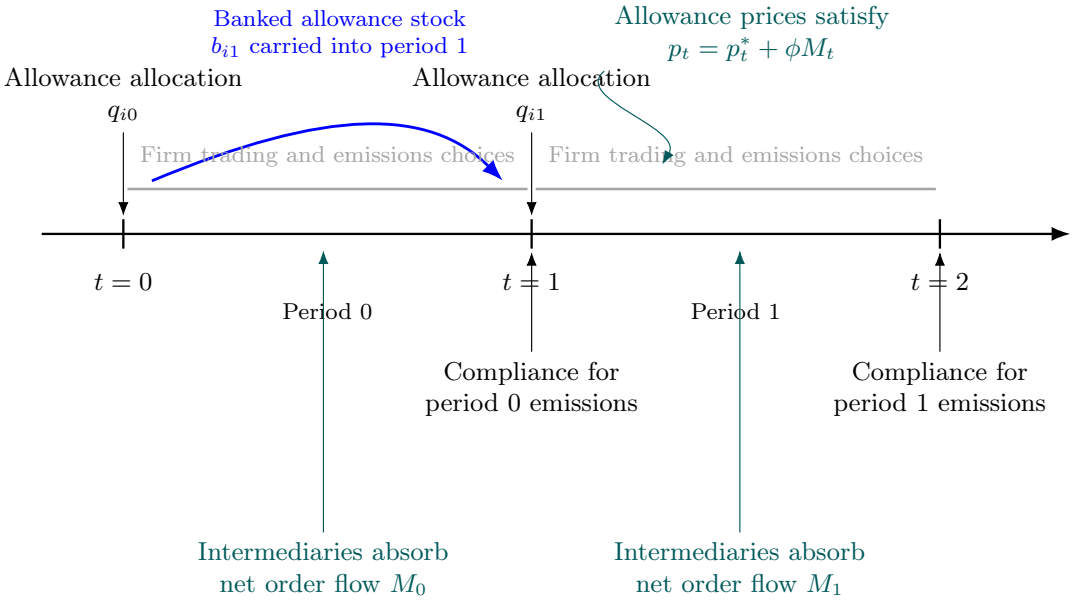
subject to

$$b_{i1} = b_{i0} + q_{i0} + m_{i0} - x_{i0}, \qquad x_{i1} = b_{i1} + q_{i1} + m_{i1}, \qquad b_{i1} \geq 0. \tag{A.21}$$

Figure A.2 illustrates the timing of the model. Allowances are allocated at the beginning of each period, compliance occurs at the end of each period, and unused allowances can be banked from period 0 to period 1. The figure also highlights that intermediaries absorb aggregate order flow and that prices can depart from frictionless benchmark prices when inventory capacity is limited.

Figure A.2: Timeline of the Two-Period Model with Intermediation

This figure illustrates the timeline of the two-period model. Allowances are allocated at the beginning of periods 0 and 1, compliance occurs at the end of periods 0 and 1, unused allowances can be banked from period 0 to period 1, and intermediaries absorb net order flow in each period. The figure also highlights the price-impact relation $p_t = p_t^* + \phi M_t$.



Let λ_{i0} and λ_{i1} denote the shadow values of one additional allowance in periods 0 and 1. The first-order conditions for an interior solution are

$$\theta_{i0} - x_{i0} = \lambda_{i0}, \tag{A.22}$$

$$\theta_{i1} - x_{i1} = \lambda_{i1}, \tag{A.23}$$

$$p_0^* = \lambda_{i0}, \tag{A.24}$$

$$p_1^* = \lambda_{i1}, \tag{A.25}$$

$$\lambda_{i0} = \delta E_0[\lambda_{i1}]. \tag{A.26}$$

Let $M_t = \sum_i m_{it}$ denote aggregate operator net demand. A representative competitive intermediary sector supplies immediacy: at price p_t it is willing to absorb a flow M_t^s that maximizes $(p_t - p_t^*)M_t^s - \frac{\phi}{2}(M_t^s)^2$,

taking $p_t - p_t^*$ as given. The associated inverse supply curve for immediacy is

$$p_t - p_t^* = \phi M_t^s, \tag{A.27}$$

where ϕM_t^s is the marginal cost of absorbing an additional unit of order flow. Market clearing at operator demand $M_t^s = M_t$ then yields the equilibrium pricing relation

$$p_t = p_t^* + \phi M_t. \tag{A.28}$$

The intermediary does not choose operator order flow; rather, ϕ traces the slope of the dealers' inverse supply curve, and the equilibrium condition $p_t = p_t^* + \phi M_t$ obtains by clearing. Combining these conditions yields the benchmark and price-impact equations reported in the main text:

$$\theta_{i0} - x_{i0} = \lambda_{i0}, \qquad \theta_{i1} - x_{i1} = \lambda_{i1}, \tag{A.29}$$

$$p_0^* = \lambda_{i0}, \qquad p_1^* = \lambda_{i1}, \tag{A.30}$$

$$p_0^* = \delta E_0[p_1^*], \tag{A.31}$$

$$p_t = p_t^* + \phi M_t. \tag{A.32}$$

Thus, in the absence of frictions, firms trade until marginal abatement costs are equalized across firms and over time. Extending the model to many periods is straightforward and yields the same Euler equation between adjacent dates, together with a standard transversality condition ruling out unbounded accumulation of unused allowances.

Participation Costs and Surrender-Month Trading

To derive the participation threshold, consider a firm expecting to be short by d_i allowances at the surrender date. If the firm becomes active at the early trading date, it pays $p_e d_i + K_i$. If instead it delays until April, its expected cost is $E_0[p_A]d_i$. Early trading is therefore optimal if and only if

$$p_e d_i + K_i \leq E_0[p_A]d_i, \tag{A.33}$$

which is equivalent to

$$d_i \left(E_0[p_A] - p_e \right) \geq K_i. \tag{A.34}$$

At each within-cycle date $\tau \in \{e, A\}$, intermediary absorption costs imply

$$p_\tau = p^*_\tau + \phi M_\tau. \tag{A.35}$$

Hence, if delayed compliance demand is concentrated at the surrender date, then M_A is large and the April price can exceed the early price even when the compliance wave is predictable.

Private Information and Speculative Trading

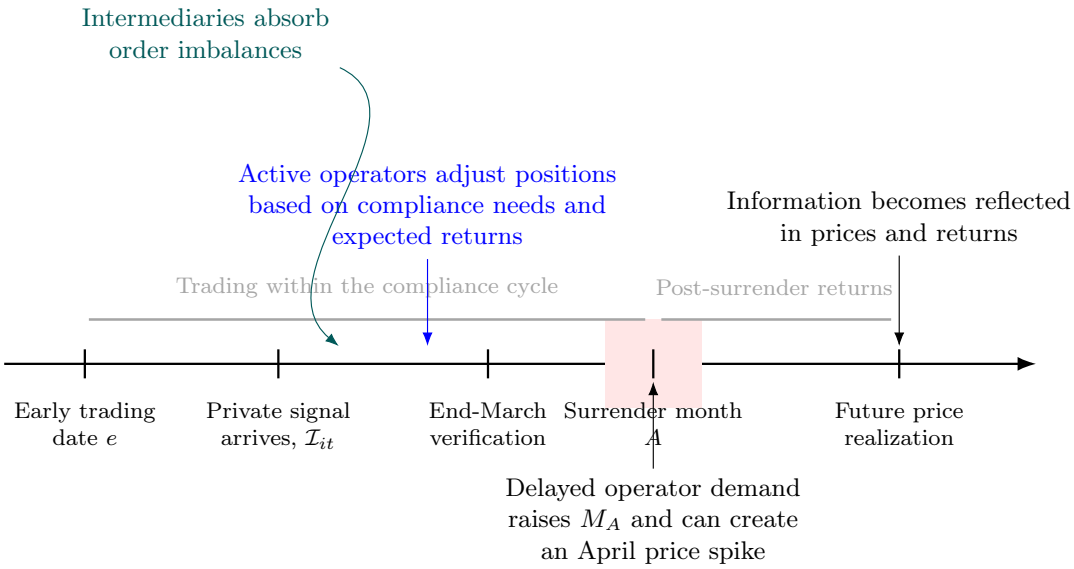
Suppose an active firm receives a private signal summarized by the information set $\mathcal{I}_{it}$ and forms the conditional expected excess return

$$\mu_{it} = E_t \left[p_{t+1} - p_t \mid \mathcal{I}_{it} \right]. \tag{A.36}$$

Figure A.3 illustrates the within-period timeline between the early trading date e and the surrender month A within a compliance cycle. It highlights the arrival of private information, the verification and surrender dates, the intermediary absorption of order imbalances, and the distinction between trading during the compliance cycle and subsequent return realization. This timing is particularly relevant for interpreting both the predictable April price pressure and the return patterns documented in the main text.

Figure A.3: Within-Period Timeline with Private Information

This figure illustrates the within-period timeline from the early trading date e to the surrender month A within a compliance cycle. It highlights the arrival of private information, the end-March verification date, the surrender date, intermediary absorption of order imbalances, and the subsequent realization of returns. Delayed operator demand raises the order imbalance absorbed by intermediaries at the surrender date and can generate an April price spike.



The firm chooses a speculative inventory position n_{it} to solve

$$\max_{n_{it}} \left\{ \mu_{it} n_{it} - \frac{\gamma_i}{2} n_{it}^2 \right\}. \tag{A.37}$$

The first-order condition is

$$\mu_{it} - \gamma_i n_{it} = 0, \tag{A.38}$$

so the optimal speculative position is

$$n_{it} = \frac{\mu_{it}}{\gamma_i}. \tag{A.39}$$

Aggregate operator order flow can therefore predict future returns when a subset of firms trades on private information.

Private Compliance Costs and Participation Frictions

This subsection is not intended as a full welfare analysis. Its purpose is to clarify how the private compliance-cost measure used in the empirical section differs from aggregate resource costs. The main text distinguishes between private compliance costs borne by regulated firms and the underlying real frictions. Private costs from delayed purchases are given by

$$\text{Cost}_i^{\text{delay}} = d_i(p_A - p_e), \tag{A.40}$$

which need not coincide with aggregate resource costs because the higher April price partly reflects a transfer to the counterparties selling allowances. When those counterparties are intermediaries, these private costs also represent transfers from regulated firms to financial participants outside the manufacturing sector. Because allowance payments are partly transfers, the resource-cost object in the main text strips them out and subtracts only the real immediacy-provision (or order-flow absorption) cost.

A simple quadratic benchmark illustrates the resource-cost components associated with deviations from the frictionless allocation. Let $\{x_{i0}^*, x_{i1}^*\}$ denote the frictionless allocation and $\{x_{i0}, x_{i1}\}$ the actual allocation. Assume that the cap fixes aggregate emissions in each period, so that $\sum_i x_{it} = \sum_i x_{it}^*$ for $t = 0, 1$. Under the quadratic benchmark, period-t abatement costs are

$$\frac{1}{2}\sum_i (\theta_{it} - x_{it})^2 . \tag{A.41}$$

Writing $x_{it} = x_{it}^* + \Delta x_{it}$ and using $\sum_i \Delta x_{it} = 0$, the difference in aggregate abatement costs between the actual and frictionless allocations is exactly

$$\frac{1}{2}\sum_i (\theta_{it} - x_{it})^2 - \frac{1}{2}\sum_i \left(\theta_{it} - x_{it}^*\right)^2 = \frac{1}{2}\sum_i (\Delta x_{it})^2 . \tag{A.42}$$

Summing over the two periods and adding the real participation costs and immediacy-provision costs gives the resource-cost components corresponding to the loss relative to the frictionless allocation reported in the main text:

$$L = \frac{1}{2}\sum_i (x_{i0} - x_{i0}^*)^2 + \frac{\delta}{2} E_0\left[\sum_i (x_{i1} - x_{i1}^*)^2\right] + \sum_i a_i K_i + \frac{\phi}{2}\left(M_0^2 + \delta E_0[M_1^2]\right). \tag{A.43}$$

Finally, the participation condition implies that if a firm delays trading, then

$$K_i \geq d_i \left(E_0[p_A] - p_e \right), \tag{A.44}$$

so the observed April trading costs provide a lower bound on the attention or participation frictions needed to rationalize delayed trading.