

NBER WORKING PAPER SERIES

TRADING FRICTIONS IN DYNAMIC CAP-AND-TRADE MARKETS

Nicola Borri
Yukun Liu
Aleh Tsyvinski
Xi Wu

Working Paper 35356
<http://www.nber.org/papers/w35356>

NATIONAL BUREAU OF ECONOMIC RESEARCH
1050 Massachusetts Avenue
Cambridge, MA 02138
June 2026

The empirical analysis in this paper subsumes that of “Inefficiencies of Carbon Trading Markets” first posted on arXiv on August 12, 2024. We thank Bruno Biais, Felix Bierbrauer, and Lasse Pedersen for comments. The views expressed herein are those of the authors and do not necessarily reflect the views of the National Bureau of Economic Research.

NBER working papers are circulated for discussion and comment purposes. They have not been peer-reviewed or been subject to the review by the NBER Board of Directors that accompanies official NBER publications.

© 2026 by Nicola Borri, Yukun Liu, Aleh Tsyvinski, and Xi Wu. All rights reserved. Short sections of text, not to exceed two paragraphs, may be quoted without explicit permission provided that full credit, including © notice, is given to the source.

Trading Frictions in Dynamic Cap-and-Trade Markets
Nicola Borri, Yukun Liu, Aleh Tsyvinski, and Xi Wu
NBER Working Paper No. 35356
June 2026
JEL No. H0

ABSTRACT

We develop a dynamic stochastic model of markets with an externality and multiple trading frictions, and cap-and-trade as the leading application. Slow participation, limited intermediation, and heterogeneous information interact in equilibrium: agents choose costly market access, access determines residual compliance demand, intermediary constraints translate residual demand into a surrender-month premium, and the premium feeds back into access incentives. These interactions shape how effectively the market corrects the externality. We characterize access choices in closed form, prove that the equilibrium premium is unique, and show that endogenous access dampens the response to each friction in isolation, while the interaction of multiple frictions is non-additive and can amplify the price response. We quantify the model using 2.7 million EU ETS registry transactions and compliance records from 2005–2021. About 40% of operators do not trade annually, purchases concentrate in April when returns are systematically high, and operator flow predicts future returns.

Nicola Borri
Luiss University
nborri@luiss.it

Yukun Liu
University of Rochester
yliu229@simon.rochester.edu

Aleh Tsyvinski
Yale University
Department of Economics
and NBER
a.tsyvinski@yale.edu

Xi Wu
University of California, Berkeley
Haas School of Business
xiwu@haas.berkeley.edu

1 Introduction

Carbon emissions are the classic example of a climate externality. The leading real-world implementation of a market-based corrective instrument is the cap-and-trade system, in which agents can trade allowances to emit carbon (Dales, 1968; Montgomery, 1972; Rubin, 1996). By making emission rights tradable and bankable, a cap-and-trade system creates an asset market: the cap determines aggregate scarcity and the secondary allowance market must reallocate compliance obligations across firms and over time and aggregate information about scarcity. The central question of this paper is how an externality can be corrected with a market-based policy when the market has multiple frictions. And more specifically, we study how multiple interacting frictions shape one of the most prominent practical mechanism to curb a climate externality.

In canonical Pigouvian settings, of which climate is a leading example, an optimal corrective instrument can target the externality directly (Weitzman, 1974; Nordhaus, 1992). Golosov, Hassler, Krusell and Tsyvinski (2014) show that in a wide class of climate models the optimal correction admits a simple closed-form solution which can be implemented either as a carbon tax or as a cap-and-trade system. Barrage (2020) and Kübler (2025) characterize how this unconstrained benchmark is modified by distortionary fiscal policy and by incomplete financial markets, respectively. When multiple frictions are present, as in cap-and-trade markets, designing and analyzing a corrective instrument is substantially harder. The main challenge is that the joint effect of the corrective instrument on prices and quantities need not equal the sum of the separate effects of each friction, as in the classic work by Lipsey and Lancaster (1956), and the interactions of the frictions may amplify or dampen the size of the response. In this paper we study the effectiveness of the Pigouvian correction theoretically, empirically, and quantitatively in a dynamic stochastic model with multiple interacting frictions.

We study this question in the European Union Emissions Trading System (EU ETS), the leading cap-and-trade market, covering about 40% of the EU's greenhouse-gas emissions. Recent climate-finance research treats climate risk, carbon pricing, and green finance as important objects for asset pricing and market design (Engle, Giglio, Kelly, Lee and Stroebe, 2020; Giglio, Kelly and Stroebe, 2021). An important paper by Pedersen (2026) compares carbon pricing and green finance as alternative tools for the climate transition and emphasizes the central role of price-based instruments. Känzig (2023) shows that restrictive carbon policy shocks in Europe reduce emissions and affect aggregate economic activity, which is evidence that carbon prices have real effects. Hong, Wang and Yang (2023) model climate disasters as Poisson arrivals and show that adaptation and capital taxes complement carbon pricing in the optimal climate policy mix. Other recent work studies financial dimensions of cap-and-trade from complementary angles: Biaies, Hombert, Schmidt and Weill (2026) examine the spot-futures wedge under tran-

sition risk with imperfect hedging and financial constraints, and [Biais and Landier \(2026\)](#) study how emission caps interact with firms' green-technology investment under government commitment. We develop a dynamic stochastic model of cap-and-trade with three important trading frictions that interact in equilibrium—slow participation, limited intermediation, and heterogeneous information—and bring it to the data using the universe of EU ETS registry transactions and compliance records from the European Union Transaction Log (EUTL), about 2.7 million observations over 2005–2021. The paper's main contribution is to characterize and quantify the interactions of multiple frictions—endogenous market access, intermediary absorption constraints, and informative trading—and their amplification and dampening effects on the equilibrium price response.

To organize the analysis, we start from the frictionless benchmark underlying the classic case for cap-and-trade. In a market with bankable allowances and no trading frictions, firms 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.

When multiple interactions are present, the frictions themselves are endogenous in equilibrium, and the interaction among these frictions makes the analysis much more challenging. Costly market access makes participation endogenous; endogenous access determines how much compliance demand remains unresolved near the surrender deadline; residual terminal demand creates price pressure when intermediary capacity is limited; and expected price pressure feeds back into firms' incentives to acquire access before the deadline. The same access margin determines which operators can act on private information, so the model links participation, price pressure, and return predictability through a single dynamic mechanism. The interaction runs in two directions. Dampening occurs when agents' endogenous access responses partially offset the direct effect of a policy that reduces one friction, so the equilibrium premium moves less than the policy's mechanical impact would imply. Amplification occurs when one friction makes the equilibrium premium more responsive to another, so the price effect of a given friction is larger when other frictions are more severe. A static analysis would predict price responses that reflect each friction's direct effect alone. We show that in our quantitative model, calibrated to the universe of transactions of the market, the dominant effect is dampening: halving intermediary return impact lowers the equilibrium premium by about 38% rather than the 50% a static demand-pressure model would predict, because endogenous access responses raise residual terminal demand by about 24%. The interaction can also amplify the price effect of one friction in the presence of another: when access costs are higher, the access response is weaker, and the equilibrium premium becomes more sensitive to intermediary capacity. In our calibration, doubling access costs raises the elasticity of the

equilibrium premium to intermediary return impact from about 0.69 to 0.72.

We first illustrate the mechanism of endogenous interacting frictions in a simple two-period model and then generalize it in a continuous-time framework. The continuous-time model remains analytically tractable despite the interaction of the three frictions, delivering a closed-form access rule, a unique fixed point for the surrender-month premium, and a comparative static showing that endogenous access dampens the price effects of intermediary constraints, with the elasticity of the equilibrium premium to intermediary return impact bounded strictly in $(1/2, 1)$: dampening is strongest when access frictions are small and weakest when they are large. The model yields three empirical predictions—persistent non-trading, surrender-month demand pressure, and informative operator flow—which we test with the EU ETS data.

The EUTL data cover the universe of transactions and compliance information currently available, spanning February 2005 to September 2021. They cover 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 this market: regulated firms, intermediaries, and regulators. They are largely represented in the EU ETS 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 allow us to observe how regulated firms interact with the secondary allowance market: whether they participate, when they trade, with whom they trade, and whether their order flow predicts future returns.

The model also allows us to quantify the magnitude of these frictions and to evaluate market-design counterfactuals. Trading-frequency moments discipline firms' access intensities, and the relation between April return premia and model-implied terminal demand disciplines surrender-window return impact. The resulting model-implied terminal demand has a correlation of 0.95 with observed April compliance purchases at the year level. Because access is endogenous, the counterfactual effects involve the interaction of these frictions: changing intermediary capacity or the timing of surrender obligations changes both the direct price impact of terminal demand and firms' incentives to access the market before the deadline. Under the one-shot access specification used in the quantitative exercise, our main policy counterfactual in which surrender deadlines are staggered across four equally sized groups reduces the surrender-month premium by about 59% and the implied April premium paid by delayed buyers (the additional cost of purchasing allowances in April rather than earlier in the cycle) by about 42%, with no change in aggregate emissions or the cap.

The model also quantifies how the access and intermediary frictions interact. Halving intermediary return impact alone lowers the equilibrium surrender-month premium by about 38%, rather than the 50% a static demand-pressure model would predict, because endogenous access responses raise residual terminal demand. Halving the access cost alone lowers the premium by about 20%,

but halving both frictions jointly lowers it by exactly 50%: the joint effect is not the sum of the separate effects. When the access-cost friction is doubled relative to the baseline calibration, the elasticity of the equilibrium surrender-month premium with respect to intermediary capacity rises from 0.69 to 0.72, illustrating how the interaction can amplify the response to a single friction. In euro terms, the baseline implied April premium paid by delayed buyers is about €176 million per cycle and falls to roughly €88 million under the joint counterfactual and €101 million under the staggered-deadline policy.

We then test the three model predictions in the data. 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 falls to 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, Convery and Di Maria, 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.

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 a sizable April premium paid by delayed buyers. The EU ETS market operates on a fixed verification and surrender schedule each year. By the end of April, firms must surrender a sufficient number of emission allowances to the administrative 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 roughly ten times larger than in typical non-April months. This concentration of compliance demand is associated with systematically high April returns, averaging about 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 due imminently and the price of emission allowances is predictably high. We estimate the implied April premium paid by delayed buyers from this surrender pattern at about €5 billion, relative to purchasing allowances earlier at lower prices. This number should be interpreted as a financial premium paid in April rather than as a direct measure of aggregate resource costs, because part of the higher April price is paid to counterparties. 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 trades display 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. Their behavior is consistent with informational advantages about future scarcity and prices: they buy emission allowances ahead of price increases and sell them ahead of price declines. Comparing their net purchases and sales to subsequent price movements implies 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—a proxy for operators with higher chosen market-access intensity and lower effective access costs—whose order flow reflects informational advantages not yet fully impounded in prices. This return-predictability pattern implies roughly €8 billion in cumulative ex post trading performance over the sample period.

The paper contributes to three strands of the literature. The first studies climate finance, sustainable investing, and the pricing of climate-related risks. Recent work shows that climate exposures and carbon-transition risks are reflected in asset prices and portfolio choices (Sautner, Van Lent, Vilkov and Zhang, 2023; Bolton and Kacperczyk, 2023; Croce, Guinez, Inzunza Méndez, Nguyen and Tebaldi, 2025). Related work studies how sustainability preferences and ESG characteristics enter portfolio choice and equilibrium expected returns (Pástor, Stambaugh and Taylor, 2021; Pedersen, Fitzgibbons and Pomorski, 2021). Our paper differs by studying the secondary allowance market at the core of cap-and-trade—the leading real-world setting in which carbon prices are formed through trading. Rather than asking how climate risk is priced in equities or portfolios, we study how trading frictions and their interaction shape participation, intermediation, and return predictability in the secondary market for emission allowances.

The second strand studies climate economics, carbon pricing, and market-based environmental policy. Modern climate-macro models characterize optimal carbon taxation and the general-equilibrium effects of fossil-fuel use (Nordhaus, 1992; Golosov, Hassler, Krusell and Tsyvinski, 2014; Barrage, 2020; Kübler, 2025; Bierbrauer, Polborn, Ritterrath and Weizsäcker, 2026). Recent empirical work studies the real effects of carbon pricing: Colmer, Martin, Muûls and Wagner (2025) show that the EU ETS reduced regulated firms’ emissions without detectable contractions in economic activity. Recent theoretical and quantitative work studies climate damages, social costs of carbon, macroeconomic climate modeling, and directed innovation in the energy transi-

tion (Acemoglu, Aghion, Barrage and Hémous, 2024; Barrage and Nordhaus, 2024; Fernández-Villaverde, Gillingham and Scheidegger, 2025; Folini, Friedl, Kübler and Scheidegger, 2025). In cap-and-trade, Biais, Hombert, Schmidt and Weill (2026) study imperfect hedging, financial constraints, delayed spot purchases, and the spot-futures wedge under transition risk, while Biais and Landier (2026) study how emission caps interact with firms’ investment in green technologies and government commitment. We take the cap-and-trade system as given and study the secondary market mechanism through which permits are traded, compliance demand is reallocated, and scarcity information is incorporated into prices.

The third strand studies trading frictions and information in asset markets. We build on models of slow-moving capital and search frictions (Duffie, 2010; Duffie, Gârleanu and Pedersen, 2007), constrained intermediation and demand pressure (Acharya, Lochstoer and Ramadorai, 2013; Gârleanu, Pedersen and Poteshman, 2009; Hendershott and Menkveld, 2014), and heterogeneous information and informative order flow (Kyle, 1985; Glosten and Milgrom, 1985; Subrahmanyam, 1991). Our contribution is to characterize how these frictions interact in a dynamic cap-and-trade setting: endogenous market access determines participation and residual compliance demand, limited intermediary capacity turns residual demand into surrender-month price pressure, and heterogeneous information makes operator order flow predictive of future allowance returns. Combined in a single equilibrium, these mechanisms generate the feedback between access choices and surrender-month pressure that organizes our empirical results. Our intermediation channel parallels the funding-liquidity mechanism of Brunnermeier and Pedersen (2009), in which intermediary capacity and asset prices are jointly determined in equilibrium.

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

This section develops a simple two-period model of a cap-and-trade market with trading 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 following Harstad and Eskeland (2010), combined with ingredients motivated by finance models of constrained intermediation, order imbalances, and informed trading (Acharya, Lochstoer and Ramadorai, 2013; Gârleanu, Pedersen and Poteshman, 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

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

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 Online 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

$$\Pi_{it} = -\frac{1}{2}(\theta_{it} - x_{it})^2 - p_t m_{it}, \quad (1)$$

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

$$b_{i1} = b_{i0} + q_{i0} + m_{i0} - x_{i0}, \quad 0 = b_{i1} + q_{i1} + m_{i1} - x_{i1}, \quad b_{i1} \geq 0, \quad (2)$$

where q_{it} denotes the quantity of allowances allocated to the firm. Let $M_t = \sum_i m_{it}$ denote aggregate operator net order flow. Thus, $M_t > 0$ 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. There exist intermediaries that provide immediacy by absorbing this order flow. We represent their costs in reduced form by 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, following [Hendershott and Menkveld \(2014\)](#), rather than as a literal inventory-holding cost.

In the frictionless benchmark, $\phi = 0$, participation costs are zero, and no trader possesses private information². Let ξ_{i0} and ξ_{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

$$\theta_{i0} - x_{i0} = \xi_{i0}, \quad \theta_{i1} - x_{i1} = \xi_{i1}, \quad (3)$$

$$p_0^* = \xi_{i0}, \quad p_1^* = \xi_{i1}, \quad (4)$$

$$p_0^* = \delta E_0[p_1^*]. \quad (5)$$

Equations (3)–(5) imply that, absent frictions, marginal abatement costs are equalized across firms

²Throughout 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.

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)–(??) hold for an interior banking solution. When the no-borrowing corner $b_{i1} = 0$ binds, the Euler equation (5) holds with inequality because the firm would like to shift additional allowance resources from the future to the present but cannot run a negative bank. Current allowances are then especially valuable at the margin because the firm has no spare bank to draw down. Empirically, this asymmetry is most relevant for firms that approach surrender with a positive net compliance gap after accounting for any banked allowances: they may use existing banked permits to smooth compliance, but they cannot borrow from future allocations to avoid deadline-month purchases. In that sense, the no-borrowing constraint can reinforce the surrender-month demand pressure formalized below. This force is distinct from costly market access: a binding no-borrowing corner raises the value of current-period allowances for firms with positive compliance gaps. Surrender-month pressure arises when such gaps remain unresolved because access is costly or adjustment is delayed. 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

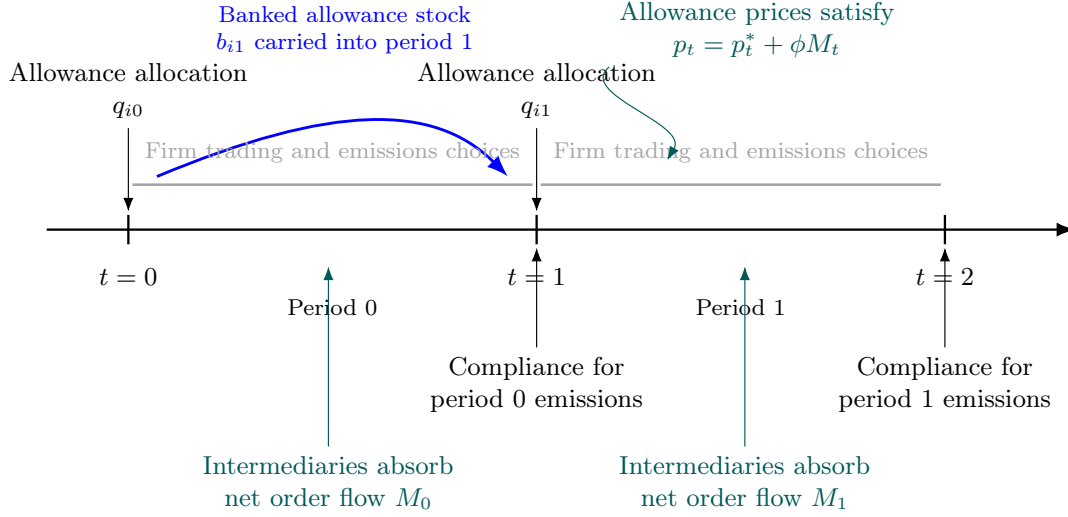
$$p_t = p_t^* + \phi M_t. \quad (6)$$

Equation (6) says that prices can depart from frictionless benchmark prices when intermediaries must absorb 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 mechanism 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. Figure 1 provides a high-level summary of the timing of the two-period setup and the role of intermediation.

Figure 1: Timeline of the Two-Period Model with Intermediation.

This figure illustrates the timing of the two-period model used in the main text. 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. In the EU ETS, the compliance date corresponds to the April compliance deadline. The figure also highlights the price-impact relation $p_t = p_t^* + \phi M_t$.



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 , which corresponds to the April compliance deadline in the EU ETS. Figure A.2 in the Online Appendix provides a more detailed within-cycle timeline, including early trading, verification, and surrender. Let $d_i \equiv \max\{\text{expected allowance shortfall}_i, 0\}$ denote firm i 's expected positive shortfall at the surrender date, that is, the expected quantity of allowances it still needs to acquire to meet compliance after accounting for allocations and any banked permits. 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

$$d_i(E_0[p_A] - p_e) \geq K_i. \quad (7)$$

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.³ 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. Because intermediaries absorb net order flow subject to immediacy-provision costs, prices at each within-cycle date $\tau \in \{e, A\}$ satisfy

$$p_\tau = p_\tau^* + \phi M_\tau. \quad (8)$$

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, Pedersen and Poteshman, 2009; Hendershott and Menkveld, 2014; Gabaix and Maggiori, 2015). This within-cycle mechanism generates the delayed-purchase prediction: firms with positive expected shortfalls may wait until the surrender month. Together with costly market access, it also helps rationalize persistent non-trading by firms whose expected shortfalls are zero, covered by banked allowances, or too costly to rebalance through discretionary trading. The continuous-time model below endogenizes this costly market access and links low chosen access intensity to persistent non-trading and residual terminal demand. Either way, the resulting compliance buying in April can be mapped into the April premium paid by delayed buyers, which we define in Section 2.3 and quantify later in the empirical analysis.

The fixed cost K_i should be interpreted as a reduced-form representation of costly market access. In the continuous-time model below, we represent this friction by allowing operators to choose a Poisson market-access intensity at a convex attention or treasury-management cost. The two-period condition captures the early-versus-surrender timing tradeoff, while the continuous-time model endogenizes the speed at which firms access the market during the compliance cycle.

³Equivalently, 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. 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.

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 $\mathcal{F}_t$ denote the public information set at time t , and let $\mathcal{I}_{it} \supseteq \mathcal{F}_t$ denote firm i 's information set, which may also contain private information. Define μ_{it} as the private component of expected benchmark returns,

$$\mu_{it} = E_t[p_{t+1}^* - p_t^* | \mathcal{I}_{it}] - E_t[p_{t+1}^* - p_t^* | \mathcal{F}_t], \quad (9)$$

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. The same appendix figure (Figure A.2) also indicates the arrival of private information and the subsequent realization of returns. For tractability, this speculative component uses a maintained price-taking simplification: firms optimize with respect to expected benchmark returns, rather than internalizing any effect of their own speculative orders on actual prices through the intermediary-supply specification. Taking prices as given, an active firm can, in addition to its compliance-motivated position, take a speculative inventory position n_{it} and solves

$$\max_{n_{it}} \left\{ \mu_{it} n_{it} - \frac{\gamma_i}{2} n_{it}^2 \right\}, \quad (10)$$

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

$$n_{it} = \frac{\mu_{it}}{\gamma_i}. \quad (11)$$

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 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 April Premium Paid by Delayed Buyers

The two-period model clarifies the April premium paid by regulated firms that delay compliance purchases. If firm i expects to be short by d_i allowances at the surrender date, its expected

additional payment from delaying purchases until April is

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

Aggregating across delayed buyers yields the back-of-the-envelope 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 April premium paid 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 of Cap-and-Trade Market with Trading Frictions

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 generalizes the same mechanism in a continuous-time framework designed to discipline the timing and return-predictability implications. 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 results in a seasonal benchmark and transparent comparative statics beyond that benchmark.⁴

The relationship between the two-period model and the continuous time is direct. The high participation cost of the two-period model becomes an endogenous Poisson market-access intensity: in simple terms, firms do not rebalance continuously, but instead choose how often to receive trading opportunities, balancing the value of earlier adjustment against attention or treasury costs. A lower chosen 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.

⁴Full derivations and proofs are in the Online Appendix.

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

$$X_t = (B_t, z_t, s_t), \quad (12)$$

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,

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

which obtains under affine state dynamics and linear marginal compliance value, as implied by the quadratic-cost benchmark in Section 2, 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 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 Endogenous Poisson Market Access

The key departure from the frictionless benchmark is that operators adjust compliance positions only intermittently. Rather than treating market-access intensity as exogenous, we represent slow participation endogenously by letting operators choose, at a convex attention or treasury-management cost, how often they receive discretionary trading opportunities during the compliance cycle. Let c_{it} denote operator i 's *compliance allowance holdings* and h_{it}^* the operator's target compliance holdings, which reflect expected end-of-cycle obligations and any optimal banking motive.

Endogenous access choice. At the beginning of each compliance cycle, operator i chooses a market-access intensity λ_i for the cycle. Conditional on this choice, discretionary trading opportunities arrive according to a Poisson process with intensity λ_i . Let τ denote the time between the access-choice date and the surrender deadline, $d_i \geq 0$ the operator's expected compliance shortfall, and let π_A denote the expected surrender-month return premium relative to the frictionless benchmark. For expositional simplicity, we normalize the expected frictionless return between the early trading date and the surrender month to zero and work in return units relative to the early-trading

price level:

$$\pi_A \equiv E_0 \left[\frac{P_A - P^*(X_A)}{P_e} \right] = E_0 \left[\frac{P_A - P_e}{P_e} \right] \approx E_0 [\log P_A - \log P_e], \quad (14)$$

where $X_A = (B_A, z_A, s_A)$ denotes the public state in the surrender month, $P^*(X_A)$ is the corresponding frictionless benchmark price, P_e is the relevant early-trading price, and P_A is the surrender-month price. To keep the access-choice problem on the same scale, let $\tilde{\psi}_i \equiv \psi_i/P_e$ denote the normalized access-cost parameter. The private value of obtaining at least one discretionary access opportunity before surrender—and so avoiding terminal compliance buying for the expected shortfall—is

$$S_i = d_i \pi_A. \quad (15)$$

The expression $S_i = d_i \pi_A$ treats any access opportunity during $[0, \tau]$ as enabling early purchase at the frictionless benchmark price level. A richer timing model would allow the value of access to decline as the access arrival approaches the surrender date; we leave that extension to future work. The operator chooses λ_i to solve

$$\max_{\lambda_i \geq 0} \left\{ S_i (1 - e^{-\lambda_i \tau}) - \frac{\tilde{\psi}_i}{2} \lambda_i^2 \right\}, \quad (16)$$

where $\tilde{\psi}_i > 0$ indexes the operator's normalized cost of market access, attention, and treasury management. The first-order condition is

$$S_i \tau e^{-\lambda_i \tau} = \tilde{\psi}_i \lambda_i, \quad (17)$$

and the unique solution is, using the Lambert W function⁵,

$$\lambda_i^* = \frac{1}{\tau} W \left(\frac{S_i \tau^2}{\tilde{\psi}_i} \right) = \frac{1}{\tau} W \left(\frac{d_i \pi_A \tau^2}{\tilde{\psi}_i} \right). \quad (18)$$

Since W is increasing on $\mathbb{R}_+$, the chosen intensity λ_i^* is increasing in d_i and π_A , and decreasing in $\tilde{\psi}_i$.

The solution makes slow participation endogenous. Firms with larger expected shortfalls or higher expected surrender-month premia choose higher access intensity, while firms facing a higher normalized access-cost parameter choose lower access intensity. Low-access firms are more likely to leave compliance needs unresolved until the surrender month. The chosen intensity λ_i^* is the

⁵The Lambert W function is the inverse of $x \mapsto x e^x$ on $\mathbb{R}_+$, so if $y e^y = z$, then $y = W(z)$.

dynamic counterpart of the fixed participation cost K_i in the two-period model: a higher $\tilde{\psi}_i$ generates a larger effective K_i , whereas a higher value of access $S_i = d_i \pi_A$ generates a smaller effective K_i . Empirically, however, λ_i^* should be interpreted as a reduced-form effective-access index rather than as a literal structural access parameter for every operator-year: outside the positive-shortfall buyer subset emphasized here, observed non-trading can reflect both limited access and a weak motive to trade.

Compliance-holdings dynamics under endogenous access. Compliance positions can be re-balanced only when operator i 's Poisson counting process for access opportunities, N_{it} , jumps. Upon access, the operator resets compliance holdings to the desired level:

$$dc_{it} = (h_{it}^* - c_{it-}) dN_{it}, \quad E_t[dN_{it}] = \lambda_i^* dt. \quad (19)$$

Here c_{it-} denotes compliance holdings immediately before the access jump at time t . The intensity λ_i^* is constant within a compliance cycle and is reset at the beginning of each cycle through the access-choice problem in (16), as $(d_i, \pi_A, \tilde{\psi}_i)$ realize for that cycle.⁶

Define the operator-level compliance gap as

$$g_{it} = h_{it}^* - c_{it}. \quad (20)$$

Speculative or information-motivated positions, introduced in Section 3.5 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, away from the terminal surrender date, unresolved compliance demand at the operator level obeys the linear adjustment equation

$$E_t[dg_{it}] = E_t[dh_{it}^*] - \lambda_i^* g_{it} dt. \quad (21)$$

⁶Allowing operators to close only a fraction $\alpha_i \in (0,1]$ of the gap upon access replaces λ_i^* by $\alpha_i \lambda_i^*$ in the expected gap dynamics and residual-demand formula. We use full reset in the main text to keep the endogenous-access mechanism transparent.

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

$$E_t[dG_t] = A(s_t, X_t) dt - \sum_i \lambda_i^* g_{it} dt = A(s_t, X_t) dt - \bar{\lambda}_t^* G_t dt, \quad \bar{\lambda}_t^* \equiv \frac{\sum_i \lambda_i^* g_{it}}{G_t}, \quad (22)$$

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 chosen access intensity. Aggregate adjustment is slower when many operators have high normalized access costs $\tilde{\psi}_i$, small expected shortfalls d_i , or low perceived value of early trading π_A . When desired compliance demand rises toward surrender and operators adjust only gradually, unresolved demand accumulates and is largest near surrender. For the closed-form fixed-point specification and the quantitative implementation, $\bar{\lambda}_t^*$ is evaluated on the positive-shortfall buyer subset $\{i : d_i > 0\}$, so that the denominator G_t is the aggregate positive gap and the gap-weighted average behaves as a proper mean. Outside that subset, the signed-gap aggregation should be read as motivating notation rather than as the object used in the calibration.

Terminal compliance and residual demand. Conditional on the cycle-level choice λ_i^* , the realized compliance gap evolves in continuous time during the cycle. Because surrender is mandatory, any remaining positive compliance shortfall must be closed at the surrender date through a terminal compliance trade. Under full reset upon access (that is, $\alpha_i = 1$), the probability that operator i receives no discretionary access opportunity before surrender is $e^{-\lambda_i^* \tau}$, and the corresponding expected residual terminal demand for operator i is

$$R_{iA} = d_i e^{-\lambda_i^* \tau}. \quad (23)$$

For the closed-form fixed-point specification and the quantitative calibration below, we specialize the richer dynamic environment to a one-shot known-shortfall object at the cycle level: each operator enters the cycle with a single expected shortfall d_i , and terminal demand depends on whether at least one discretionary access opportunity arrives before surrender. Aggregating across operators then yields the expected residual terminal demand

$$D_A(\pi_A) = \sum_i d_i e^{-\lambda_i^*(\pi_A) \tau}, \quad (24)$$

where $\lambda_i^*(\pi_A)$ is the function defined by (18). In the broader stochastic target-process environment, expected residual gaps would also depend on the initial gap and the drift of desired holdings; the appendix aggregation discussion reports that more general law of motion. In that richer environment, aggregate unresolved gaps can be signed because some operators may temporarily hold

excess allowances, but terminal compliance demand is based on positive residual gaps only. If instead one tracks the realized residual gap $g_{iA-} = h_{iA}^* - c_{iA-}$ immediately before surrender, the terminal compliance trade is $\Delta c_{iA} = g_{iA-}^+$ and aggregate terminal order flow is

$$M_A^{\text{term}} = \sum_i g_{iA-}^+, \quad (25)$$

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

$$P_A - P^*(X_A) = \frac{\phi}{h_A} \left(M_A^{\text{term}} + M_A^{\text{info}} \right). \quad (26)$$

We treat the surrender date as a windowed event of length h_A : M_A^{term} and M_A^{info} in (26) are stock-equivalent integrated flows over the surrender window, and ϕ in (26) is the same instantaneous price-level impact parameter as in (32). Dividing by h_A converts the integrated terminal flow into the average flow rate over the window, against which ϕ acts as in (32). Endogenous access determines how much demand remains unresolved at the surrender date: lower normalized access costs $\tilde{\psi}_i$ or larger benefits of early trading S_i raise λ_i^* , reduce expected residual terminal demand R_{iA} , and weaken surrender-month price pressure. To connect this residual demand to the return-based object used in the quantitative section, let $\varphi > 0$ denote the reduced-form surrender-window return-impact parameter, distinct from the instantaneous price-level impact parameter ϕ in (32). Specifically, $\varphi = \phi / (P_e h_A)$, where P_e is the early-cycle price level used to define π_A and h_A is the surrender-window length introduced after (26). The equilibrium surrender-month return premium then satisfies $\pi_A = \varphi D_A(\pi_A)$. 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. This fixed-point specification is deliberately narrower than the richer stochastic target-process environment introduced above: it is a buyer-side reduced-form representation for operators with expected positive shortfalls $d_i > 0$, used to characterize residual surrender demand and to organize the quantitative calibration.

Endogenous access and surrender-month pressure. The endogenous-access mechanism delivers the following closed-form characterization, which states the optimal access intensity, its comparative statics, and the existence and uniqueness of the equilibrium surrender-month premium.

Proposition 1. *Suppose operator i has expected compliance shortfall $d_i \geq 0$, normalized access-cost parameter $\tilde{\psi}_i > 0$, and chooses a cycle-level Poisson market-access intensity λ_i over horizon*

$\tau > 0$. If the expected surrender-month return premium is $\pi_A \geq 0$, the optimal access intensity is

$$\lambda_i^*(\pi_A) = \frac{1}{\tau} W\left(\frac{d_i \pi_A \tau^2}{\tilde{\psi}_i}\right), \quad (27)$$

where $W(\cdot)$ is the Lambert W function. The intensity $\lambda_i^*(\pi_A)$ is increasing in d_i and π_A , and decreasing in $\tilde{\psi}_i$. Under full reset upon access (that is, $\alpha_i = 1$), expected aggregate residual terminal demand is

$$D_A(\pi_A) = \sum_i d_i e^{-\lambda_i^*(\pi_A) \tau}. \quad (28)$$

If the surrender-month return premium satisfies the reduced-form return-impact relation

$$\pi_A = \varphi D_A(\pi_A), \quad (29)$$

with $\varphi > 0$, then whenever $\sum_i d_i > 0$ the map $\pi_A \mapsto \varphi D_A(\pi_A)$ is continuous and strictly decreasing, satisfies $\varphi D_A(0) = \varphi \sum_i d_i > 0$, and converges to zero as $\pi_A \rightarrow \infty$. Hence there exists a unique positive equilibrium return premium solving (29).

Corollary 1. *At the equilibrium of Proposition 1, whenever $\sum_i d_i > 0$ the elasticity of the surrender-month return premium with respect to the surrender-window return-impact parameter is bounded strictly below unity:*

$$\frac{d \log \pi_A}{d \log \varphi} = \frac{1}{1 + \varphi |D'_A(\pi_A)|} \in (0, 1), \quad (30)$$

where $D'_A(\pi_A) < 0$ because endogenous access $\lambda_i^*(\pi_A)$ is strictly increasing in π_A .

Corollary 1 formalizes the *dampening* feature emphasized in the introduction: endogenous access reabsorbs a fraction of any policy change in φ through adjustment of access intensities and hence of residual terminal demand, so the equilibrium premium responds less than one-for-one to a change in intermediary return impact. A static price-impact model with exogenous λ_i would instead deliver $d \log \pi_A / d \log \varphi = 1$; the shortfall from unity is the elasticity gap absorbed by endogenous access.

Corollary 2. *In the limit $\tilde{\psi}_i \rightarrow \infty$ for all i , $\lambda_i^*(\pi_A) \rightarrow 0$, $|D'_A(\pi_A)| \rightarrow 0$, and the elasticity (30) converges to unity. The endogenous-access dampening of the equilibrium premium therefore vanishes in the static-access limit, and the equilibrium response to a change in φ depends on the level of normalized access costs.*

Corollary 2 formalizes the second comparative-static feature emphasized in the introduction:

the two frictions interact, and the joint effect of φ and $\tilde{\psi}$ on π_A is not the sum of their separate effects. In the calibrated quantitative exercise of Section 4, this non-additivity is illustrated numerically: starting from a doubled- $\tilde{\psi}$ economy, halving φ implies an elasticity of 0.72, slightly above the 0.69 realized at baseline; and halving both frictions jointly lowers the equilibrium premium by exactly 50%, restoring the static-additive prediction.

Corollary 3. *At the equilibrium of Proposition 1, whenever $\sum_i d_i > 0$, the elasticity in (30) admits the closed form*

$$\frac{d \log \pi_A}{d \log \varphi} = \frac{1}{1 + \omega}, \quad \omega \equiv \sum_i s_i \frac{W_i}{1 + W_i}, \quad (31)$$

where $W_i \equiv W(d_i \pi_A \tau^2 / \tilde{\psi}_i)$, $s_i \equiv d_i e^{-W_i} / \sum_j d_j e^{-W_j}$ is operator i 's share of aggregate residual terminal demand, and $W(\cdot)$ is the Lambert W function. Because $W_i / (1 + W_i) \in (0, 1)$ pointwise, $\omega \in (0, 1)$ and the elasticity is bounded strictly in $(\frac{1}{2}, 1)$. The elasticity converges to $1/2$ in the limit $\tilde{\psi}_i \rightarrow 0$ for all i .

Corollary 3 sharpens the dampening result of Corollary 1: the equilibrium-premium elasticity to intermediary return impact is bounded strictly above $1/2$ as well as strictly below 1. The lower limit is attained when access frictions are negligible, in which case the equilibrium premium scales as $\sqrt{\varphi}$; the upper limit was characterized in Corollary 2.

Corollary 4. *For any $k > 0$, the joint rescaling $(\varphi, \tilde{\psi}_i) \mapsto (k\varphi, k\tilde{\psi}_i)$ for all i leaves the Lambert- W argument $d_i \pi_A \tau^2 / \tilde{\psi}_i$ in (27) invariant and scales the equilibrium of Proposition 1 by the same factor: $\pi_A \mapsto k\pi_A$, with each λ_i^* and $D_A(\pi_A)$ unchanged. In particular, halving both frictions ($k = 1/2$) halves the equilibrium premium exactly.*

Corollary 4 explains why the joint half- φ and half- $\tilde{\psi}$ row in Table 2 lowers the equilibrium premium by exactly 50%: under the joint rescaling, the Lambert- W argument is invariant, so access intensities and residual demand do not respond, and the equilibrium response coincides with the prediction of a static price-impact model in which λ_i is held fixed and only φ varies. Proofs of Proposition 1 and Corollaries 1–4 are in the Online Appendix.

3.3 Intermediation and Affine Price Impact

A representative reduced-form intermediary supply schedule provides 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

$$P_t = P_t^*(X_t) + \phi M_t, \quad (32)$$

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:

$$M_t = M_t^{\text{comp}} + M_t^{\text{info}}. \quad (33)$$

For the compliance component, under full reset upon access, the expected contemporaneous compliance-related order-flow rate is

$$E_t[M_t^{\text{comp}}] = \sum_i \lambda_i^* g_{it} = \bar{\lambda}_t^* G_t, \quad \bar{\lambda}_t^* \equiv \frac{\sum_i \lambda_i^* g_{it}}{G_t}, \quad (34)$$

whenever $G_t > 0$, where λ_i^* is the cycle-level access intensity from (18). So contemporaneous compliance order flow is proportional to the aggregate unresolved compliance gap with the endogenous gap-weighted access intensity $\bar{\lambda}_t^*$. The rate $\bar{\lambda}_t^*$ is not a free parameter unrelated to participation: it is the gap-weighted speed at which unresolved compliance gaps are converted into order flow, and it inherits the comparative statics of the underlying λ_i^* in the value of access $S_i = d_i \pi_A$ and the normalized access-cost parameter $\tilde{\psi}_i$. Equation (34) 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 counterpart 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. In the Online Appendix's resource-cost decomposition, the term $\frac{\phi}{2} M_t^2$ 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 Equilibrium Surrender-Month Premium

The expected surrender-month return premium π_A enters individual access choices through the value of access $S_i = d_i \pi_A$ in (18), and it is in turn determined by aggregate residual terminal demand $D_A(\pi_A)$ in (24). We close the model's quantitative specification with the reduced-form

return-impact condition $\pi_A = \varphi D_A(\pi_A)$, a price-pressure relation rather than a no-arbitrage condition. Each operator takes π_A as given when choosing λ_i ; in a continuum economy, individual operators do not internalize their own effect on aggregate terminal demand. Proposition 1 establishes that this fixed-point equation has a unique positive solution whenever aggregate expected short-fall $\sum_i d_i$ is positive, so that the surrender-month return premium is an equilibrium object rather than an assumed primitive: it reflects the joint outcome of operators' endogenous access choices and residual terminal demand. A higher premium raises the private value of early access, while costly access leaves residual terminal demand that generates the premium through the reduced-form surrender-window return-impact relation. This parameter φ is distinct from the instantaneous price-level impact parameter ϕ in (32).

3.5 Informational Advantages and Operator Flow

To preserve a return-predictability prediction, we introduce a parsimonious private-signal component. A subset $\mathcal{I}$ of operators receives a private signal about short-run innovations to the frictionless benchmark price. The private-information component 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 (21). Define operator i 's informational advantage in return form as

$$\mu_{it} = E_t[P_{t+\Delta}^* - P_t^* | \mathcal{I}_{it}] - E_t[P_{t+\Delta}^* - P_t^* | \mathcal{F}_t], \quad (35)$$

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

$$\max_{q_{it}^{\text{info}}} \left\{ \mu_{it} q_{it}^{\text{info}} - \frac{\gamma_i}{2} (q_{it}^{\text{info}})^2 \right\}, \quad q_{it}^{\text{info}} = \frac{\mu_{it}}{\gamma_i}, \quad (36)$$

where $\gamma_i > 0$ captures inventory, risk-management, or balance-sheet costs. Because access arrivals are Poisson with chosen 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,

$$M_t^{\text{info}} = \sum_{i \in \mathcal{I}} \lambda_i^* \frac{\mu_{it}}{\gamma_i}, \quad (37)$$

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 a higher chosen λ_i^* implies faster and more informative flow. Operator flow imbalance 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 who choose high access intensity, that is, those with low normalized access costs $\tilde{\psi}_i$ 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

$$E_t[P_{t+\Delta}^* - P_t^* | M_t^{\text{info}}] = \beta M_t^{\text{info}}, \quad (38)$$

then actual returns are positively predictable when the signal-loading coefficient β exceeds the expected reversal of contemporaneous price impact. The private-information component is therefore reduced form: it isolates the empirical content of informative order flow without solving an underlying filtering or Kyle-style problem.

3.6 Model Predictions

The continuous-time model delivers the same three core empirical predictions as the two-period model and one additional structural prediction about how the trading frictions interact, in a richer endogenous setting that also generates quantitative objects for the analysis below.

First, the model predicts endogenous slow participation. By Proposition 1, operators choose lower cycle-level market-access intensity λ_i^* when expected shortfalls d_i are small, expected surrender-month return premia π_A are low, or normalized access costs $\tilde{\psi}_i$ are high. These firms are less likely to receive discretionary trading opportunities during the compliance cycle and trade less often. As a result, unresolved compliance gaps g_{it} are more persistent and adjust more slowly, concentrating 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: a high $\tilde{\psi}_i$ generates a large effective K_i , while a high $S_i = d_i \pi_A$ generates a small effective K_i . Empirically, this prediction maps into persistent non-trading and the concentration of purchases near surrender, especially among operators with high attention or treasury costs.

Second, the model predicts surrender-month price pressure. Endogenous access choices determine how much compliance demand remains unresolved before the deadline. During the compliance cycle, the unresolved gap G_t generates compliance-related order flow with expected rate

$E_t[M_t^{\text{comp}}] = \bar{\lambda}_i^* G_t$, where $\bar{\lambda}_i^*$ is the gap-weighted access intensity defined in (34). At the surrender date, any remaining positive gap is converted into terminal compliance order flow M_A^{term} as in (25), with aggregate expected residual demand $D_A(\pi_A) = \sum_i d_i e^{-\lambda_i^*(\pi_A)\tau}$. The resulting surrender-month return premium is the unique solution to the reduced-form fixed-point condition $\pi_A = \varphi D_A(\pi_A)$ in Proposition 1. Within this closed-form specification, the premium is stronger when normalized access costs $\tilde{\psi}_i$ are higher, the surrender-window return impact φ is larger, or expected shortfalls d_i are larger. More generally, higher scarcity or a smaller allowance bank can also raise residual terminal demand and hence the premium. 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. Conditional on receiving market access, a subset $\mathcal{I}$ of operators trades on private signals about short-run innovations to the frictionless benchmark price, and the informational component of operator order flow, M_t^{info} , forecasts future allowance returns. Because access intensity is endogenous, operators with higher chosen access intensity λ_i^* contribute more frequently to informative order flow: the expected contribution of operator i to M_t^{info} scales with λ_i^* . Higher chosen access intensities make informative order flow appear more quickly and more strongly in the market, so the predictive content of operator flow should be stronger among operators who trade more actively. In the data, however, that frequent-trader pattern is reduced form: it can reflect lower access frictions, higher informational value of access, or both. Empirically, this prediction maps into Operator Flow Imbalance forecasting future allowance returns, especially for frequent traders. The empirical analysis below tests the signs and cross-sectional patterns implied by the model, while the quantitative exercise recovers reduced-form model-implied access and return-impact magnitudes rather than fully structural estimates of EU ETS primitives.

Fourth, the model implies that the two trading frictions interact non-additively in their effect on the equilibrium premium. By Corollary 1, the elasticity of π_A with respect to φ is bounded strictly below unity: endogenous access reabsorbs a fraction of any policy change in φ through adjustment of λ_i^* and hence of residual terminal demand D_A , so a static demand-pressure model overstates the price response to changes in intermediary capacity. By Corollary 2, this dampening vanishes in the static-access limit, so the equilibrium response to a change in φ depends on the level of $\tilde{\psi}$. Empirically, this prediction maps into the quantitative counterfactuals of Section 4: halving φ alone lowers the equilibrium premium by less than 50% because endogenous access raises D_A ; halving both φ and $\tilde{\psi}$ jointly lowers it by exactly 50%, restoring the static-additive prediction; and the elasticity of π_A with respect to φ is larger when access costs start at a doubled baseline level.

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, and that the two underlying trading frictions interact non-additively in their effect on equilibrium prices. We next quantify the implied frictions and run simple counterfactuals using the closed-form objects of the model, before turning to formal empirical tests of the three predictions.

4 Quantifying Trading Frictions

This section uses the model to benchmark the magnitude of the trading frictions and to evaluate simple market-design counterfactuals. The exercise is deliberately partial-equilibrium: we do not estimate a full dynamic model of abatement, banking, and permit prices. Instead, we use the closed-form access-choice condition (18), the residual terminal-demand expression (24), and the reduced-form surrender-window return-impact relation $\pi_A = \varphi D_A(\pi_A)$ to calibrate normalized access frictions and a return-impact sensitivity parameter, and then evaluate counterfactual changes to φ , $\tilde{\psi}_i$, and the timing of surrender obligations. We treat the recovered objects as model-implied magnitudes rather than as fully structural estimates of EU ETS primitives. Detailed construction choices, banking-adjusted shortfall measures, alternative premium definitions, and leave-one-out diagnostics are reported in Appendix A.6.

4.1 Construction of the Quantitative Objects

We define each compliance cycle as running from May of year $t - 1$ through April of year t , with April as the surrender month. For each operator i and cycle t , we use the transparent allocation-based shortfall proxy

$$d_{it}^{\text{alloc}} = \max\{\text{surrender}_{it} - \text{allocation}_{it}, 0\}, \quad (39)$$

and report bank-adjusted, cumulative-surplus, and lagged-surplus alternatives in Appendix A.6.

Throughout this section, we use the one-shot known-shortfall specialization of Proposition 1, rather than the broader stochastic h_{it}^* process. We back out a firm-specific access intensity $\hat{\lambda}_i$ from observed pre-surrender trading frequency. Under the model, the probability that operator i receives no discretionary access opportunity over a cycle of length τ is $e^{-\lambda_i \tau}$, so a descriptive shrinkage estimator inverts this relation using each operator's empirical frequency of no-trade years. The

model-implied residual terminal demand is then

$$D_{A,t}^{\text{model}} = \sum_i d_{it} e^{-\hat{\lambda}_i \tau}, \quad (40)$$

and we calibrate the surrender-window return-impact sensitivity parameter φ from the relation between observed April return premia $\pi_{A,t}$ and $D_{A,t}^{\text{model}}$. In this quantitative implementation, $\hat{\lambda}_i$ is best read as a reduced-form index of effective pre-surrender market access among positive-shortfall operators, rather than as a literal structural access parameter for every operator-year, since observed no-trade can also reflect little motive to rebalance. Operationally, the shortfall input d_{it} is measured from annual compliance-cycle data rather than inferred from the richer continuous-time target process. We use the raw April log return as the headline premium and report non-April-adjusted and residualized variants in the appendix. Implied normalized access-cost indices $\hat{\psi}_i$ are recovered from the access-choice first-order condition $S_i \tau e^{-\lambda_i \tau} = \tilde{\psi}_i \lambda_i$ at the firm-cycle level and aggregated to the firm level via the median across cycles.

4.2 Headline Distribution and Model Fit

The headline sample uses 109,247 operator-year observations for 10,696 operators over 2005–2021, where a year is defined as the May-to-April compliance cycle. The descriptive estimator implies a median $\hat{\lambda}_i$ of 0.568, with 10th-to-90th percentile range [0.170, 1.482]. With the cycle normalized to unit length, the median estimate corresponds to roughly one pre-surrender trading opportunity every 1.8 compliance years; equivalently, the Poisson benchmark implies a 43.3% probability of at least one such opportunity in a given compliance year. Table 1 reports the access-intensity distribution together with the main validation moments. Within Panel B, the pre-surrender no-trade share is the directly disciplined moment, while the remaining rows are untargeted validation moments. The model matches the pre-surrender no-trade share closely (data 0.560, model 0.546)⁷ but not exactly, because the descriptive $\hat{\lambda}_i$ uses a shrinkage-adjusted firm-level no-trade probability rather than forcing the pooled sample share to match by construction. The model also matches the one-year persistence of non-trading reasonably well (data 0.658, model 0.637). The remaining timing-of-trade moments in Table 1 highlight the main limits of the single-parameter Poisson benchmark, which we discuss next.

⁷This pre-surrender no-trade share differs from the roughly 40% annual no-trade fact documented in Figure 4. The annual measure counts an operator as trading if it makes any market trade during the year, including purchases in April, whereas the quantitative model’s Poisson access object is the probability of no discretionary market access before surrender. Operators who wait until April count as traders in the annual Figure 4 measure but as no pre-surrender trade in the model-aligned moment.

Table 1: Headline access-intensity distribution and core model fit

Moment	Data/Value	Model
<i>Panel A: Distribution of $\hat{\lambda}_i$ and implied no-access probabilities</i>		
$\hat{\lambda}_i$ p10	0.170	
$\hat{\lambda}_i$ p25	0.325	
Median $\hat{\lambda}_i$	0.568	
Mean $\hat{\lambda}_i$	0.724	
$\hat{\lambda}_i$ p75	0.961	
$\hat{\lambda}_i$ p90	1.482	
Mean no-access probability $e^{-\hat{\lambda}_i \tau}$	0.548	
Median no-access probability	0.567	
<i>Panel B: Validation moments</i>		
Pre-surrender no-trade share	0.560	0.546
One-year no-trade persistence	0.658	0.637
Median pre-surrender trade count	0.000	1.474
Share of annual volume in April purchases	0.231	0.131
Corr($D_{A,t}^{\text{model}}, D_{A,t}^{\text{data}}$)	0.949	
Slope of $D_{A,t}^{\text{data}}$ on $D_{A,t}^{\text{model}}$	1.642	
R^2 of terminal-demand validation	0.901	

Notes: Headline sample 2005–2021. Panel A reports the descriptive access-intensity estimator (shrinkage-adjusted) and the implied no-access probabilities used in the quantitative exercise, so no separate model counterpart is shown there. Panel B compares headline data moments to model objects used for the quantitative exercise. Within Panel B, the pre-surrender no-trade share is the directly disciplined moment, while the remaining rows are untargeted validation moments. The no-trade and trade-count moments are defined over pre-surrender trading opportunities within the compliance cycle. The model trade-count row reports the median of cycle-specific expected pre-surrender trades under the Poisson benchmark, not the median of simulated realized counts.

The descriptive Poisson estimator matches the pre-surrender no-trade share and one-year persistence well, but is calibrated on the probability of zero pre-surrender trades rather than on the within-cycle timing of trade. Two related moments illustrate the limits of the single-parameter description. First, the model-implied median expected pre-surrender trade count exceeds the realized median of zero, because operator trading is lumpy and zero-inflated rather than time-homogeneous Poisson. Second, the model under-predicts the share of annual volume that occurs in April. Both gaps point to a common extension: a calendar-timed access process with intensity that rises around verification, reflecting either an end-March h_{it}^* jump or a seasonal increase in effective access. Still, model-implied terminal demand tracks observed April purchases closely at the year level, so the cross-year variation used to calibrate $\hat{\phi}$ and to organize the counterfactual changes is well-captured. The level mismatch on the April-share moment makes the baseline euro measure of the April premium paid by delayed buyers a conservative lower bound, and likely under-states the gains from a

market-design intervention that staggers the deadline.

The terminal-demand validation is the central diagnostic for the calibration of φ , since the return-impact relation is meaningful only if the model-implied demand object tracks observed surrender-month buying pressure. In the headline sample, the correlation between $D_{A,t}^{\text{model}}$ and observed April net operator purchases is 0.949, with $R^2 = 0.901$ in a yearly regression of observed on model-implied terminal demand. The fitted slope of 1.64 indicates that the model captures year-to-year variation well but understates the magnitude of observed April buying in levels, consistent with the April-share under-prediction noted above. Figure 2 reports the year-by-year scatter.

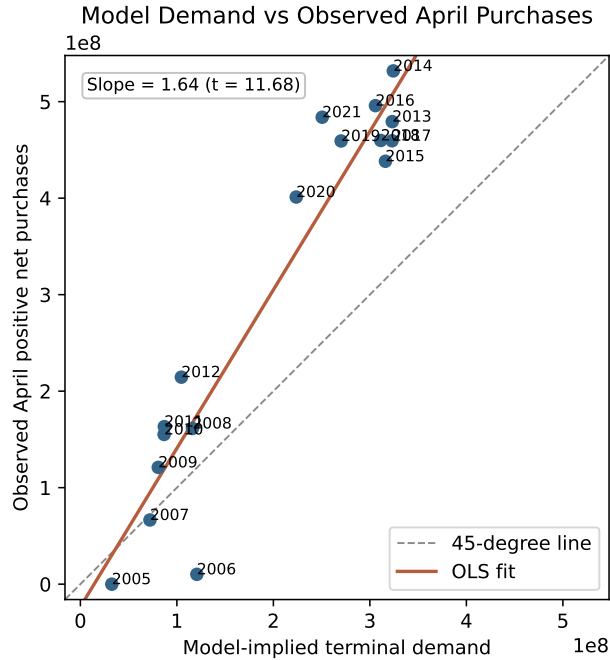


Figure 2: Model-implied terminal demand and observed April net operator purchases, by year. Headline 2005–2021 sample. The 45-degree line marks one-to-one correspondence; the fitted line is the OLS regression of observed on model-implied terminal demand.

4.3 Calibrated Return Impact

We calibrate φ from the relation $\pi_{A,t} = \varphi D_{A,t}^{\text{model}} + \varepsilon_t$ using the headline raw April premium. Restricting to positive-premium years yields a median $\hat{\varphi}$ of 7.83×10^{-10} across 11 years; using all 17 years with OLS through the origin yields $\hat{\varphi} = 4.10 \times 10^{-10}$. The order of magnitude is stable across non-April-adjusted and residualized premium definitions and across the sample-period robustness checks reported in Table A.5 in the Online Appendix. Leave-one-out diagnostics on the headline raw-April OLS specification produce a tight range of $[3.71 \times 10^{-10}, 4.67 \times 10^{-10}]$. The all-year OLS R^2 is small, reflecting the volatility of April returns; we therefore present the

positive-year median as the headline calibration and use the all-year OLS as a conservative lower bound. Throughout, we interpret $\hat{\varphi}$ as a calibrated return-impact sensitivity parameter rather than as a tightly identified structural primitive, so relative counterfactual changes are more informative than exact level magnitudes in euros.

4.4 Counterfactuals

We use the calibrated objects to solve the equilibrium fixed point $\pi_A = \varphi D_A(\pi_A)$ under alternative scenarios. Because each operator's chosen λ_i^* depends on the equilibrium premium, the counterfactuals incorporate the feedback emphasized in Proposition 1: a lower premium reduces the private value of early access and partially reverses the direct effect of any policy change. The baseline fixed-point premium of 0.072 is lower than the raw April sample mean of 0.104 because the counterfactual exercise isolates the compliance-demand component captured by $\pi_A = \varphi D_A(\pi_A)$ using a single calibrated $\hat{\varphi}$ and model-implied terminal demand, rather than the full April return including residual shocks. This same level understatement is reflected in the terminal-demand slope above and in the under-predicted April-volume share.

Table 2 reports the headline counterfactuals. Halving surrender-window return impact lowers the equilibrium premium by 38.2% rather than the 50% that a static return-impact model would predict, because residual terminal demand D_A rises by 23.7% once operators internalize the lower premium. The realized elasticity is $\log(0.618)/\log(0.5) \approx 0.69$, consistent with Corollary 1: the model-implied $|D'_A(\pi_A)|$ at the calibrated equilibrium absorbs roughly 30% of any direct change in φ . Halving access costs lowers the premium by 20.3% and the April premium paid by delayed buyers by 36.9%, working through both lower D_A and a lower equilibrium premium. Under the one-shot access specification, the largest illustrative effects come from staggering the surrender deadline across K size-balanced groups: $K = 2$ lowers the premium by 36.9% and the April premium paid by delayed buyers by 22.8%, while $K = 4$ lowers the premium by 59.4% and the April premium paid by delayed buyers by 42.2%. In the one-shot access specification, splitting terminal demand into two size-balanced surrender windows reduces the order-flow pressure faced in each window and therefore resembles a reduction in effective surrender-window return impact; this explains why the $K = 2$ staggered-deadline counterfactual is numerically close to the half- φ counterfactual. The non-linear gain from $K = 2$ to $K = 4$ reflects the convex absorption-cost technology combined with the endogenous-access feedback. For this reason, we emphasize the direction and relative ranking of the counterfactual changes more than the exact euro levels.

The joint half- φ and half- $\tilde{\psi}$ row in Table 2 shows that when both frictions are reduced at once, the equilibrium premium falls by exactly 50% and the April premium paid by delayed buyers also falls by 50%, as established analytically by the scale invariance of Corollary 4: the +23.7% rise in

D_A from a lower φ is exactly offset by the -20.3% fall in D_A from a lower $\tilde{\psi}$, so the static-additive prediction is restored when both frictions move together. A separate sensitivity comparison, not reported in the table, illustrates the amplification channel. Doubling $\tilde{\psi}$ —making market access more costly—raises the equilibrium premium to 0.089, and from that higher-friction starting point halving φ lowers the premium to 0.054. The implied elasticity, $\log(0.054/0.089)/\log(0.5) \approx 0.72$, is slightly above the 0.69 obtained in the baseline calibration, consistent with the comparative-static implication of Corollary 1: when access frictions are larger, the endogenous-access feedback is weaker, the dampening channel attenuates, and the equilibrium premium becomes more sensitive to changes in intermediary capacity. The amplification effect is small in magnitude but operates in the direction predicted by the model.

Table 2: Headline counterfactuals from the quantitative access model

Scenario	Premium	D_A change	April premium paid (€ mn)	Premium-paid change	Absorption cost change
Baseline	0.072	—	175.6	—	—
Half φ	0.044	+23.7%	135.5	-22.9%	-22.9%
Half normalized access cost ($\tilde{\psi}$)	0.057	-20.3%	110.8	-36.9%	-36.9%
Half φ and half $\tilde{\psi}$ jointly	0.036	0.0%	87.8	-50.0%	-50.0%
Stagger $K = 2$	0.045	+23.7%	135.5	-22.8%	-22.8%
Stagger $K = 4$	0.029	+50.3%	101.5	-42.2%	-42.2%

Notes: The baseline row reports the level of the equilibrium premium and the April premium paid by delayed buyers in millions of euros. Counterfactual rows report changes relative to the baseline. Premium is the equilibrium fixed-point solution to $\pi_A = \varphi D_A(\pi_A)$; φ is the surrender-window return-impact parameter; D_A is aggregate residual terminal demand under the counterfactual; the April premium paid by delayed buyers is $P_e \pi_A D_A$ at the within-cycle early price level P_e ; absorption cost is $\frac{1}{2} P_e \pi_A D_A$; and $\tilde{\psi}$ denotes the access-cost shifter used in the quantitative exercise. Because $\hat{\varphi}$ is a calibrated sensitivity parameter inferred from noisy annual premium variation, the directional and relative counterfactual changes are more robust than the euro levels. Under the one-shot access specification, the staggered-deadline counterfactual splits operators into K equally sized surrender groups, solves the fixed point separately for each group, and should be interpreted as illustrative.

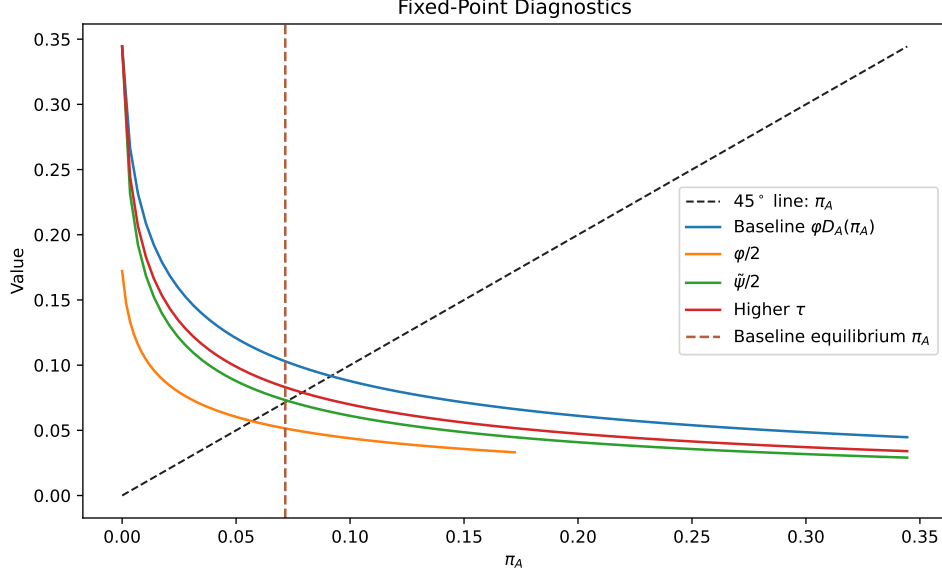


Figure 3: Fixed-point logic behind the quantitative counterfactuals. The 45-degree line plots the equilibrium identity $\pi_A = \pi_A$; the curves plot $\varphi D_A(\pi_A)$ under the baseline calibration and under three counterfactuals. Equilibrium premia are the intersections of each curve with the 45-degree line. Halving φ rotates the curve down but raises D_A at any given premium, generating less than a one-for-one fall in π_A .

The counterfactuals share two features that should temper the headline magnitudes. First, the exercise is partial-equilibrium: it holds $\hat{\lambda}_i$ moments and D_A shapes fixed in the data and re-solves only the access-choice and return-impact components. Aggregate scarcity, the allowance bank, and the abatement decision are unchanged. Second, the calibration uses a single scalar $\hat{\varphi}$ inferred from noisy yearly variation in April return premia, so the counterfactual levels inherit the sampling uncertainty in $\hat{\varphi}$ documented in the Online Appendix. Within these bounds, the directional and relative comparisons are more informative than the exact euro magnitudes. The staggered-deadline result should be read as an illustrative implication of the one-shot access specification: splitting compliance demand across four equal groups reduces the implied April premium paid by delayed buyers by roughly 42%, with no change in aggregate emissions or the cap.

5 Empirical Tests

We now test the three empirical predictions generated by the model with frictions. Subsection 5.1 describes the institutional setting and transaction data. Subsections 5.2–5.5 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.

5.1 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 Online 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.

5.2 Prediction 1: Slow Participation

The first prediction of the model is that endogenous market-access choice generates persistent gaps between desired and actual allowance holdings, implying delayed trading and non-participation for the subset of operators with high normalized access costs $\tilde{\psi}_i$ or low value of access S_i . 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 operators choosing low effective λ_i^* in the continuous-time model.

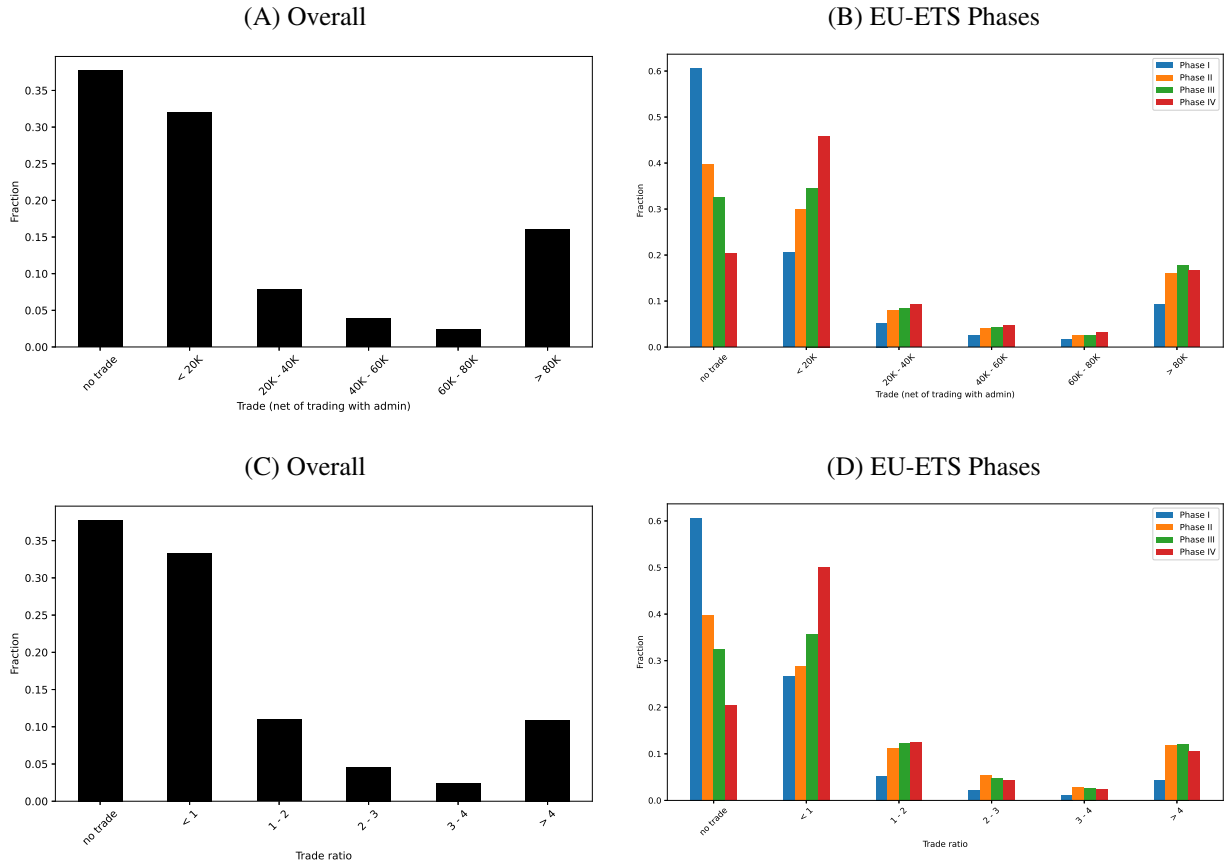
In any cap-and-trade system, cost-effective reallocation of allowances through the secondary market requires firms to actively participate and trade. We analyze the trading behavior of firms with operator holding accounts in the EU ETS. The top panels of Figure 4 show the yearly share of firms by trading 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 operators choosing low effective market-access intensity λ_i^* in the continuous-time model—because of high normalized access costs $\tilde{\psi}_i$ or low value of access S_i —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.

Figure 4 also reports the trading-volume bins and trading-ratio categories used in the analysis.

Figure 4: 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; $0 \leq \text{trades} < 20\text{k}$; $20\text{k} \leq \text{trades} < 40\text{k}$; $40\text{k} \leq \text{trades} < 60\text{k}$; $60\text{k} \leq \text{trades} < 80\text{k}$; and $80\text{k} \leq \text{trades}$. 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 ; $1 \leq \text{trade ratios} < 2$; $2 \leq \text{trade ratios} < 3$; $3 \leq \text{trade ratios} < 4$; $4 \leq \text{trade ratios}$. Data is for the period from February 2005 to September 2021.



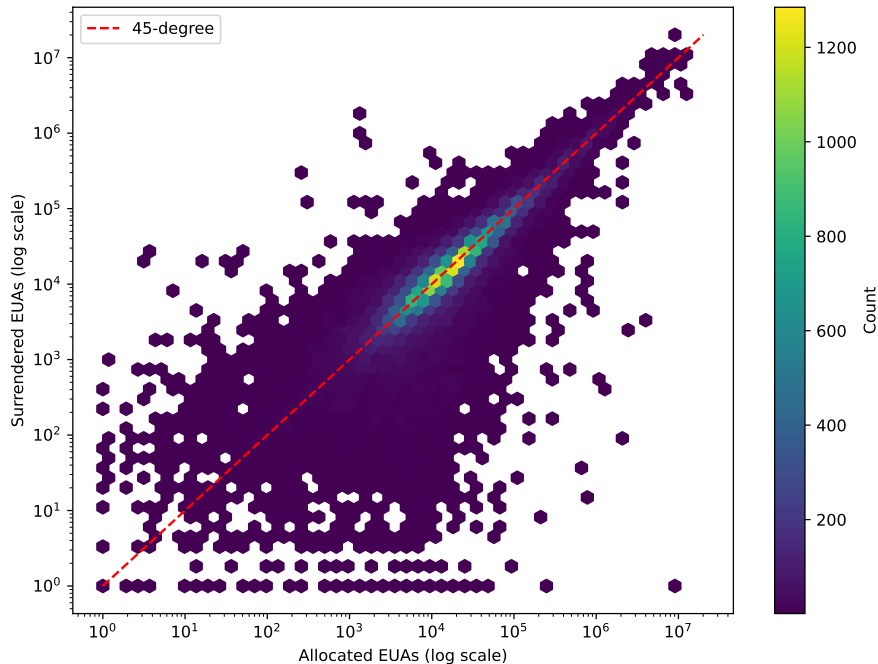
Among firms that do trade, average annual volume is approximately 95,000 contracts, but the distribution is highly skewed. The median is about 3,700; 10% of firms trade fewer than 100 contracts, while another 10% trade more than 86,000. To control for firm size, we normalize trading volume by the surrendering amount. The bottom panels of Figure 4 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 operators choosing low effective λ_i^* because of high normalized access costs $\tilde{\psi}_i$ or low value of access S_i . Second, over 10% of firms trade far more than needed for compliance, consistent with informational advantages and speculative positions taken by operators that choose high access intensity; we return to this group in Sections 5.4–5.5.

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 induce trade. The patterns in the data are difficult to reconcile with this benchmark.

Figure 5: 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 4 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 5 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 3 presents panel regression results where the dependent variable is an operator-level dummy indicating no trade in year $t + h$ (for $h = 1, \dots, 5$), 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.⁸

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 chooses low effective access intensity λ_i^* because of high normalized access costs $\tilde{\psi}_i$ or low value of access S_i , so that expected compliance gaps remain unresolved and trading is systematically delayed. The gradual decline in non-trading rates over time points to a fall in effective access costs and hence higher chosen λ_i^* ,

⁸Banking was not allowed between Phase I and Phase II, so unused Phase I allowances expired at the end of 2007. The Phase III/IV transition was different: existing Phase III allowances remained valid into Phase IV, although newly issued Phase IV allowances could not be used for 2020 compliance in April 2021. 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 adjacent 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%.

plausibly driven by market development and policy changes across phases, but a substantial share of firms remains effectively inactive even in the current phase.

Table 3: 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), covering the mature market and the Phase III/IV transition and excluding the non-bankable Phase I break. Both baseline specifications and models with year fixed effects are reported. Standard errors are adjusted for overlapping-horizon serial correlation using Newey-West ([Newey and West, 1987](#)) 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.

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

5.3 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 administrative account to match their emissions. April is 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 6. 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. Although 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.

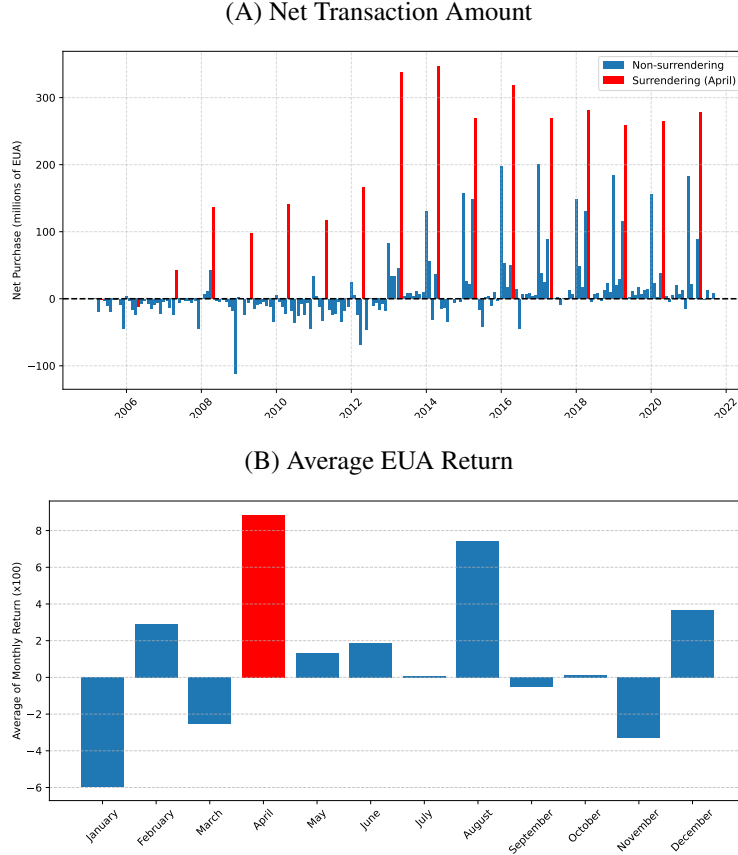
Next, we examine the average monthly emission allowance return, which we document in Panel B of Figure 6. The emission allowance return is computed as the average monthly log change in emission allowance prices. We observe that the average return is highest in April, at about 10%. This monthly return is significantly higher than returns in other months of the year. The concentrated demand in April drives a predictable price surge in emission allowances. The April premium is sizeable but does not imply unbounded arbitrage: profitable positions are bounded by finite aggregate residual demand D_A , and arbitrageurs face risk from end-of-March verification uncertainty and from operators' endogenous banking choices.

To quantify the implied April premium paid by 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 about 10% predictable appreciation for the surrender month, we estimate that the total additional payment 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 measure of the April premium paid by 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 6: 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.



5.4 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 chosen access intensity λ_i^* —reflecting low normalized access costs $\tilde{\psi}_i$ or large benefits of early trading—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 4 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, although slower-moving inventory rebalancing or persistent client-order imbalances could also contribute to the predictability pattern. The horizon pattern is informative about interpretation. Pure temporary liquidity pressure would predict short-run price impact followed by reversal. Instead, the coefficients remain positive over longer horizons, consistent with operator flow containing information about future scarcity rather than only transitory demand pressure.

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 Online 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 interpret this figure as an illustrative magnitude consistent with informational advantages among a subset of active operators, not as a direct structural estimate of abnormal returns or informed trading.

Table 4: 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.

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

5.5 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. This mapping is reduced form: in the data, infrequent trading can reflect both limited access and a weak motive to trade, while frequent trading can reflect lower access frictions, higher informational value of access, or both. 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*^{high} and *Operator Flow Imbalance*^{low} as the imbalance for firms above and below the sample median, respectively. This high-versus-low split uses full-sample trading frequency and is intended as a descriptive cross-sectional classification rather than as a real-time trading rule.

Table 5 presents the results from time-series return predictability regressions using both *Operator Flow Imbalance*^{high} and *Operator Flow Imbalance*^{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*^{high} are always positive, with or without month fixed effects. Without month fixed effects, the coefficient estimates of *Operator Flow Imbalance*^{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*^{low} are never significant in any specification. Return predictability is driven by trading by frequent operators, as implied by the third model prediction. In the model, frequent traders are a reduced-form proxy for operators with higher chosen market-access intensity λ_i^* , lower effective normalized access costs $\tilde{\psi}_i$, higher informational value of market access, or some combination of these forces, rather than a direct estimate of the structural objects λ_i^* , $\tilde{\psi}_i$, or γ_i . The empirical analysis tests the signs and cross-sectional patterns implied by the model rather than estimating these structural parameters.

Table 5: 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*^{low} and *Operator Flow Imbalance*^{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 over the full sample. 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.

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

6 Conclusions

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 in which operators endogenously choose how often to access the market at a convex attention cost, intermediaries supply immediacy at a quadratic absorption cost, and a subset of operators trade on private information. The framework implies that gradual participation and limited intermediary capacity can generate delayed trading, surrender-month price pressure, and informative operator flow, and yields a strict comparative-static feature: when intermediary capacity changes, operators' access choices respond in a way that partially offsets the direct effect, so the equilibrium price reaction is smaller than a static demand-pressure model would predict. The two frictions also interact non-additively. Endogenous access dampens the response of the equilibrium premium to intermediary capacity, but the dampening is itself weaker when access costs are larger, so the equilibrium response can also be amplified by interacting frictions. Halving intermediary return impact alone lowers the equilibrium premium by about 38%, halving access costs alone lowers it by about 20%, but halving both frictions jointly lowers it by exactly 50%—show that the joint effect of the two frictions need not equal the sum of their separate effects ([Lipsey and Lancaster, 1956](#)). We then quantify the magnitude of these frictions and evaluate simple market-design counterfactuals. Model-implied terminal demand has a correlation of 0.95 with observed April compliance purchases at the year level, and an illustrative counterfactual that staggers surrender deadlines across four equally sized groups would reduce the surrender-month premium by about 59% and the implied April premium paid by delayed buyers by about 42%, with no change in aggregate emissions or the cap.

References

- Abrell, Jan**, “Database for the European Union Transaction Log,” 2023. Available at [Euets.info](https://euets.info).
- Acemoglu, Daron, Philippe Aghion, Lint Barrage, and David Hémous**, “Climate Change, Directed Innovation, and Energy Transition: The Long-run Consequences of the Shale Gas Revolution,” 2024. Working paper.
- Acharya, Viral V, Lars A Lochstoer, and Tarun Ramadorai**, “Limits to arbitrage and hedging: Evidence from commodity markets,” *Journal of Financial Economics*, 2013, 109 (2), 441–465.
- Barrage, Lint**, “Optimal Dynamic Carbon Taxes in a Climate–Economy Model with Distortionary Fiscal Policy,” *The Review of Economic Studies*, 2020, 87 (1), 1–39.
- **and William D. Nordhaus**, “Policies, Projections, and the Social Cost of Carbon: Results from the DICE-2023 Model,” *Proceedings of the National Academy of Sciences*, 2024, 121 (13), e2312030121.
- Betz, Regina and Tobias S Schmidt**, “Transfer patterns in Phase I of the European Union Emissions Trading System: a first reality check based on cluster analysis,” *Climate Policy*, 2016, 16 (4), 474–495.
- Biais, Bruno and Augustin Landier**, “Emission Caps and Investment in Green Technologies,” *Review of Financial Studies*, 2026. Forthcoming.
- **, Johan Hombert, Daniel Schmidt, and Pierre-Olivier Weill**, “Cap and Trade with Imperfect Hedging,” 2026. Working paper, February 26, 2026. HEC Research Papers Series No. 1574. Available at SSRN.
- Bierbrauer, Felix J., Mattias Polborn, Marten Ritterrath, and Georg Weizsäcker**, “Morals and the Political Economy of Corrective Taxes,” 2026. Working Paper.
- Bolton, Patrick and Marcin Kacperczyk**, “Global Pricing of Carbon-Transition Risk,” *Journal of Finance*, 2023, 78 (6), 3677–3754.
- Brunnermeier, Markus K. and Lasse Heje Pedersen**, “Market Liquidity and Funding Liquidity,” *Review of Financial Studies*, 2009, 22 (6), 2201–2238.
- Cantillon, Estelle and Aurélie Slechten**, “Market design for the environment,” in “Handbook of the Economics of Corporate Finance” 2023.
- Colmer, Jonathan, Ralf Martin, Mirabelle Muûls, and Ulrich J. Wagner**, “Does Pricing Carbon Mitigate Climate Change? Firm-Level Evidence from the European Union Emissions Trading System,” *Review of Economic Studies*, 2025, 92 (3), 1625–1660.
- Croce, Mariano Max Massimiliano, Nicolas Guinez, Alejandra Inzunza Méndez, Thien T. Nguyen, and Claudio Tebaldi**, “Green Coins,” *Available at SSRN 5139356*, 2025.
- Dales, John Harkness**, *Pollution, property, and prices*, University of Toronto Press, 1968.

- Duffie, Darrell**, “Presidential Address: Asset Price Dynamics with Slow-Moving Capital,” *Journal of Finance*, 2010, 65 (4), 1237–1267.
- , **Nicolae Gârleanu**, and **Lasse Heje Pedersen**, “Valuation in Over-the-Counter Markets,” *Review of Financial Studies*, 2007, 20 (6), 1865–1900.
- Engle, Robert F., Stefano Giglio, Bryan Kelly, Heebum Lee, and Johannes Stroebe**, “Hedging Climate Change News,” *Review of Financial Studies*, 2020, 33 (3), 1184–1216.
- Fernández-Villaverde, Jesús, Kenneth T. Gillingham, and Simon Scheidegger**, “Climate Change Through the Lens of Macroeconomic Modeling,” *Annual Review of Economics*, 2025, 17, 125–150.
- Folini, Doris, Aleksandra Friedl, Felix Kübler, and Simon Scheidegger**, “The Climate in Climate Economics,” *The Review of Economic Studies*, 01 2025, 92 (1), 299–338.
- Gabaix, Xavier and Matteo Maggiori**, “International liquidity and exchange rate dynamics,” *The Quarterly Journal of Economics*, 2015, 130 (3), 1369–1420.
- Gârleanu, Nicolae, Lasse Heje Pedersen, and Allen M Poteshman**, “Demand-based option pricing,” *Review of Financial Studies*, 2009, 22 (10), 4259–4299.
- Giglio, Stefano, Bryan Kelly, and Johannes Stroebe**, “Climate Finance,” *Annual Review of Financial Economics*, 2021, 13, 15–36.
- Glosten, Lawrence R. and Paul R. Milgrom**, “Bid, Ask and Transaction Prices in a Specialist Market with Heterogeneously Informed Traders,” *Journal of Financial Economics*, 1985, 14 (1), 71–100.
- Golosov, Mikhail, John Hassler, Per Krusell, and Aleh Tsyvinski**, “Optimal Taxes on Fossil Fuel in General Equilibrium,” *Econometrica*, 2014, 82 (1), 41–88.
- Goulder, Lawrence H.**, “Markets for pollution allowances: what are the (new) lessons?,” *Journal of Economic Perspectives*, 2013, 27 (1), 87–102.
- Harstad, Bård and Gunnar S Eskeland**, “Trading for the future: Signaling in permit markets,” *Journal of Public Economics*, 2010, 94 (9-10), 749–760.
- Hendershott, Terrence and Albert J Menkveld**, “Price pressures,” *Journal of Financial Economics*, 2014, 114 (3), 405–423.
- Hong, Harrison, Neng Wang, and Jinqiang Yang**, “Mitigating Disaster Risks in the Age of Climate Change,” *Econometrica*, 2023, 91 (5), 1763–1802.
- Jaraitė, Jūratė, Frank Convery, and Corrado Di Maria**, “Transaction costs for firms in the EU ETS: lessons from Ireland,” *Climate Policy*, 2010, 10 (2), 190–215.
- Känzig, Diego R.**, “The unequal economic consequences of carbon pricing,” Technical Report, National Bureau of Economic Research 2023.

- Kübler, Felix**, “Incomplete Financial Markets, the Social Cost of Carbon and Constrained Efficient Carbon Pricing,” *Journal of Economic Theory*, 2025, 230, 106105.
- Kyle, Albert S.**, “Continuous Auctions and Insider Trading,” *Econometrica*, 1985, 53 (6), 1315–1335.
- Lipsey, R. G. and Kelvin Lancaster**, “The General Theory of Second Best,” *Review of Economic Studies*, 1956, 24 (1), 11–32.
- Montgomery, W David**, “Markets in licenses and efficient pollution control programs,” *Journal of Economic Theory*, 1972, 5 (3), 395–418.
- Newey, Whitney and Kenneth West**, “A simple, positive semi-definite, heteroskedasticity and autocorrelation consistent covariance matrix,” *Econometrica*, 1987, 55 (3), 703–708.
- Nordhaus, William D.**, “An optimal transition path for controlling greenhouse gases,” *Science*, 1992, 258 (5086), 1315–1319.
- Pástor, L’uboš, Robert F. Stambaugh, and Lucian A. Taylor**, “Sustainable Investing in Equilibrium,” *Journal of Financial Economics*, 2021, 142 (2), 550–571.
- Pedersen, Lasse Heje**, “Carbon Pricing versus Green Finance,” *Journal of Finance*, 2026, 81 (2), 561–602.
- , **Shaun Fitzgibbons, and Lukasz Pomorski**, “Responsible Investing: The ESG-Efficient Frontier,” *Journal of Financial Economics*, 2021, 142 (2), 572–597.
- Rubin, Jonathan D.**, “A Model of Intertemporal Emission Trading, Banking, and Borrowing,” *Journal of Environmental Economics and Management*, 1996, 31 (3), 269–286.
- Sautner, Zacharias, Laurence Van Lent, Grigory Vilkov, and Ruishen Zhang**, “Firm-Level Climate Change Exposure,” *Journal of Finance*, 2023, 78 (3), 1449–1498.
- Schmalensee, Richard and Robert N Stavins**, “The SO₂ allowance trading system: The ironic history of a grand policy experiment,” *Journal of Economic Perspectives*, 2013, 27 (1), 103–122.
- Stambaugh, Robert F.**, “Predictive regressions,” *Journal of Financial Economics*, 1999, 54 (3), 375–421.
- Subrahmanyam, Avaniidhar**, “Risk aversion, market liquidity, and price efficiency,” *Review of Financial Studies*, 1991, 4 (3), 417–441.
- Valkanov, Rossen**, “Long-horizon regressions: theoretical results and applications,” *Journal of Financial Economics*, 2003, 68 (2), 201–232.
- Weitzman, Martin L.**, “Prices vs. Quantities,” *Review of Economic Studies*, 1974, 41 (4), 477–491.

A Online 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 surrender month 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).⁹ The EUTL serves as the primary platform for reporting and monitoring within the EU ETS. This tool enables the European Commission to transparently disclose data regarding compliance by regulated entities, the participation of stakeholders within the system, 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, covering the first three phases of the EU ETS and part of the current

⁹The 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>.

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 is the account through which an installation—the regulated participant—receives, transfers, and surrenders emission allowances. Non-regulated participants (or financial intermediaries) can participate in the EU ETS by using person 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. 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. The next subsection reports account-type evolution and transaction volumes, followed by operator-level trading statistics.

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 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) sold and €700 billion (45 billion 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.¹⁰

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. 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 proportions 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.

¹⁰We 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.

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.

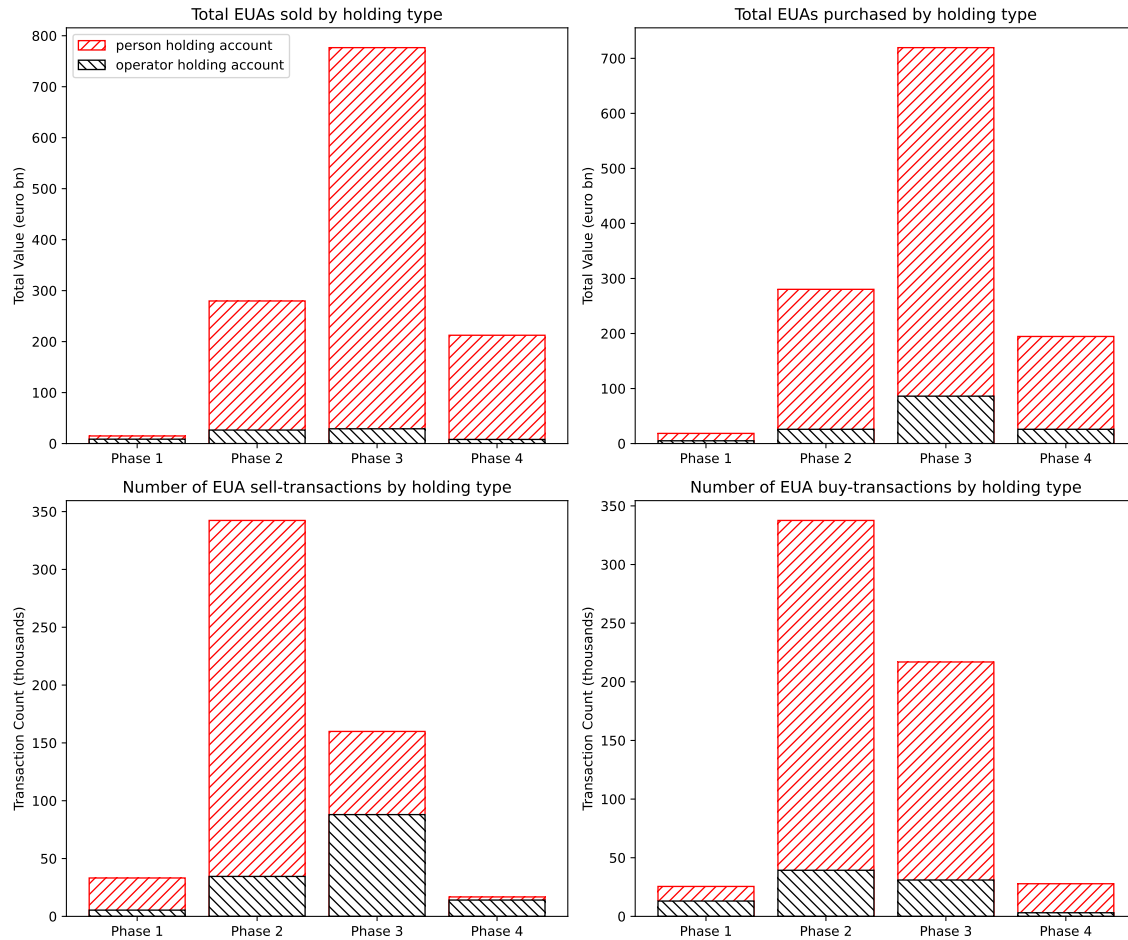


Table A.1: Operator Trading in the EU ETS Market

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). Under this procedure, the coefficient estimates remain significant at the 5% level from the two-month-ahead horizon onward. 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

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

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

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

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.

Endogenous Market Access

This subsection proves Proposition 1 for the one-shot known-shortfall specialization used in the quantitative section. Consider a compliance cycle with remaining horizon τ . Operator i has expected positive shortfall d_i and takes the expected surrender-month return premium π_A as given. If the operator receives at least one discretionary access opportunity before surrender, it can avoid buying the shortfall at the surrender premium. Normalizing the per-cycle access cost by the early-cycle price level, the value of access is $S_i = d_i \pi_A$ and the normalized access-cost parameter is $\tilde{\psi}_i \equiv \psi_i / P_e$. Choosing access intensity $\lambda_i \geq 0$ delivers access probability $1 - e^{-\lambda_i \tau}$ and convex

per-cycle cost $\tilde{\psi}_i \lambda_i^2 / 2$. The operator's problem is

$$\max_{\lambda_i \geq 0} \left\{ S_i \left(1 - e^{-\lambda_i \tau} \right) - \frac{\tilde{\psi}_i}{2} \lambda_i^2 \right\}, \quad (\text{A.3})$$

which is strictly concave in λ_i because its second derivative is $-S_i \tau^2 e^{-\lambda_i \tau} - \tilde{\psi}_i < 0$. The first-order condition is therefore sufficient and given by $S_i \tau e^{-\lambda_i \tau} = \tilde{\psi}_i \lambda_i$. Multiplying through by $\tau e^{\lambda_i \tau}$ gives $(\lambda_i \tau) e^{\lambda_i \tau} = S_i \tau^2 / \tilde{\psi}_i$, so

$$\lambda_i^* = \frac{1}{\tau} W \left(\frac{S_i \tau^2}{\tilde{\psi}_i} \right), \quad (\text{A.4})$$

where $W(\cdot)$ denotes the Lambert W function on $\mathbb{R}_+$, that is, the inverse of $y \mapsto y e^y$ for nonnegative arguments. Since W is increasing on $\mathbb{R}_+$, λ_i^* is increasing in $S_i = d_i \pi_A$ and decreasing in $\tilde{\psi}_i$. The probability of receiving no discretionary access opportunity before surrender is $e^{-\lambda_i^* \tau}$, so expected residual terminal demand for operator i is

$$R_{iA} = d_i e^{-\lambda_i^* \tau}. \quad (\text{A.5})$$

Aggregating across operators,

$$D_A(\pi_A) = \sum_i d_i e^{-\lambda_i^*(\pi_A) \tau}. \quad (\text{A.6})$$

The equilibrium surrender-month return premium then satisfies the fixed-point condition

$$\pi_A = \varphi D_A(\pi_A) = \varphi \sum_i d_i e^{-\lambda_i^*(\pi_A) \tau}. \quad (\text{A.7})$$

The right-hand side of (A.7) is continuous and strictly decreasing in π_A whenever $\sum_i d_i > 0$: a higher expected premium raises every $\lambda_i^*(\pi_A)$ for which $d_i > 0$ and lowers each corresponding term $d_i e^{-\lambda_i^*(\pi_A) \tau}$. At $\pi_A = 0$, the right-hand side equals $\varphi \sum_i d_i > 0$. As $\pi_A \rightarrow \infty$, each $\lambda_i^*(\pi_A) \rightarrow \infty$, so each term $d_i e^{-\lambda_i^*(\pi_A) \tau} \rightarrow 0$ and therefore the right-hand side converges to zero. The left-hand side is increasing in π_A and equals zero at the origin. Hence there exists a unique positive fixed point whenever $\sum_i d_i > 0$, so the equilibrium surrender-month return premium is a uniquely determined object rather than an assumed primitive. This establishes Proposition 1.

To establish Corollary 1, define $F(\pi, \varphi) \equiv \pi - \varphi D_A(\pi)$ so that the equilibrium is the implicit

solution to $F(\pi_A, \varphi) = 0$. Differentiating the fixed point yields, by the implicit function theorem,

$$\frac{d\pi_A}{d\varphi} = \frac{D_A(\pi_A)}{1 - \varphi D'_A(\pi_A)}. \quad (\text{A.8})$$

The differentiability of $D_A(\cdot)$ on $\mathbb{R}_+$ follows from differentiability of the Lambert W function on $\mathbb{R}_+$. Because $\lambda_i^*(\pi_A)$ is strictly increasing in π_A for each i with $d_i > 0$, the function $D_A(\pi_A) = \sum_i d_i e^{-\lambda_i^*(\pi_A)\tau}$ is strictly decreasing in π_A , that is, $D'_A(\pi_A) < 0$ whenever $\sum_i d_i > 0$. Hence $1 - \varphi D'_A(\pi_A) > 1$. At the equilibrium, $\varphi D_A(\pi_A) = \pi_A$, so dividing both sides of (A.8) by π_A/φ gives

$$\frac{d \log \pi_A}{d \log \varphi} = \frac{1}{1 - \varphi D'_A(\pi_A)} = \frac{1}{1 + \varphi |D'_A(\pi_A)|}, \quad (\text{A.9})$$

which lies strictly in $(0, 1)$ whenever $\sum_i d_i > 0$. The elasticity approaches one in the static-access limit $|D'_A(\pi_A)| \rightarrow 0$, which obtains when normalized access costs $\tilde{\psi}_i$ are uniformly large enough that every λ_i^* becomes insensitive to π_A at the equilibrium. This establishes Corollaries 1 and 2.

To establish Corollary 3, we compute $|D'_A(\pi_A)|$ explicitly. The chain rule gives $dW_i/d\pi_A = W'(x_i) dx_i/d\pi_A$, where $x_i \equiv d_i \pi_A \tau^2 / \tilde{\psi}_i$ and $dx_i/d\pi_A = x_i/\pi_A$. Differentiating the defining identity $W(x)e^{W(x)} = x$ yields $W'(x) = e^{-W(x)} / (1 + W(x)) = W(x) / [x(1 + W(x))]$, where the second equality uses the exact identity $e^{-W(x)} = W(x)/x$. Combining,

$$\frac{dW_i}{d\pi_A} = \frac{W_i}{(1 + W_i) \pi_A}. \quad (\text{A.10})$$

Differentiating $D_A(\pi_A) = \sum_i d_i e^{-W_i}$ then yields

$$D'_A(\pi_A) = -\frac{1}{\pi_A} \sum_i d_i e^{-W_i} \frac{W_i}{1 + W_i}. \quad (\text{A.11})$$

Substituting into (A.9) and using $\varphi/\pi_A = 1/D_A(\pi_A)$ at the equilibrium gives

$$\frac{d \log \pi_A}{d \log \varphi} = \frac{1}{1 + \omega}, \quad \omega = \sum_i s_i \frac{W_i}{1 + W_i}, \quad (\text{A.12})$$

where $s_i \equiv d_i e^{-W_i} / \sum_j d_j e^{-W_j}$ is operator i 's share of aggregate residual terminal demand. Since $W_i/(1 + W_i) \in (0, 1)$ pointwise whenever $d_i, \pi_A, \tau, \tilde{\psi}_i > 0$, ω is a convex combination of values in $(0, 1)$ and therefore $\omega \in (0, 1)$. Hence the elasticity is bounded strictly in $(1/2, 1)$. As $\tilde{\psi}_i \rightarrow 0$ uniformly, $W_i \rightarrow \infty$ for each i with $d_i > 0$, $W_i/(1 + W_i) \rightarrow 1$, $\omega \rightarrow 1$, and the elasticity converges to $1/2$. This establishes Corollary 3.

To establish Corollary 4, suppose $(\varphi, \tilde{\psi}_i) \mapsto (k\varphi, k\tilde{\psi}_i)$ for all i with $k > 0$, and conjecture that the new equilibrium premium is $\pi_A^{\text{new}} = k\pi_A$, where π_A is the equilibrium under the original

parameters. Under the scaled parameters, the Lambert- W argument evaluated at the conjectured equilibrium is

$$\frac{d_i \pi_A^{\text{new}} \tau^2}{k \tilde{\psi}_i} = \frac{d_i (k \pi_A) \tau^2}{k \tilde{\psi}_i} = \frac{d_i \pi_A \tau^2}{\tilde{\psi}_i}, \quad (\text{A.13})$$

identical to the original Lambert- W argument. Hence each scaled access intensity $\lambda_i^{*,\text{new}}(\pi_A^{\text{new}}) = \lambda_i^*(\pi_A)$, and the scaled residual terminal demand $D_A^{\text{new}}(\pi_A^{\text{new}}) = D_A(\pi_A)$. The scaled fixed-point condition becomes

$$\pi_A^{\text{new}} = k \varphi D_A^{\text{new}}(\pi_A^{\text{new}}) = k \varphi D_A(\pi_A) = k \pi_A, \quad (\text{A.14})$$

where the last equality uses the original fixed point $\pi_A = \varphi D_A(\pi_A)$. The conjecture is verified, and uniqueness from Proposition 1 guarantees this is the equilibrium of the scaled system. This establishes Corollary 4.

Aggregation of Compliance Demand

Let c_{it} denote operator i 's compliance allowance holdings and h_{it}^* target compliance holdings. Under the endogenous Poisson access specification (19), the compliance gap $g_{it} = h_{it}^* - c_{it}$ satisfies the reduced-form adjustment equation in the main text,

$$E_t[dg_{it}] = E_t[dh_{it}^*] - \lambda_i^* g_{it} dt. \quad (\text{A.15})$$

Speculative positions are tracked separately from c_{it} in Section 3.5 of the main text and do not enter (A.15). 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 ($s_t \rightarrow 0$). 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 component in Section 3.5 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, starting from date t_0 , the conditional expected gap satisfies

$$E_{t_0}[g_{it}] = e^{-\lambda_i^*(t-t_0)} g_{i,t_0} + \int_{t_0}^t e^{-\lambda_i^*(t-u)} a_i(s_u, X_u) du. \quad (\text{A.16})$$

Aggregating across operators yields

$$E_t[dG_t] = A(s_t, X_t) dt - \sum_i \lambda_i^* g_{it} dt = A(s_t, X_t) dt - \bar{\lambda}_t^* G_t dt, \quad G_t = \sum_i g_{it}, \quad (\text{A.17})$$

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 average of cycle-level chosen access intensities. Equation (A.17) is the aggregate analog of (A.15) and is the law of motion for unresolved aggregate compliance demand. This richer dynamic object motivates the general seasonal discussion in the main text, but it is not the exact object calibrated in Section 4, which instead uses the one-shot specialization $D_A(\pi_A) = \sum_i d_i e^{-\lambda_i^*(\pi_A)\tau}$. A closed-form expression for $E[G_t]$ under a simple seasonal profile and a locally constant $\bar{\lambda}_t^*$ is derived below.

Intermediary Price Impact

There exist intermediaries that provide immediacy by absorbing the contemporaneous net operator order-flow rate M_t . We represent their instantaneous immediacy-provision (or balance-sheet) cost in reduced form by $\frac{\phi}{2} M_t^2$, where $\phi > 0$ indexes limited intermediary capacity. Taking $P_t - P_t^*$ as given, intermediary supply M_t^s satisfies the inverse-supply condition $P_t - P_t^* = \phi M_t^s$. Market clearing at operator demand $M_t^s = M_t$ then yields the equilibrium pricing relation

$$P_t = P_t^*(X_t) + \phi M_t, \quad (\text{A.18})$$

which is equation (32) 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.18) 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 component, under full reset upon access (that is, $\alpha_i = 1$), the expected compliance order-flow rate is $E_t[M_t^{\text{comp}}] = \sum_i \lambda_i^* g_{it} = \bar{\lambda}_t^* G_t$ stated in (34) of the main text; $\bar{\lambda}_t^*$ is the gap-weighted access intensity and inherits the comparative statics of the underlying λ_i^* in $S_i = d_i \pi_A$ and $\tilde{\psi}_i$.

Informed Demand

Let $\mathcal{F}_t = \sigma(X_t)$ denote the public information set generated by the public state X_t , and let $\mathcal{I}_{it} \supseteq \mathcal{F}_t$ denote operator i 's information set, including any private signal. Define the informational advantage μ_{it} in return form as in (35). 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 chosen 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,

$$M_t^{\text{info}} = \sum_{i \in \mathcal{I}} \lambda_i^* \frac{\mu_{it}}{\gamma_i}, \quad (\text{A.19})$$

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 a higher chosen λ_i^* implies faster and more informative flow.

Derivation of Prediction 1 (Slow Participation)

From (A.16), $E_{t_0}[g_{it}]$ is decreasing in λ_i^* pointwise in t when $a_i(\cdot) \geq 0$; in particular, a lower chosen intensity λ_i^* implies a strictly larger expected compliance gap at every point in the compliance cycle. By (A.4), λ_i^* is increasing in $S_i = d_i \pi_A$ and decreasing in $\tilde{\psi}_i$, so operators with high normalized attention or treasury costs $\tilde{\psi}_i$ or small expected shortfalls d_i optimally choose lower access intensity. 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 chosen 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. $\square$

Derivation of Prediction 2 (Seasonal Price Pressure)

Combining (A.18) with the decomposition $M_t = M_t^{\text{comp}} + M_t^{\text{info}}$, and using the full-reset access-weighted aggregation $E_t[M_t^{\text{comp}}] = \bar{\lambda}_t^* G_t$, where $\bar{\lambda}_t^* \equiv \sum_i \lambda_i^* g_{it} / G_t$ is the gap-weighted chosen access intensity,

$$E_t[P_t - P_t^*(X_t)] = \phi \bar{\lambda}_t^* G_t + \phi E_t[M_t^{\text{info}}]. \quad (\text{A.20})$$

Equation (A.17) implies that $E_t[G_t]$ is strictly larger in calendar subintervals where $A(s_t, X_t)$ is larger, in particular as $s_t \rightarrow 0$ from above. Hence $E_t[P_t - P_t^*(X_t)]$ is strictly positive near surrender whenever $\phi > 0$ and $\bar{\lambda}_t^* > 0$. Comparative statics in φ are immediate from the fixed-point relation (A.7). Comparative statics in access costs follow from (A.4): higher normalized access costs $\tilde{\psi}_i$ lower each λ_i^* , raise expected residual terminal demand D_A , and—through the equilibrium fixed point (A.7)—raise π_A . 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. $\square$

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 $E[P_{t+\Delta}^* - P_t^* | \mu_{it}] = c \mu_{it}$ with $c > 0$ whenever $\mathcal{I}_{it}$ contains a non-trivial signal about the future benchmark price. Because M_t^{info} in (A.19) is a linear combination of $\lambda_i^* \mu_{it} / \gamma_i$ across informed operators, $E[P_{t+\Delta}^* - P_t^* | M_t^{\text{info}}]$ is strictly increasing in M_t^{info} , and hence operator flow imbalance forecasts future benchmark-price innovations. 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 chosen access intensity. Therefore the predictive content of M_t^{info} is larger for operators who choose high λ_i^* , that is, those with low normalized access costs $\tilde{\psi}_i$ and large benefits of early trading. Cross-sectional comparative statics in γ_i follow from $q_{it}^{\text{info}} = \mu_{it} / \gamma_i$. This establishes the benchmark-price-innovation form of Prediction 3; the actual-return extension follows from (A.21)–(A.22) below under the maintained reduced-form condition that contemporaneous price impact under-reveals the private signal. $\square$

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

$$P_t = P_t^*(X_t) + \phi M_t, \quad (\text{A.21})$$

so that

$$P_{t+\Delta} - P_t = (P_{t+\Delta}^* - P_t^*) + \phi (M_{t+\Delta} - M_t). \quad (\text{A.22})$$

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

$$E_t[P_{t+\Delta}^* - P_t^* | M_t^{\text{info}}] = \beta M_t^{\text{info}}, \quad (\text{A.23})$$

and the expected reversal of contemporaneous price impact is smaller than the signal-revelation component, then $E_t[P_{t+\Delta} - P_t | M_t^{\text{info}}]$ is increasing in M_t^{info} . In the special case in which $E_t[M_{t+\Delta} | M_t^{\text{info}}]$ 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.

The frequent-trader split used in the empirical section should therefore also be read in reduced-form terms. In a richer extension, the value of market access could include both compliance and speculative components, for example $S_i = d_i \pi_A + V_i^{\text{info}}$ in return units, so that frequent trading could reflect lower access frictions, higher informational value of access, or both. We do not estimate that extension here.

April Premium Paid by Delayed Buyers and Real Frictions

This subsection is not intended as a full resource-cost accounting exercise. Its purpose is to clarify how the empirical measure of the April premium paid by delayed buyers differs from aggregate resource costs. The empirical calculation measures a financial premium paid in April 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:

$$W = E_0 \int_0^\infty e^{-\rho t} \left[- \sum_i \frac{1}{2} (\theta_{it} - x_{it})^2 - \sum_i \frac{\psi_i}{2} \lambda_i^2 - \frac{\phi}{2} M_t^2 \right] dt, \quad (\text{A.24})$$

where $\rho > 0$ 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 access-cost term $\sum_i \frac{\psi_i}{2} \lambda_i^2$ in Equation (A.24) should be read as a continuous-time analogue of the per-cycle access costs used in the main model: in the quantitative implementation, access costs are imposed at the compliance-cycle level rather than as continuously paid flow 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 April premium paid by delayed buyers is measurable, but it is not a direct measure of aggregate resource costs because allowance payments are partly transfers across market participants.

Closed-Form Seasonal Compliance Gap and Price Premium

This subsection displays a simple closed-form object that illustrates 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

$$A(u) = \begin{cases} a_0, & 0 \leq u < T - \tau, \\ a_0 + a_1, & T - \tau \leq u \leq T, \end{cases} \quad (\text{A.25})$$

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

$$\frac{dG(u)}{du} = A(u) - \bar{\lambda} G(u), \quad (\text{A.26})$$

where $\bar{\lambda}$ is treated as a constant proxy for the endogenous access-adjustment rate $\bar{\lambda}_t^*$ for transparency in this displayed example, and, for expositional simplicity, take $G(0) = 0$. Equation (A.26) is a linear first-order ODE with piecewise-constant forcing, and its solution is

$$G(u) = \frac{a_0}{\bar{\lambda}} \left(1 - e^{-\bar{\lambda}u} \right), \quad 0 \leq u < T - \tau, \quad (\text{A.27})$$

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

$$G(u) = G(T - \tau) e^{-\bar{\lambda}(u - (T - \tau))} + \frac{a_0 + a_1}{\bar{\lambda}} \left(1 - e^{-\bar{\lambda}(u - (T - \tau))} \right), \quad (\text{A.28})$$

with $G(T - \tau) = \frac{a_0}{\bar{\lambda}} (1 - e^{-\bar{\lambda}(T - \tau)})$. Combining (A.18) with $M_t^{\text{comp}} = \bar{\lambda} G_t$ and abstracting from M_t^{info} in expectation delivers the seasonal-premium representation

$$P(u) - P^*(u) = \phi \bar{\lambda} G(u), \quad (\text{A.29})$$

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

Equations (A.27)–(A.29) 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:

$$M_T^{\text{term}} = \sum_i g_{iT-}^+. \quad (\text{A.30})$$

Under the same reduced-form immediacy-supply relation (A.18), and treating the surrender date as a windowed event of length h_T over which M_T^{term} and M_T^{info} are stock-equivalent integrated flows,

$$P_T - P_T^*(X_T) = \frac{\phi}{h_T} \left(M_T^{\text{term}} + M_T^{\text{info}} \right). \quad (\text{A.31})$$

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.27)–(A.29) 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 $\phi\bar{\lambda}$, which combines intermediary price impact and the access-weighted closing rate at which active operators close unresolved gaps. In this displayed example we treat $\bar{\lambda}$ as a constant proxy for $\bar{\lambda}_t^*$ for transparency; in the general formulation, the gap-weighted chosen access intensity $\bar{\lambda}_t^*$ depends on the cross-sectional distribution of g_{it} and on the chosen access intensities λ_i^* .

We emphasize the scope of this closed-form object. The affine benchmark price P_t^* and the aggregate compliance-gap law (A.17) are standard. The continuous-time model provides a dynamic analogue of the participation cost in the two-period model through the endogenous access choice in (A.4), but $E_t[M_t^{\text{comp}}] = \bar{\lambda}_t^* G_t$ remains an access-weighted aggregation step rather than a fully micro-founded market-clearing condition, and the informed-flow component (A.19) is kept in conditional expected-flow form for tractability rather than derived from a filtering or Kyle-style equilibrium. The full periodic affine system with time-varying coefficients in B_t and z_t admits closed-form solutions in special limits, but we do not pursue a general exact solution here. The displayed example is the core tractable object 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 Quantitative Robustness for Section 4

This appendix documents the construction details, sample-period harmonization, shortfall robustness, premium definitions, and price-impact diagnostics underlying Section 4.

Sample harmonization

Table A.2 reports operator-year counts (one observation per operator and compliance cycle) for the headline 2005–2021 sample and for two robustness samples: one excluding the non-bankable Phase I pilot and one restricted to mature-market Phase III years.

Table A.2: Sample harmonization for the quantitative analysis

Sample	Years	N operator-years	N operators	Notes
Main sample	2005– 2021	109,247	10,696	Baseline
Post-Phase-I	2008– 2021	94,810	10,593	Excludes non-bankable pilot phase
Phase III onward	2013– 2021	62,828	9,448	Mature market robustness

Shortfall measures

Table A.3 reports four alternative shortfall constructions. The allocation baseline d_{it}^{alloc} is the headline measure used in Section 4; the lagged-bank measure is the preferred bank-adjusted robustness; cumulative-surplus and lagged-surplus measures are simpler fallbacks. Aggregate shortfalls move within a band of approximately five percent across the four constructions, but the share of operator-cycles with $d_{it} > 0$ is meaningfully smaller under the bank-adjusted measures.

Table A.3: Headline shortfall measures

Shortfall measure	Mean d_{it}	Median d_{it}	Share $d_{it} > 0$	Aggregate d_{it}	Notes
Allocation baseline	101,621	0	46.4%	10,831,543,430	Main
Lagged bank	97,208	0	31.5%	10,361,253,174	Preferred robustness
Cumulative surplus	96,356	0	31.0%	10,270,378,475	Fallback robustness
Lagged surplus	99,542	0	43.1%	10,609,984,018	Simple fallback

April premium definitions

Table A.4 summarizes the four premium measures used to discipline the surrender-window return-impact calibration. The raw April log return is the headline; the non-April-adjusted measure subtracts the average non-April monthly return; the residualized measure is the residual from a return regression on lagged returns, lagged volatility, and phase indicators; the raw signed series retains negative-premium years.

Table A.4: Headline April premium definitions

Premium measure	Mean	Median	Share positive	Std. dev.	N years
Raw April	0.104	0.112	64.7%	0.087	17
Non-April adjusted	0.055	0.130	76.5%	0.224	17
Residualized	0.059	0.112	76.5%	0.213	17
Raw signed April	0.046	0.112	64.7%	0.228	17

Return-impact calibration

Table A.5 reports headline $\hat{\varphi}$ across premium definitions and estimation samples. The positive-only rows restrict to years with positive premia and report the median across years; the all-year OLS rows estimate $\pi_{A,t} = \varphi D_{A,t}^{\text{model}} + \varepsilon_t$ on all available years; the NNLS rows impose $\varphi \geq 0$. Order-of-magnitude estimates are stable across specifications. The all-year OLS R^2 s are small because annual April returns are noisy relative to the slow variation in $D_{A,t}^{\text{model}}$; we therefore prefer the positive-year median as the headline calibration. As in the main text, these values should be read as calibrated return-impact sensitivities rather than as tightly identified structural primitives.

Table A.5: Headline return-impact calibration across premium definitions

Premium measure	Sample	$\hat{\varphi}$	Bootstrap CI	R^2	N years
Raw April	Positive only	7.83×10^{-10}			11
Raw April	All years OLS	4.10×10^{-10}	$[3.00, 5.61] \times 10^{-10}$	0.0006	17
Raw April	All years NNLS	4.10×10^{-10}	$[3.00, 5.61] \times 10^{-10}$	0.0006	17
Non-April adjusted	Positive only	8.62×10^{-10}			13
Non-April adjusted	All years OLS	1.84×10^{-10}	$[-0.84, 4.45] \times 10^{-10}$	0.0042	17
Non-April adjusted	All years NNLS	1.84×10^{-10}	$[0, 4.45] \times 10^{-10}$	0.0042	17
Residualized	Positive only	7.22×10^{-10}			13
Residualized	All years OLS	1.71×10^{-10}	$[-0.87, 4.26] \times 10^{-10}$	0.0173	17
Residualized	All years NNLS	1.71×10^{-10}	$[0, 4.26] \times 10^{-10}$	0.0173	17

Notes: All rows use the headline 2005–2021 sample and the allocation shortfall proxy d_{it}^{alloc} . Positive-only rows use positive-premium years only. OLS and NNLS rows use all available years for the selected premium measure. Bootstrap CIs use 1,000 replications drawing yearly observations with replacement. The reported $\hat{\varphi}$ values are calibration targets for return-impact sensitivity.

Table A.6: Leave-one-out robustness for the headline $\hat{\varphi}$

Specification	Median LOO $\hat{\varphi}$	Min	Max	Median R^2
Raw April, OLS	4.05×10^{-10}	3.71×10^{-10}	4.67×10^{-10}	0.0035
Raw April, NNLS	4.05×10^{-10}	3.71×10^{-10}	4.67×10^{-10}	0.0035

Notes: The leave-one-out exercise drops one calendar year at a time and re-estimates the headline φ specification on the remaining 16 years. We report the median, minimum, and maximum across the 17 leave-one-out runs.

A.7 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 :

$$\begin{aligned} \max_{x_{i0}, x_{i1}, m_{i0}, m_{i1}, b_{i1}} & \left[-\frac{1}{2} (\theta_{i0} - x_{i0})^2 - p_0^* m_{i0} \right] \\ & + \delta E_0 \left[-\frac{1}{2} (\theta_{i1} - x_{i1})^2 - p_1^* m_{i1} \right] \end{aligned} \quad (\text{A.32})$$

subject to

$$b_{i1} = b_{i0} + q_{i0} + m_{i0} - x_{i0}, \quad x_{i1} = b_{i1} + q_{i1} + m_{i1}, \quad b_{i1} \geq 0. \quad (\text{A.33})$$

Figure 1 in the main text summarizes the timing of the two-period model and the role of intermediation.

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

$$\theta_{i0} - x_{i0} = \xi_{i0}, \quad (\text{A.34})$$

$$\theta_{i1} - x_{i1} = \xi_{i1}, \quad (\text{A.35})$$

$$p_0^* = \xi_{i0}, \quad (\text{A.36})$$

$$p_1^* = \xi_{i1}, \quad (\text{A.37})$$

$$\xi_{i0} = \delta E_0[\xi_{i1}]. \quad (\text{A.38})$$

Let $M_t = \sum_i m_{it}$ denote aggregate operator net order flow. There exist intermediaries that provide immediacy. In reduced form, at price p_t they are 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

$$p_t - p_t^* = \phi M_t^s, \quad (\text{A.39})$$

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

$$p_t = p_t^* + \phi M_t. \quad (\text{A.40})$$

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:

$$\theta_{i0} - x_{i0} = \xi_{i0}, \quad \theta_{i1} - x_{i1} = \xi_{i1}, \quad (\text{A.41})$$

$$p_0^* = \xi_{i0}, \quad p_1^* = \xi_{i1}, \quad (\text{A.42})$$

$$p_0^* = \delta E_0[p_1^*], \quad (\text{A.43})$$

$$p_t = p_t^* + \phi M_t. \quad (\text{A.44})$$

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

$$p_e d_i + K_i \leq E_0[p_A] d_i, \quad (\text{A.45})$$

which is equivalent to

$$d_i (E_0[p_A] - p_e) \geq K_i. \quad (\text{A.46})$$

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

$$p_\tau = p_\tau^* + \phi M_\tau. \quad (\text{A.47})$$

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

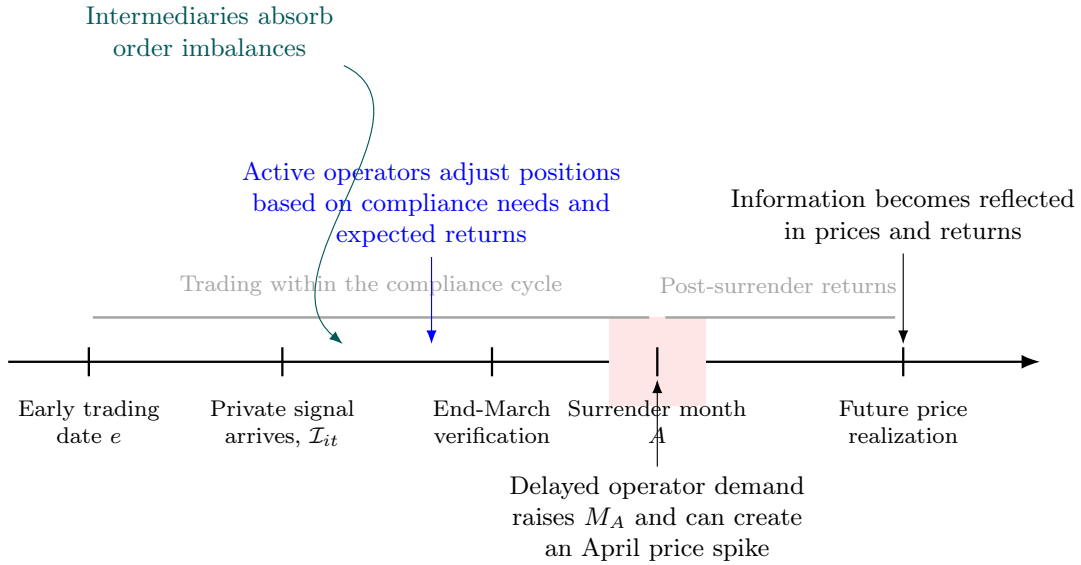
Suppose an active firm receives a private signal summarized by the information set $\mathcal{I}_{it}$. Let $\mathcal{F}_t$ denote the public information set. Consistent with the main text, define

$$\mu_{it} = E_t[p_{t+1}^* - p_t^* | \mathcal{I}_{it}] - E_t[p_{t+1}^* - p_t^* | \mathcal{F}_t]. \quad (\text{A.48})$$

Figure A.2 summarizes the within-period timeline between the early trading date e and the surrender month A , including private-signal arrival, verification, surrender, and subsequent return realization. This timing is useful for interpreting both the predictable April price pressure and the return patterns documented in the main text.

Figure A.2: 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

$$\max_{n_{it}} \left\{ \mu_{it} n_{it} - \frac{\gamma_i}{2} n_{it}^2 \right\}. \quad (\text{A.49})$$

The first-order condition is

$$\mu_{it} - \gamma_i n_{it} = 0, \quad (\text{A.50})$$

so the optimal speculative position is

$$n_{it} = \frac{\mu_{it}}{\gamma_i}. \quad (\text{A.51})$$

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

private information.

April Premium Paid by Delayed Buyers and Participation Frictions

This subsection gives the two-period version of the distinction between the April premium paid by delayed buyers and underlying real frictions. The object of interest is the financial premium paid by a delayed buyer, not a direct resource-cost measure. The additional payment from delayed purchases is given by

$$\text{Premium}_i^{\text{delay}} = d_i(p_A - p_e), \quad (\text{A.52})$$

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 payments 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

$$\frac{1}{2} \sum_i (\theta_{it} - x_{it})^2. \quad (\text{A.53})$$

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

$$\frac{1}{2} \sum_i (\theta_{it} - x_{it})^2 - \frac{1}{2} \sum_i (\theta_{it} - x_{it}^*)^2 = \frac{1}{2} \sum_i (\Delta x_{it})^2. \quad (\text{A.54})$$

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:

$$L = \frac{1}{2} \sum_i (x_{i0} - x_{i0}^*)^2 + \frac{\delta}{2} E_0 \left[\sum_i (x_{i1} - x_{i1}^*)^2 \right] + \sum_i a_i K_i + \frac{\phi}{2} (M_0^2 + \delta E_0[M_1^2]). \quad (\text{A.55})$$

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

$$K_i \geq d_i (E_0[p_A] - p_e), \quad (\text{A.56})$$

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