Overlapping Climate Policies*

Grischa Perino[1] Robert A. Ritz[2] Arthur A. van Benthem[3,4]

June 2021

Abstract

Major carbon-pricing systems in Europe and North America involve multiple countries or states. Individual jurisdictions often pursue additional initiatives—such as unilateral carbon price floors, legislation to phase out coal, aviation taxes or support programs for renewable energy—that overlap with the wider carbon-pricing system. We develop a general framework to study how the climate benefit of such overlapping policies depends on their design, location and timing. Some policies leverage additional climate benefits elsewhere in the system while others backfire by raising aggregate emissions. Our model encompasses almost every type of carbon-pricing system used in practice.

Keywords: overlapping policy, internal carbon leakage, waterbed effect, cap-and-trade, carbon pricing, hybrid regulation

JEL codes: H23 (externalities), Q54 (climate)

*We thank Severin Borenstein, Dallas Burtraw, Jim Bushnell, Reyer Gerlach, Marten Ovaere, Sebastian Rausch, Mar Reguant, Knut Einar Rosendahl, Herman Vollebergh, Maximilian Willner and Yuting Yang for helpful comments and suggestions. We also thank seminar and conference participants at the Cambridge EPRG 2019 Spring Seminar, EAERE 2019 Annual Meeting, AERE 2021 Summer Conference, the 9th Mannheim Conference on Energy and the Environment, the Toulouse School of Economics, the SWEEEP seminar at ZEW Mannheim, Florence School of Regulation at the European University Institute and the University of Pennsylvania. Perino's research was funded by the DFG (German Research Foundation) under Germany's Excellence Strategy, cluster EXC 2037 "Climate, Climatic Change, and Society" (project 390683824) and ARIADNE (BMBF project 03SFK5S0). Van Benthem thanks the National Science Foundation (award SES1530494), the Kleinman Center for Energy Policy at the University of Pennsylvania, the Analytics at Wharton Data Science and Business Analytics Fund, and the Wharton Dean's Research Fund for support.

[1]Department of Socioeconomics and Center for Earth System Research and Sustainability (CEN), University of Hamburg, Germany. Email: Grischa.Perino@uni-hamburg.de. [2]Energy Policy Research Group (EPRG) and Judge Business School, University of Cambridge, United Kingdom. Email: rar36@cam.ac.uk. [3]The Wharton School, University of Pennsylvania, 327 Vance Hall, 3733 Spruce Street, Philadelphia, PA 19104, United States. Phone: +1 (215) 898-3013. Fax: +1 (215) 898-7635. Email: arthurv@wharton.upenn.edu. [4]National Bureau of Economic Research.

1 Introduction

The world is under increasing pressure to deliver on the ambition of the 2015 Paris Climate Agreement, and over 60 national and sub-national jurisdictions are putting a price on carbon emissions (World Bank, 2020). Two features of the carbon-pricing landscape are striking. First, by using hybrid designs that combine elements of price and quantity regulation, practice has run far ahead of the simple carbon tax and cap-and-trade policies emphasised in economics textbooks. North American carbon markets such as the Regional Greenhouse Gas Initiative (RGGI) use price floors and ceilings as a flexibility mechanism to contain the variability of the allowance price. Since its 2018 reform, the European Union's Emissions Trading System (EU ETS) features a complex mechanism that cancels allowances under certain market circumstances. Second, major carbon-pricing systems involve multiple jurisdictions: the EU ETS spans 30 countries while RGGI covers power generation in eleven states in the northeastern United States.

Individual jurisdictions, in turn, often pursue unilateral climate initiatives that overlap with the wider carbon-pricing system. The EU is a classic example, with individual countries "doing more" than what is centrally provided by the EU ETS. The UK introduced in 2013 a fee that added £18/tCO_2 to the allowance price faced by its power generators under the EU ETS; the Netherlands introduced in 2021 a unilateral carbon price floor for its industrial sectors and has also announced one for electricity.[1] A plethora of national policies exists to support renewable energy, phase out coal-fired power, and levy extra carbon taxes on air travel.[2] These examples share a common feature: they are policies by an individual jurisdiction that operate alongside a wider carbon-pricing system.

Our question in this paper is simple: What is the climate benefit of such overlapping policies? As it is a global public good, any mitigation of climate change will be driven solely by changes in aggregate emissions. For a cap-and-trade system with a fixed emissions cap, like the pre-2018 EU ETS, the answer is clear: if an overlapping policy reduces EU-wide emissions demand (say, from power generation) by 1 ton of CO_2, this will be precisely offset by increased demand of 1 tCO_2 elsewhere in the system—the "waterbed effect" is 100%. At the opposite end, a simple carbon tax does not have an emissions cap and so the waterbed effect is zero. Our main interest is in real-world hybrid carbon-market designs which typically feature dynamic "punctured" waterbeds that lie between

[1] The EU ETS includes power generation, industrial sectors, and domestic aviation and is the world's largest carbon-pricing system. The UK's Carbon Price Support has been hailed as "perhaps the clearest example in the world of a carbon tax leading to a significant cut in emissions" (New York Times, 2019).

[2] Under the EU's 2009 Renewables Directive, each member state developed a national action plan aimed at increasing the share of renewables in its energy mix. The Powering Past Coal Alliance groups national and sub-national governments, including twelve EU member states, committed to phasing out coal. Motivations for overlapping policies range from climate benefits to concerns about low or volatile carbon prices to other market failures such as innovation externalities (Newbery et al., 2019).

these two extremes. This enables overlapping policies to have a global climate benefit.

Yet this chain of reasoning still has a missing link which we refer to as "internal carbon leakage". Suppose that a unilateral Dutch carbon price on power generation reduces its domestic emissions demand by 1 tCO_2 but, within an integrated European electricity market, this leads to an increase in Dutch electricity imports which in turn raises emissions demand by 1 tCO_2 in other EU ETS countries. This overlapping policy has no climate benefit either: its rate of internal carbon leakage is 100%. This conclusion, in turn, applies irrespective of the extent of the waterbed effect. In sum, the answer to our question must be driven by a combination of the waterbed effect and internal leakage.

This paper provides a novel integrated approach through which to understand and quantify the overall emissions impact of an overlapping policy that applies only to part of a multi-jurisdiction carbon-pricing system. Section 2 presents a model-independent conceptual framework that provides a mapping from the "local" emissions reduction the overlapping policy achieves to its "global" impact which includes any knock-on effects elsewhere in the system. Internal carbon leakage captures emissions displacement within the system (e.g., greater product imports from a neighbouring country) for a given system-wide carbon price. The waterbed effect endogenises the policy's interaction with the system's carbon price (and any emissions cap). A distinguishing feature of this paper is to combine both leakage and waterbed effects within an integrated framework.

Section 3 presents a theory of internal carbon leakage that focuses on emissions displacement between different jurisdictions in the same sector. We consider two groups of overlapping policies: "supply-side" policies that unilaterally raise the carbon price or directly limit emissions-intensive production, and "demand-side" policies that reduce the demand for emissions-intensive production, e.g., by promoting renewables or energy efficiency.[3] We show that supply-side policies have positive internal leakage—sometimes in excess of 100%—as they raise emissions demand from other parts of the system to "fill the gap" due to lower domestic production. By contrast, demand-side policies have negative internal leakage as they also displace imported emissions. While some recent empirical work has estimated internal leakage for specific overlapping policies (Vollebergh, 2018; Abrell et al., 2019; Gerarden et al., 2020), our first contribution is to provide new theoretical insight into its economics across a wide range of commonly-used policies.[4]

Section 4 introduces a general two-period analysis of the waterbed effect. While the literature has studied the waterbed effect in specific circumstances, notably the EU ETS

[3]Our use of the term "supply-side" policy differs from the literature which focuses on the market for fossil resources (Sinn, 2008; Harstad, 2012)—our reference point is the market for goods produced by a polluting industry. We discuss broader connections with this strand of research in the conclusion.

[4]Internal carbon leakage as a result of overlapping policies has also been studied outside of the context of a carbon-pricing system; see, e.g., Goulder and Stavins (2011) and Goulder et al. (2012) on interactions between federal and state-level policies in the United States.

(Fankhauser et al., 2010; Böhringer, 2014; Perino, 2018; Gerlagh et al., 2021), there is still very limited understanding of its operation across different types of hybrid carbon-market designs that puncture the waterbed. Our model encompasses price-based flexibility mechanisms based on past allowance prices (including price ceilings and floors) (Roberts and Spence, 1976; Pizer, 2002; Newell et al., 2005; Borenstein et al., 2019; Burtraw et al., 2020), quantity-based flexibility mechanisms based on past allowance banking, and a simple carbon tax and cap-and-trade.[5] We uncover a natural connection between the extent of the waterbed and classic principles from the literature on tax incidence (Jenkin, 1872; Weyl and Fabinger, 2013). Our second contribution, therefore, is to bring together waterbed-effect results from prior literature in a unifying framework that covers almost every type of carbon-pricing system used in practice, and connect them to simple economic principles.

Section 5 illustrates the empirical usefulness of the modelling framework. It derives values for internal leakage and the waterbed effect using a combination of simple formulae from our theory results and prior empirical work. We cover overlapping policies in Europe and in North American carbon-pricing systems such as RGGI, the California-Québec carbon market, and Canada's federal minimum carbon price (see Figure 4). Consistent with our theory, we find that supply-side (demand-side) overlapping policies have positive (negative) internal leakage. Our results illustrate how a policy's overall climate benefit varies widely depending on its design, location and timing. Section 6 concludes.

We hope that our analysis will also be of value to policymakers. It provides practical guidance on the climate benefits of 25 different combinations of overlapping policy instruments (see Figure 1) and types of carbon-pricing designs (see Figure 3). The introduction of flexibility mechanisms in cap-and-trade systems has, in part, been motivated by a desire to make unilateral policies more effective. For example, in designing the EU ETS's new Market Stability Reserve, the European Union noted that "the reserve will also enhance synergy with other climate and energy policies" (European Parliament and Council, 2015)—thus alluding to what are often termed "complementary" policies. Our analysis highlights greater subtlety: with a punctured waterbed, some policies are truly complementary in that they induce further emissions reductions elsewhere but those with very high internal carbon leakage can now backfire by raising aggregate emissions. Our results can inform a cost-benefit analysis of a new overlapping policy as well as how a change in market design alters the economics of pre-existing policies.

Finally, our focus in this paper differs from external carbon leakage to jurisdictions *outside* a carbon-pricing system. Prior literature has examined the global impacts of unilateral policy in sectors such as cement and steel where the scope of the product market

[5] A two-period model is necessary to be able to incorporate banking of allowances in a cap-and-trade system which, in turn, can interact with the extent of the waterbed. We also derive the waterbed effect of the multi-period EU ETS Market Stability Reserve.

is wider than that of the carbon price.[6] We here explore leakage among jurisdictions *inside* the system because (i) it is less well-understood in the literature, in part because it did not matter in systems with an 100% waterbed effect like the pre-2018 EU ETS; and (ii) it has received much less policy attention, despite likely being more important than its external cousin for sectors such as airlines and electricity.[7]

2 Conceptual framework

We begin by setting out a simple conceptual framework that encompasses a wide range of carbon-market designs and highlights the dual role of internal carbon leakage and the waterbed effect in determining the climate benefit of different overlapping policies.

Consider a multi-jurisdiction carbon-pricing system that may cover a single sector (like RGGI) or multiple sectors (like the EU ETS). An "overlapping policy", in general, is any unilateral policy that applies only to part of the system; our leading example is a policy by a single jurisdiction that hence applies only to a subset of competing firms in a sector. For simplicity, we consider two time periods, $t = 1, 2$, and think of the first period as the short run and the second period as the long run. Denote by $\boldsymbol{\tau} = (\tau_1, \tau_2)$ the system-wide carbon price at each time, which is determined by the carbon-market design.

We are interested in unilateral policies by country (jurisdiction) i that, holding fixed the carbon price path $\boldsymbol{\tau}$, are successful at reducing i's domestic demand for emissions in each period, $\Delta e_{it} < 0$, and hence also $\Delta e_i \equiv \Delta e_{i1} + \Delta e_{i2} < 0$ over time. Let Δe_t^* denote the policy's impact on aggregate emissions across all countries at time t at equilibrium carbon prices (relative to a baseline without the unilateral policy). Our main question is, what is the policy's impact on cumulative equilibrium emissions, $\Delta e^* \equiv \Delta e_1^* + \Delta e_2^*$? This is the critical issue for the policy's effectiveness in combating climate change.

Our framework answers this question using two concepts. First, internal leakage captures emissions displacement within the system (e.g., greater product imports from a neighbouring country) for a given system-wide carbon price. We define the rate of internal carbon leakage associated with i's policy at time t as:

$$L_{it} \equiv -\Delta e_{-it}/\Delta e_{it}, \tag{1}$$

where Δe_{-it} is the change induced by i's policy in the emissions demand of other countries that are part of the carbon-pricing system.[8] Therefore $\Delta e_t \equiv [1 - L_{it}]\Delta e_{it}$ represents the

[6] See Martin et al. (2014); Aldy and Pizer (2015); Fowlie et al. (2016); Fowlie and Reguant (2018).

[7] Yet another form of carbon leakage occurs when, in the same jurisdiction, some sectors are not covered by the carbon-pricing system (Baylis et al., 2013; Jarke and Perino, 2017); an example is leakage from covered EU ETS sectors like electricity to uncovered sectors such as transport.

[8] Notice that this is akin to the standard definition of "external" carbon leakage (e.g., IPCC, 2007)

(net) system-wide change in emissions demand at time t so $\Delta e \equiv \Delta e_1 + \Delta e_2$ is the cumulative system-wide change in emissions demand due to the policy (for fixed $\boldsymbol{\tau}$).

Second, the waterbed effect then captures the system-wide impacts arising from any induced changes to the equilibrium path of the system-wide carbon price:

$$W \equiv 1 - \Delta e^*/\Delta e. \tag{2}$$

This translates the system-wide change in emissions demand due to i's policy into an equilibrium change in cumulative emissions that incorporates any induced changes to the carbon price path. A textbook cap-and-trade system with a fixed emissions cap has $W = 1$ while a carbon tax has $W = 0$; real-world hybrid carbon-market designs typically feature punctured waterbeds: $W \in (0, 1)$.

We can now state the central equation of our conceptual framework.

Lemma 1 *The equilibrium change in cumulative emissions due to an overlapping policy is:*

$$\Delta e^* = [1 - L_i][1 - W]\Delta e_i, \tag{3}$$

where $L_i \equiv \beta_i L_{i1} + (1 - \beta_i)L_{i2}$ *is the average internal leakage rate and* $\beta_i \equiv \Delta e_{i1}/\Delta e_i \in [0, 1]$ *is the share of the policy's impact on emissions demand that materialises in the first period.*

Lemma 1 incorporates the equilibrium carbon price path via the waterbed effect. It shows how internal carbon leakage and the waterbed effect together drive the sign and magnitude of the overlapping policy's impact on cumulative equilibrium emissions. Letting $R_i \equiv [1 - L_i][1 - W]$, we can think of policies for which leakage and waterbed effects are such that $R_i \geq 1$ as complementary (or super-additive) policies while those for which $R_i < 1$ are substitutes (or sub-additive). If $R_i < 0$, substitutability is so strong that "global" emissions rise ($\Delta e^* > 0$) even though "local" emissions fall ($\Delta e_i < 0$).

We do not attempt to explain the policy's impact on i's domestic emissions demand, $\Delta e_i < 0$; rather we are interested in the mapping from a given policy-driven local impact Δe_i to the equilibrium global impact Δe^*. We also do not attempt to endogenise the extent to which the policy operates in the short run or the long run, β_i; rather we will explore later the impact of β_i on the extent of the waterbed effect. In short, we take the unilateral policy's size (Δe_i) and its period-by-period timing (β_i) as given, and ask to what extent it translates into a long-run global climate gain.

Leveraging this conceptual framework, the remainder of the paper proceeds in three steps. First, we derive the rate of internal carbon leakage L_{it} between different jurisdictions in the same sector for a range of overlapping policies. Given the policy-timing

that relates to shifting of emissions to jurisdictions *outside* the system.

parameter β_i, these can be aggregated to give the average leakage rate L_i. Second, given the policy's induced changes to aggregate emissions demand $(\Delta e_1, \Delta e_2)$, we derive the extent of the waterbed effect W under different carbon-market designs. Third, we illustrate the empirical usefulness of the framework by deriving values for (L_i, W) for real-world overlapping policies in Europe and North America.

3 A model of internal carbon leakage

We next present a new theory of internal carbon leakage at the sectoral level. We consider two groups of overlapping policies. First, "supply-side" policies that unilaterally raise the carbon price for emissions-intensive production or directly reduce dirty production as in a coal phase-out. Second, "demand-side" policies that reduce the (residual) demand for emissions-intensive production, e.g., by promoting renewables or energy efficiency.

We derive simple intuitive formulae for the equilibrium rate of period-by-period internal carbon leakage L_{it}. We show that the economics of internal carbon leakage is similar for policies within each group but differs markedly across the two groups. In particular, leakage is always positive—and can exceed 100%—for the former group but is negative for the latter. (The model is static so time subscripts are omitted to simplify notation.)

3.1 Model setup

There are two countries, i and j, where the latter can be interpreted as an aggregate of all countries except i. A representative firm in each country k produces output x_k $(k = i, j)$ and the firms face a demand function $p(X)$ for their product, where $X \equiv x_i + x_j$ is total output and $p'(\cdot) < 0$. This formulation assumes, for simplicity, that countries' products are perfect substitutes. This is a standard assumption for electricity markets, and often a reasonable approximation for goods produced by other carbon-intensive industries.[9] Firm k's emissions are $e_k = e_k^0 - a_k$ where a_k is abatement and $e_k^0 = \theta_k x_k$ is baseline emissions, for which the emissions intensity θ_k is the "dirtiness" of k's marginal plant.

In general, firm k's cost function $G_k(x_k, a_k)$ depends on its output and abatement. For expositional convenience, we focus in the main text on the case in which these costs are separable, $G_k(x_k, a_k) \equiv [C_k(x_k) + \phi_k(a_k)]$.[10] (In Appendix A.1, we solve the model

[9] The qualitative nature of our results—notably the sign of internal leakage under different overlapping policies—does not hinge on the assumption of perfect substitutes. For example, a model with imperfect substitution between i and j due to "iceberg costs" (Samuelson, 1954) would yield similar findings. Compared with the literature on *external* carbon leakage, the assumption of perfect substitutes is likely more accurate in our context. Simply put, substitutability between, say, French and German products within the EU ETS is likely higher than between France and China or the US.

[10] A separable cost function can be interpreted as an end-of-pipe technology which cleans up production *ex post*. Examples include carbon capture and storage (CCS) and the purchase of carbon offsets.

with general cost functions.) For a well-behaved solution, we assume $C_k(0) = C_k'(0) = 0$, $C_k'(x_k) > 0$ for $x_k > 0$, and $C_k''(x_k) > 0$ as well as $\phi_k'(a_k) > 0$ for $a_k > 0$ and $\phi_k''(a_k) > 0$.

It will be useful to have a simple metric for the extent of abatement opportunity for firm k. We can think of its cost function in terms of output and emissions, $G_k(x_k, e_k) \equiv [C_k(x_k) + \phi_k(\theta_k x_k - e_k)]$ and define the following:

$$A_k \equiv \left(1 - \frac{G_k^{xe}}{G_k^{xx}} \frac{G_k^{ex}}{G_k^{ee}}\right) = \frac{C_k''}{[C_k'' + \theta_k^2 \phi_k'']} \in [0, 1). \tag{4}$$

The limiting case with $A_k \to 1$ corresponds to abatement costs becoming linear ($\phi_k'' \to 0$) so emissions can be reduced without resorting to production cuts. The case in which $A_k \to 0$ corresponds to a Leontief technology: emissions are proportional to output, as abatement is infeasible ($\phi_k'' \to \infty$) so emissions remain at their baseline level, $e_k = \theta_k x_k$.[11]

Firm k faces a carbon price τ_k on each unit of emissions, which depends on the carbon price τ that is common to both countries as part of a wider carbon-pricing system. To maximise profits, firm k solves $\max_{x_k, a_k} \Pi_k = p x_k - G_k(x_k, a_k) - \tau_k e_k$. Note it is equivalent for a firm to choose its emissions or abatement. With perfect competition in the product market, the two first-order conditions for profit-maximisation are:

$$p = C_k'(x_k) + \theta_k \phi_k'(a_k) \text{ and } \tau_k = \phi_k'(a_k). \tag{5}$$

The product price equals the firm's total marginal cost of output, and the carbon price equals the marginal abatement cost. Putting these together yields a combined condition:

$$p = C_k'(x_k) + \tau_k \theta_k \tag{6}$$

so the product price is equal to marginal cost plus per-unit carbon costs based on its baseline emissions intensity of output.[12] Due to cost separability, the extent of abatement does not affect the product-market outcome. The abatement incentive rises with the domestic carbon price, $da_k/d\tau_k = 1/\phi_k''(\cdot) > 0$ which, in turn, is independent of output.

Our main interest is the rate of internal carbon leakage for different kinds of unilateral policy by country i, denoted as λ_i. These reduce i's domestic emissions, $de_i/d\lambda_i < 0$, but may also induce a change in j's emissions. To obtain simple formulae, we focus on a "marginal" policy change (that is, a small deviation from the initial equilibrium), for which internal carbon leakage is given by $L_i = (-de_j/d\lambda_i)/(de_i/d\lambda_i)$. In the benchmark case

[11] The stability condition $G_k^{xx} G_k^{ee} - G_k^{xe} G_k^{ex} > 0$ is equivalent to $A_k < 1$. Firm k's cost function satisfies the following standard assumptions. Written in terms of output and emissions, it increases in output, $G_k^x(x_k, e_k) = C_k' + \theta_k \phi_k' > 0$, decreases in emissions, $G_k^e(x_k, e_k) = -\phi_k' < 0$, and is convex in both output and emissions, with $G_k^{xx}(x_k, e_k) = C_k'' + \theta_k^2 \phi_k'' > 0$ and $G_k^{ee}(x_k, e_k) = \phi_k'' > 0$.

[12] To guarantee an interior solution, assume $p(0) > \max_k \{C_k'(0) + \tau_k \theta_k\}$.

without abatement, $L_i = (\theta_j/\theta_i)(-dx_j/d\lambda_i)/(dx_i/d\lambda_i)$, where the first term is countries' "relative dirtiness" and the second term is output leakage $L_i^{\mathrm{O}} \equiv (-dx_j/dx_i)$.

Some equilibrium definitions will prove useful to cast our formulae in familiar terms. First, let $\varepsilon^D \equiv -p(\cdot)/Xp'(\cdot) > 0$ be the price elasticity of demand. Second, let $\sigma_k \equiv x_k/X \in (0,1)$ be the market share of country k's firm (so $\sigma_i + \sigma_j = 1$). Third, let $\widehat{C}_k'(x_k) \equiv [C_k'(x_k) + \tau_k\theta_k]$ be k's total marginal cost of output and define $\eta_k^S \equiv x_k\widehat{C}_k''(x_k)/\widehat{C}_k'(x_k) > 0$ as its elasticity, also noting that $\widehat{C}_k''(x_k) \equiv C_k''(x_k)$. By k's first-order condition, $x_k'(p) = 1/C_k''(x_k) > 0$, i.e., its supply curve is upward-sloping. So $\varepsilon_k^S \equiv px_k'(p)/x_k(p) > 0$ is k's price elasticity of supply and, at the firm's optimum, $\eta_k^S = 1/\varepsilon_k^S$. These expressions are all evaluated at the output levels of the initial equilibrium.

3.2 Supply-side unilateral policies

We begin with two "supply-side" policies that unilaterally raise the carbon price for emissions-intensive production or directly reduce production, e.g., via a coal phase-out.

For concreteness, we can think of the demand curve $p(X)$ as that of consumers in country i who are served partly by domestic production and partly by imports from j. This is a natural interpretation for an integrated electricity market or where i's consumers can choose to fly via an airport in i or j. Internal leakage then captures the extent to which i's consumers are increasingly served by j's producers.

Our first overlapping policy λ_i imposes an additional carbon price only in country i. Formally, i's firm now faces a carbon price $\tau_i = \tau_i(\tau, \lambda_i)$, where $\frac{d}{d\tau}\tau_i(\tau,\lambda_i), \frac{d}{d\lambda_i}\tau_i(\tau,\lambda_i) > 0$. A leading example is a unilateral carbon price floor that "tops up" the system-wide carbon price, $\tau_i = \tau + \lambda_i$, like Great Britain's Carbon Price Support for power generation that ran alongside the EU ETS (and continues in the UK ETS). Another possibility is a policy that lifts i's carbon price towards a higher target level $\widehat{\tau}_i$, say with $\tau_i = \tau + \lambda_i(\widehat{\tau}_i - \tau)$ where $\lambda_i \in [0,1]$. Firm j continues to be subject to the system-wide carbon price, $\tau_j = \tau$.

This policy leads to an asymmetric cost shock, inducing i's firm to cut output and emissions, $dx_i/d\lambda_i < 0$ and $de_i/d\lambda_i < 0$, but raising the "competitiveness" of its rival in j. Since j's carbon price remains unchanged, its abatement decision also stays unchanged so $de_j/d\lambda_i = \theta_j(dx_j/d\lambda_i)$, and any change in its emissions is driven solely by output. Hence the policy's rate of internal leakage will be signed by j's output response.

Our second policy has country i institute a unilateral reduction in carbon-intensive production. A topical example is the phase-out of coal-fired power generation, which a number of European countries have individually committed to—alongside these plants being covered by the EU ETS. Formally, we suppose that i's policy λ_i directly imposes a (marginal) reduction in i's output, $dx_i/d\lambda_i < 0$. In contrast to the previous policy, the carbon price faced by i's firm remains unchanged, so $\tau_k = \tau$ for $k = i, j$, and so i's

abatement decision here also is unchanged.

Proposition 1 *A supply-side unilateral policy by country i has internal carbon leakage to country j of:*

$$L_i = \frac{\theta_j}{\theta_i}\left[\frac{\sigma_j}{(\sigma_j + \varepsilon^D/\varepsilon_j^S)}\right]\frac{1}{(1+\gamma Z_i)} > 0,$$

where $\gamma \in \{0,1\}$ equals zero (one) for a unilateral reduction in carbon-intensive production (unilateral carbon price), and $Z_i \equiv \frac{A_i}{(1-A_i)}\left(1 + \frac{(1-\sigma_j)\varepsilon_i^S/\varepsilon_j^S}{(\sigma_j+\varepsilon^D/\varepsilon_j^S)}\right) \geq 0$ is an abatement effect.

Proposition 1 provides a simple formula to quantify internal carbon leakage. For both supply-side policies, carbon leakage is always positive as the underlying output leakage is positive: i's firm loses market share to j's either because it incurs an asymmetric cost shock or has its production directly reduced. While output leakage is always less than 100%—as i's policy raises the product market price—carbon leakage can exceed 100% if j's firm is sufficiently dirtier, that is, θ_j/θ_i is sufficiently large.

To understand the result, consider the unilateral cut in carbon-intensive production ($\gamma = 0$). The comparative statics are intuitive: output leakage $L_i^{\text{O}} = \sigma_j/(\sigma_j + \varepsilon^D/\varepsilon_j^S)$ is more pronounced where: (i) j's market share is larger (higher σ_j), (ii) demand is relatively inelastic (lower ε^D), and (iii) j's firm is more supply-responsive, e.g., because of significant spare capacity (higher ε_j^S). In short, j's firm more aggressively "fills the gap" in market supply due to the policy when it is larger and more responsive. Output leakage then maps into carbon leakage by way of the relative emissions intensity θ_j/θ_i.

For a unilateral carbon price ($\gamma = 1$), internal carbon leakage is mitigated by the induced abatement A_i. Abatement breaks the direct link between output and emissions: for a given output contraction by i—and resulting competitive gain by j—domestic emissions fall by more. With near-costless additional abatement, carbon leakage tends to zero, $L_i \to 0$ as $Z_i \to \infty$ ($A_i \to 1$ as $\phi_i''(\cdot) \to 0$).

From a policy perspective, Proposition 1 formalises the rationale for a regional coalition within the EU introducing a carbon price floor (Newbery et al., 2019): this combines greater market share than single-country action and thereby contains internal leakage. Note also that the formula for L_i does not depend on the precise functional form of i's policy $\tau_i = \tau_i(\tau, \lambda_i)$; at the margin, this matters for the *absolute* output and emissions impacts but not for the *relative* effects—which is what our leakage rate captures.

To get a sense for magnitudes, suppose that the demand elasticity $\varepsilon^D = .5$ and that j has market share $\sigma_j = 20\%$ with a supply-responsiveness $\eta_j^S = .2 \Leftrightarrow \varepsilon_j^S = 5$. These parameter values might be plausible in the context of an electricity market in which i imports some of its consumption from generators located in j. If emissions intensities are identical, $\theta_i = \theta_j$, and there is no abatement ($A_i = 0$ or $\gamma = 0$), then $L_i = 67\%$,

driven solely by output leakage. If instead j's technology is less responsive with $\eta_j^S = 1 \Leftrightarrow \varepsilon_j^S = 1$ or market demand is more elastic with $\varepsilon^D = 2.5$, then leakage falls to $L_i = 28\%$. Conversely, if instead j's firm is twice as dirty—which approximates coal vs gas-fired power generation—then leakage doubles to $L_i = 133\%$. If i has significant abatement opportunity faced with a unilateral carbon price, as implied by $A_i = .25$, this yields $L_i = 60\%$, illustrating how abatement can help bring forth an overall emissions reduction—and how policies that trigger abatement (e.g., a unilateral carbon price) have lower leakage rates than policies that do not (e.g., a coal phase-out.)

3.3 Demand-side unilateral policies

We now turn to three "demand-side" policies that reduce the (residual) demand for emissions-intensive production: promoting zero-carbon renewables, an energy-efficiency program, and a carbon-consumption tax. We retain the interpretation that the demand curve reflects that of consumers in country i who are served partly by domestic production and partly by imports from j. Formally, we model a unilateral policy λ_i by country i and write $p(X;\lambda_i)$ where $\frac{\partial}{\partial\lambda_i}p(X;\lambda_i) < 0$ so the unilateral policy reduces demand for both i and j's firms. Both firms continue to face the common carbon price τ.

The policies fit into this setup as follows. First, for the renewables program, we write demand as $p(X;\lambda_i) = p(X + \lambda_i)$ where λ_i is the volume of zero-carbon production supported by the policy. Second, for the energy-efficiency program, we write direct demand as $D(p;\lambda_i) = (1-\lambda_i)D(p)$ so it reduces demand by a fraction λ_i (for a given p) and hence $p(X;\lambda_i) = D^{-1}\left(X/(1-\lambda_i)\right)$. Third, for the carbon-consumption tax, we write $p(X;\lambda_i) = [p(X) - \lambda_i\theta_i]$ where the tax λ_i is levied on consumption according to i's baseline emissions intensity θ_i. In all three cases, $\frac{\partial}{\partial\lambda_i}p(X;\lambda_i) < 0$ at an interior equilibrium.

Proposition 2 *A demand-side unilateral policy by country i of (i) a renewables support program that brings in additional zero-carbon production, or (ii) an energy-efficiency program that reduces demand for carbon-intensive production, or (iii) a carbon-consumption tax has internal carbon leakage to country j of:*

$$L_i = -\frac{\theta_j}{\theta_i}\frac{\sigma_j}{(1-\sigma_j)}\frac{\varepsilon_j^S}{\varepsilon_i^S} < 0.$$

Internal carbon leakage is always negative: j's firm is now directly affected by the policy and responds by also cutting output and emissions. This means that the "global" emissions reduction here is more pronounced than the "local" reduction. Akin to Proposition 1, leakage is more strongly negative where j's firm is dirtier, more supply-responsive and has greater market share. In addition, it is more pronounced if i's own supply-responsiveness

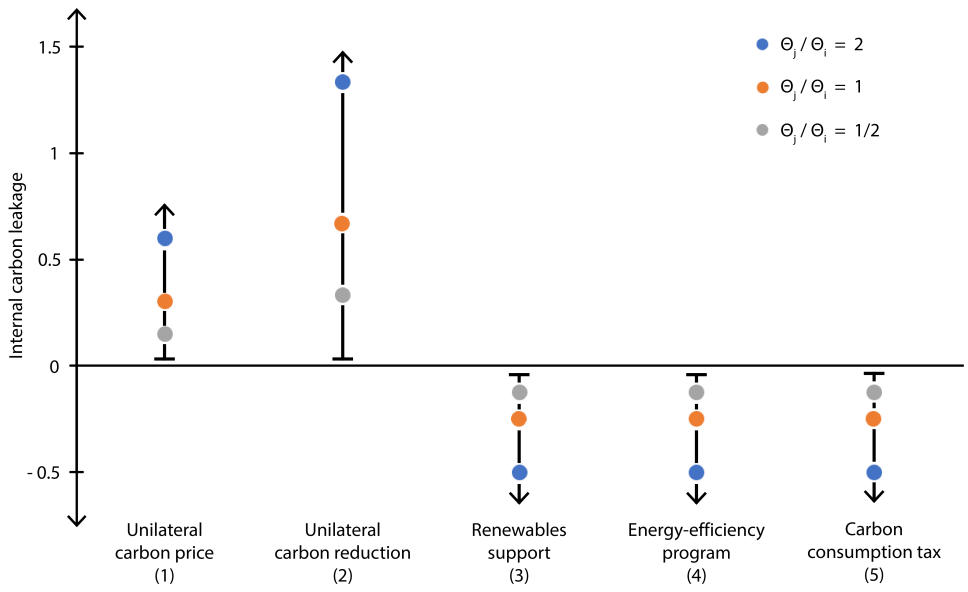
is weaker; then i's output contraction is smaller relative to j's. As the carbon price remains fixed, unilateral action here brings no extra abatement ($da_k/d\lambda_i = 0$ for $k = i, j$).

Proposition 2's internal leakage rate does not depend on any demand characteristics, including the precise form of $p(X; \lambda_i)$ and the demand elasticity ε^D. To first order, for a marginal policy, the reduction in i's production—and hence also of i's emissions—is proportional to $\frac{\partial}{\partial \lambda_i} p(X; \lambda_i)$. To first order, this is also true for the changes in j's production and emissions. So the relative magnitude of emissions changes, as captured by the leakage rate, does not depend on $\frac{\partial}{\partial \lambda_i} p(X; \lambda_i)$—and so all three demand-side unilateral policies have identical leakage properties.

To illustrate magnitudes, again using $\sigma_j = 20\%$, $\theta_i = \theta_j$, and $\varepsilon_i^S = \varepsilon_j^S$, yields internal carbon leakage of $L_i = -25\%$. If, instead, j's firms are twice as dirty or twice as supply-responsive than i's, leakage doubles in absolute terms to $L_i = -50\%$. With both $\theta_j/\theta_i = 2$ and $\varepsilon_j^S/\varepsilon_i^S = 2$, internal leakage becomes $L_i = -100\%$, and so the "global" reduction in emissions demand is now twice the size of the "local" reduction.

Figure 1 summarises the key differences in internal carbon leakage across different overlapping policies. Appendix A.2 shows that these insights from Propositions 1 and 2 are robust in the case with general cost functions $G_k(x_k, a_k)$. Appendix A.3 shows that they also apply for larger, non-marginal changes in policy.

Figure 1: Internal carbon leakage for different types of unilateral climate policies



Notes: Using parameter values $\sigma_j = 20\%$, $\varepsilon^D = .5$, $\varepsilon_i^S = \varepsilon_j^S = 5$, $A_i = .25$ in Propositions 1 (for supply-side policies (1) ($\gamma = 1$) and (2) ($\gamma = 0$)) and Proposition 2 (for demand-side policies (3)-(5)), with three different values of the relative emissions intensity θ_j/θ_i.

4 A model of the waterbed effect

We now turn to the second building block of our conceptual framework: the waterbed effect $W = 1 - \Delta e^*/\Delta e$, for which the carbon price path $\boldsymbol{\tau}$ is now derived endogenously. Again we represent the overlapping policy by λ_i and focus on a "marginal" policy change so the waterbed effect is $W = 1 - (de^*/d\lambda_i)/(de/d\lambda_i)$.[13]

Consider a stylised two-period model of an intertemporal allowance market. By design, the allowance market is geographically blind. We assume inverse aggregate demand functions for allowances $\rho_t(e_t, \lambda_i)$ where e_t are aggregate emissions in period $t = 1, 2$ and $\partial\rho_t/\partial e_t < 0$ and $\partial\rho_t/\partial\lambda_i \leq 0$. We assume that any overlapping policy is announced at the beginning of period 1, regardless of the timing of its impacts. Hence, policy impacts are perfectly anticipated by all market participants. We restrict attention to markets with perfect intertemporal arbitrage in which any borrowing and banking constraints do not bind, i.e., $\tau_2 = (1 + r)\tau_1$. For a carbon tax or a binding price corridor (i.e., a combination of a price floor and a price ceiling), the interest rate r instead reflects an exogenous increase in the carbon price.

We first analyse how an anticipated shift in allowance demand affects total emissions and the equilibrium price of allowances when the carbon-pricing scheme features a (weakly) increasing allowance supply function. Then we consider a design where the cumulative cap is adjusted based on banked allowances as has been the case in the EU ETS since the 2018 reform (Perino, 2018). Due to particular design features that cannot be fully captured by a two-period model, we conclude this section by taking a closer look at the EU ETS's Market Stability Reserve.

4.1 Flexibility mechanisms based on past allowances prices

Most real-world carbon-pricing designs—such as the California-Québec scheme and RGGI (Burtraw et al., 2020), the pre-2018 EU ETS (Perino and Willner, 2016), the new UK-ETS and all carbon taxes—feature flexibility mechanisms based on past allowance prices.

Allowance supply is given by a fixed number of allowances s_1 issued in period 1 and a flexible number of allowances, $s_2(\tau_1)$, issued in period 2. For the remainder of this subsection we restrict attention to weakly upward-sloping allowance supply curves ($\partial s_2/\partial\tau_1 \geq 0$).[14] A cap-and-trade scheme with a fixed cap or any vertical section of an allowance supply curve are represented by $\partial s_2/\partial\tau_1 = 0$. A carbon tax or a horizontal

[13] Qualitatively all findings of this section extend to non-marginal changes in policy (see Appendix B.2.)

[14] If overlapping policies are anticipated at the beginning of period 1, firms are forward looking and $\tau_2 = (1 + r)\tau_1$, then the timing of price-based supply adjustments is irrelevant, i.e., all $s_1(\tau_1) + s_2(\tau_2)$ that yield the same $s(\tau_1)$ are equivalent. The assumption that s_1 is fixed is therefore without loss of generality.

section of an allowance supply curve, such as a price floor, are represented by $s_2(\tau_1)$ being perfectly price elastic at a particular $\bar{\tau}_1$.

The three equilibrium conditions of this carbon-market design are:

$$\begin{aligned} \rho_1(e_1, \lambda_i) - \tau_1 &= 0 && (7) \\ \rho_2(e_2, \lambda_i) - (1+r)\tau_1 &= 0 && (8) \\ e_1 + e_2 - s_1 - s_2(\tau_1) &= 0, && (9) \end{aligned}$$

where (7) and (8) balance marginal costs of abatement with the carbon price for periods 1 and 2, respectively, while (9) is the market-clearing condition for the allowance market.

The equilibrium conditions yield the impact of the unilateral policy on the system-wide equilibrium carbon price (see Appendix B):

$$\frac{\partial \tau_1}{\partial \lambda_i} = \frac{\frac{de}{d\lambda_i}}{\frac{\partial s_2}{\partial \tau_1} - \frac{\partial e}{\partial \tau_1}} > 0 \tag{10}$$

where $\partial e/\partial \tau_1 < 0$ is the slope of the total allowance demand curve. The change in the equilibrium allowance price is key to identifying the waterbed effect. Temporal and geographical distributions of the change in allowance demand are irrelevant. Adjustments in total equilibrium emissions e^* are also independent of how the unilateral policy is spread over time and space:

$$\frac{de^*}{d\lambda_i} = \frac{\partial s_2}{\partial \tau_1}\frac{\partial \tau_1}{\partial \lambda_i} = \frac{de}{d\lambda_i}\frac{\frac{\partial s_2}{\partial \tau_1}}{\frac{\partial s_2}{\partial \tau_1} - \frac{\partial e}{\partial \tau_1}} = \frac{de}{d\lambda_i}\frac{\kappa^S}{\kappa^S + \kappa^D} \tag{11}$$

where $\kappa^D > 0$ and $\kappa^S \geq 0$ are the long-run elasticities of allowance demand and supply.

Proposition 3 *The waterbed effect for a marginal policy overlapping a carbon-pricing scheme with a (weakly) increasing allowance supply and strictly decreasing allowance demand is:*

$$W = \frac{-\frac{\partial e}{\partial \tau_1}}{\frac{\partial s_2}{\partial \tau_1} - \frac{\partial e}{\partial \tau_1}} = \frac{\kappa^D}{\kappa^S + \kappa^D} \in [0, 1], \tag{12}$$

which is independent of the specifics of the overlapping policy (λ_i) and leakage rates (L_{it}).

Proposition 3 shows that the waterbed effect depends only on elasticities of total allowance demand and supply—and is independent of the type of overlapping policy, its temporal and geographical impacts, and its internal carbon leakage.

We thus uncover a natural connection between the waterbed effect and classic principles from the literature on tax incidence (Jenkin, 1872; Weyl and Fabinger, 2013). In

particular, note that Proposition 3 corresponds to the cost pass-through rate from the tax-incidence literature. Since the allowance supply is assumed to be (weakly) monotonically increasing, it mimics a supply curve. The drop of producer prices in response to a tax-induced shift in inverse demand in the tax-incidence literature exactly mimics the impact of an overlapping policy on the carbon price in a carbon market with a weakly upward-sloping allowance supply curve.

Equation (12) has at opposite ends a zero waterbed for a carbon tax ($\partial s_2/\partial \tau_1 \to \infty$) and a 100% waterbed effect under a plain cap-and-trade system ($\partial s_2/\partial \tau_1 = \kappa^S = 0$). For marginal changes, i.e., policies inducing relatively small shifts in allowance demand, this conclusion applies also to step-wise allowance supply functions featured in the California-Québec scheme, RGGI and the new UK ETS. If the initial equilibrium is in a vertical (horizontal) section of the supply curve, the waterbed effect is 100% (zero).

The expected waterbed effect of marginal changes is in the intermediate range if at the time of passing legislation for an overlapping policy future market outcomes are still uncertain (Borenstein et al., 2019). If the probability that the equilibrium is in any of the horizontal sections of the allowance supply curve is π, then $E(W) = 1 - \pi$. Ex-post the waterbed effect is either zero or 100%.

Once larger interventions are considered—such that allowance demand moves across one or several kinks in the step-wise supply schedule—none of the extreme cases appropriately capture the impact on supply. The average waterbed effect of a large-scale policy can be computed by integrating over the marginal effects.

4.2 Flexibility mechanisms based on past allowance banking

With the 2018 reform of the EU ETS, namely the introduction of cancellations within the Market Stability Reserve, an entirely new form of flexibility mechanism entered the scene. Here we present a stylised two-period version of such a mechanism. It adjusts the cumulative cap based on the number of allowances banked for future use in earlier periods, $s_2(b)$, where $b = s_1 - e_1$ is banking at the end of period 1 and $\partial s_2/\partial b \in [-1, 0]$.[15] A plain cap-and-trade scheme is again nested as a special case ($\partial s_2/\partial b = 0$).

Analogous to the price-based flexibility mechanism, the equilibrium conditions are:

$$\rho_1(e_1, \lambda_i) - \tau_1 = 0 \tag{13}$$

$$\rho_2(e_2, \lambda_i) - (1 + r)\tau_1 = 0 \tag{14}$$

$$e_1 + e_2 - s_1 - s_2(s_1 - e_1) = 0. \tag{15}$$

[15] Restricting $\partial s_2/\partial b > -1$ is somewhat arbitrary as one could imagine schemes with more responsive rules. However, imposing this lower bound simplifies the analysis and includes the entire range of values relevant for the EU ETS. For details see Lemma 2 below.

Cramer's rule and the implicit function theorem (see Appendix B) yield the response of short-run equilibrium emissions to the overall change in allowance demand:

$$\frac{\partial e_1^*}{\partial \lambda_i} = -\frac{\frac{de}{d\lambda_i}}{1+\frac{\partial s_2}{\partial b}\frac{\frac{\partial e_1}{\partial \tau_1}}{\frac{\partial e}{\partial \tau_1}}} \cdot \left[\frac{\frac{\partial e_1}{\partial \tau_1}}{\frac{\partial e}{\partial \tau_1}} - \frac{\frac{de_1}{d\lambda_i}}{\frac{de}{d\lambda_i}}\right]. \tag{16}$$

Given the impact of the overlapping policy on overall allowance demand $de/d\lambda_i$, the direction of the policy's impact on equilibrium emissions in period 1 (e_1^*) depends on the relative size of the two terms in brackets in equation (16). Both have an intuitive economic interpretation. The first term is the ratio of the slopes of first-period and cumulative allowance demand. It captures how much of the total change in emissions induced by the price response materialises in period 1. The second term is the percentage of the shift in the cumulative demand curve occurring in the first period. Shifting the allowance demand curve to the left in period 1 ($de_1/d\lambda_i$), ceteris paribus, reduces first-period equilibrium emissions. The price drop triggered by the decrease in overall scarcity induces a movement along the demand curve and, ceteris paribus, increases first-period equilibrium emissions. The direct demand-shifting (Perino, 2018) and the indirect price-mediated effect (Rosendahl, 2019) are hence antagonistic (see Figure 2).

Whether an overlapping policy increases or decreases first-period emissions in equilibrium depends on the timing of its impacts. If the policy is front-loaded in that most of the shift in allowance demand occurs early on, then first-period emissions decrease. By contrast, a policy that mainly affects allowance demand in the future raises first-period emissions. In terms of short and long-run elasticities, the term in brackets is more likely to be positive the smaller the difference between the price elasticity of short-run (κ_1^D) and cumulative (κ^D) allowance demand.

This dependence on timing directly translates to the change in total equilibrium emissions e^* via adjustment of the cumulative cap, where the banking of allowances $b = s_1 - e_1^*$ mirrors the change in first-period emissions as the first-period cap s_1 is fixed:

$$\frac{de^*}{d\lambda_i} = \frac{ds_2}{d\lambda_i} = \frac{\partial s_2}{\partial b}\frac{\partial b}{\partial e_1^*}\frac{\partial e_1^*}{\partial \lambda_i} = \frac{\frac{de}{d\lambda_i}\frac{\partial s_2}{\partial b}}{1+\frac{\partial s_2}{\partial b}\frac{\frac{\partial e_1}{\partial \tau_1}}{\frac{\partial e}{\partial \tau_1}}} \cdot \left[\frac{\frac{\partial e_1}{\partial \tau_1}}{\frac{\partial e}{\partial \tau_1}} - \frac{\frac{de_1}{d\lambda_i}}{\frac{de}{d\lambda_i}}\right]. \tag{17}$$

Policies that mainly reduce allowance demand early on reduce the cumulative cap as firms respond to the shift in the first-period demand curve by emitting less and banking more. The increase in the bank induces additional cancellations. Policies reducing allowance demand in the distant future tend to increase the cumulative cap. As firms anticipate the drop in demand, they have less incentives to bank allowances and therefore emit more in

the first period. The reduction in the bank results in fewer cancellations.

Proposition 4 *The waterbed effect for a marginal, anticipated policy overlapping a carbon-pricing scheme with a flexibility mechanism based on past allowance banking is:*

$$W = \frac{1+\frac{\partial s_2}{\partial b}\frac{\frac{de_1}{d\lambda_i}}{\frac{de}{d\lambda_i}}}{1+\frac{\partial s_2}{\partial b}\frac{\frac{\partial e_1}{\partial \tau_1}}{\frac{\partial e}{\partial \tau_1}}} \tag{18}$$

where the numerator captures the direct impact of the overlapping policy and the denominator the indirect effect mediated through the price response. Defining $\beta = (de_1/d\lambda_i)/(de/d\lambda_i)$*, the following holds:*

(i) A unilateral policy effective only in period 1 ($\beta = 1$*) has a waterbed effect weakly smaller than 1 (because* $\partial s_2/\partial b \in [-1, 0]$ *and* $(\partial e_1/\partial \tau_1)/(\partial e/\partial \tau_1) \in (0, 1)$*)*

$$W = \frac{1+\frac{\partial s_2}{\partial b}}{1+\frac{\partial s_2}{\partial b}\frac{\frac{\partial e_1}{\partial \tau_1}}{\frac{\partial e}{\partial \tau_1}}} \in (0, 1]. \tag{19}$$

(ii) A unilateral policy effective only in period 2 ($\beta = 0$*) has a waterbed effect weakly larger than 1 (because* $\partial s_2/\partial b \in [-1, 0]$ *and* $(\partial e_1/\partial \tau_1)/(\partial e/\partial \tau_1) \in (0, 1)$*)*

$$W = \frac{1}{1+\frac{\partial s_2}{\partial b}\frac{\frac{\partial e_1}{\partial \tau_1}}{\frac{\partial e}{\partial \tau_1}}} \geq 1. \tag{20}$$

(iii) For any given change in total allowance demand $(de/d\lambda_i) < 0$*, there exists a threshold* $\underline{\beta} = (\partial e_1/\partial \tau_1)/(\partial e/\partial \tau_1) \in (0, 1)$ *for which* $W = 1$*. For all* $\beta < \underline{\beta}$*,* $W > 1$ *and vice versa. The larger the difference between short-run and cumulative responsiveness (price elasticity) of allowance demand, the lower* $\underline{\beta}$*.*

(iv) An overlapping policy features a negative waterbed effect, $W < 0$*, if it reduces aggregate allowance demand in period 1 and across both periods but sufficiently increases it in period 2 (*$\beta > 1, de/d\lambda_i < 0, de_1/d\lambda_i < 0, de_2/d\lambda_i > 0$*) according to:*

$$\beta > \bar{\beta} = -\frac{1}{\frac{\partial s_2}{\partial b}} \geq 1. \tag{21}$$

In sum, there are three regimes for the waterbed effect:

$$W \quad = \quad \begin{cases} < 0 & \textit{if} \quad \beta > \bar{\beta} \geq 1 \\ \in [0,1] & \textit{if} \quad \underline{\beta} \leq \beta \leq \bar{\beta} \\ > 1 & \textit{if} \quad \beta < \underline{\beta} \in (0,1) \end{cases} .$$

The special cases presented in parts (i) and (ii) highlight the direct and the price-mediated indirect effect on cumulative emissions. For policies affecting aggregate demand early on (see Equation (19)), the price effect is always of second order and the waterbed is punctured. However, policies affecting aggregate demand only in the far future (Equation (20)) such as an anticipated coal phase-out have no immediate emissions-demand impact—and hence the price-driven effect dominates. Ceteris paribus, such policies increase the supply of allowances. Anticipation of a future reduction in relative scarcity leads to lower carbon-price expectations for period 2 and reduces the incentives to bank allowances. The resulting drop in the bank reduces the number of allowances cancelled by the flexibility mechanism. In other words, such policies refill the waterbed—the "Rosendahl effect".[16]

In sum, for a given quantity-based flexibility mechanism ($\partial s_2/\partial b$) and given market characteristics ($\rho_t(e_t, \lambda_i)$), an overlapping policy with a given impact on total allowance demand ($de/d\lambda_i$), can increase total emissions ($W > 1$), leave them unaffected ($W = 1$), decrease them ($W < 1$), and even decrease them by more than the initial shift in aggregate demand ($W < 0$)—all driven exclusively by the timing of its impact on aggregate allowance demand (β). Hence the waterbed effect is also a function of changes in internal carbon leakage over time. Both points are in stark contrast to price-based flexibility mechanisms where policy timing and internal leakage were irrelevant (Proposition 3). An example for a policy featuring a negative waterbed (case (iv) in Proposition 4) is an amendment of a previously enacted coal phase-out plan that shuts down old inefficient plants earlier but grants new, highly-efficient plants a longer grace period.

Next we show that there is an economically intuitive link between Propositions 3 and 4. While the cap adjustment in the flexibility mechanisms based on past banking does not explicitly refer to prices, one can construct an equilibrium expansion path for overlapping policies of different stringency but identical timing of impacts that mimics an effective supply curve for allowances.

Corollary 1 *Propositions 3 and 4 are equivalent when considering the equilibrium expansion path as an effective allowance supply function that is specific to the overlapping*

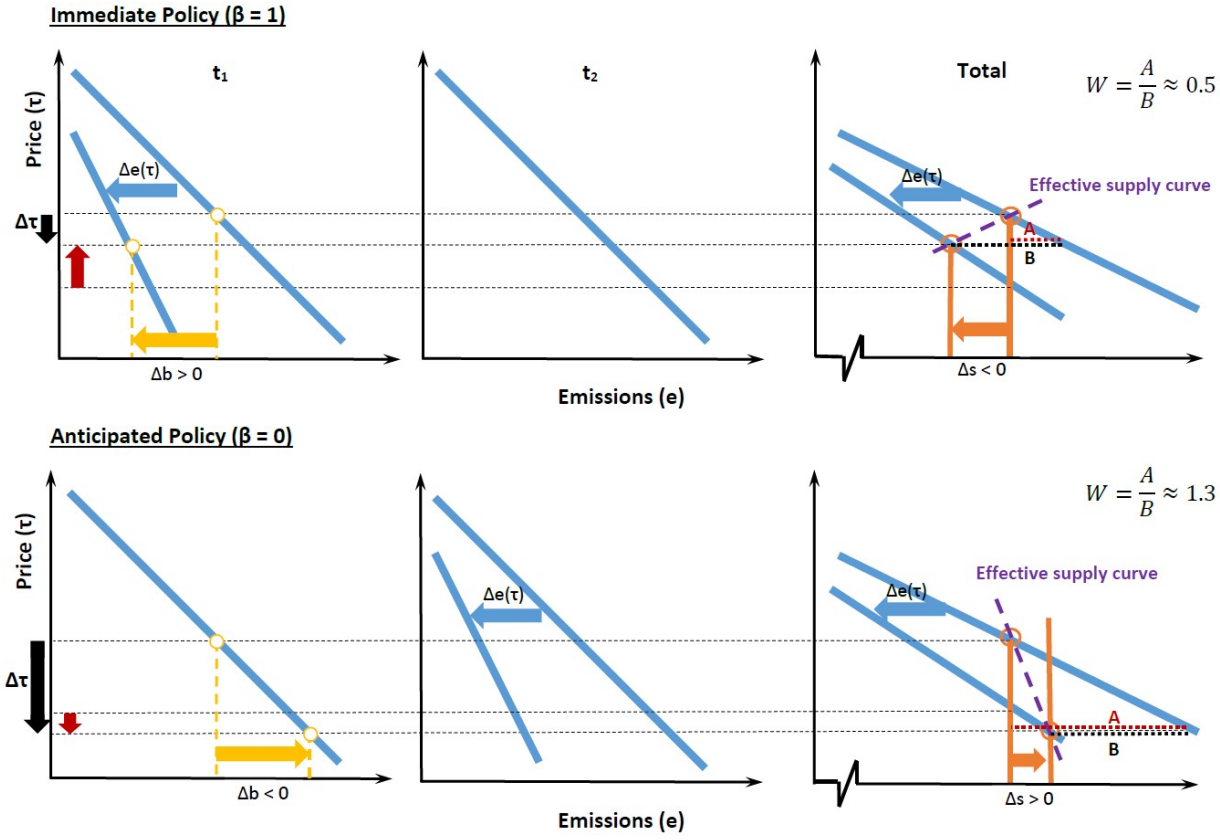
[16]This effect was first described by Rosendahl (2019) and confirmed by Bruninx et al. (2019); Gerlagh et al. (2021); Pahle et al. (2019).

policy under consideration:

$$\left.\frac{ds_2}{d\tau_1}\right|_{equilibrium} = \frac{\frac{ds_2}{d\lambda_i}}{\frac{\partial \tau_1}{\partial \lambda_i}} = \frac{\frac{\partial s_2}{\partial b}\frac{\partial e}{\partial \tau_1}}{1+\frac{\partial s_2}{\partial b}\beta} \cdot \left[\beta - \frac{\frac{\partial e_1}{\partial \tau_1}}{\frac{\partial e}{\partial \tau_1}}\right]. \tag{22}$$

The equilibrium expansion path (22) is highly instrument-specific so in contrast to the allowance supply function specified in Section 4.1 it is not a common and defining feature of the carbon-pricing scheme but specific to the overlapping policy under consideration. (See Appendix B.1 for a proof: Plugging Equation (22) into Equation (12) yields the same W as using (18).) Effective supply curves are strictly downward-sloping whenever the waterbed effect of the policy is either above 100% or negative.

Figure 2: The timing of overlapping policies, waterbed effects and effective supply curves



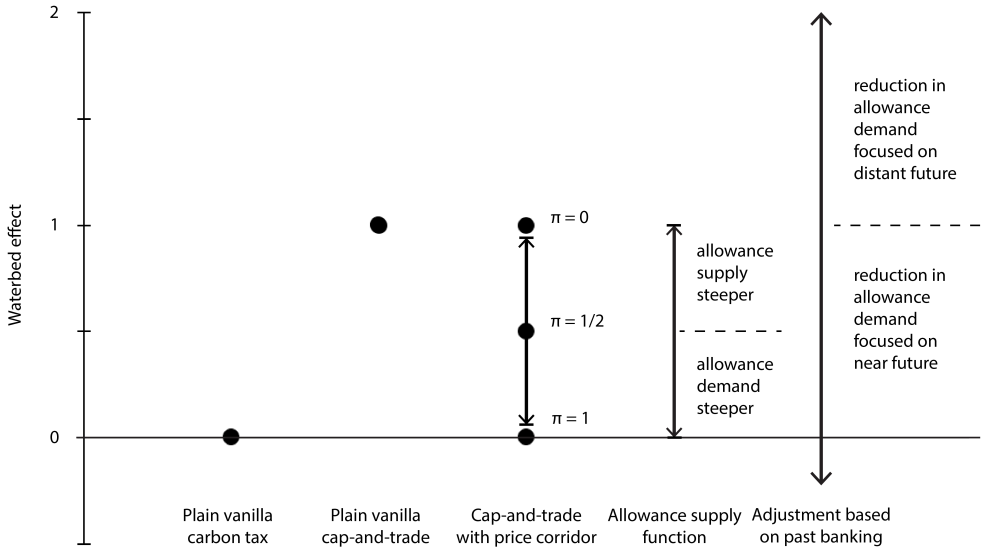
Notes: An identical shift in total allowance demand ($\Delta e(\tau) = de/d\lambda_i$, blue) induced by an overlapping policy occurs either at $t = 1$ (upper panel) or at $t = 2$ (lower panel). Direction of change in equilibrium emissions at $t = 1$ (yellow) and hence supply response (orange) depends on timing of policy. Effective supply curve (dashed purple) represents response to a continuum of demand shifts of the same timing. Total demand and effective supply curve jointly determine the waterbed effect W in line with the tax incidence literature. Price response to policy (black arrow) is reduced (upper panel) or increased (lower panel) relative to a fixed cap (red arrow).

Figure 2 illustrates Proposition 4. The direct impact of the overlapping policy that shifts the demand curve for allowances. The indirect effect is represented by movements

along a given allowance demand curve and mediated by changes in prices. These two effects jointly drive first-period equilibrium emissions, and their interaction directly determines the direction of the supply adjustment. The decomposition illustrated in Figure 2 is new to the literature, as is the connection between the waterbed effect under flexibility mechanisms based on past prices and past banking identified by Corollary 1.

Summarising our analysis, Figure 3 presents the possible ranges of the waterbed effect for typical carbon-market designs captured by Propositions 3 and 4.[17] While the extent of the waterbed is unique for a carbon tax ($W = 0$) and cap-and-trade ($W = 1$), all hybrid policies and flexibility mechanisms yield ranges that depend on the specifics of the carbon-market design, the probability that any price bounds are binding (π), the timing of the overlapping policy, or the long and short-run elasticity of emissions demand.

Figure 3: Waterbed effects for typical carbon-pricing policies



Notes: The expected (but not the ex-post) waterbed effect of a marginal overlapping policy in a cap-and-trade scheme with a price corridor depends on the probability that price bounds are binding (π).

4.3 The reformed EU ETS

To further address subtleties originating from the timing of overlapping policies, we take a closer look at the reformed EU ETS in a multi-period context. We show how to compute $\partial s/\partial b$ when a policy impacts the bank in more than one period.

[17] Note that Propositions 3 and 4 apply to policies announced at the beginning of period 1. Due to the design of both types of flexibility mechanisms—both respond to market outcomes in period 1—an unanticipated policy implemented in period 2 has a waterbed effect of $W = 1$. It simply escapes the radar of the flexibility mechanism.

The EU ETS's flexibility mechanism, the Market Stability Reserve (MSR), works as follows.[18] If the bank, known as the "total number of allowances in circulation" (TNAC) in the legal language of the EU ETS, exceeds 833 million at the end of a given year (in 2017 or later), then the number of allowances auctioned in the 12 months following October of the following year is reduced by a certain percentage of the size of the bank (see Table 1). Allowances withheld are placed in the MSR and released in installments of 100 million/year once the bank has dropped below 400 million. We label $t_{B=833}$ the year in which the bank drops below the 833 million threshold and the MSR hence stops taking in allowances.

Year (if bank $>$ 833 million on Dec. 31^{st})	Intake rate (%)
2017	16
2018 - 2021	24
2021 - $t_{B=833}$	12

Table 1: Intake rates for the EU ETS Market Stability Reserve (MSR)

Starting in 2023, the maximum number of allowances held in the MSR is limited to the number auctioned in the previous year.[19] Allowances in excess of this upper bound are permanently cancelled. Given that the MSR is seeded with a large quantity of allowances and that the threshold for cancellations is decreasing along with the number of auctioned allowances, any additional allowance drawn into the MSR is eventually cancelled.[20]

Computing the waterbed effect for the EU ETS faces several challenges. First, the MSR's intake rate changes over time (Table 1). Second, the MSR is active over multiple periods so the cumulative effect of an early shift in allowance demand depends on its impact on the TNAC in all periods up to $t_{B=833}$. Third, $t_{B=833}$ is itself determined by market outcomes and hence by the overlapping policy. Fourth, the price-mediated Rosendahl effect of anticipated future changes in allowance demand depends on the same dynamics. These complexities mean that W can only be estimated by numerical simulation.[21]

Next, we derive the sensitivity of the cumulative cap to changes in the bank $\partial s/\partial b$ as an explicit function of time. Based on this we derive an instantaneous waterbed effect that captures the first two complexities: the MSR's time-varying intake rate and its multi-

[18]The rules are laid down in European Parliament and Council (2018) and discussed by Perino (2018).

[19]The target share of auctioning in Phase 4 is 57% (European Parliament and Council, 2018) with the remaining allowances being freely allocated.

[20]At the end of 2020 the MSR contained 1.9 billion allowances with a further 379 million being added before September 2022 (European Commission, 2021). The cancellation threshold in 2023 will be below 1 billion.

[21]See Bruninx et al. (2019); Gerlagh et al. (2021); Pahle et al. (2019) for simulation results and Rosendahl (2019); Perino (2019) for informal discussions.

period nature. An instantaneous change in the number of banked allowances triggers a sequence of transfers to the MSR.[22] This implies (see Appendix B.3):

Lemma 2 *Adding one allowance to the bank in year t and with the bank dropping below 833 million allowances in year $t_{B=833}$, the effective sensitivity of the cumulative cap in the EU ETS is given by:*

$$\begin{aligned}\frac{\partial}{\partial b}s\left(t,t_{B=833}\right) = &-(1-.16)^{\max[0,\min[2018,t_{B=833}]-\max[2017,t]]} \\ &\times (1-.24)^{\max[0,\min[2022,t_{B=833}]-\max[2018,t]]} \\ &\times (1-.12)^{\max[0,\max[2022,t_{B=833}]-\max[2022,t]]}.\end{aligned} \tag{23}$$

The instantaneous waterbed effect $\hat{W}\left(t_a,t,t_{B=833}\right)$ in response to a one-off reduction in aggregate allowance demand in year t that is announced in year $t_a \leq t$ is thus:

$$\hat{W}\left(t_a,t,t_{B=833}\right) = \frac{1+\frac{\partial}{\partial b}s\left(t,t_{B=833}\right)}{1+\frac{\partial}{\partial b}s\left(t_a,t_{B=833}\right)\frac{\frac{\partial e_{t_a}}{\partial \tau_{t_a}}}{\frac{\partial e}{\partial \tau_{t_a}}}}. \tag{24}$$

Abstracting from changes in the carbon price, this simplifies to:

$$\hat{W}\left(t,t_{B=833}\right)|_{\tau fixed} = 1+\frac{\partial}{\partial b}s\left(t,t_{B=833}\right). \tag{25}$$

Lemma 2 highlights the triple importance of timing: the year an overlapping policy is announced, t_a, the year it shifts allowance demand, t, and the year the carbon-pricing scheme stops responding to past market outcomes, t_{833}, jointly determine the size of the instantaneous waterbed effect. Note that this still ignores the endogeneity of t_{833}.

5 Illustrations of unilateral overlapping policies

There are many real-world unilateral policies that overlap with wider carbon-pricing systems, leading to different degrees of waterbed effects and internal carbon leakage. We now illustrate how several such overlapping policies fit into our conceptual framework from Section 2. The equilibrium change in cumulative emissions is $\Delta e^* = [1-L_i][1-W]\Delta e_i$ (Lemma 1), and our main outcome of interest here is the effective emissions reduction rate $R_i \equiv [1-L_i][1-W]$. We use a combination of sources to quantify leakage and waterbed effects, which allows us to compute R_i for a range of policies. (A limitation is that our

[22] A share ν_t of the increase in the bank is transferred in the first year, the remainder $(1-\nu_t)$ adds to the bank in the following year and again induces a transfer at rate ν_{t+1}, i.e., $(1-\nu_t)\nu_{t+1}$, and so on.

sources do not provide time-varying estimates so we leave out the t subscript for L_i—and sometimes drop subscripts altogether for ease of exposition.[23])

Figure 4 is the visual summary of this section. It plots the contour lines of R in (L, W)-space along with various policy examples for which we have found estimates of L and W using existing literature. This is a novel way to graphically summarise the climate-effectiveness of a rich array of overlapping policies. Policies in the green regions are highly effective; policies in the orange regions have little effect, or worse, increase aggregate emissions. The evidence is consistent with the predictions from our theory of internal carbon leakage: a unilateral carbon price floor, aviation tax, and coal phase-out have positive leakage (Proposition 1) while renewables support has negative leakage (Proposition 2). The following subsections explain the various overlapping policies in more detail, first for those overlapping the EU ETS and then for those in North America.

5.1 Overlapping policies in the EU ETS

We first consider policies overlapping the reformed EU ETS for which the waterbed effect depends on the timing of the policies—a result of the flexibility mechanism based on past allowance banking (the MSR). As discussed in Section 4.3, the eventual impact of a marginal change in the allowance bank in year t on overall EU ETS emissions—and thus the "instantaneous waterbed effect" $\hat{W}_t$ for a fixed carbon price path in Equation (25) in Lemma 2—changes over time. Therefore, the effective emissions reduction rate for policies in the EU ETS changes over time, and we refer to it as $\hat{R}_{it} = (1 - L_i)(1 - \hat{W}_t)$.

As given by Equation (23), $\hat{W}_t$ depends on the year t in which the policy takes effect and the number of years until the bank drops below 833 million allowances, $t_{B=833}$. We use $t_{B=833} = 2030$ as a lower-end mid-range value[24] and contrast policies acting in years $t = 2020$, 2025 and 2030. In a sensitivity analysis, we also consider $t_{B=833} = 2048$, as estimated in Gerlagh et al. (2021) (see Appendix C). As time moves on, $\hat{W}_t$ increases from 0.21 to 0.53 to 1 and all European policies in Figure 4 move north, as indicated by

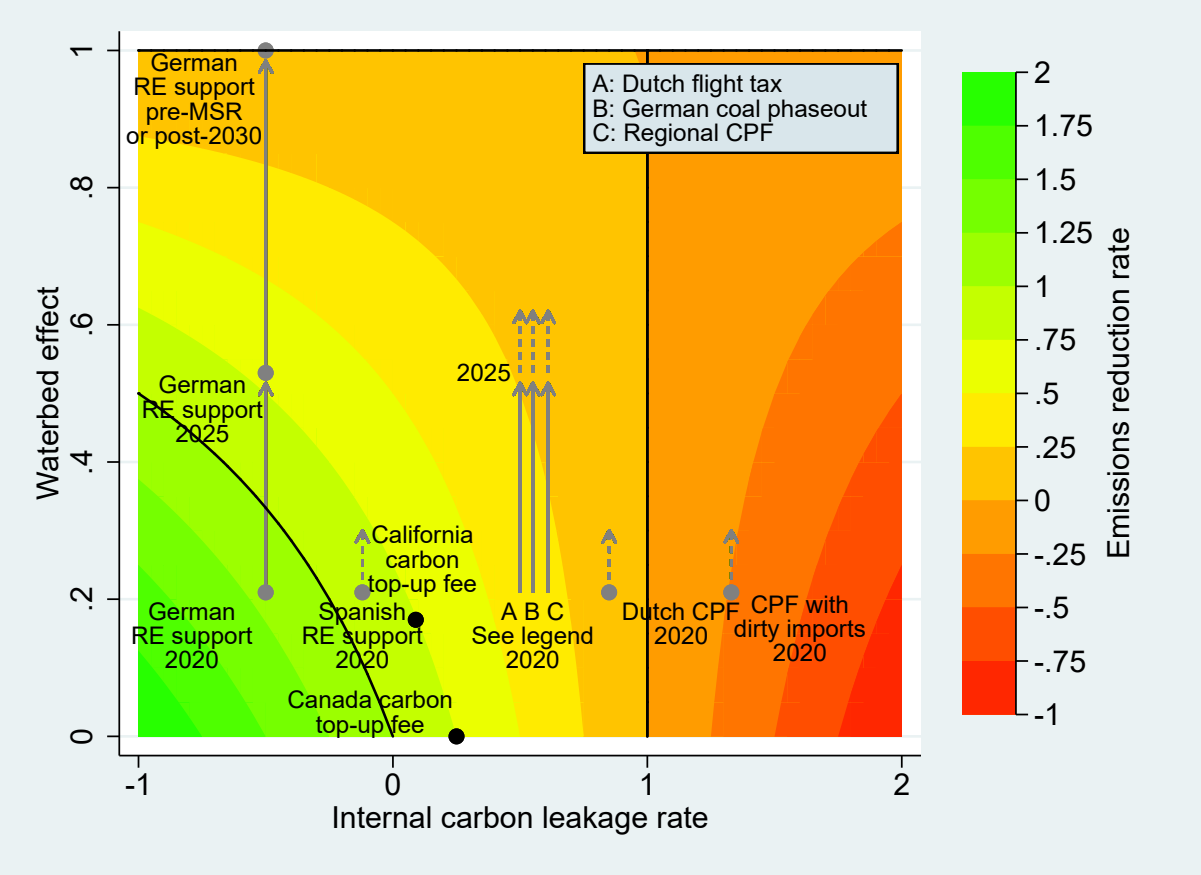
[23]Empirical estimates of internal carbon leakage do not always make clear whether the system-wide carbon is held fixed (as in our framework) or not. An alternative definition of "total" internal carbon leakage directly features any induced change:

$$L_i^{\mathrm{T}} = \frac{\frac{de_j}{d\lambda_i} + \frac{de_j}{d\tau}\frac{d\tau}{d\lambda_i}}{-\left(\frac{de_i}{d\lambda_i} + \frac{de_i}{d\tau}\frac{d\tau}{d\lambda_i}\right)}.$$

where typically $de_k/d\tau < 0$ for $k = i, j$. For a carbon tax, the two concepts are equivalent, $L_i^{\mathrm{T}} = L_i$, as then $d\tau/d\lambda_i = 0$ (zero waterbed). For cap-and-trade systems, $d\tau/d\lambda_i \neq 0$ so there can be a wedge between L_i^{T} and L_i. However, as long as $d\tau/d\lambda_i \simeq 0$ for i's policy, we expect that the two concepts give similar results, $L_i^{\mathrm{T}} \simeq L_i$, and rely on this in our empirical calibrations unless stated otherwise.

[24]This date is subject to substantial uncertainty, with estimates ranging from 2022 (Perino and Willner, 2017) to the second half of the 2030s (Quemin and Trotignon, 2021), and $t_{B=833} = 2030$ as a mid-range value (Vollebergh, 2018).

Figure 4: Unilateral policies facing internal carbon leakage and a waterbed effect



Notes: Figure shows the contour plot of the effective emissions reduction rate $R_{it} = (1 - L_{it})(1 - W)$ of various policies discussed in this section. Solid black lines indicate the contour lines where $R_{it} = 0$ (when $L = 1$ or $W = 1$) and $R_{it} = 1$ (bottom left). For EU ETS policies, we plot the instantaneous waterbed effect $\hat{W}_t$ for a fixed carbon-price path. Dashed grey arrows indicate that, in the EU ETS, a policy's $\hat{R}_{it}$ moves towards zero as t approaches $t_{B=833}$ and $\hat{W}_t \to 1$. We assume $t_{B=833} = 2030$. Solid grey arrows show specific shifts in time for the German renewable energy support schemes and for a proposed regional carbon price floor.

the dotted lines. The values for $\hat{W}_t$ can be calculated using Equation (23) evaluated at $t_{B=833} = 2030$ and $t = 2020, 2025, 2030$. The internal leakage rate L_i is policy specific and we discuss empirical estimates for various policies below. Note that Figure 4 shows a sequence of emissions reduction rates for policies operating in different years. The overall performance of a policy that is in effect for multiple years can be summarised by the emissions-reduction weighted average over the values of $\hat{R}_{it}$ along the grey lines for the relevant time period.

We finally note that the policies that we highlight below (e.g., a carbon price floor, a coal phase-out, an aviation tax, or renewables support) likely have negligible external carbon leakage to regions outside the EU ETS, justifying our focus on internal leakage.

"Supply-side" unilateral policies

Following the structure of Section 3.1, we now discuss "supply-side" unilateral policies such as national carbon price floors, aviation taxes, and low-carbon mandates.

Electricity

We first consider unilateral cost-raising policies such as a national carbon price floor (CPF) for electricity generation. For example, the Dutch government announced a national CPF for the electricity sector in 2018 and is awaiting a final vote in parliament as of the spring of 2021. It is slated to increase from EUR 12.30/tCO_2 in 2020 to EUR 31.90/tCO_2 in 2030. In 2013, Great Britain introduced a carbon fee for its power sector. Proposition 1 shows that such policies, if binding, suffer from intra-EU leakage as domestic generation gets replaced with imports. We expect high leakage for small countries (high σ_j) that are strongly interconnected to neighbours with flexible yet dirty supply (high $\varepsilon_j^S, \theta_j/\theta_i$).

Consistent with this, quantitative estimates find $L \simeq 0.85$ for the Dutch CPF, while a regional CPF including the Benelux, France and Germany faces $L = 0.61$ (Frontier Economics, 2018; Vollebergh, 2018).[25] Such CPFs in small interconnected countries are unlikely to reduce EU-wide emissions by much, with $\hat{R}_{2020} = 0.12$ ($\hat{W}_{2020} = 0.21, L = 0.85$) even under the punctured waterbed (see Figure 4).[26] As more countries join the CPF, $\hat{R}_{2020}$ rises to 0.31 ($\hat{W}_{2020} = 0.21, L = 0.61$). Furthermore, the solid grey arrow shows that the regional CPF's $\hat{R}$ decreases to 0.18 by 2025 when $\hat{W}_{2025} = 0.53$, so early action is preferable.

Cost-raising policies can backfire if imports are substantially dirtier than domestic production (see Proposition 1). We plot a hypothetical "CPF with dirty imports" for which $L = 1.33$ such that EU-wide emissions *increase*, $R < 0$.[27] Since this policy lies to the right of the $R = 0$ contour line, the negative effect gets *weaker* over time as the waterbed effect gets stronger.

Mandates to reduce carbon-intensive production in the electricity sector are also supply-side policies (Proposition 1). Examples include the British and Dutch policies to close their remaining coal-fired power plants by 2025 and 2030, respectively. Germany has also passed regulation to phase out coal by 2038.[28] This would lead to reduced de-

[25]Table 1 in Frontier Economics (2018) estimates that the Dutch price floor will reduce domestic emissions by 26 million tCO_2 in 2030, but the net EU-wide emissions reduction is only 4 million tCO_2, implying $L = 0.85$. Vollebergh (2018) estimates internal carbon leakage to be 85% for the Dutch price floor and 61% for a regional CPF including the Benelux, France and Germany.

[26]We expect internal carbon leakage to be lower for Great Britain's carbon fee as import supply is more inelastic due to interconnection constraints, but we are not aware of any empirical estimates.

[27]We assume $\theta_j/\theta_i = 2$, $\varepsilon_j^S = 5 \Leftrightarrow \eta_j^S = 0.2$, $\sigma_j = 0.2$, $\varepsilon^D = 0.5$ and $A_i = 0$.

[28]Sources: `https://www.bmwi.de/Redaktion/EN/Pressemitteilungen/2020/20200703-final-decision-to-launch-the-coal-phase-out.html` (press release), `http://www.bgbl.de/xaver/bgbl/start.xav?startbk=Bundesanzeiger_BGBl&jumpTo=bgbl120s1818.pdf` (coal phase-out law),

mand for allowances both before and after this date, relative to the counterfactual. The policy has been estimated to have an internal carbon leakage rate of 55% in 2020 (Pahle et al., 2019), so $\hat{R}_{2020} = 0.36$ ($\hat{W}_{2020} = 0.21, L = 0.55$) and decreasing to zero by 2030.

Post-2030, $\hat{W}_t = 1$, so all unilateral policies within the EU ETS end up at $R = 0$.

Aviation

Several European countries, such as Austria, Germany, Norway and Sweden, have aviation taxes. Others, such as Denmark, Ireland and the Netherlands, abolished them after initial implementation. Such policies are prone to leakage: when the Netherlands adopted an aviation tax in July 2008 at a rate of EUR 11.25 for short-haul flights and EUR 45 for long-haul flights, about 50% of the decline in passengers at Dutch airports was offset by increased passenger volumes at nearby airports in Belgium and Germany (Gordijn and Kolkman, 2011).[29] This intra-EU leakage rate of 50% is in line with Proposition 1. As a result, the Dutch government abolished the tax in July 2009—but then reintroduced a modest ticket tax of EUR 7 on all flights starting in 2021 (Forbes, 2020a). Assuming the same internal leakage rate as in 2008-9, we estimate $\hat{R}_{2020} = 0.40$ ($\hat{W}_{2020} = 0.21, L = 0.50$).

There is some broader evidence that aviation taxes are most likely in countries where leakage is mitigated—e.g., in high-population countries such as France, Germany, Italy and the United Kingdom (low σ_j) as well as countries such as Norway and Sweden whose population is far away from low-tax airports abroad (high ε_j^S) (PricewaterhouseCoopers, 2017). Austria is an exception given the proximity of Vienna to Bratislava. Greece, Croatia and Latvia—countries that also have aviation taxes—are relatively small, though their geographies are such that leakage may be less severe than for the Netherlands.

"Demand-side" unilateral policies

We now look at unilateral "demand-side" policies such as renewables support. Germany and Spain have adopted some of the world's most ambitious incentives for wind and solar energy, which include feed-in tariffs and market premium programs. Consistent with Proposition 2, Abrell et al. (2019) estimate *negative* carbon (and output) leakage as additional zero-carbon energy depresses wholesale electricity prices and offsets imported gas- and coal-fired electricity in Germany ($L = -0.50$) and Spain ($L = -0.12$).[30] Similarly,

`https://www.wired.co.uk/article/germany-coal-electricity-spremberg` (press coverage).

[29] Gordijn and Kolkman (2011) estimate that the tax accounted for nearly two million fewer passengers from Amsterdam's Schiphol Airport during the period over which the tax was in effect, while an extra one million Dutch passengers flew from foreign airports.

[30] In their Table 3, Abrell et al. (2019) report d(import quantity)/d(policy) and d(domestic quantity)/d(policy), from which we calculate output leakage as -78%, -77%, -7% and -21% for German wind, German solar, Spanish wind and Spanish solar, respectively. Similarly, we compute carbon leakage from their Table 5: -49%, -50%, -6% and -19%, respectively. Averaged over wind and solar, we use $L = -0.50$ for Germany and $L = -0.12$ for Spain in Figure 4. Schnaars (2019) provides an even more negative

a German government report finds $L = -0.65$ (Klobasa and Sensfuss, 2016). Figure 4 shows that, at least in the year 2020, the renewable support scheme in Germany reduces system-wide emissions considerably ($\hat{W}_{2020} = 0.21, L = -0.50, \hat{R}_{2020} = 1.19$); in fact, by *more* than the domestic emissions reduction in Germany. As time passes, W increases and eventually the puncture is sealed, reducing R to zero from 2030 onwards.

Proposition 2 shows equivalence between renewables support and other demand-side policies such as energy-efficiency programs and a carbon-consumption tax. Therefore, we expect negative internal leakage also for these policies but are not aware of any empirical estimates, so do not include them in Figure 4.

Sensitivity analysis

In Appendix C, we show the sensitivity of our results to $t_{B=833}$, the year in which the MSR will stop taking in allowances. The instantaneous waterbed effect decreases substantially when $t_{B=833}$ lies further in the future. We also show how the performance of two key policies—renewable energy support and a coal phase-out in Germany—changes when we allow for the Rosendahl effect (see Equation (24) in Lemma 2). This increases $\hat{W}_t$, especially for years close to $t_{B=833}$. Until the mid-2030s, the waterbed effect is still relatively limited (below 0.5) but in or after the year 2048, the waterbed effect is *larger* than 1—consistent with Proposition 4. This highlights the potential unintended consequences of announcing policies that reduce emissions demand far into the future.

5.2 Overlapping policies in North America

We now turn to discussing examples of unilateral carbon policies in North America, two of which are plotted in Figure 4. Recall from Section 4.1 that the waterbed can also be punctured due to the stochastic nature of when a carbon price corridor is binding or when a larger policy moves across steps in the allowance supply curve. A cap-and-trade system in which the carbon price trades at an auction price floor or cap has $W = 0$ while in the intermediate price range $W = 1$. For a marginal policy, the expected waterbed effect that applies to an overlapping policy thus depends on the probability that the auction price floor or cap is binding in a given year. The higher the probability that the system will trade at the price floor or cap, the more effective the puncture (Figure 3, policy 3). This feature is relevant for the two carbon markets in the United States.

carbon leakage rate of -73%, further bolstering the case for negative leakage. The differences between output and emissions leakage in Germany and Spain suggest that the marginal unit of output reduction in Germany is approximately 50% more carbon intensive than the marginal reduction for its trading partners; for Spain the emissions intensity of these marginal units are about equal. Abrell et al. (2019) show that the German power mix is indeed dirtier than Spain's.

California-Québec carbon market

California and Québec have a joint carbon market with a hybrid design. There is an auction price floor ($17.71 in 2021)[31] and a price ceiling ($65 in 2021) (Politico, 2018). Before the hard price cap is reached, there are two soft price caps that create horizontal segments in the allowance supply function: up to some limit, allowances will be offered at $41.40 and at $53.20 before the market could reach the hard price cap. Borenstein et al. (2017) estimate that, by 2030, the probability that the equilibrium will occur on any of the horizontal sections of the allowance supply curve equals $\pi = 0.83$—therefore, the expected waterbed effect $W = 1 - \pi = 0.17$ (Figure 3).[32]

The California-Québec carbon market is known to cause *external* leakage to neighbouring states that are interconnected in the electricity market (Fowlie, 2009; Caron et al., 2015). We now consider a counterfactual Western Climate Initiative (WCI) in which states surrounding California join the carbon market.[33] If California then imposed a unilateral carbon top-up fee, this would lead to "intra-WCI" carbon leakage to neighbouring states. Thus *external* leakage under the current system gets transformed into *internal* leakage under a counterfactual WCI, allowing us to rely on existing estimates from the literature. Fowlie (2009) finds that a carbon price in California that exempts out-of-state producers achieves only 25-35% of the total emissions reductions achieved under complete regulation (Arizona, Nevada, New Mexico, Oregon, Utah and Washington) so that $L = 0.65$-0.75. Caron et al. (2015) provide a relevant leakage estimate of $L = 0.09$ for California's cap-and-trade program assuming that—as the current market rules specify—there is a border-tax adjustment and "resource shuffling" is banned.[34] Figure 4 plots the hypothetical California carbon top-up fee using $L = 0.09$, as this estimate corresponds most closely to California's current market rules. Given these values, the overlapping policy would be reasonably climate effective: for every ton of carbon saved in California, system-wide emissions decrease by $R = 0.76$ tons ($W = 0.17, L = 0.09$).

[31] The auction price floor was binding in various auctions in the year 2016. In addition, in many other quarterly auctions, the markets cleared only slightly above the price ceiling. See `https://ww3.arb.ca.gov/cc/capandtrade/capandtrade.htm` for details.

[32] Borenstein et al. (2017)'s calculation is based on values of the price floor, steps, and cap that differ somewhat from the eventually-implemented level, but we expect this to have a minor impact on their estimate of π.

[33] The WCI (`http://www.wci-inc.org/`) started in 2007 as an initiative by the governors of Arizona, California, New Mexico, Oregon and Washington with a goal to develop a regional multi-sector cap-and-trade market. Most states left during the economic downturn in the early 2010s but the idea of regional carbon trading has resurfaced in discussions among states.

[34] Resource shuffling is defined as "any plan, scheme, or artifice to receive credit based on emissions reductions that have not occurred, involving the delivery of electricity to the California grid" (Caron et al., 2015). For example, out-of-state generators could reconfigure transmission so that low-carbon electricity is diverted to California and high-carbon electricity is sold to other states.

Regional Greenhouse Gas Initiative

The Regional Greenhouse Gas Initiative (RGGI) caps CO_2 emissions from electricity in eleven Northeastern states. It has a flexibility mechanism based on past allowances prices, with a 'hard' price floor and a 'soft' price cap that offers up to 10 million allowances at a fixed price ($13 in 2021; increasing at 7% per year). Once these allowances are exhausted then prices would continue to rise. The price floor was binding during 2010-2012;[35] the states decided to retire unsold allowances. The soft price cap was triggered in 2014 and 2015. Effectively, this produces an upward-sloping step-function allowance supply function which fits the discussion in Section 4.1. Once larger, non-marginal interventions are considered—such that allowance demand moves across one or several steps in the supply schedule—the effective waterbed effect lies between zero and one.

Several RGGI states have floated the idea of unilateral policies. Most notably, New York has proposed an additional carbon fee equal to the difference between the social cost of carbon and the RGGI allowance price (Forbes, 2020b). Shawhan et al. (2019) estimate the emissions leakage to other RGGI states that results from New York's policy at $L = 0.58$.[36] We do not plot New York's carbon fee in Figure 4 as we are not aware of an empirical estimate of the fraction of the time that the system is expected to trade at the price floor or ceiling, so W is missing.

Canada's national minimum carbon tax

Canada adopted a national minimum carbon tax of $20 per ton starting in 2019, increasing to $50 by 2022. Some provinces, such as Alberta and British Columbia, already had in place carbon taxes with a price above the national minimum level. Such unilateral carbon taxes face no waterbed effect (Proposition 3) but may suffer from internal leakage to other provinces. Though we are not aware of direct leakage estimates, Murray and Rivers (2015) and Yamazaki (2017) find that British Columbia's carbon tax has had negligible or modest effects on the aggregate economy, suggesting leakage is modest, and so Figure 4 plots this policy assuming $L = 0.25$ and $W = 0$, leaving a higher carbon tax in British Columbia reasonably climate-effective ($R = 0.75$).

[35] See `https://fas.org/sgp/crs/misc/R41836.pdf`

[36] New York's carbon-pricing policy differs somewhat from our theory. First, a border tax applies to imported electricity from other RGGI states. Second, there is scope for nontrivial *external* leakage to non-RGGI states. Shawhan et al. (2019) estimate this external carbon leakage to be substantially negative—an increase in renewable power in New York reduces dirty imports from non-RGGI to RGGI states. This underscores that external and internal leakage are distinct phenomena that can even have different signs. Fell and Maniloff (2018) find positive *external* leakage of 51% from the introduction of RGGI as a whole. As this is a very different policy than New York's proposed carbon price we have no reason to expect that external leakage rates would be similar.

6 Conclusion

This paper has presented a new modelling framework—combining internal carbon leakage and waterbed effects—to understand overlapping climate policies within a wider carbon-pricing system. Design matters in that different policy types have very different leakage properties. Space matters as leakages can differ substantially across industries and countries. Time matters as it affects the magnitude of the waterbed. Our results provide policy-relevant guidance on the climate benefits of 25 different combinations of unilateral policies and types of carbon-pricing systems. The issues we have highlighted extend beyond policy-making in Europe and North America and are critical for the design of new climate policies like the ongoing design of China's national cap-and-trade system. Future research on hybrid carbon-market designs should pay close attention to internal carbon leakage—more empirical estimates could help improve policy-making substantially.

Our framework has deeper connections to the wider literature on environmental economics. A fixed resource stock that will be exhausted sooner or later (Sinn, 2008; Eichner and Pethig, 2011; Van der Ploeg, 2016) impedes the climate effectiveness of policies that reduce fossil demand—and corresponds to a 100% waterbed effect in cap-and-trade.[37] Moreover, there is (external) carbon leakage if non-coalition countries increase their emissions in response to sub-global action. This has close parallels with the design of allowance supply functions in carbon markets in our framework. For example, Harstad (2012) shows how a coalition, by buying but then not exploiting specific non-coalition fossil resources, can create a vertical section in the global resource supply function—which then makes fully effective its domestic resource-conservation policy. This becomes equivalent to individual countries inside a multi-jurisdiction emissions trading system with a fixed cap (like the pre-2018 EU ETS) pursuing unilateral policies that involve cancelling allowances.

[37] The green paradox is concerned with carbon entering the economy (with the burning of fossil resources) while the waterbed effect is concerned with carbon leaving it (due to carbon pricing).

References

Abrell, Jan, Mirjam Kosch, and Sebastian Rausch, "Carbon Abatement with Renewables: Evaluating Wind and Solar Subsidies in Germany and Spain," *Journal of Public Economics*, 2019, *169*, 172–202.

Aldy, Joseph E. and William A. Pizer, "The Competitiveness Impacts of Climate Change Mitigation Policies," *Journal of the Association of Environmental and Resource Economists*, 2015, *2* (4), 565–595.

Baylis, Kathy, Don Fullerton, and Daniel H. Karney, "Leakage, Welfare and Cost-Effectiveness of Carbon Policy," *American Economic Review: Papers & Proceedings*, 2013, *103* (3), 332–337.

Böhringer, Christoph, "Two Decades of European Climate Policy: A Critical Appraisal," *Review of Environmental Economics and Policy*, 2014, *8* (1), 1–17.

Borenstein, Severin, James Bushnell, and Frank A. Wolak, "California's Cap-and-Trade Market Through 2030: A Preliminary Supply/Demand Analysis," *Energy Institute at Haas Working Paper 281*, 2017.

_ , _ , _ , and Matthew Zaragoza-Watkins, "Expecting the Unexpected: Emissions Uncertainty and Environmental Market Design," *American Economic Review*, 2019, *109* (11), 3953–3977.

Bruninx, Kenneth, Marten Ovaere, Kenneth Gillingham, and Erik Delarue, "The Unintended Consequences of the EU ETS Cancellation Policy," 2019. MPRA Paper No. 96437. Available at `https://mpra.ub.uni-muenchen.de/96437/`.

Burtraw, Dallas, Charlie Holt, Karen Palmer, and William Shobe, "Quantities with Prices: Price-Responsive Allowance Supply in Environmental Markets," *Resources for the Future Working Paper*, 2020, pp. 20–17. Available at `https://www.rff.org/publications/working-papers/quantities-prices-price-responsive-allowance-supply-environmental-markets/`.

Caron, Justin, Sebastian Rausch, and Niven Winchester, "Leakage from Sub-National Climate Policy: The Case of California's Cap-and-Trade Program," *The Energy Journal*, 2015, *36* (2), 167–190.

der Ploeg, Frederick Van, "Second-best carbon taxation in the global economy: the Green Paradox and carbon leakage revisited," *Journal of Environmental Economics and Management*, 2016, *78*, 85–105.

Eichner, Thomas and Rüdiger Pethig, "Carbon leakage, the green paradox, and perfect future markets," *International Economic Review*, 2011, *52* (3), 767–805.

European Commission, "Publication of the Total Number of Allowances in Circulation in 2020 for the Purposes of the Market Stability Reserve under the EU Emissions Trading System Established by Directive 2003/87/EC," Technical Report C(2021) 3266 final, European Commission 2021.

European Parliament and Council, "DECISION (EU) 2015/1814 OF THE EUROPEAN PARLIAMENT AND OF THE COUNCIL of 6 October 2015 concerning the establishment and operation of a market stability reserve for the Union greenhouse gas emission trading scheme and amending Directive 2003/87/EC," *Official Journal of the European Union*, 2015, *9.10.2015*, L264/1 – L264/5.

_ , "Directive (EU) 2018/410 of the European Parliament and of the Council of 14 March 2018 Amending Directive 2003/87/EC to Enhance Cost-Effective Emission Reductions and Low-Carbon Investments, and Decision (EU) 2015/1814," *Official Journal of the European Union*, 2018, *19.3.2018*, L76/4 – L76/27.

Fankhauser, Samuel, Cameron Hepburn, and Jisung Park, "Combining Multiple Climate Policy Instruments: How Not to Do It," *Climate Change Economics*, 2010, *1* (3), 209–225.

Fell, Harrison and Peter Maniloff, "Leakage in Regional Environmental Policy: The Case of the Regional Greenhouse Gas Initiative," *Journal of Environmental Economics and Management*, 2018, *87*, 1–23.

Forbes, "The Netherlands Keeps Going Upstream When It Comes to Aviation Tax," April 18th 2020. Available at `https://www.forbes.com/sites/emanuelabarbiroglio/2020/04/18/the-netherlands-keeps-going-upstream-when-it-comes-to-aviation-tax/#552d54136e86`.

_ , "New York Power Grid Proposes Adding Carbon Costs to Market Prices for Electricity," February 20th 2020. Available at `https://www.forbes.com/sites/peterdetwiler/2020/02/20/in-a-path-breaking-approach-new-yorks-grid-operator-proposes-inclusion-of-carbon-costs-in-market-prices/`.

Fowlie, Meredith L., "Incomplete Environmental Regulation, Imperfect Competition, and Emissions Leakage," *American Economic Journal: Economic Policy*, 2009, *1* (2), 72–112.

_ **and Mar Reguant**, "Challenges in the Measurement of Leakage Risk," *AEA Papers & Proceedings*, 2018, *108*, 124–129.

_ **,** _ **, and Stephen P. Ryan**, "Market-Based Emissions Regulation and Industry Dynamics," *Journal of Political Economy*, 2016, *124* (1), 249–302.

Frontier Economics, "Research on the Effects of a Carbon Price Floor," 2018. Available at `https://www.frontier-economics.com/media/2240/reasearch-effects-carbon-price-floor.pdf`.

Gerarden, Todd D., W. Spencer Reeder, and James H. Stock, "Federal Coal Program Reform, the Clean Power Plan, and the Interaction of Upstream and Downstream Climate Policies," *American Economic Journal: Economic Policy*, 2020, *12* (1), 167–199.

Gerlagh, Reyer, Roweno JRK Heijmans, and Knut Einar Rosendahl, "An Endogenous Emissions Cap Produces a Green Paradox," *Economic Policy*, 2021, *eiab011.*

Gordijn, Hugo and Joost Kolkman, "Effects of the Air Passenger Tax: Behavioral Responses of Passengers, Airlines and Airports," 2011. KiM Netherlands Institute for Transport Policy Analysis. Available at `http://publicaties.minienm.nl/documenten/effects-of-the-air-passenger-tax-behavourial-responses-of-passen`.

Goulder, Lawrence H. and Robert N. Stavins, "Challenges from State-Federal Interactions in US Climate Change Policy," *American Economic Review: Papers & Proceedings*, 2011, *101* (3), 253–257.

_ , Mark R. Jacobsen, and Arthur A. van Benthem, "Unintended Consequences from Nested State and Federal Regulations: The Case of the Pavley Greenhouse-Gas-per-Mile Limits," *Journal of Environmental Economics and Management*, 2012, *63* (2), 187–207.

Harstad, Bard, "Buy Coal! A Case for Supply-Side Environmental Policy," *Journal of Political Economy*, 2012, *120* (1), 77–115.

IPCC, "Fourth Assessment Report: Climate Change 2007—Working Group III: Mitigation of Climate Change," 2007. Chapter 11.7: International Spillover Effects, Intergovernmental Panel on Climate Change: Geneva, Switzerland.

Jarke, Johannes and Grischa Perino, "Do Renewable Energy Policies Reduce Carbon Emissions? On Caps and Inter-Industry Leakage," *Journal of Environmental Economics and Management*, 2017, *84*, 102–124.

Jenkin, Fleeming, "3. On the Principles which Regulate the Incidence of Taxes," *Proceedings of the Royal Society of Edinburgh*, 1872, *7*, 618–631.

Klobasa, Marian and Frank Sensfuss, "CO2-Minderung im Stromsektor durch den Einsatz Erneuerbarer Energien in den Jahren 2012 und 2013," 2016. Umweltbundesamt. Available at `https://www.umweltbundesamt.de/publikationen/co2-minderung-im-stromsektor-durch-den-einsatz`.

Martin, Ralf, Mirabelle Muûls, Laure B. de Preux, and Ulrich J. Wagner, "Industry Compensation under Relocation Risk: A Firm-Level Analysis of the EU Emissions Trading Scheme," *American Economic Review*, 2014, *104* (8), 2482–2508.

Murray, Brian and Nicholas Rivers, "British Columbia's Revenue-Neutral Carbon Tax: A Review of the Latest "Grand Experiment" in Environmental Policy," *Energy Policy*, 2015, *86*, 674–683.

New York Times, "These Countries Have Prices on Carbon. Are They Working?," April 2nd 2019. Available at `https://www.nytimes.com/interactive/2019/04/02/climate/pricing-carbon-emissions.html`.

Newbery, David M., David M. Reiner, and Robert A. Ritz, "The Political Economy of a Carbon Price Floor for Power Generation," *The Energy Journal*, 2019, *40.*

Newell, Richard, William Pizer, and Jiangfeng Zhang, "Managing Permit Markets to Stabilize Prices," *Environmental and Resource Economics*, 2005, *31* (2), 133–157.

Pahle, Michael, Ottmar Edenhofer, Robert Pietzcker, Oliver Tietjen, Sebastian Osorio, and Christian Flachsland, "Die Unterschätzten Risiken des Kohleausstiegs," *Energiewirtschaftliche Tagesfragen*, 2019, *69* (6), 1–4.

Perino, Grischa, "New EU ETS Phase 4 Rules Temporarily Puncture Waterbed," *Nature Climate Change*, 2018, *8* (4), 262–264.

_ , "Reply: EU ETS and the Waterbed Effect," *Nature Climate Change*, 2019, *9* (10), 736.

_ **and Maximilian Willner**, "Procrastinating Reform: The Impact of the Market Stability Reserve on the EU ETS," *Journal of Environmental Economics and Management*, 2016, *80*, 37–52.

_ **and** _ , "EU-ETS Phase IV: Allowance Prices, Design Choices and the Market Stability Reserve," *Climate Policy*, 2017, *17* (7), 936–946.

Pizer, William A., "Combining Price and Quantity Controls to Mitigate Global Climate Change," *Journal of Public Economics*, 2002, *85* (3), 409–434.

Politico, "California Pro Preview: Carbon Prices," November 16th 2018. Available at `https://www.politico.com/states/california/newsletters/politico-california-pro-preview/2018/11/16/carbon-prices-136174`.

PricewaterhouseCoopers, "The Economic Impact of Air Taxes in Europe: Germany," 2017. Available at `https://www.bdl.aero/wp-content/uploads/2018/10/The-economic-impact-of-air-taxes-in-Europe-Germany-004.pdf`.

Quemin, Simon and Raphaël Trotignon, "Emissions trading with rolling horizons," *Journal of Economic Dynamics and Control*, 2021, *125*, 104099.

Roberts, Marc J. and Michael Spence, "Effluent Charges and Licenses under Uncertainty," *Journal of Public Economics*, 1976, *5* (3-4), 193–208.

Rosendahl, Knut Einar, "EU ETS and the Waterbed Effect," *Nature Climate Change*, 2019, *9* (10), 734–735.

Samuelson, Paul, "The Transfer Problem and Transport Costs, II: Analysis of Effects of Trade Impediments," *Economic Journal*, 1954, *64*, 264–289.

Schnaars, Philip, "The Real Substitution Effect of Renewable Electricity: An Empirical Analysis for Germany," 2019. Working Paper, University of Hamburg. Available at `https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3411782`.

Shawhan, Daniel, Paul Picciano, and Karen Palmer, "Benefits and Costs of Power Plant Carbon Emissions Pricing in New York," 2019. Resources for the Future. Available at `https://www.rff.org/publications/reports/benefits-and-costs-of-the-new-york-independent-system-operators-carbon-pricing-initiative/`.

Sinn, Hans-Werner, "Public policies against global warming: a supply side approach," *International Tax and Public Finance*, 2008, *15* (4), 360–394.

Vollebergh, Herman, "National Measures Complementary to EU ETS," 2018. PBL Netherlands Environmental Assessment Agency. Available at https://www.eprg.group.cam.ac.uk/wp-content/uploads/2018/12/H.-Vollebergh_final.pdf.

Weyl, E. Glen and Michal Fabinger, "Pass-Through as an Economic Tool: Principles of Incidence under Imperfect Competition," *Journal of Political Economy*, 2013, *121* (3), 528–583.

World Bank, "State and Trends of Carbon Pricing 2020," 2020. Washington, DC: World Bank. Available at https://openknowledge.worldbank.org/handle/10986/33809. License: CC BY 3.0 IGO.

Yamazaki, Akio, "Jobs and Climate Policy: Evidence from British Columbia's Revenue-Neutral Carbon Tax," *Journal of Environmental Economics and Management*, 2017, *83*, 197–216.

Appendix A: Proofs of results and robustness discussion for internal carbon leakage

First, we derive two general results, Propositions 1A–2A, on internal carbon leakage for supply-side and demand-side unilateral policies using a general cost function $G_k(x_k, a_k)$ for $k = i, j$. Second, we then obtain Propositions 1–2 from the main text, which use the separable cost function $G_k(x_k, a_k) \equiv [C_k(x_k) + \phi_k(a_k)]$, as simple corollaries (with $G_k^{xa}(x_k, a_k = 0)$ and discuss how the key insights from this simplified model are robust. Third, we discuss robustness for unilateral policies that are not marginal.

A.1. General results with non-separable cost functions

As in the main text, firm k's emissions are $e_k = e_k^0 - a_k$ where a_k is abatement and $e_k^0 = \theta_k x_k$ is baseline emissions. Standing assumptions are $G_k^x, G_k^a > 0$ and $G_k^{xx}, G_k^{aa} > 0$ so $G_k^{aa} \to \infty$ means that additional abatement is infeasible (corresponding to $A_k = 0$ in our simple model). To maximise profits, firm k solves $\max_{x_k, a_k} \Pi_k = px_k - G_k(x_k, a_k) - \tau_k e_k$. The two first-order conditions are:

$$p = G_k^x + \tau_k \theta_k \text{ and } \tau_k = G_k^a.$$

Let $M_k(x_k; a_k) \equiv [G_k^x + \theta_k G_k^a]$ be k's optimal marginal cost of output, given its optimal choice of abatement with $\tau_k = G_k^a$. We assume that this optimised cost increases with abatement, $M_k^a(x_k; a_k) \equiv [G_k^{xa} + \theta_k G_k^{aa}] > 0$, or equivalently that:

$$\delta_k \equiv \left(1 + \frac{G_k^{ax}}{\theta_k G_k^{aa}}\right) > 0.$$

This condition is trivially met for a separable cost function ($G_k^{xa} = 0$) and, more generally, is satisfied if $G_k^{xa} \geq 0$ or $G_k^{xa} < 0$ but not too negative. Intuitively, it limits the degree of cost complementarity between output and abatement so there is "no free lunch."

It will also be useful to define an index of non-separability of k's cost function:

$$\psi_k \equiv \frac{G_k^{xa}}{G_k^{xx}} \frac{G_k^{ax}}{G_k^{aa}} \in [0, 1).$$

The separable case is nested where $\psi_k = 0$ while $\psi_k < 1$ again follows by stability. Finally, a key metric to characterise output responses in the general model will be:

$$\mu_k \equiv \frac{-p'}{[-p' + G_k^{xx}(1 - \psi_k)]} \in (0, 1)$$

where $\mu_k < 1$ is satisfied because of stability of equilibrium, $\psi_k < 1$. Armed with these preliminaries, we now derive generalisations of the results from the main text.

Supply-side unilateral policies

Proposition 1A. *With general cost functions, a supply-side unilateral policy by by country i has internal carbon leakage to country j of:*

$$L_i = \frac{\theta_j}{\theta_i}\mu_j\frac{\delta_j}{\delta_i}\frac{1}{[1+\gamma Z_i^G]} > 0,$$

where the rate of output leakage is $L_i^O = \mu_j \in (0,1)$, $\gamma = 0$ for a unilateral reduction in carbon-intensive production, $\gamma = 1$ for a unilateral carbon price, and $Z_i^G \equiv \frac{G_i^{aa}}{M_i^a}\frac{G_i^{xx}}{M_i^a}\left[(1-\psi_i)+\mu_j(1-\psi_j)\frac{G_j^{xx}}{G_i^{xx}}\right] \geq 0$ is an abatement effect.

Proof of Proposition 1A. We begin with i's unilateral carbon price for which $\tau_i = \tau_i(\tau, \lambda_i)$, and then obtain the unilateral reduction in carbon-intensive production as a special case. Differentiating i's first-order conditions yields:

$$p'(X)\left(\frac{dx_i}{d\lambda_i}+\frac{dx_j}{d\lambda_i}\right) - G_i^{xx}\frac{dx_i}{d\lambda_i} - G_i^{xa}\frac{da_i}{d\lambda_i} - \theta_i\frac{d\tau_i}{d\lambda_i} = 0$$

$$\frac{d\tau_i}{d\lambda_i} - G_i^{ax}\frac{dx_i}{d\lambda_i} - G_i^{aa}\frac{da_i}{d\lambda_i} = 0 \Longrightarrow \frac{da_i}{d\lambda_i} = \frac{1}{G_i^{aa}}\left[\frac{d\tau_i}{d\lambda_i} - G_i^{ax}\frac{dx_i}{d\lambda_i}\right].$$

As j's carbon price remains fixed, $\tau_j = \tau$, differentiating j's first-order conditions yields:

$$p'(X)\left(\frac{dx_i}{d\lambda_i}+\frac{dx_j}{d\lambda_i}\right) - G_j^{xx}\frac{dx_j}{d\lambda_i} - G_j^{xa}\frac{da_j}{d\lambda_i} = 0$$

$$-G_j^{ax}\frac{dx_j}{d\lambda_i} - G_j^{aa}\frac{da_j}{d\lambda_i} = 0 \Longrightarrow \frac{da_j}{d\lambda_i} = -\frac{G_j^{ax}}{G_j^{aa}}\frac{dx_j}{d\lambda_i}.$$

We proceed in two main steps, first deriving equilibrium changes in output levels, and then deriving changes in emissions—and hence internal carbon leakage. First, combining j's first-order conditions shows that the firms' output changes are related according to:

$$p'\frac{dx_i}{d\lambda_i} = \left[-p' + G_j^{xx}(1-\psi_j)\right]\frac{dx_j}{d\lambda_i}.$$

The same approach for i yields:

$$p'\frac{dx_j}{d\lambda_i} = \theta_i\delta_i\frac{d\tau_i}{d\lambda_i} + \left[-p' + G_i^{xx}(1-\psi_i)\right]\frac{dx_i}{d\lambda_i}.$$

using the definitions of ψ_k and δ_k. Writing this two-equation system in more compact

form using the definition of μ_k gives:

$$-\mu_j \frac{dx_i}{d\lambda_i} = \frac{dx_j}{d\lambda_i} \text{ and } -\mu_i \frac{dx_j}{d\lambda_i} = \mu_i \frac{\theta_i \delta_i}{(-p')} \frac{d\tau_i}{d\lambda_i} + \frac{dx_i}{d\lambda_i}.$$

Hence solving for the equilibrium output responses yields:

$$\frac{dx_i}{d\lambda_i} = -\left[\frac{\mu_i}{(1-\mu_i\mu_j)} \frac{\theta_i \delta_i}{(-p')}\right] \frac{d\tau_i}{d\lambda_i} < 0 \text{ and } \frac{dx_j}{d\lambda_i} = \left[\frac{\mu_i\mu_j}{(1-\mu_i\mu_j)} \frac{\theta_i \delta_i}{(-p')}\right] \frac{d\tau_i}{d\lambda_i} > 0.$$

Therefore the rate of internal output leakage is:

$$L_i^{\mathrm{O}} \equiv \frac{dx_j/d\lambda_i}{-dx_j/d\lambda_i} = \mu_j \in (0,1),$$

which is always positive but less than 100% by stability. Second, recall that emissions changes and output changes are related according to:

$$\frac{de_k}{d\lambda_i} = \theta_k \frac{dx_k}{d\lambda_i} - \frac{da_k}{d\lambda_i}.$$

Using j's equilibrium output response and its first-order condition for abatement we obtain:

$$\frac{de_j}{d\lambda_i} = \theta_j \delta_j \frac{dx_j}{d\lambda_i} = \theta_i \theta_j \frac{\mu_i \mu_j}{(1-\mu_i\mu_j)} \frac{\delta_i \delta_j}{(-p')} \frac{d\tau_i}{d\lambda_i} > 0.$$

We similarly obtain for i:

$$\frac{de_i}{d\lambda_i} = \theta_i \delta_i \frac{dx_i}{d\lambda_i} - \frac{1}{G_i^{aa}} \frac{d\tau_i}{d\lambda_i} = -\theta_i^2 \left[\frac{\mu_i}{(1-\mu_i\mu_j)} \frac{\delta_i^2}{(-p')} + \frac{1}{\theta_i^2 G_i^{aa}}\right] \frac{d\tau_i}{d\lambda_i} < 0.$$

Therefore the rate of internal carbon leakage due to the unilateral carbon price satisfies:

$$L_i = \frac{\theta_j}{\theta_i} \mu_j \frac{\frac{\mu_i}{(1-\mu_i\mu_j)} \frac{\delta_i \delta_j}{(-p')}}{\left[\frac{\mu_i}{(1-\mu_i\mu_j)} \frac{\delta_i^2}{(-p')} + \frac{1}{\theta_i^2 G_i^{aa}}\right]} = \frac{\theta_j}{\theta_i} \mu_j \frac{\delta_j}{\delta_i} \frac{1}{\left[1 + \frac{(-p')}{\delta_i^2 \theta_i^2 G_i^{aa}} \left[\frac{1}{\mu_i} - \mu_j\right]\right]}.$$

It will be useful to rewrite the last term as follows. Recalling the definition $\mu_k \equiv (-p')/\left[-p' + G_k^{xx}(1-\psi_k)\right]$, observe that:

$$(-p')\left[\frac{1}{\mu_i} - \mu_j\right] = \left[\frac{G_i^{xx}(1-\psi_i)}{-p'} + \frac{G_j^{xx}(1-\psi_j)}{\left[-p' + G_j^{xx}(1-\psi_j)\right]}\right](-p') = G_i^{xx}\left[(1-\psi_i) + \mu_j(1-\psi_j)\frac{G_j^{xx}}{G_i^{xx}}\right].$$

Also recalling that $M_i^a(x_k; a_k) \equiv [G_i^{xa} + \theta_i G_i^{aa}] > 0$, we have:

$$\frac{G_i^{xx}}{\delta_i^2 \theta_i^2 G_i^{aa}} = \frac{1}{\left(\theta_i + \frac{G_i^{ax}}{G_i^{aa}}\right)^2 \frac{G_i^{aa}}{G_i^{xx}}} = \frac{G_i^{aa} G_i^{xx}}{(G_i^{ax} + \theta_i G_i^{aa})^2} = \frac{G_i^{aa}}{M_i^a} \frac{G_i^{xx}}{M_i^a}.$$

Using these terms in the expression for internal carbon leakage yields the result as claimed for the unilateral carbon prices.

Now consider the unilateral reduction in carbon-intensive production by country i, represented as $dx_i/d\lambda_i < 0$, where the common carbon price remains unchanged, $\tau_i = \tau_j = \tau$. This problem has the same structure as before—except that i's output change $dx_i/d\lambda_i < 0$ is determined by policy directly rather than induced in equilibrium by a unilateral carbon price. The remaining choices—abatement by i and output and abatement by j—remain optimal by the respective first-order conditions.

Hence differentiating i's first-order condition for abatement yields:

$$-G_i^{ax} \frac{dx_i}{d\lambda_i} - G_i^{aa} \frac{da_i}{d\lambda_i} = 0 \Longrightarrow \frac{da_i}{d\lambda_i} = -\frac{G_i^{ax}}{G_i^{aa}} \frac{dx_i}{d\lambda_i}.$$

Differentiating j's first-order conditions yields:

$$p'(X) \left(\frac{dx_i}{d\lambda_i} + \frac{dx_j}{d\lambda_i} \right) - G_j^{xx} \frac{dx_j}{d\lambda_i} - G_j^{xa} \frac{da_j}{d\lambda_i} = 0$$

$$-G_j^{ax} \frac{dx_j}{d\lambda_i} - G_j^{aa} \frac{da_j}{d\lambda_i} = 0 \Longrightarrow \frac{da_j}{d\lambda_i} = -\frac{G_j^{ax}}{G_j^{aa}} \frac{dx_j}{d\lambda_i}.$$

Writing these conditions in more compact form, using the definitions of ψ_j and μ_j, gives:

$$-\mu_j \frac{dx_i}{d\lambda_i} = \frac{dx_j}{d\lambda_i} > 0 \Longrightarrow L_i^{\mathrm{O}} \equiv \frac{dx_j/d\lambda_i}{-dx_j/d\lambda_i} = \mu_j \in (0, 1),$$

which is exactly as for the unilateral carbon price.

Emissions changes and output changes are again related according to:

$$\frac{de_k}{d\lambda_i} = \theta_k \frac{dx_k}{d\lambda_i} - \frac{da_k}{d\lambda_i}.$$

Using firms' equilibrium output responses and first-order conditions for abatement, we obtain:

$$\frac{de_i}{d\lambda_i} = \theta_i \delta_i \frac{dx_i}{d\lambda_i} < 0 \text{ and } \frac{de_j}{d\lambda_i} = \theta_j \delta_j \frac{dx_j}{d\lambda_i} = -\theta_j \delta_j \mu_j \frac{dx_i}{d\lambda_i} > 0$$

So the equilibrium rate of internal carbon leakage is as claimed:

$$L_i = \frac{\theta_j}{\theta_i}\mu_j\frac{\delta_j}{\delta_i} > 0.$$

Demand-side unilateral policies

Proposition 2A. *With general cost functions, a demand-side unilateral policy by country i of (i) a renewables support program that brings in additional zero-carbon production, or (ii) an energy-efficiency program that reduces demand for carbon-intensive production, or (iii) a carbon-consumption tax has internal carbon leakage to country j of:*

$$L_i = -\frac{\theta_j}{\theta_i}\left[\frac{\mu_j/(1-\mu_j)}{\mu_i/(1-\mu_i)}\right]\frac{\delta_j}{\delta_i} < 0,$$

where the rate of output leakage is $L_i^O = -\frac{\mu_j/(1-\mu_j)}{\mu_i/(1-\mu_i)} < 0$.

Proof of Proposition 2A. As explained in the main text, all three of these demand-side unilateral policies are modeled via their impact on the demand curve, with $\frac{\partial}{\partial\lambda_i}p(X;\lambda_i) < 0$. The common carbon price remains unchanged, $\tau_i = \tau_j = \tau$. Thus differentiating i's first-order conditions for the impact of the unilateral policy yields:

$$\frac{\partial}{\partial\lambda_i}p(X;\lambda_i) + p'(X;\lambda_i)\left(\frac{dx_i}{d\lambda_i} + \frac{dx_j}{d\lambda_i}\right) - G_i^{xx}\frac{dx_i}{d\lambda_i} - G_i^{xa}\frac{da_i}{d\lambda_i} = 0$$

$$-G_i^{ax}\frac{dx_i}{d\lambda_i} - G_i^{aa}\frac{da_i}{d\lambda_i} = 0 \Longrightarrow \frac{da_i}{d\lambda_i} = -\frac{G_i^{ax}}{G_i^{aa}}\frac{dx_i}{d\lambda_i}.$$

Differentiating j's first-order conditions yields symmetrically:

$$\frac{\partial}{\partial\lambda_i}p(X;\lambda_i) + p'(X;\lambda_i)\left(\frac{dx_i}{d\lambda_i} + \frac{dx_j}{d\lambda_i}\right) - G_j^{xx}\frac{dx_j}{d\lambda_i} - G_j^{xa}\frac{da_j}{d\lambda_i} = 0$$

$$-G_j^{ax}\frac{dx_j}{d\lambda_i} - G_j^{aa}\frac{da_j}{d\lambda_i} = 0 \Longrightarrow \frac{da_j}{d\lambda_i} = -\frac{G_j^{ax}}{G_j^{aa}}\frac{dx_j}{d\lambda_i}.$$

We again proceed in two main steps, first deriving equilibrium output responses, and then emissions responses. First, combining j's first-order conditions shows that firms' output changes are related according to:

$$\frac{\partial}{\partial\lambda_i}p(X;\lambda_i) + p'\frac{dx_i}{d\lambda_i} = \frac{dx_j}{d\lambda_i}\left[-p' + G_j^{xx}(1-\psi_j)\right].$$

The same approach for i yields:

$$\frac{\partial}{\partial \lambda_i} p(X;\lambda_i) + p' \frac{dx_j}{d\lambda_i} = \frac{dx_i}{d\lambda_i}\left[-p' + G_i^{xx}(1-\psi_i)\right]$$

using the definition of ψ_k. Writing this two-equation system using the definition of μ_k gives:

$$\frac{dx_i}{d\lambda_i} = -\mu_i \left[\frac{dx_j}{d\lambda_i} - \frac{1}{(-p')}\frac{\partial}{\partial \lambda_i} p(X;\lambda_i)\right]$$

$$\frac{dx_j}{d\lambda_i} = -\mu_j \left[\frac{dx_i}{d\lambda_i} - \frac{1}{(-p')}\frac{\partial}{\partial \lambda_i} p(X;\lambda_i)\right].$$

Solving for equilibrium output responses yields:

$$\frac{dx_i}{d\lambda_i} = \frac{\mu_i(1-\mu_j)}{(1-\mu_i\mu_j)}\frac{1}{(-p')}\frac{\partial}{\partial \lambda_i} p(X;\lambda_i) < 0$$

$$\frac{dx_j}{d\lambda_i} = \frac{\mu_j(1-\mu_i)}{(1-\mu_i\mu_j)}\frac{1}{(-p')}\frac{\partial}{\partial \lambda_i} p(X;\lambda_i) < 0.$$

So the rate of internal output leakage is:

$$L_i^{\mathrm{O}} \equiv \frac{dx_j/d\lambda_i}{-dx_j/d\lambda_i} = -\frac{\mu_j(1-\mu_i)}{\mu_i(1-\mu_j)} < 0$$

which is always negative. Second, emissions changes and output changes are here related according to:

$$\frac{de_k}{d\lambda_i} = \theta_k \frac{dx_k}{d\lambda_i} - \frac{da_k}{d\lambda_i}.$$

Using j's equilibrium output response, its first-order condition for abatement, and the definition of δ_k, we obtain:

$$\frac{de_j}{d\lambda_i} = \left(\theta_j + \frac{G_j^{ax}}{G_j^{aa}}\right)\frac{dx_j}{d\lambda_i} = \theta_j\delta_j \frac{\mu_j(1-\mu_i)}{(1-\mu_i\mu_j)}\frac{1}{(-p')}\frac{\partial}{\partial \lambda_i} p(X;\lambda_i) < 0.$$

We similarly obtain for i:

$$\frac{de_i}{d\lambda_i} = \left(\theta_i + \frac{G_i^{xa}}{G_i^{aa}}\right)\frac{dx_i}{d\lambda_i} = \theta_i\delta_i \frac{\mu_i(1-\mu_j)}{(1-\mu_i\mu_j)}\frac{1}{(-p')}\frac{\partial}{\partial \lambda_i} p(X;\lambda_i) < 0.$$

Therefore the equilibrium rate of internal carbon leakage is:

$$L_i = -\frac{\theta_j}{\theta_i}\left[\frac{\mu_j/(1-\mu_j)}{\mu_i/(1-\mu_i)}\right]\frac{\delta_j}{\delta_i},$$

as claimed in the result.

A.2. Robustness of results with separable cost functions

We now derive Propositions 1–2 from the main text, for separable cost functions, as direct corollaries of Propositions 1A–2A. The key difference is that, in the separable model, output and abatement decisions are independent while, in the general model, abatement can also be induced by changes to output. In comparing the two sets of results, we discuss how the key insights from the simplified model in the main text are nonetheless robust.

The separable cost function $G_k(x_k, a_k) \equiv [C_k(x_k) + \phi_k(a_k)]$ is nested within the general model where $G_k^{xa}(x_k, a_k) = 0$. The general model then simplifies with $\delta_k = 1$, $\psi_k = 0$ as well as $\mu_k = (-p')/(-p' + C_k'') \in (0, 1)$ for $k = i, j$.

To present leakage formulae in terms of simple demand and supply elasticities, we begin by recording two preliminary results. First, using the price elasticity of demand $\varepsilon^D \equiv -p(\cdot)/Xp'(\cdot) > 0$ and k's elasticity of total marginal cost $\eta_k^S \equiv x_k \widehat{C}_k''(x_k)/\widehat{C}_k'(x_k) > 0$, where $\widehat{C}_k'(x_k) \equiv [C_k'(x_k) + \tau_k \theta_k] = p(X)$ and $\widehat{C}_k''(x_k) \equiv C_k''(x_k)$, we can rewrite the cost term as follows:

$$C_k''(x_k) = \frac{x_k C_k''(x_k)}{\widehat{C}_k'(x_k)} \frac{\widehat{C}_k'(x_k)}{x_k} = \frac{x_k \widehat{C}_k''(x_k)}{\widehat{C}_k'(x_k)} \frac{\widehat{C}_k'(x_k)}{x_k} = \eta_k^S \frac{p(X)}{X} \frac{1}{\sigma_k} = \frac{p(X)}{X} \frac{1}{\sigma_k \varepsilon_k^S}$$

where the last expression uses the definition of k's market share, $\sigma_k \equiv x_k/X \in (0, 1)$, and $\eta_k^S = 1/\varepsilon_k^S$ by its first-order condition. Second, using the same ingredients, we also obtain that:

$$\mu_k \equiv \frac{-p'}{(-p' + C_k'')} = \frac{\sigma_k}{(\sigma_k + \varepsilon^D/\varepsilon_k^S)} > 0,$$

which will again be the key driver of firms' equilibrium output responses.

Supply-side unilateral policies

Proposition 1. A supply-side unilateral policy by by country i has internal carbon leakage to country j of:

$$L_i = \frac{\theta_j}{\theta_i} \left[\frac{\sigma_j}{(\sigma_j + \varepsilon^D/\varepsilon_j^S)} \right] \frac{1}{(1 + \gamma Z_i)} > 0,$$

where $\gamma = 0$ for a unilateral reduction in carbon-intensive production, $\gamma = 1$ for a unilateral carbon price, and $Z_i \equiv \frac{A_i}{(1-A_i)} \left(1 + \frac{(1-\sigma_j)\varepsilon_i^S/\varepsilon_j^S}{(\sigma_j + \varepsilon^D/\varepsilon_j^S)} \right) \geq 0$ is an abatement effect.

Proof of Proposition 1. For the unilateral carbon price ($\gamma = 1$), the leakage formula from Proposition 1A simplifies to:

$$L_i = \frac{\theta_j}{\theta_i} \mu_j \frac{1}{\left[1 + \frac{G_i^{aa}}{M_i^a} \frac{G_i^{xx}}{M_i^a} \left[1 + \mu_j \frac{G_j^{xx}}{G_i^{xx}} \right] \right]} = \frac{\theta_j}{\theta_i} \mu_j \frac{1}{\left[1 + \frac{C_i''}{\theta_i^2 \phi_i''} \left[1 + \mu_j \frac{C_j''}{C_i''} \right] \right]}.$$

Using the two preliminary results, including that $C_j''/C_i'' = \sigma_i\varepsilon_i^S/\sigma_j\varepsilon_j^S$, and the definition of k's abatement opportunity from the main text, $A_k = C_k''/[C_k''+\theta_k^2\phi_k'']$, yields the result. For the unilateral reduction in carbon-intensive production ($\phi = 0$), Proposition 1A simplifies directly to:

$$L_i = \frac{\theta_j}{\theta_i}\mu_j\frac{\delta_j}{\delta_i} = \frac{\theta_j}{\theta_i}\mu_j.$$

Comparing this with the general result from Proposition 1A, an obvious difference is the absence of the term δ_j/δ_i, where $\delta_k = (1 + G_k^{ax}/\theta_k G_k^{aa}) > 0$ captures the extent of non-separability in k's cost function. There are two immediate observations. First, all else equal, the two results will be similar—even identical—if non-separability plays out similarly for both firms, with $\delta_i \simeq \delta_j \neq 1$. Second, there is no obvious bias: the simplified result is an overestimate of internal leakage if $\delta_j < \delta_i$ and an underestimate otherwise.

To understand the economics, observe that, if $\delta_k < 1 \Leftrightarrow G_k^{xa} < 0$ ($k = i, j$), j tends to abate more for a given output increase—which pushes downwards the internal leakage of i's policy. By the same token, however, i's output reduction then undermines its own abatement incentive—which pushes internal leakage upwards. The net effect is therefore ambiguous. The reverse logic applies where $\delta_k > 1 \Leftrightarrow G_k^{xa} > 0$.

A second difference between the two results arises via the rate of output leakage. In particular, recall that $L_i^{\mathrm{O}} = \mu_j \equiv (-p')/\left[-p' + G_j^{xx}(1-\psi_j)\right] \in (0,1)$ in the general case. Hence, from the same starting point, output leakage is more pronounced in the general case ($\psi_j > 0$) than in the separable case ($\psi_j = 0$). Intuitively, if $G_j^{xa} < 0$, then abatement raises the marginal return to output, and vice versa, so, all else equal, j's output increase is more pronounced. The same logic applies in reverse for $G_j^{xa} > 0$: abatement makes output less attractive, and vice versa. Hence, across both cases, non-separability raises j's marginal return to output—so output leakage L_i^{O} is higher for $G_j^{xa} \neq 0$, all else equal, than for $G_j^{xa} = 0$.

The relative emissions intensity θ_j/θ_i plays exactly the same role in both results, and internal carbon leakage exceeds 100% if it is sufficiently pronounced. Finally, the abatement effect also plays a similar role in the general (Z_i^G) and separable (Z_i) models for the unilateral carbon price—but is irrelevant for the unilateral production cut.

In sum, while the precise numbers may differ, the main insights from the case with separable cost functions hold more generally—most notably that internal leakage from supply-side policies is always positive.

Demand-side unilateral policies

Proposition 2. A demand-side unilateral policy by country i of (i) a renewables support program that brings in additional zero-carbon production, or (ii) an energy-efficiency pro-

gram that reduces demand for carbon-intensive production, or (iii) a carbon consumption tax has internal carbon leakage to country j of:

$$L_i = -\frac{\theta_j}{\theta_i}\frac{\sigma_j}{(1-\sigma_j)}\frac{\varepsilon_j^S}{\varepsilon_i^S} < 0.$$

Proof of Proposition 2. The expression for internal carbon leakage from Proposition 2A simplifies as:

$$L_i = -\frac{\theta_j}{\theta_i}\left[\frac{\mu_j/(1-\mu_j)}{\mu_i/(1-\mu_i)}\right]\frac{\delta_j}{\delta_i} = -\frac{\theta_j}{\theta_i}\frac{C_i''}{C_j''}.$$

Using the relationship $C_k''(x_k) = \frac{p(X)}{X}\frac{1}{\sigma_k \varepsilon_k^S}$ yields the result as claimed.

Comparing this with the general result from Proposition 2A, similar effects are at work as for supply-side policies. A simplification is that demand-side policies do not lead to a carbon price-induced abatement effect, neither in the separable nor in the general case.

First, exactly as for supply-side policies, the term δ_j/δ_i is absent in the separable case. However, by the same arguments as before, this effect (i) becomes negligible if non-separability plays out similarly for both firms, with $\delta_i \simeq \delta_j \neq 1$ and (ii) does not lead to any clear-cut bias in the result on internal leakage for the separable case.

Second, for demand-side policies, by contrast, the impact of separability on output leakage is now ambiguous as firms in both countries experience a direct change on their marginal return to output. In particular, note that $L_i^{\mathrm{O}} = -[\mu_j/(1-\mu_j)]/[\mu_i/(1-\mu_i)] = -G_i^{xx}(1-\psi_i)/G_j^{xx}(1-\psi_j)$ in the general case. This makes clear that, very similar to the previous point, this non-separability additional effect from the general case may be negligible and does not lead to any clear-cut bias in Proposition 2.

Third, the relative emissions intensity θ_j/θ_i again plays an identical role in both results.

In sum, the main insights from the separable case again hold more generally—most notably that internal leakage from demand-side policies is always negative.

A.3. Robustness of results with non-marginal unilateral policies

Our results so far have focused on marginal unilateral policies, with $d\lambda_i > 0$, that shift equilibrium emissions by small amounts, de_i and de_j. This yields a rate of internal carbon leakage $L_i = \frac{de_j}{-de_i}$ that can be seen as an approximation to a non-marginal rate $L_i = \frac{\triangle e_j}{-\triangle e_i}$. More generally, a unilateral policy tightens from an initial level $\underline{\lambda}_i \geq 0$ to a new level $\overline{\lambda}_i$ where $\triangle\lambda_i \equiv (\overline{\lambda}_i - \underline{\lambda}_i)$ is a discrete change. We here make two points on the robustness of the results from the first-order approximation.

The first point is that the insight that supply-side policies have positive internal leakage while it is negative for demand-side policies also holds for non-marginal policy. To see

why, write the non-marginal leakage rate as:

$$L_i \equiv \frac{\triangle e_j}{-\triangle e_i} = \frac{\int_{\underline{\lambda}_i}^{\underline{\lambda}_i+\triangle\lambda_i} \left(\frac{de_j}{-de_i}\right)\left(\frac{de_i}{d\lambda_i}\right) d\lambda_i}{\int_{\underline{\lambda}_i}^{\underline{\lambda}_i+\triangle\lambda_i} \left(\frac{de_i}{d\lambda_i}\right) d\lambda_i},$$

showing that it is a weighted average of marginal leakage rates, where $de_i/d\lambda_i < 0$ for all $\lambda_i \in [\underline{\lambda}_i, \underline{\lambda}_i + \triangle\lambda_i]$. Hence $\text{sign}(\frac{\triangle e_j}{-\triangle e_i}) = \text{sign}(\frac{de_j}{-de_i})$, which is unambiguously positive (negative) for supply-side (demand-side) policies, as shown in Propositions 1(A) and 2(A).

The second point is that the marginal approximation implies no obvious bias in the magnitude and, in an important special case, yields an exact result. As we have seen, marginal rates of internal leakage in general depend on first-order derivatives of demand (via the demand elasticity) and second-order derivatives of cost functions (via supply elasticities and abatement opportunities). So the non-marginal leakage rate will be quantitatively similar to marginal leakage as long as any second-order demand terms and third-order cost terms are negligible as needed. For supply-side policies, this obtains exactly if the demand curve is linear ($p'(X)$ is constant) and the cost functions are quadratic in output and abatement (in the general case, $G_k^{xx}, G_k^{aa}, G_k^{xa}$ all constant). Then $L_i \equiv \frac{\triangle e_j}{-\triangle e_i} = \frac{de_j}{-de_i}$ since marginal leakage $\frac{de_j}{-de_i}$ is constant over $\lambda_i \in [\underline{\lambda}_i, \underline{\lambda}_i + \triangle\lambda_i]$. By contrast, for demand-side policies, the exact result does not require any restrictions on the demand curve, recalling that output leakage in the general case $L_i^{\text{O}} = -[\mu_j/(1-\mu_j)]/[\mu_i/(1-\mu_i)] = -G_i^{xx}(1-\psi_i)/G_j^{xx}(1-\psi_j)$ does not depend on any demand-side properties. Moreover, the simple marginal formulae contain no obvious bias: they could be an over- or underestimate depending on the precise higher-order properties of cost functions, and on whether demand is convex or concave.

In sum, the main insights on internal leakage from marginal analysis also hold for potentially much more complex non-marginal policies.

Appendix B: Proofs of results and robustness discussion for waterbed effects

First, we present proofs for several results from Section 4. Second, we discuss the robustness of our results to overlapping policies that are non-marginal. Third, we derive Lemma 2 with the instantaneous waterbed effect in the reformed EU ETS.

B.1. Proofs of results on waterbed effects

Derivation of Equation (10)

Application of Cramer's rule to conditions (7)-(9) yields:

$$\frac{\partial \tau_1}{\partial \lambda_i} = \frac{\frac{\partial \rho_2}{\partial e_2}\frac{\partial \rho_1}{\partial \lambda_i} + \frac{\partial \rho_1}{\partial e_1}\frac{\partial \rho_2}{\partial \lambda_i}}{\frac{\partial \rho_2}{\partial e_2} + (1+r)\frac{\partial \rho_1}{\partial e_1} - \frac{\partial \rho_1}{\partial e_1}\frac{\partial \rho_2}{\partial e_2}\frac{\partial s_2}{\partial \tau_1}}. \tag{A.1}$$

Total differentiation of $\tau_t = \rho_t(e_t, \lambda_i)$ (see first-order conditions (7) and (8)) yields

$$\frac{de_t}{d\lambda_i} = -\frac{\frac{\partial \rho_t}{\partial \lambda_i}}{\frac{\partial \rho_t}{\partial e_t}}. \tag{A.2}$$

Cancelling $-\frac{\partial \rho_1}{\partial e_1}\frac{\partial \rho_2}{\partial e_2}$ from (A.1), using (A.2), and substituting the slope of the allowance demand curve $\frac{\partial e_t}{\partial \tau_t}$ for the inverse of the slope of the inverse allowance demand curve $(\partial \rho_t/\partial e_t)^{-1}$ in the denominator yields:

$$\frac{\partial \tau_1}{\partial \lambda_i} = \frac{\frac{de_1}{d\lambda_i} + \frac{de_2}{d\lambda_i}}{\frac{\partial s_2}{\partial \tau_1} - \frac{\partial e_1}{\partial \tau_1} - (1+r)\frac{\partial e_2}{\partial \tau_2}}.$$

Using $de/d\lambda_i = de_1/d\lambda_i + de_2/d\lambda_i$, $\frac{\partial e}{\partial \tau_1} = \frac{\partial e_1}{\partial \tau_1} + \frac{\partial e_2}{\partial \tau_1}$ and $\frac{\partial e_2}{\partial \tau_1} = \frac{1+r}{1+r}\frac{\partial e_2}{\partial \tau_1} = (1+r)\frac{\partial e_2}{\partial \tau_2}$ yields Equation (10).

Derivation of Equation (16)

Application of Cramer's rule to conditions (13)-(15) yields:

$$\frac{\partial e_1^*}{\partial \lambda_i} = \frac{\frac{\partial \rho_2}{\partial \lambda_i} - (1+r)\frac{\partial \rho_1}{\partial \lambda_i}}{(1+r)\frac{\partial \rho_1}{\partial e_1} - \frac{\partial \rho_2}{\partial e_2}\left(1 + \frac{\partial s_2}{\partial b}\right)}.$$

Cancelling $-\frac{\partial \rho_1}{\partial e_1}\frac{\partial \rho_2}{\partial e_2}$, using (A.2), and substituting the slope of the allowance demand curve $\frac{\partial e_t}{\partial \tau_t}$ for the inverse of the slope of the inverse allowance demand curve $(\partial \rho_t/\partial e_t)^{-1}$ yields:

$$\frac{\partial e_1^*}{\partial \lambda_i} = -\frac{\frac{de_2}{d\lambda_i}\frac{\partial e_1}{\partial \tau_1} - (1+r)\frac{de_1}{d\lambda_i}\frac{\partial e_2}{\partial \tau_2}}{(1+r)\frac{\partial e_2}{\partial \tau_2} - \frac{\partial e_1}{\partial \tau_1}\left(1 + \frac{\partial s_2}{\partial b}\right)}.$$

Using $de/d\lambda_i = de_1/d\lambda_i + de_2/d\lambda_i$, $\frac{\partial e}{\partial \tau_1} = \frac{\partial e_1}{\partial \tau_1} + \frac{\partial e_2}{\partial \tau_1}$ and $\frac{\partial e_2}{\partial \tau_1} = \frac{1+r}{1+r}\frac{\partial e_2}{\partial \tau_1} = (1+r)\frac{\partial e_2}{\partial \tau_2}$ yields

$$\frac{\partial e_1^*}{\partial \lambda_i} = -\frac{\frac{de}{d\lambda_i}\frac{\partial e}{\partial \tau_1} - \frac{de_1}{d\lambda_i}\frac{\partial e_2}{\partial \tau_2}}{\frac{\partial e}{\partial \tau_1} + \frac{\partial e_1}{\partial \tau_1}\frac{\partial s_2}{\partial b}}.$$

Factoring out $de/d\lambda_i$ in the numerator and dividing both numerator and denominator by $\partial e/\partial \tau_1$ yields (16).

Proof of Corollary 1

First use conditions (13)-(15) and Cramer's rule to compute:

$$\frac{\partial \tau_1}{\partial \lambda_i} = -\frac{\frac{\partial \rho_1}{\partial \lambda_i}\frac{\partial \rho_2}{\partial e_2}\left(1+\frac{\partial s_2}{\partial b}\right)+\frac{\partial \rho_1}{\partial e_1}\frac{\partial \rho_2}{\partial \lambda_i}}{(1+r)\frac{\partial \rho_1}{\partial e_1}-\frac{\partial \rho_2}{\partial e_2}\left(1+\frac{\partial s_2}{\partial b}\right)}.$$

Cancelling $-\frac{\partial \rho_1}{\partial e_1}\frac{\partial \rho_2}{\partial e_2}$, using (A.2), and substituting the slope of the allowance demand curve $\frac{\partial e_t}{\partial \tau_t}$ for the inverse of the slope of the inverse allowance demand curve $(\partial \rho_t/\partial e_t)^{-1}$ yields:

$$\frac{\partial \tau_1}{\partial \lambda_i} = -\frac{\frac{de_1}{d\lambda_i}\left(1+\frac{\partial s_2}{\partial b}\right)+\frac{de_2}{d\lambda_i}}{(1+r)\frac{\partial e_2}{\partial \tau_2}-\frac{\partial e_1}{\partial \tau_1}\left(1+\frac{\partial s_2}{\partial b}\right)}.$$

Using $de/d\lambda_i = de_1/d\lambda_i + de_2/d\lambda_i$, $\frac{\partial e}{\partial \tau_1} = \frac{\partial e_1}{\partial \tau_1} + \frac{\partial e_2}{\partial \tau_1}$ and $\frac{\partial e_2}{\partial \tau_1} = \frac{1+r}{1+r}\frac{\partial e_2}{\partial \tau_1} = (1+r)\frac{\partial e_2}{\partial \tau_2}$ yields

$$\frac{\partial \tau_1}{\partial \lambda_i} = -\frac{\frac{de}{d\lambda_i}+\frac{de_1}{d\lambda_i}\frac{\partial s_2}{\partial b}}{\frac{\partial e}{\partial \tau_1}+\frac{\partial e_1}{\partial \tau_1}\frac{\partial s_2}{\partial b}}. \tag{A.3}$$

Next we derive the equilibrium expansion path by relating changes in the equilibrium allowance supply to changes in the equilibrium allowance price that are induced by the shift in total allowance demand:

$$\left.\frac{ds_2}{d\tau_1}\right|_{equ} = \frac{\frac{\partial s_2}{\partial \lambda_i}}{\frac{\partial \tau_1}{\partial \lambda_i}} = \frac{\frac{\partial s_2}{\partial b}}{1+\frac{\partial s_2}{\partial b}\frac{\frac{\partial e_1}{\partial \tau_1}}{\frac{\partial e}{\partial \tau_1}}}\cdot\left[\frac{\frac{\partial e_1}{\partial \tau_1}}{\frac{\partial e}{\partial \tau_1}}-\beta\right]\cdot(-1)\frac{\frac{\partial e}{\partial \tau_1}\left[1+\frac{\partial s_2}{\partial b}\frac{\frac{\partial e_1}{\partial \tau_1}}{\frac{\partial e}{\partial \tau_1}}\right]}{\left[1+\frac{\partial s_2}{\partial b}\beta\right]}.$$

Cancelling $1+\frac{\partial s_2}{\partial b}\frac{\frac{\partial E_1}{\partial \tau_1}}{\frac{\partial E}{\partial \tau_1}}$ yields Equation (22).

It now remains to be shown that Propositions 3 and 4 are equivalent. To see this substitute (22) into (12) to get:

$$W = \frac{-\frac{\partial e}{\partial \tau_1}}{\left[\frac{\frac{\partial s_2}{\partial b}\frac{\partial e}{\partial \tau_1}}{1+\frac{\partial s_2}{\partial b}\beta}\cdot\left(\beta-\frac{\frac{\partial e_1}{\partial \tau_1}}{\frac{\partial e}{\partial \tau_1}}\right)\right]-\frac{\partial e}{\partial \tau_1}}.$$

Note that $W \in [0,1]$ only holds for weakly upward-sloping allowance supply curves. This is no longer guaranteed once we substitute in Equation (22). Both values below 0 and above 1 are now possible. Dividing the above equation by $-\frac{\partial e}{\partial \tau_1}$ and multiplying it by

$1 + \frac{\partial s_2}{\partial b}\beta$ obtains:

$$W = \frac{1 + \frac{\partial s_2}{\partial b}\beta}{\frac{\partial s_2}{\partial b} \cdot \left[\frac{\frac{\partial e_1}{\partial \tau_1}}{\frac{\partial e}{\partial \tau_1}} - \beta\right] + 1 + \frac{\partial s_2}{\partial b}\beta}.$$

Cancel $\frac{\partial s_2}{\partial b}\beta$ in the denominator to obtain Equation (18).

B.2. Robustness of results with non-marginal unilateral policies

We have so far focused on marginal overlapping policies, with $d\lambda_i > 0$, that shift emissions demand at fixed carbon prices by a small amount, de. This yields a waterbed effect $W = 1 - \frac{de^*}{de}$ that can be seen as an approximation to a non-marginal rate $W = 1 - \frac{\triangle e^*}{\triangle e}$. More generally, we now consider a unilateral policy that tightens from an initial level $\underline{\lambda}_i \geq 0$ to a new level $\overline{\lambda}_i$ where $\triangle\lambda_i \equiv (\overline{\lambda}_i - \underline{\lambda}_i)$ is a discrete change.

Here we show that Proposition 3 and 4 extend to non-marginal changes in policy. To see why, write the non-marginal waterbed effect as:

$$W = 1 - \frac{\triangle e^*}{\triangle e} = 1 - \frac{\int_{\underline{\lambda}_i}^{\underline{\lambda}_i + \triangle\lambda_i} \left(\frac{de}{d\lambda_i} + \frac{de^*}{d\tau_1}\frac{d\tau_1}{d\lambda_i}\right) d\lambda_i}{\int_{\underline{\lambda}_i}^{\underline{\lambda}_i + \triangle\lambda_i} \left(\frac{de}{d\lambda_i}\right) d\lambda_i} = -\frac{\int_{\underline{\lambda}_i}^{\underline{\lambda}_i + \triangle\lambda_i} \left(\frac{de^*}{d\tau_1}\frac{d\tau_1}{d\lambda_i}\right) d\lambda_i}{\int_{\underline{\lambda}_i}^{\underline{\lambda}_i + \triangle\lambda_i} \left(\frac{de}{d\lambda_i}\right) d\lambda_i}. \tag{A.4}$$

Plugging equation (10) into (A.4) we get the non-marginal version of (12):

$$W = \frac{\int_{\underline{\lambda}_i}^{\underline{\lambda}_i + \triangle\lambda_i} \left(\frac{de}{d\lambda_i}\frac{-\frac{\partial e}{\partial \tau_1}}{\frac{\partial s_2}{\partial \tau_1} - \frac{\partial e}{\partial \tau_1}}\right) d\lambda_i}{\int_{\underline{\lambda}_i}^{\underline{\lambda}_i + \triangle\lambda_i} \left(\frac{de}{d\lambda_i}\right) d\lambda_i}$$

Since the slope of the allowances supply curve is assumed to be weakly positive and that of the cumulative allowance demand curve to be strictly negative this expression is always between zero and one. Hence, Proposition 3 extends to non-marginal changes in overlapping policies.

Moving to quantity-based flexibility mechanisms, we plug (A.3) into (A.4) and simplify:

$$W = \frac{\int_{\underline{\lambda}_i}^{\underline{\lambda}_i + \triangle\lambda_i} \left(\frac{de}{d\lambda_i}\frac{1 + \frac{\partial s_2}{\partial b}\frac{\frac{de_1}{d\lambda_i}}{\frac{de}{d\lambda_i}}}{1 + \frac{\partial s_2}{\partial b}\frac{\frac{\partial e_1}{\partial \tau_1}}{\frac{\partial e}{\partial \tau_1}}}\right) d\lambda_i}{\int_{\underline{\lambda}_i}^{\underline{\lambda}_i + \triangle\lambda_i} \left(\frac{de}{d\lambda_i}\right) d\lambda_i}$$

If the non-marginal policy change does not affect the demand for allowances in period 2,

i.e., if $de_1/d\lambda_i = de/d\lambda_i$ for all $\lambda_i \in [\underline{\lambda}_i, \overline{\lambda}_i]$, then the waterbed effect is weakly smaller than 100% confirming that part (i) of Proposition 4 extends to non-marginal changes in overlapping policies. If the non-marginal policy change does not affect the demand for allowances in period 1, i.e., if $de_1/d\lambda_i = 0$ for all $\lambda_i \in [\underline{\lambda}_i, \overline{\lambda}_i]$, then the waterbed effect is weakly larger than 100%, i.e. part (ii) of Proposition 4 applies to non-marginal policies, too. By using a continuity argument, part (iii) has to extend to non-marginal changes as well. If $(\partial s_2/\partial b)(de_1/d\lambda_i)/(de/d\lambda_i) < 0$ for all $\lambda_i \in [\underline{\lambda}_i, \overline{\lambda}_i]$, then the waterbed effect is negative. Hence, part (iv) also applies to non-marginal policies.

The results from the marginal analysis are quantitatively equivalent to the non-marginal analysis if allowance demand in all periods and allowance supply adjustments (either in prices or in banks) are linear. In this case $\partial e_1/\partial \tau_1$, $\partial e/\partial \tau_1$, $\partial s_2/\partial \tau_1$ and $\partial s_2/\partial b$ are all constants.

B.3. Proof of Lemma 2 for the EU ETS

Equation (23) follows directly from the parameters presented in Table 1 and the explanation given in the paragraph preceding the Lemma. See also Perino (2018).

The instantaneous waterbed effect $\hat{W}(t_a, t, t_{B=833})$ measures the waterbed effect of a reduction in allowance demand in a single year (t). Hence, the β measuring the temporal distribution of a policy's impact that appears in Equation (18) is equal to 1. Since we now consider a setting with more than two periods and the time of announcement of the overlapping policy is no longer fixed, we need to explicitly take this into account. In a market with perfect intertemporal arbitrage prices will respond to the announcement of a policy t_a. The denominator of Equation (18) capturing the price effect is therefore adjusted accordingly.

Appendix C: Additional numerical illustrations

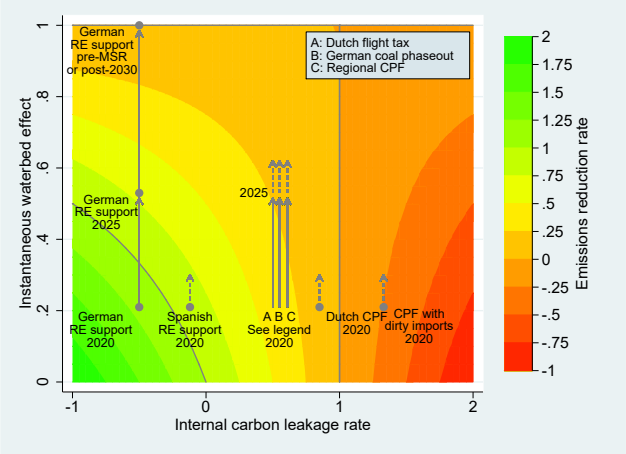
Sensitivity to $t_{B=833}$ and the Rosendahl effect

In the main text, we assume the MSR will stop taking in allowances in 2030 ($t_{B=833} = 2030$) (Figure A1, Panel (a)). In Figure A1, Panel (b), we investigate how the effective emissions reduction rate changes when we assume $t_{B=833} = 2048$ (following Gerlagh et al. (2021)). Panel (c) shows the performance of two key policies—renewable energy support and a coal phase-out in Germany—when we consider the instantaneous waterbed effect without holding carbon prices fixed and thus allowing for the Rosendahl effect (see Equation (24) in Lemma 2). We use Gerlagh et al. (2021)'s estimates of the Rosendahl effect but note that estimates in the literature differ and this is a highly active area of research.

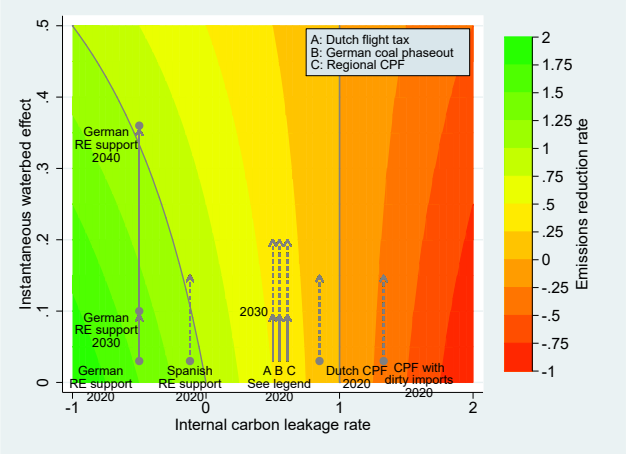
Panel (b) of Figure A1 shows that, compared to our original estimates in Panel (a), the instantaneous waterbed effect decreases substantially when $t_{B=833}$ lies further in the future. The waterbed effect can only go below 100% if the MSR takes in allowances; if allowances still flow into the MSR in the 2030s and 2040s, then $\hat{W}_t < 1$ for many more years over which policies operate. In Panel (a), $\hat{W}_{2030} = 1$; in Panel (b), $\hat{W}_{2030}$ falls by an order of magnitude.

Panel (c) compares $\hat{W}_t$ holding carbon prices fixed (grey arrows and dots) with endogenous allowance prices (black arrows and dots). A black dot should be interpreted as a policy announced in 2020 but expected to reduce the demand for emissions allowances in year $t \geq 2020$. The Rosendahl effect increases $\hat{W}_t$ substantially, especially for years close to $t_{B=833}$. Until the mid-2030s, the waterbed effect is still relatively limited (below 0.5) but in or after the year 2048, the waterbed effect is *larger* than 1. This is consistent with Proposition 4 and highlights the potential unintended consequences of announcing policies that reduce emissions demand far into the future.

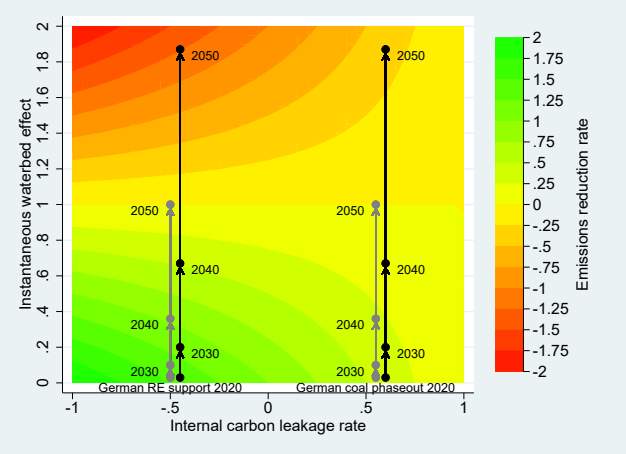
Figure A1: Leakage and waterbed effects in the EU ETS under varying assumptions



(a) $t_{B=833} = 2030$; no Rosendahl effect



(b) $t_{B=833} = 2048$; no Rosendahl effect



(c) $t_{B=833} = 2048$; with Rosendahl effect

Notes: Panel (a) presents Figure 4 excluding policies outside the EU ETS. Panel (b) plots the same policies assuming $t_{B=833} = 2048$ instead of $t_{B=833} = 2030$. Panel (c) adds the Rosendahl effect as estimated in Gerlagh et al. (2021), together with their estimate of $t_{B=833} = 2048$.