Tiered Climate Clubs:
Global Abatement Without Global Agreement*

Terrence Iverson†‡

May 30, 2024

Abstract

The paper proposes a novel policy structure with the potential to effectively reduce global carbon emissions without the need for broad multilateral cooperation. It extends Nordhaus's (2015) climate club by adding a second tier. Countries in the second tier must price carbon at a fixed fraction of the average carbon price within the first tier, or face tariffs. Tier 1 countries abate more since doing so induces matching abatement in the second tier. The unique stable coalition consists of the US, EU, and UK. It optimally sets the first tier carbon price near the global Social Cost of Carbon, achieving global abatement over half the efficient level.

KEYWORDS: international environmental agreement, climate club, trade penalties, bottom-up coalition

JEL CLASSIFICATIONS: Q54, F18, F53

*An earlier version of this paper was circulated under the title "Tiered Climate Clubs: A Bottom-Up Approach to Abate Global Emission" (Iverson, 2022).

†Department of Economics; Colorado State University. *Address:* 1771 Campus Delivery, Fort Collins CO, 80523, USA. terry.iverson@colostate.edu.

‡*Acknowledgments.* The author thanks Ed Barbier, Lint Barrage, Jesse Burkhardt, Jared Carbone, Chris Costello, Carolyn Fischer, Mikhail Golosov, Achim Hagan, Larry Karp, David Kelly, David Mushinski, Alessandro Peri, Robert Schmidt, Christian Traeger and seminar participants at CU Boulder, Colorado Schoool of Mines, ETH Zurich, SURED 2022, the Front Range Energy Economics Workshop (2022), and the UC Santa Barbara Workshop on Natural Resource Economics Theory (2023) for helpful conversations and feedback. The author is grateful for funding from the SoGES Resident Fellows Program at Colorado State University.

1 Introduction

> We should bring to the table the smallest possible number of countries needed to have the largest possible impact on solving a particular problem, whether trade or AIDS.
>
> *Moisés Naím*, Minilateralism (Naim, 2009)

The ideal climate policy, a carbon price equal to the marginal global damage, would induce consumers and firms to reduce carbon emissions whenever the global benefit exceeds the cost. Unfortunately, no actual government has an incentive to adopt such a policy. A government would have to care about the global impact of emissions, but self-interest inclines them to care only for the portion that affects their constituents. This is the free-rider problem. It goes away if countries are forced to move together, jointly accounting for the global impact of action. Each country then receives its portion of the average global payoff, and the typical country gains. But globally coordinated climate regulation appears increasingly beyond reach, raising an essential question: If countries must be forced to move together to realize the benefits of climate action, where is the coercion to come from?

Lacking a supranational authority to enforce compliance, a burgeoning body of research proposes leveraging trade tariffs to introduce enforcement mechanisms into international climate agreements (for instance, Barrett, 1997; Helm and Schmidt, 2015; Böhringer, Carbone and Rutherford, 2016). A compelling proposal is Nordhaus (2015)'s climate club. Members agree to adopt a target carbon price and to impose tariffs on imports from countries that fail to do the same. The arrangement can achieve substantial global abatement when participation is high (Nordhaus, 2015). However, it is designed to work as a top-down global agreement, and it runs into major problems when the club is small.

First, the limited trading leverage of a small climate club may trigger retaliatory tariffs (Hagen and Schneider, 2021). Second, since only club countries abate, a small club must resort to high marginal cost abatement while leaving cheap options in other countries untapped (Weyant, 1999). Third, a weaker version of the free-rider problem still applies—while a group of countries can jointly gain by internalizing climate damages within their collective borders, they still don't have reason to care about damages outside. Finally, by moving ahead of others, a small climate club puts its industry at a competitive disadvantage and induces carbon leakage, where domestically curbed emissions reemerge elsewhere (Felder and Rutherford, 1993). Taken together, these hurdles suggest that a small climate club has limited potential to stand in for an effective global climate agreement. Meanwhile, as three decades of failed international climate negotiations have shown,[1] achieving the kind of broad multilateral cooperation needed to operationalize a successful top-down climate club will be immensely challenging.

This paper proposes an alternative approach: a Tiered Climate Club (TCC). Rather than try to make a climate club work through high participation, a TCC changes the agreement terms to ensure effectiveness with low participation. It is more likely to be adopted since cooperation is easier with

[1] The main outcome of these negotiations, the 2015 Paris Agreement, relies on voluntary and non-binding pledges.

fewer participants. The approach focuses lead country efforts on using trade leverage to partly bind actions across nations. The countries that join (the "first tier") use tariff threats to coerce other nations (the "second tier") to price carbon at a fixed fraction of the average carbon price adopted in the first tier. This fraction, called the "match rate," is typically well below one, so the agreement defines more lenient terms for non-signatory countries than a traditional climate club. In addition, as in a traditional climate club, Tier 1 countries use tariff threats with each other to increase policy ambition in the first tier.

The design enables a TCC to overcome the noted pitfalls of a small climate club. First, it reduces the likelihood of trade retaliation because the first tier can strategically set the match rate just low enough to ensure compliance while avoiding trade retaliation. Second, by taking advantage of the cheapest abatement in the previously unregulated region, it replaces the highest marginal cost activities in the first tier with the cheapest remaining outside it (Section 3.1.2). Third, it reduces free-rider incentives—through a mechanism that resembles the match concept in charitable fundraising, Tier 1 countries factor in that their policy will be magnified through matching abatement in the second tier (Section 3.1.1). Finally, by reducing energy price differences between regions, competitiveness concerns and carbon leakage both decline (Section 3.1.3).

The model follows Nordhaus (2015) with a few adjustments.[2] Most importantly, the first tier chooses policy endogenously; in contrast, Nordhaus (2015) assumes the target carbon price is fixed before countries choose to join. The agreement is modelled as a two-stage game. In the first stage, countries decide whether to join or not. In the second stage, the first tier acts as a Stackelberg leader with the second tier. It sets the match rate as high as it can without triggering a trade war—given assumptions about how the threshold for triggering retaliation depends on the combined economic leverage of the Tier 1 countries. The first tier then sets policy endogenously.

The paper is divided into analytical and quantitative sections. The analytical section studies the second stage of the agreement game with the first tier fixed. It derives expressions for the optimal carbon price and global abatement as functions of the model parameters. The analysis shows how a TCC navigates the challenges associated with sub-global abatement, elucidates its capability to attain global efficiency, and considers the impact of Tier 1 cooperation.

The quantitative results are as follows:

1. The set of Tier 1 countries that satisfy the internal and external stability conditions for coalition formation (d'Aspremont et al., 1983; Barrett, 1994)[3] is unique. It consists of the US, the EU, and the UK.

2. If the US, EU, and UK adopt a conventional climate club,[4] choosing policy to maximize club surplus, their average carbon price increases by a factor of two and a half compared with the Nash

[2] See Section 2.4 for a detailed explanation of model differences.

[3] See Section 2.3.3 for a discussion of the stability concept.

[4] In this comparison, I treat the club as fixed, so the US, EU, and UK do not succeed in bringing in more members.

equilibrium (where all countries act independently).[5] Carbon prices outside the club remain at the Nash level. Global abatement increases from 8 percent of the efficient level under Nash to 10 percent of the efficient level. In contrast, with a TCC, a US-EU-UK first tier sets carbon prices four times higher than with a traditional climate club. The large increase in Tier 1 effort reflects amplified incentives due to matching abatement in a second tier that accounts for over three quarters of global carbon emissions. The second tier adopts a carbon price 46 percent as high, leading to global abatement 55 percent of the efficient level. Shifting from a conventional climate club to a TCC results in a more than five-fold increase in global abatement.

3. If the match rate were hypothetically increased to 67 percent, a US-EU-UK led TCC would achieve the efficient level of global abatement. The first tier prices carbon above the efficient level, over twice as high as the second tier. The first tier is willing to do this because it has lower carbon intensity, being more developed, which lowers the cost of a carbon price.

4. If the first tier is non-cooperative, perhaps because a participant is "too big to punish",[6] the TCC abates 25 percent of the efficient level—less than half the cooperative outcome, but more than twice that of a conventional climate club (where member countries do cooperate).

With a TCC, a small group of economically powerful nations drives the coercion needed to coordinate carbon abatement across countries. The approach risks being dismissed as a form of "neocolonial bullying," with wealthier countries exerting pressure on poorer ones to address a problem largely of the former's making. Although this perspective acknowledges a valid concern, I contend that a different viewpoint is more fitting. Once we recognize the necessity for some form of coercion to overcome the incentive issues at the root of the climate problem, then it is just a question of where this coercion should come from. Since the most powerful economies are in the best position to provide it, one could argue that they bear an ethical responsibility to play this role. Meanwhile, the Tier 2 countries stand to gain the most (see Section 5.1). Not only do they benefit from a large reduction in climate damages while incurring lower costs, but they are also shielded from competitiveness concerns, as the arrangement ensures their position at the bottom of the global distribution of carbon prices.

1.1 Related literature

The paper contributes to the broad game theoretic literature on International Environmental Agreements (IEAs) (for comprehensive reviews, see Barrett, 2003; Marrouch, Chaudhuri et al., 2016). A core strand of this literature considers pure IEA games. Countries negotiate an environmental treaty

[5] Under Nash, the US, EU, and UK price carbon at about 10 percent of the globally efficient level. When they use a conventional climate club to achieve the cooperative outcome, they price carbon at about 25 percent of the globally efficient level.

[6] Böhringer and Rutherford (2017) use a global trade model to contemplate the potential for the EU and China to use the threat of trade sanctions to force the US back into the Paris Agreement after President Trump committed to withdraw. They concluded that the US may be "too big to punish".

(like a climate agreement) without linking the problem with other issues. Contributions generate non-excludable positive externalities. Because the benefits are non-excludable, countries have incentive to free ride. Applied to climate change, the main conclusion of this literature is that stable bottom-up coalitions are unlikely to achieve anything close to the efficient level of global abatement. If a coalition aims to extract the collective gains from cooperation, the number of participants is small—typically three at most (Carraro and Siniscalco 1993, and Barrett 1994, Bosetti et al. 2012). Large coalitions only form when the gains from cooperation are small (Barrett 1994, Kolstad 2007, Pavlova and de Zeeuw 2013). This pessimistic conclusion has been called the "small coalition paradox" (Nordhaus 2015).[7]

While the current paper engages the core challenge identified in the "small coalition paradox," it contributes more directly to the accompanying literature that aims to surmount the noted challenges by linking IEA negotiation with other issues of value to nations. The basic idea is to combine the non-excludable-benefits-producing IEA game with a separate club-like game that generates excludable benefits (Folmer, Mouche and Ragland, 1993). While several versions of this linkage have been suggested—including side payments (Petrakis and Xepapadeas, 1996), reputation (Hoel and Schneider, 1997), and R&D spillovers (Carraro, Siniscalco et al., 1995)—the most promising and widely explored option is trade.

Two main strands in the trade-based climate policy literature are noteworthy. The first encompasses papers that use the threat of trade-based punishments to increase the size of a stable coalition (Barrett, 1997; Lessmann, Marschinski and Edenhofer, 2009; Helm and Schmidt, 2015; Nordhaus, 2015; Böhringer, Carbone and Rutherford, 2016; Al Khourdajie and Finus, 2020; Farrokhi and Lashkaripour, 2020; Hagen and Schneider, 2021; Nordhaus, 2021). This literature primarily focuses on designing trade-based policies for countering free-rider incentives to support a top-down agreement with broad participation. The second strand of literature involves papers that utilize Border Carbon Adjustments (BCAs) to enhance outcomes in bottom-up settings. BCAs entail carbon tariffs that effectively tax the carbon content of imports, thus leveling the playing field for domestic producers. The primary focus in this area is the potential for BCAs to reduce carbon leakage (Felder and Rutherford, 1993; Fischer and Fox, 2012; Böhringer, Balistreri and Rutherford, 2012; Branger and Quirion, 2014; Böhringer et al., 2022). However, the literature also explores the potential for BCAs to induce non-coalition countries to price carbon preemptively to avoid tariffs (Helm and Schmidt, 2015; Böhringer, Carbone and Rutherford, 2016; Al Khourdajie and Finus, 2020). This literature finds that BCAs cannot support a high level of aggregate abatement unless they result in broad participation in a climate agreement (Farrokhi and Lashkaripour, 2020).

The current paper adopts an approach that combines elements from both strands. Like the second one, a TCC is intended to work in a setting in which the lead coalition is potentially small. But

[7] Two settings in which pure IEA games admit more hopeful outcomes are repeated games (e.g., Barrett 1999) and farsighted equilibria (e.g., de Zeeuw 2008). However, both results hinge on strong assumptions that are unlikely to apply in realistic settings (e.g., Carraro and Siniscalco, 1998).

because Tier 2 countries participate to some extent, even if they are only passive participants in the agreement, the agreement moves beyond the case of purely unilateral abatement, and achieves some of the advantages of a top-down agreement. To my knowledge, this is the first paper to demonstrate that a significant portion of the globally efficient level of aggregate abatement can be achieved through the initiative of a relative small coalition of countries that prioritize national-self-interest.

This paper is also related to a recent proposal by the IMF for implementing an international carbon price floor for major carbon emitters (Parry, Black and Roaf, 2021). The IMF proposal advocates for a negotiation process among large emitters wherein a moderate-sized group of countries aims to agree upon a differentiated minimum carbon price floor based on development levels. In one version, middle-income countries adopt a carbon price that is two-thirds that of high-income countries, while low-income countries adopt a carbon price one-third as high. Parry, Black and Roaf (2021) do not model country payoffs or optimal decision-making, but they do simulate the effects of the implied distribution of carbon prices on global emissions. The present study bolsters the case for a global carbon price floor by demonstrating that this approach would significantly reduce the global abatement cost inefficiency linked to sub-global climate policy, and by proposing an agreement structure under which differentiated carbon price floors are the endogenous outcome of an explicit treaty formation game in which countries pursue self-interest.

Finally, although several papers have emphasized the importance of having the largest economies—such as the EU, US, and China—take the lead in developing a global climate agreement (e.g., Gwatipedza and Barbier, 2014; Falkner, 2016; Tagliapietra and Wolff, 2021), to my knowledge, this is the first paper to show that leadership by major economies follows strictly from self-interest in a stable coalition. The analysis also finds—in the context of implementing a TCC—that while the US and EU strictly benefit from leading the process, China does not.

2 Model and proposed agreement

The model follows Nordhaus (2015) with a few important changes. I explain and justify the changes at the end of the section. First, I present the model, agreement structure, and optimal policy problem.

2.1 Model

Setup and notation The model is static with N heterogeneous "countries" (or regions). Each country acts like a unitary decision maker (in matters of climate policy). There is a single composite commodity traded between countries. CO2 emissions arise as a negative externality from production, and countries can reduce emissions by adopting costly abatement.

Each country i is characterized by exogenous output, Q_i, an exogenous flow of baseline CO2 emissions, E_i, and constant marginal domestic damages from CO2 emissions, γ_i.[8] Similarly, Q is

[8]Constant marginal damages is a plausible assumption in a static model since the flow-stock ratio for CO2 is low within a given policy period. The assumption is also consistent with analytically tractable IAMs, such as Golosov et al.

global GDP, E global CO2 emissions, and γ global marginal damages from CO2 emissions (the Social Cost of Carbon, or SCC). In addition, baseline trade flows between countries are defined by a bilateral trade matrix $\mathbf{X}$.

To indicate variables as a fraction of the global total, I reserve the letter ϕ. For example, ϕ_i^Q is country i's GDP as a fraction of global GDP. I also sometimes replace the country subscript with a set of countries, Ω. Thus, $\phi_\Omega^Q \equiv \sum_{k\in\Omega} \phi_k^Q$ is the fraction of global GDP comprised by the countries in Ω. Finally, to indicate how big a country is as a fraction of the set total, I add a "hat". For instance, if $i \in \Omega$ then $\hat{\phi}_i^Q \equiv \frac{\phi_i^Q}{\phi_\Omega^Q}$ is i's GDP as a fraction of the total GDP of Ω.

Policy cost The regional abatement cost functions from Nordhaus (2015) are given by

$$\Psi_i(\mu_i) = Q_i \theta_{1,i} (\mu_i)^{\theta_2}, \tag{1}$$

where μ_i is the abatement rate in i, while $\theta_{1,i}$ and θ_2 are parameters. Since countries set carbon prices, not abatement rates, it is useful to express the cost functions in terms of the carbon price, τ_i. By equating marginal abatement costs (in emission units) with the carbon price, we get the following relationship between the abatement rate and the carbon price:[9]

$$\mu_i = a_i (\tau_i)^b \equiv G_i(\tau_i). \tag{2}$$

where

$$a_i = \left[\frac{\sigma_i}{\theta_1 \theta_2}\right]^{\frac{1}{\theta_2 - 1}} \tag{3}$$

and

$$b = \frac{1}{\theta_2 - 1}. \tag{4}$$

$\sigma_i = \frac{E_i}{Q_i}$ is the emissions intensity of output in i. Combining (1) and (2) gives abatement costs in terms of the carbon price:

$$C_i(\tau_i) = Q_i \theta_1 a_i^{\theta_2} (\tau_i)^{b\theta_2}. \tag{5}$$

Climate benefits The climate benefits from policy depend on the reduction in global emissions, which depends on policy in all countries. In writing the policy problem, it is sometimes useful to separate the carbon price in i from that in the rest of the world. For this purpose, I use the notation $\tau_{-i} \equiv \{\tau_j\}_{j\neq i}$. The global abatement rate can then be written

$$\mu(\tau_i, \tau_{-i}) = \sum_j \phi_j^E G_j(\tau_j), \tag{6}$$

where $G_j(\cdot)$, defined in (2), gives the domestic abatement rate in j as a function of its carbon price. The policy benefit to i is the domestic value of global abatement:

$$B_i(\tau_i, \tau_{-i}) = \gamma_i E \mu(\tau_i, \tau_{-i}). \tag{7}$$

(2014) and Traeger (2023).

[9] Details are in Appendix A.1.

Conditional trade incentives The policies in the paper use conditional trade threats that consist of uniform tariffs on a country's exports to the punishing region if it fails to oblige a minimum carbon price target.

To quantify the economic impact of tariffs, I follow Nordhaus (2015) who uses simulations of a global general equilibrium trade war model (Ossa, 2014) to estimate reduced form tariff benefit/cost functions for each bilateral country/region pair. The tariff benefit function is linear-quadratic with a linear term that captures the terms-of-trade effect and a quadratic term that captures the efficiency loss. The resulting net income loss in country i when j applies tariff t_{ji} on imports from i is

$$P_{ji}(t_{ji}) = \mathbf{X}_{i,j}(\alpha_j t_{ji} + \beta_j t_{ji}^2), \tag{8}$$

where $\mathbf{X}_{i,j}$ (the i,j-th element of the bilateral trade matrix $\mathbf{X}$) is the flow of imports into j from i, and α_j and β_j are the estimated parameters for a uniform tariff in country j on imports from any country i.[10]

From the perspective of country i, what matters is the total penalty (net income loss) caused by the tariffs applied by all countries in the punishing region. Suppose each country $j \in \Omega$ imposes uniform tariff t_{ji} on imports from i. Then the total penalty, normalized as a fraction of i's GDP, is

$$\omega_{i,\Omega} = \frac{1}{Q_i} \sum_{j \in \Omega} P_{ji}(t_{ji}). \tag{9}$$

For convenience, I often suppress the dependence of the total penalty on the underlying tariffs. In that case, the conditional trade incentive is a pair $(\omega,\underline{\tau})$, where ωQ_i is the (directly specified) trade penalty imposed on i if it does not adhere to the minimum carbon price $\underline{\tau}$. In the quantitative section, I use actual bilateral trade flows to model the relationship between tariffs from the penalizing region and total net income losses to the exporting country.

National payoffs Given the conditional trade incentive $(\omega,\underline{\tau})$, payoffs for country i are just the domestic climate benefit less the domestic policy cost less the trade-based penalty, if triggered:

$$\Pi_i(\tau_i, \tau_{-i}; \omega, \underline{\tau}) = B_i(\tau_i, \tau_{-i}) - C_i(\tau_i) - \omega Q_i \mathbf{1}_{\tau_i < \underline{\tau}}, \tag{10}$$

where the indicator function $\mathbf{1}_{\tau_i < \underline{\tau}}$ is one if $\tau_i < \underline{\tau}$ and zero otherwise.

2.2 A Tiered Climate Club

A Tiered Climate Club (TCC) divides countries into two tiers. Countries that choose to join constitute the first tier—indicated by the set Ω. All other countries are in the second tier. Tier 1 countries establish conditional trade incentives for each tier, and they use uniform import tariffs to mete out punishments.

[10] Nordhaus (2015) initially allows for the possibility that the α and β parameters vary for all country pairs, but he later assumes that the parameters for each importing country are independent of the country from which the imports originate. Since I employ his calibration, I adopt this assumption from the start.

Tier 1 incentives Tier 1 countries use conditional trade incentives with each other to increase cooperation amongst themselves. These incentives have the effect of increasing the ambition of Tier 1 policy, and they operate much like the conditional trade incentives used in a conventional climate club. The Tier 1 incentives consist of a pair $(\omega_{1,i},\underline{\tau}_i)$ for each $i \in \Omega$. It requires country i to implement a domestic carbon price at or above $\underline{\tau}_i$ or face uniform tariffs on exports to the other Tier 1 countries with total economic (as a fraction of i's GDP) equal to $\omega_{1,i}$, as determined through Eq. 9. When Tier 1 countries set the Tier 1 incentives, a range of behavioral assumptions are possible, as explored in the next subsection.

Tier 2 incentives The first tier also sets conditional trade incentives for countries in the second tier. The terms depend on the match rate, which must be chosen at the same time that the Tier 2 incentives are defined. I explain below how this choice is modeled. The minimum carbon price required of each Tier 2 country is $\alpha\tau^{AVG}$, where the match rate $\alpha \in [0,1]$ is fixed and the (size-weighted) average Tier 1 carbon price is given by

$$\tau^{AVG} \equiv \sum_{i\in\Omega} \hat{\phi}_i \tau_i, \tag{11}$$

where each τ_i is the carbon price actually implemented (and observed) in Tier 1 country i. The definition of the Tier 2 incentive assumes that decisions are made sequentially. Tier 1 countries choose their own carbon prices first and these are observed. Next, Tier 2 countries adopt carbon prices.

The match rate Unconstrained, the first tier would choose to set the match rate as high as possible, since doing so pushes more abatement onto the second tier. However, since a higher match rate imposes higher costs onto Tier 2 countries, the first tier is constrained by the need to get Tier 2 countries to comply without triggering trade retaliation.

To model this constraint, I assume there is a reduced form relationship between the Tier 1 set Ω and the highest match rate that the first tier can get away with before triggering a trade war. Since the first tier will optimally choose this highest feasible rate, I simply refer to it as the match rate. In modelling "the match rate," I aim to capture the most important determinant of general trading leverage, which is the size of the domestic market that another country may hope to get access to.[11] I assume that the (highest achievable) match rate is proportional to the combined GDP of the first tier:

$$\alpha(\Omega) = b\phi_{\Omega}^{Q}, \tag{12}$$

where $b > 0$ captures how responsive the match rate is to coalition GDP.

Under this assumption, a group of countries with combined GDP that adds up to the GDP of a single large country has the same bargaining power as the large country. In reality, the large country would likely have more bargaining power, since the coalition of small countries has to overcome a collective action problem to fight a trade war in a unified way. If so, then the assumption understates

[11] In reality, other factors would matter, including the size distribution of Tier 1 countries, goodwill, and global buy-in to the idea of using trade threats to bolster climate policy.

the potential role of large countries in a TCC. Since the quantitative analysis—which maintains Eq. 12—finds that the biggest economies are most likely to join the first tier anyway, the assumption can be viewed as conservative.

2.3 The agreement game

The game is modeled in two stages (as in, e.g., Barrett, 1994). In the first stage, countries decide on their membership status—whether or not to join the first tier. In the second stage, the first tier acts as a Stackelberg leader vis-à-vis the second tier. The game is solved with backward induction.

2.3.1 Second stage

Once formed, the first tier—denoted Ω—optimally sets the match rate at the highest rate it can without triggering a trade war. By assumption, this is the match rate given in Eq. 12, written $\alpha(\Omega)$. The first tier then defines conditional trade incentives for each tier.

For these trade threats to be credible, the Tier 1 countries must be willing to follow through with them when called for. The trade literature generally finds that countries imposing small to moderate tariffs benefit economically as long as the tariffs don't incite retaliation (e.g., Ossa, 2014). Indeed, this fact is often used to defend the use of trade tariffs as a punishing devise in climate agreements (e.g., Nordhaus, 2015). I thus assume that Tier 1 countries always follow through with tariff threats as long as the agreement terms are lenient enough to avoid a trade war.

The Tier 2 incentive requires the Tier 2 countries to price carbon at or above $\alpha(\Omega)\tau^{AVG}$, or face the Tier 2 penalty ω_2. Since a higher match rate (α) requires a higher Tier 2 penalty (ω_2) to make it in j's best interest to follow through with the carbon price target, the assumption that $\alpha(\Omega)$ is feasible (without triggering retaliation) implies that setting ω_2 high enough to ensure incentive compatibility is also feasible.[12] Hence, to model the Tier 2 incentive, I directly assume that, given $\alpha(\Omega)$, ω_2 is always set high enough to ensure compliance from all Tier 2 countries. If the implied ω_2 is implausibly high, one would conclude that the originally assumed match rate was not politically feasible. In Section 5.3, I back out the uniform import tariffs needed to support match rates of different levels in the equilibrium of the quantitative model.

The first tier also chooses the minimum carbon price targets for the Tier 1 countries. When thinking about this endogenous choice of Tier 1 policy, there are three plausible scenarios.

Cooperative policy In the first scenario, Tier 1 countries act cooperatively, choosing policy to maximize Tier 1 surplus. This policy is a natural upper bound on how much abatement a given set of Tier 1 countries could be expected to achieve since it entails the Tier 1 countries jointly taking into

[12] The match rate $\alpha(\Omega)$ and the Tier 2 penalty ω_2 are linked through the participation constraint for Tier 2 country j:

$$B_j(\alpha\tau^{AVG}, \tau_{-j}) - C_j(\alpha\tau^{AVG}) \geq B_j(\gamma_j, \tau_{-j}) - C_j(\gamma_j) - \omega_2 Q_j \mathbf{1}_{\gamma_j < \alpha\tau^{AVG}},$$

which takes into account that the unilateral best response for j, ignoring the penalty, is to set the domestic carbon price at the domestic marginal damage from emissions, γ_j.

account the negative spillovers they impose on each other. The assumption is common in the IEA literature (see Marrouch, Chaudhuri et al. (2016) for a review).

The cooperative policy problem is easy to state if we assume the Tier 1 and Tier 2 penalties are always set high enough to ensure compliance. We argued above that $\alpha(\Omega)$ being feasible means the Tier 2 penalty can be set high enough to ensure compliance for Tier 2 countries. Similarly, if the match rate model presumes cooperative behavior by the first tier (as I do in the quantitative model) then $\alpha(\Omega)$ being feasible also implies that the Tier 2 penalty can be set high enough to ensure compliance by the Tier 1 countries in implementing the cooperative policy. Under these assumptions, penalties are never imposed in equilibrium and global abatement is a deterministic function of the Tier 1 carbon price targets and the match rate.

The cooperative policy solves

$$\max_{\{\hat{\tau}_i \geq 0\}_{i\in\Omega}} \sum_{i\in\Omega} \left[\gamma_i E \mu(\{\hat{\tau}_i\}_{i\in\Omega}; \alpha(\Omega)) - C_i(\tau_i)\right] \tag{\boldsymbol{P1}}$$

where $C_i(\cdot)$ is defined in Eq. 5 and global abatement is given by

$$\mu(\{\hat{\tau}_i\}_{i\in\Omega}; \alpha) = \sum_{i\in\Omega} \phi_i^E G_i(\hat{\tau}_i) + \sum_{j\notin\Omega} \phi_j^E G_j\left[\max(\alpha\tau^{AVG}, \gamma_j)\right],$$

with $G_i(\cdot)$ defined in Eq. 2 and

$$\tau^{AVG} = \sum_{i\in\Omega} \hat{\phi}_i \hat{\tau}_i.$$

The maximum operators within the $G_j(\cdot)$ expressions account for the possibility that optimal policy for Tier 2 country j could exceed the carbon price target if the match rate is low enough. In that case, j optimally sets its domestic carbon price at the domestic marginal climate damage, γ_j.

The solution to ***P1*** entails the same (harmonized) carbon price applied to all Tier 1 countries (Proposition 1). I denote this by $\tau^C(\Omega)$. In Proposition 1, I back out the Tier 1 penalty needed to support the cooperative outcome as a function of the model parameters. In addition, in Section 5.3, I back out the Tier 1 and Tier 2 tariff threats needed to support the cooperative outcome in the quantitative model given observed trade flows between nations.

Non-cooperative policy In the second case, Tier 1 countries interact non-cooperatively. While the cooperative case is a natural upper bound on Tier 1 policy, the non-cooperative case is the natural lower bound. It coincides with a Tier 1 penalty of zero. Each Tier 1 country i takes the policy targets of the other Tier 1 countries as given, then chooses its own carbon price τ_i to solve

$$\max_{\tau_i \geq 0} \left[\gamma_i E \mu(\tau_i, \{\hat{\tau}_j\}_{j\in\Omega, j\neq i}; \alpha(\Omega)) - C_i(\tau_i)\right], \text{ for each } i \in \Omega, \tag{\boldsymbol{P2}}$$

where the global abatement rate is given by

$$\begin{aligned}\mu(\tau_i, \{\hat{\tau}_j\}_{j\in\Omega, j\neq i}; \alpha) = \phi_i^E G_i(\tau_i) &+ \sum_{j\in\Omega, j\neq i} \phi_j^E G_j(\hat{\tau}_j) \\ &+ \sum_{k\notin\Omega} \phi_k^E G_k\Big[\max\Big(\alpha\underbrace{[\hat{\phi}_i^E \tau_i + \sum_{j\in\Omega, j\neq i} \hat{\phi}_j^E \hat{\tau}_j]}_{\tau^{AVG}}, \gamma_k\Big)\Big],\end{aligned}$$

with $G_i(\cdot)$ defined in Eq. 2. As in the cooperative case, the maximum operator inside the $G_k(\cdot)$ expressions in the second line capture the possibility that each Tier 2 country k might optimally price carbon above the target required by the Tier 2 incentive. ***P2*** defines a system of n equations in n Tier 1 carbon prices, where n is the number of countries in Ω.

The expression for the global abatement rate highlights the two channels through which i's choice of τ_i impacts global abatement. First, as reflected in the first line, i's abatement has a *direct effect* on aggregate abatement: the greater i's emissions, the bigger the impact of its own abatement on global abatement. Second, as reflected in the second line, by impacting the average carbon price, i's choice of τ_i has an *indirect effect* through the abatement undertaken in Tier 2 countries. The latter effect depends on the match rate α and on i's size relative to the first tier. In the full Nash Equilibrium, each country simply internalizes the direct effect. With a TCC, the indirect effect increases abatement incentives above the Nash level, even when Tier 1 countries interact non-cooperatively.

Intermediate cases In addition to the extreme cases above, the first tier could modify the Tier 1 incentives to support any Tier 1 carbon price between the cooperative and non-cooperative bounds. To capture this possibility, we can reframe the Tier 1 policy problem by assuming that it takes the Tier 1 penalty ω_1 as given—e.g., as politically constrained—then chooses carbon price targets to maximize the Tier 1 surplus subject to the constraint that no country has an incentive to deviate from the agreement. Under this framing, the non-cooperative problem falls out as the special case in which $\omega_1 = 0$, and the cooperative case coincides with the situation in which the Tier 1 penalty is just high enough to support the cooperative outcome. In the paper, I focus mainly on the cooperative and non-cooperative cases, so I have relegated this more general statement of the optimal policy problem to Appendix A.2.

2.3.2 First stage

In the first stage, countries choose whether or not to join the first tier. In doing so, they anticipate the impact of their own membership decision on the match rate—through Eq. 12—and on the incentive for Tier 1 countries to price carbon—as explained in Section 5.1.

2.3.3 Coalition stability

To close the model, the paper employs the notion of coalition stability that is most common in the noncooperative game theory literature on International Environmental Agreements (Carraro and Siniscalco, 1993; Barrett, 1994; Carraro, 2003). Building on the study of cartel formation in d'Aspremont et al. (1983), this literature assumes that a coalition is stable if it is both internally stable and externally stable, meaning that no countries can benefit from a unilateral deviation from their decision to join or not in the first stage.

When the first tier consists of the countries in Ω, let $\tau_j^*(\Omega)$ be the endogenous (incentive compatible) carbon price target for each country $j \in \Omega$. The target could be cooperative, non-cooperative, or

otherwise. Moreover, define

$$\tau^{AVG}(\Omega) \equiv \sum_{j \in \Omega} \hat{\phi}_j \tau_j^*(\Omega).$$

Then the payoff for country i if it is inside the coalition Ω is

$$\Pi_i^{IN}(\Omega) \equiv \gamma_i E\mu(\{\tau_j^*(\Omega)\}_{j\in\Omega}, \{\alpha\tau^{AVG}(\Omega)\}_{k\notin\Omega},) - C_i(\tau_i^*(\Omega)). \tag{13}$$

The payoff for country i if it is outside coalition Ω is

$$\Pi_i^{OUT}(\Omega) \equiv \gamma_i E\mu(\{\tau_k^*(\Omega)\}_{k\in\Omega}, \{\alpha\tau^{AVG}(\Omega)\}_{k\notin\Omega},) - C_i(\alpha\tau^{AVG}(\Omega)). \tag{14}$$

Coalition Ω is stable if it satisfies *internal stability* and *external stability*. It satisfies internal stability if every $i \in \Omega$ is better off staying in the coalition than it would be if it unilaterally left:

$$\Pi_i^{IN}(\Omega) \geq \Pi_i^{OUT}(\Omega \setminus i), \tag{15}$$

where the symbol $\setminus$ denotes set subtraction. The coalition satisfies external stability if every $j \notin \Omega$ is better off staying out than it would be if it unilaterally joined:

$$\Pi_j^{OUT}(\Omega) \geq \Pi_j^{IN}(\Omega \cup j). \tag{16}$$

2.4 Model changes relative to Nordhaus (2015)

The model differs from Nordhaus (2015) in two key ways. These changes are essential to assess the TCC proposal in a satisfactory way. First, I explicitly model the intentional policy choice of a given set of Tier 1 countries. Second, I modify the equilibrium concept to enable evaluation of equilibrium uniqueness. I also adjust the calibration, as discussed in Section 4.1.

Endogenous policy In Nordhaus (2015), both the carbon price target and the tariff rate are exogenously fixed when countries decide whether to join a club. The approach allows him to make quantitative statements about about how big tariffs need to be to support different carbon price targets using a climate club structure. Nevertheless, the assumption obscures a crucial dimension of the policy problem, which is that a smaller set of lead countries would internalize a smaller portion of global climate damages when setting policy. When considering a TCC, there is also a crucial linkage between the set of countries that join the first tier and the match rate that can be imposed without triggering a trade war. A country's decision to join the first tier also impacts abatement incentives for the remaining tier-one countries in nontrivial ways, as explained in Section 5.1.

Coalition formation Another change is the stability concept. As discussed, I define a stable coalition relative to unilateral deviations. In contrast, Nordhaus (2015) studies Coalition Nash Equilibria, where a candidate stable coalition is compared with all alternative sub-coalitions. There are two reasons for the change.

The first reason is numerical. Nordhaus (2015) notes that solving for Coalition Nash Equilibria is NP-hard and probably computationally infeasible for his model. As a result, he adopts two key

shortcuts that reduce the computational burden of identifying stable clubs. First, by assuming that coalition policy is exogenous, he effectively sets up the problem in a way that avoids the need for optimization. If I were to apply the same equilibrium concept to the TCC model, it would be vastly more computationally burdensome since my model contains a considerable optimization step in order to calculate the optimal policy response for each candidate club. In addition, even without an optimization step, Nordhaus (2015) uses an approximate stochastic search algorithm to identify stable coalitions. While Nordhaus (2015) asserts that the algorithm exhibits good convergence properties, he doesn't consider the possibility of multiple stable coalitions. Nevertheless, these models are known to have multiple equilibria (e.g., Hagen and Schneider, 2021). A major advantage of the more standard equilibrium concept from the IEA literature is that I can exhaustively search all possible permutations of candidate clubs, which makes it possible to conclusively study the issue of uniqueness.

The second justification for changing the stability concept is conservatism. The unilateral deviation concept is the basis for the broadly pessimistic findings in the IEA literature that if a climate agreement is negotiated in isolation (without issue linkage) then stable climate agreements do not achieve a signficnat portion of the globally efficient level of abatmeent: either the stable coalition is small or its policy ambition is low (Carraro and Siniscalco, 1993; Barrett, 1994; Rubio and Ulph, 2007; Marrouch, Chaudhuri et al., 2016). Typically in this literature when people explore equilibrium concepts with more foresight, it is with the intention of achieving more optimistic results (i.e., more global abatement). The basic intuition is that if the deviation of one country can induce the coalition to fall apart, then the mere threat of losing all cooperative benefits can deter deviations in the first place and support cooperation. Given this intuition, I view the use of the more conventional equilibrium concept as a conservative choice.

3 Qualitative results

In this section, I study the second stage of the agreement formation game. The first tier (and thus by implication the match rate) are fixed. I use analytical arguments to show how a TCC mitigates the difficulties that undermine sub-global carbon abatement.[13] I also show how policy changes if Tier 1 countries interact non-cooperatively.

3.1 Mitigating the hurdles of a small climate club

As with any attempt to abate carbon emissions at a sub-global scale, a small (conventional) climate club is hampered by free-rider incentives, inefficiency, and carbon leakage. This section shows how a TCC addresses each of these.

[13] I use the term "sub-global abatement" to describe abatement by any coalition smaller than the grand coalition.

3.1.1 The free-rider problem

To examine free-rider incentives, I consider the problem of a fixed coalition of countries, Ω, that behaves cooperatively. The coalition's objective is the same as in the cooperative Stage Two policy problem stated in ***P1***. To simplify the exposition, I redefine terms to express the objective as follows:[14]

$$\text{Coalition Surplus} = \underbrace{\gamma_\Omega \times a_G(a_\Omega)}_{\substack{\text{Coalition} \\ \text{Climate Benefit}}} - \underbrace{C_\Omega(a_\Omega)}_{\substack{\text{Coalition} \\ \text{Abatement Cost}}} ,$$

where γ_Ω is the marginal climate damage inside the coalition, $a_G(a_\Omega)$ is global abatement in emission units written as a function of coalition abatement in emissions units, and $C_\Omega(a_\Omega)$ is the coalition abatement function cost assuming efficient policy. Here, I assume that the coalition directly chooses coalition abatement, a_Ω, but it accomplishes this efficiently using a harmonized carbon price within its borders.

The first-order condition for the coalition problem gives

$$\underbrace{\gamma_\Omega \times \frac{\mathrm{d}a_G}{\mathrm{d}a_\Omega}}_{\text{Marginal benefit}} = \underbrace{C'(a_\Omega)}_{\text{Marginal cost}} . \tag{17}$$

In the standard case—including a conventional climate club[15]—the coalition only controls coalition abatement, so $\frac{\mathrm{d}a_G}{\mathrm{d}a_\Omega} = 1$. Eq. 17 then reduces to the statement that marginal damages, γ_Ω, equal marginal cost. Since marginal damage for a coalition of modest size is well below the global marginal damage—$\gamma_\Omega << \gamma$—abatement is inefficiently low. This is a weaker version of the standard free-rider problem when all countries act unilaterally.

Next, suppose the coalition adopts a TCC, requiring the second tier to match the first tier carbon price at rate $\alpha > 0$. Global abatement induced by the policy is greater than coalition abatement, so $\frac{\mathrm{d}a_G}{\mathrm{d}a_C} > 1$. I refer to this derivative as the *amplification factor*. It increases the incentive for Tier 1 countries to price carbon because Tier 2 abatement is effectively free from the perspective of the first tier. The solution to the cooperative Stage 2 problem defined in ***P1*** shows the role of the amplification factor in determining abatement incentives under a TCC. I state the solution as a proposition.

Proposition 1. *Let the first tier consist of countries in the set Ω. These countries differ in terms of GDP, ϕ_i^Q, CO2 emissions, ϕ_i^E, and climate damages, γ_i, but they have the same emissions intensity, σ_Ω,[16] and the same abatement cost scale parameter, $\theta_{1,\Omega}$. The second tier (labeled "R" for Rest of World) has a common emissions intensity σ_R and a common abatement cost scale parameter $\theta_{1,R}$.*

[14] Specifically, $\gamma_\Omega \equiv \sum_{i\in\Omega} \gamma_i$ is the coalition marginal damage, $a_G(\tau_\Omega) \equiv E\mu(\{\hat{\tau}_i\}_{i\in\Omega}; \alpha(\Omega))$ is global abatement in emission units, and $C_\Omega(a_\Omega(\tau_\Omega))$ is the coalition abatement cost with efficient policy.

[15] The club only controls abatement in member countries, and it uses mutual trade threats to increase club abatement to the cooperative level.

[16] Emissions intensity is CO2 emissions over GDP.

Abatement costs are quadratic ($\theta_2 = 2$),[17] *and Tier 2 countries only abate with external coercion.*[18] *Then the cooperative policy—defined in* ***P1****—imposes the following harmonized carbon price on all* $i \in \Omega$*:*

$$\tau^C(\alpha) = \gamma_\Omega \times \frac{\mathrm{d}a_G}{\mathrm{d}a_C}, \tag{18}$$

where $\gamma_\Omega = \sum_{i\in\Omega} \gamma_i$ *is the coalition marginal climate damage and*

$$\frac{\mathrm{d}a_G}{\mathrm{d}a_C} = 1 + \alpha \frac{1-\phi_C^E}{\phi_C^E} \frac{\sigma_R}{\sigma_C} \tag{19}$$

is the amplification factor defined in the text.

Proof. Appendix A.3. □

The amplification factor (Eq. 19) increases in the match rate, in the ratio $\frac{1-\phi_\Omega^E}{\phi_\Omega^E}$, and in the ratio $\frac{\sigma_R}{\sigma_\Omega}$. A higher match rate raises the matching requirement for Tier 2 countries, directly increasing the global abatement induced by Tier 1 policy. The ratio $\frac{1-\phi_\Omega^E}{\phi_\Omega^E}$ captures the relative size of the second tier compared with the first (in terms of carbon emissions); the bigger this ratio, the more matching abatement occurs in the second tier for each unit of abatement in the first. Finally, $\frac{\sigma_R}{\sigma_\Omega}$ captures the relative emissions intensity of the second tier compared with the first. Higher emissions intensity in a region implies that a given carbon price generates more abatement, all else equal (Appendix A.3). Hence, a higher ratio increases the amount of Tier 2 abatement induced by Tier 1 policy.

The matching mechanism in a TCC operates like a match-funding program in charitable giving. In the charity context, a large donor matches new donations, increasing an individual's incentive to give. With a TCC, matching Tier 2 abatement amplifies the climate benefit generated by Tier 1 policy, increasing the first tier's incentive to abate.

To see the potential to offset free-rider incentives, I express the cooperative carbon price as a fraction of the globally efficient carbon price as follows:[19]

$$\frac{\tau^C(\alpha)}{\gamma} = \frac{\phi_\Omega^\gamma}{\phi_\Omega^E}\left[\phi_\Omega^E + \alpha\frac{\sigma_R}{\sigma_\Omega}(1-\phi_\Omega^E)\right]. \tag{20}$$

The two ratios on the right side, $\frac{\phi_\Omega^\gamma}{\phi_\Omega^E}$ and $\frac{\sigma_R}{\sigma_\Omega}$, reflect systematic differences between Tier 1 and Tier 2 countries. If the first ratio differs from one, then the coalition size when measured in terms of climate damages as a fraction of global damages is different then when measured in terms of carbon emissions

[17] An earlier iteration of this paper addressed the model under the broader condition $\theta_2 \geq 2$. Although this scenario includes the most recent calibration of the DICE model—$\theta_2 = 2.6$ (Nordhaus, 2017; Barrage and Nordhaus, 2023)—the insights derived from the analytical discussion are most lucid in the quadratic case. To simplify the exposition, I focus on the quadratic case in Propositions 1, 3, 4 and 5.

[18] The latter assumption simplifies the analytical solution since we don't have to worry about whether the Tier 2 incentive is binding or not.

[19] The cooperative carbon price defined in Eq. 18 implies

$$\frac{\tau^C(\alpha)}{\gamma} = \frac{\gamma_\Omega}{\gamma}\left[1 + \alpha\frac{\sigma_R}{\sigma_\Omega}\frac{1-\phi_\Omega^E}{\phi_\Omega^E}\right] = \frac{\phi_\Omega^\gamma}{\phi_\Omega^E}\left[\phi_\Omega^E + \alpha\frac{\sigma_R}{\sigma_\Omega}(1-\phi_\Omega^E)\right].$$

as a fraction of global emissions. If the second ratio differs from one, then the emissions intensity is systematically different in the first tier compared with the rest of the world.

As a reference, it is useful to start with the case in which no such differences exist: $\frac{\phi_\Omega^\gamma}{\phi_\Omega^E} = \frac{\sigma_R}{\sigma_\Omega} = 1$. This is consistent with a setting in which all countries scale "symmetrically" in size, meaning that a country that comprises fraction ϕ of global GDP also comprises fraction ϕ of global CO2 emissions and fraction ϕ global climate damages. We then have

$$\frac{\tau^C(\alpha)}{\gamma} = \phi_\Omega + \alpha(1 - \phi_\Omega), \tag{21}$$

where $\phi_\Omega \equiv \phi_\Omega^Q = \phi_\Omega^E = \phi_\Omega^\gamma$. If the match rate is zero—as with a conventional climate club—then the right-hand side equals the coalition size, ϕ. In contrast, if the match rate is one—the natural upper bound on what the first tier might attempt—then the Tier 1 carbon price equals the globally efficient carbon price. In this case, since the match rate is one, the Tier 2 countries also adopt the globally efficient carbon price. Hence, the overall outcome is globally efficient. The amplification effect is just enough to offset the fact that the coalition only internalizes fraction ϕ_Ω of global climate damages.

Importantly, the finding does not depend on the size of the first tier. As the first tier gets smaller, internalizing a smaller fraction of global damages, the second tier expands by an equal degree. Due to these offsetting effects, the first tier retains the incentive to price carbon at the globally efficient level when the match rate is one because the amplification effect exactly offsets the impact of a shrinking damage fraction. Nevertheless, increasing the match rate to one would clearly get harder as the first tier shrinks since a smaller first tier has less trading prowess.

In realistic settings, the ratios may differ from one. Considering a US-EU-UK first tier is instructive since this will turn out to be the unique stable first tier in the quantitative model. For this group, Tier 1 climate damages and Tier 1 carbon emissions are similar fractions of the global aggregate, so $\frac{\phi_C^\gamma}{\phi_C^E} \approx 1$. At the same time, the group has substantially lower emissions intensity than the rest of the world: $\frac{\sigma_R}{\sigma_C} \approx 2.8$.

If either ratio is greater than one, the cooperative carbon price is higher than in the symmetric case. Increasing the match rate to one then causes the cooperative policy to *overshoot* the globally efficient levels. The conditions for overshooting are intuitive. In the first case, $\frac{\phi_C^\gamma}{\phi_C^E} > 1$ means the coalition internalizes a larger fraction of global climate damages then its fractional contribution to global CO2 emissions. A coalition that satisfies this condition loses more than the average country from climate damages for each unit of carbon dioxide emissions on which it faces the tax; hence, it wants to do more. In the second case, $\frac{\sigma_R}{\sigma_\Omega} > 1$ means the coalition has lower emissions intensity than the ROW. Countries with lower emissions intensity face less cost from climate policy and thus prefer more climate policy than the average country.

Returning to the example of a US-EU-UK first tier, substituting the approximate ratio values into Eq. 20 implies $\tau^C(1) \approx 2.2 \times \gamma$. With a match rate of one, the US, EU, and UK would price carbon over twice as high as the globally efficient level. This is due to the group having relatively low emissions intensity, which lowers the domestic cost of a carbon price. An implication of overshooting is that the

match rate needed to achieve globally efficient abatement is less than one. This implies that achieving something close to globally efficient abatement with a TCC may be significantly easier than it might at first appear (see Section 5).

In this case, because carbon prices are different in the first and second tiers, achieving globally efficient abatement does not imply that the overall outcome is globally efficient. Nevertheless, the efficiency loss coincides with an improvement in fairness since countries with greater historical responsibility for carbon emissions are the ones who do more.

3.1.2 Global cost inefficiency

Another problem with subglobal abatement is the large increase in global abatement costs because a coalition of abating countries cannot take advantage of cheap marginal cost abatement in the rest of the world. To quantify this problem, Nordhaus (2008) analytically derives a cost penalty that captures the increase in global abatement costs when all abatement is done by a subset of countries rather than being dispersed efficiently across countries. I extend his result to consider the impact on this cost penalty if non-coalition countries are required to match coalition policy as required under a TCC.

The setting is the same as in Nordhaus (2008). It is consistent with the model presented in Section 2.1, though we only need a few pieces for the analysis. Specifically, all countries have the same "DICE-like" abatement cost function.[20] Countries may differ in size but they scale "symmetrically," meaning (as before) that each country's fraction of global GDP is the same as its fraction of global carbon emissions and its fraction of global climate damages.

Figure 1 demonstrates the inefficiency using an example in which the coalition comprises a third of global emissions ($\phi = \frac{1}{3}$). The left panel shows marginal abatement costs for the coalition. Without abatement, it produces 100 units of CO2 emissions, so abatement ranges from 0 to 100. The right panel shows marginal abatement costs for the rest of the world. Without abatement, it produces 200 units of emissions, so abatement ranges from 0 to 200. The exponent (or shape parameter), denoted θ_2, is calibrated to be 2.6 (as in Nordhaus, 2017; Barrage and Nordhaus, 2023).

The coalition's goal is to reduce global emissions by 80 units. When it abates all 80 units itself, total abatement costs equal the area under the coalition's marginal abatement cost curve between 0 and 80—the area $A + B + C$. With efficient implementation, the coalition abates 27 units, while the rest of the world abates 53 units. Both regions then have the same marginal abatement costs, equal to τ^*.[21] Global abatement costs are represented by the area $A + D + E$. The ratio of area $A + B + C$ over area $A + D + E$ is what Nordhaus (2008) calls the cost penalty from incomplete participation. When the coalition acts alone, it undertakes very high marginal cost abatement activities, while leaving very cheap abatement options unexploited in the rest of the world.

Nordhaus (2008) shows in this setting that the cost penalty is independent of the amount of

[20] Abatement costs take the form $Q_i \theta_1 \mu_i^{\theta_2}$, where Q_i is country i's GDP, μ_i is its abatement rate, and the parameters θ_1 and θ_2 are the same as in the DICE abatement cost function.

[21] This outcome could be implemented with a harmonized carbon price equal to τ^*.

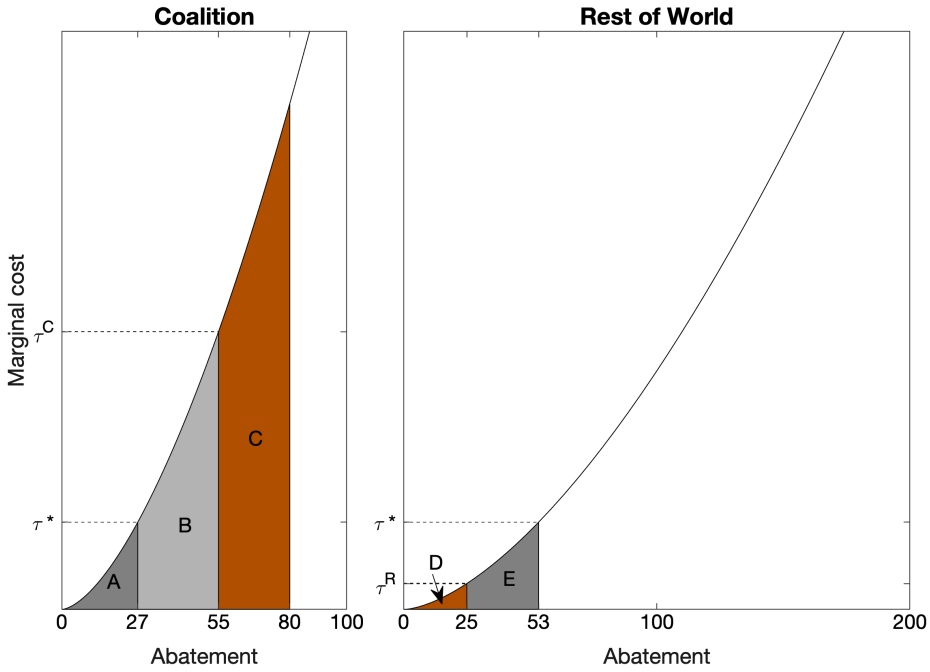


Figure 1: The figure shows marginal abatement costs for a coalition comprising a third of global emissions and for the rest of the world. The curves are generated using the abatement cost function calibrated in Barrage and Nordhaus (2023) (and other recent applications of DICE). Letters indicate colored regions whose area reflects the total cost of moving between the indicated levels of abatement.

global abatement achieved by the coalition, and it equals $\phi^{1-\theta_2}$. $\phi \in [0,1]$ is the coalition's size (e.g., CO2 emissions as a fraction of global emissions) and θ_2 is the shape parameter (or exponent) in the abatement cost function. It is easy to see from the formula that the cost penalty is exactly 1 for the grand coalition. When everyone participates, there is no penalty. We also see that the penalty rises rapidly as ϕ shrinks. If the coalition is a third of global emissions, the cost penalty is $(0.33)^{1-2.6} = 5.9$—exactly the ratio of area $A+B+C$ to area $A+D+E$. In contrast, if $\phi = 0.22$, as it would if the coalition consisted of the US and EU alone (the US and EU make up 22 percent of global CO2 emissions) then the cost penalty increases to a whopping $(0.22)^{1-2.6} = 11.3$. In the left panel of Figure 2, I plot the cost penalty as a function of coalition size. It is represented by the solid black line labeled "unilateral".[22]

To see how a partially matching carbon price in the rest of the world impacts the cost penalty, I extend Nordhaus's result in Proposition 2. In Appendix A.4, I solve the problem for the more general case with multiple non-coalition sub-regions, each with a different match rate. The setting is potentially interesting in real world applications since countries differ widely in their willingness to price carbon. The result in Proposition 2 is a special case of this more general result.

Proposition 2. *Suppose countries scale symmetrically in size (as defined in the text) and a coalition of countries comprises fraction ϕ of the global economy. The coalition adopts carbon price τ_C, while*

[22] "Unilateral" means the coalition does all the abatement while the rest of the world does nothing.

the rest of the world adopts carbon price $\tau_R = \alpha\tau_C$*, for some* $\alpha \in [0, 1]$*. Then global abatement costs can be expressed as*

$$\Psi(\mu; \phi) = P(\phi, \alpha) \times \Psi^*(\mu),$$

where μ *is the global abatement rate,* $\Psi^*(\mu) = Q\theta_1\mu^{\theta_2}$ *is the global abatement cost function in the efficient case, and* $P(\phi, \xi)$*, the cost penalty, is given by*

$$P(\phi, \alpha) = \frac{\phi + (1-\phi)\alpha^{\theta_2/(\theta_2-1)}}{\left[\phi + (1-\phi)\alpha^{1/(\theta_2-1)}\right]^{\theta_2}}. \tag{22}$$

The penalty reduces to that in Nordhaus (2008) if $\alpha = 0$*:*

$$P(\phi, 0) = \phi^{1-\theta_2}.$$

Proof. Appendix A.4. □

The cost penalty is independent of the global abatement rate, μ, and it is strictly decreasing in both coalition size and the match rate (Appendix A.4). In Figure 2, I use Eq. 41 to compute the penalty for a range of scenarios. The left panel plots the cost penalty as a function of coalition size for three different policies. The first policy (solid black line) involves a unilateral effort by the coalition, which coincides with the case considered in Nordhaus (2008). As previously noted, the penalty reaches 11.3 when $\phi = 0.22$. Requiring the rest of the world to match the coalition carbon price has a dramatic impact on the cost penalty, even for small values of the match rate. For example, if α is 10 percent, while ϕ remains at 0.22, the cost penalty drops by more than a factor of four—from 11.3 to 2.5.

The right panel of Figure 2 displays $P(\phi, \alpha)$ as a function of α for two noteworthy values of ϕ. The first corresponds to a coalition between the EU (excluding the UK) and the US, for which $\phi = 22.0\%$. The second value adds China, increasing the coalition size to $\phi = 53.1\%$. If the match rate is zero, the addition of China to the coalition has a very large impact on reducing the abatement cost penalty, from 11.3 to 2.8. This outcome affirms the widely held belief that China's inclusion in a potential climate agreement is critical. However, the figure illustrates an alternative approach to achieving a comparable outcome. Supplementing abatement by the EU and US with a 10 percent match from the rest of the world has a similar effect on the overall cost inefficiency as adding China to a unilateral agreement.

To see why the match requirement has such a large impact on the inefficiency of sub-global abatement, I continue the earlier example in Figure 1. We previously saw that moving from unilateral coalition policy to efficient policy led to a giant reduction in global abatement costs. The change had the effect of swapping the high marginal abatement cost area $B + C$ for the low marginal abatement cost area $D + E$.

In the case of a matching carbon price, a less-ambitious swap occurs, though the benefit from each unit of swapped abatement is bigger. To see this, I continue the previous example by simulating a partial match requirement in Figure 1. The coalition imposes carbon price τ^C within its borders, leading to 55 units of abatement within the coalition. Meanwhile, the rest of the world imposes the

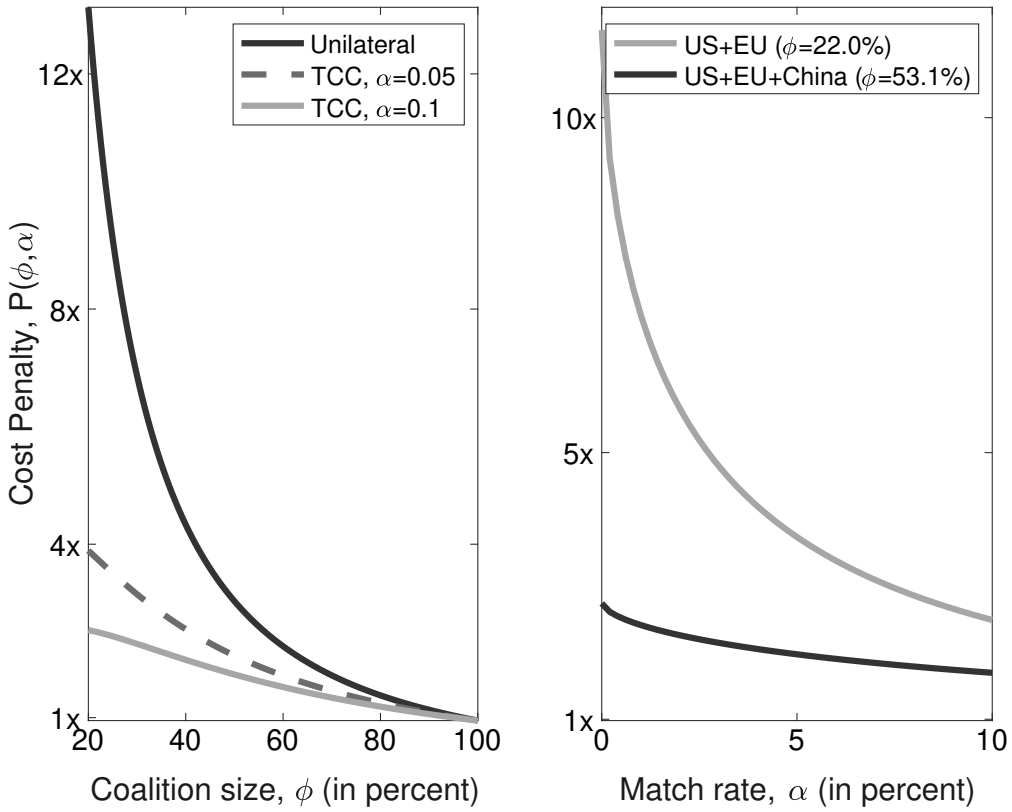


Figure 2: The left panel plots the penalty as a function of ϕ for three different policies. The right panel plots the penalty as a function of α for two values of ϕ: $\phi = 22.0\%$ (US+EU) and $\phi = 53.1\%$ (China+US+EU).

lower carbon price τ^R, leading to 25 units of abatement outside the coalition. Total abatement is 80 units. Compared to the case in which all abatement is done by the coalition, we exchange costs equal to area C for costs equal to area D. The partial match requirement takes advantage of the most valuable opportunities to swap abatement across regions. This is because the first units of abatement given up inside the coalition are the highest-cost ones, while the first units taken up outside the coalition are the cheapest.

3.1.3 Carbon leakage

The effectiveness of a small climate club is further undermined by carbon leakage. To quantify the potential for a TCC to reduce carbon leakage, one would ideally use a multi-region general equilibrium model that takes into account different leakage channels. However, such a model is beyond the scope of this paper, so I merely discuss the directional impact of a TCC on the primary drivers of carbon leakage. I also discuss how the leakage reducing impact of a TCC compares with that of a Border Carbon Adjustment (BCA) policy since this is the most common policy considered for reducing carbon leakage (e.g., Böhringer et al., 2022). A BCA taxes the carbon content of imports, effectively leveling the playing field for domestic producers within the domestic market.

Carbon leakage arises mainly through two channels: the "competitiveness channel" and the "price-effect channel." Leakage through the competitiveness channel occurs when climate policies place domestic producers at a competitive disadvantage, affecting both sales to the domestic market and sales to foreign markets. For domestic sales, a TCC mitigates leakage for Tier 1 countries by reducing energy price differences between domestic and foreign firms. However, a TCC is less effective than a BCA at curbing this leakage channel, as a BCA imposes the same carbon price on imported goods as faced by domestic producers. Nevertheless, BCAs do better at addressing the leakage that when firms sell to foreing markets. BCAs have no impact on this channel, while a TCC does since it increases energy prices in all other countries (even if only by a modest amount when the match rate is low).

Leakage through the price-effect channel takes place when decreased fossil fuel demand in abating countries causes global fossil fuel prices to drop, increasing foreign consumption of fossil fuels. BCAs are well-known for their inability to address this leakage channel (Dubey, Keen and Taxell, 2014) whereas a TCC would partly address it since the matching carbon price would decrease fossil fuel demand in the foreign market to some degree. For both channels, a TCC's impact on leakage would increase with the match rate, ultimately reaching zero if the match rate equals one. It follows that at low match rates, a TCC would be less effective than a BCA at controlling leakage for Tier 1 countries, but there would be a threshold match rate beyond which leakage for Tier 1 countries would be lower than under a BCA.

A TCC also has leakage-related advantages when we consider leakage from the perspective of Tier 2 countries. A TCC causes Tier 2 countries to adopt a positive carbon price, which usually raises concerns about carbon leakage. However, under the TCC agreement, the carbon price adopted by Tier 2 countries is guaranteed to be at the bottom of the global distribution of carbon prices, which eliminates the risk of carbon leakage. This policy feature could be crucial for garnering political support for an agreement. For instance, China, as the global manufacturing hub, has much to lose from unilateral climate policy that would reduce the competitiveness of Chinese goods in the global market. A TCC agreement in which China remains a Tier 2 country offers an environment where China can engage in abatement efforts without this concern.

3.2 The role of Tier 1 cooperation

The analysis behind Proposition 1 focused on the cooperative Tier 1 policy. To achieve this outcome, Tier 1 countries must use tariff threats with each other. However, as noted in Section 2.3, these threats may be difficult to sustain given the large size of the economies likely to select into the first tier (Section 5.1). This section shows how carbon prices and global abatement change if Tier 1 countries instead adopt the non-cooperative policy (Problem ***P2***). This policy provides a lower bound on Tier 1 policy since it reflects what Tier 1 countries would optimally do unilaterally without coordinating their action with other Tier 1 countries other than to set up and enforce the TCC terms with Tier 2 countries. At the end of the section, I solve for the minimum Tier 1 penalty needed to support the cooperative outcome.

The next proposition solves the non-cooperative problem.

Proposition 3. *Assumptions are the same as in Proposition 1. The non-cooperative policy that solves Problem* ***P2*** *is given by*[23]

$$\tau_i^N(\alpha) = \gamma_i \frac{\mathrm{d}a_G}{\mathrm{d}a_C} = \hat{\phi}_i^{\gamma} \tau^C(\alpha), \qquad i \in \Omega, \tag{23}$$

where $\hat{\phi}_i^{\gamma} \equiv \frac{\gamma_i}{\gamma_\Omega}$ *and, as in Proposition 1,*

$$\frac{\mathrm{d}a_G}{\mathrm{d}a_C} = 1 + \alpha \frac{1 - \phi_C^E}{\phi_C^E} \frac{\sigma_R}{\sigma_C} \tag{24}$$

and

$$\tau^C(\alpha) = \gamma_\Omega \times \frac{\mathrm{d}a_G}{\mathrm{d}a_C}. \tag{25}$$

Proof. Appendix A.5. □

The non-cooperative policy closely resembles the cooperative policy with the same amplification factor applied to a lower base level of climate concern. Without a TCC (i.e., $\alpha = 0$) each Tier 1 country i unilaterally prices carbon at the domestic marginal climate damage, γ_i, replicating the standard Nash result (Kotchen, 2018). When $\alpha > 0$, the same channels operate that were present in the cooperative solution (Proposition 1). The amplification factor increases in the match rate, the relative size of the second tier, and the relative emissions intensity of the second tier. For the same match rate, the extent to which each Tier 1 country i falls short of the cooperative policy when it acts unilaterally is determined by its share of Tier 1 climate damages ($\hat{\phi}_i^{\gamma}$).

Next, I consider the impact on global abatement of moving between the cooperative and non-cooperative outcomes.

Proposition 4. *Assumptions are the same as in Propositions 1 and 3. The global abatement rate in the non-cooperative case can be written*

$$\mu^N(\alpha) = \tilde{H}\big(\{\hat{\phi}_i^E, \hat{\phi}_i\}_{i\in\Omega}\big)\mu^C(\alpha), \tag{26}$$

where $\hat{\phi}_i^E \equiv \frac{E_i}{E_\Omega}$ *is i's share of Tier 1 emissions,* $\hat{\phi}_i^E \equiv \frac{\gamma_i}{\gamma_\Omega}$ *is i's share of Tier 1 damages, and*

$$\tilde{H}\big(\{\hat{\phi}_i^E, \hat{\phi}_i\}_{i\in\Omega}\big) = \sum_{i\in\Omega} \hat{\phi}_i^E \hat{\phi}_i{}^{\gamma}. \tag{27}$$

$\tilde{H}(\cdot)$ *can be regarded as an index of country-size concentration in the first tier. If Tier 1 countries scale symmetrically in emissions and damages (i.e., if* $\phi_i^E = \phi_i^{\gamma}$ *for all* $i \in \Omega$*), then the index reduces to*

$$\tilde{H}\big(\{\hat{\phi}_i^E\}_{i\in\Omega}\big) = \sum_{i\in\Omega} \big(\hat{\phi}_i^E\big)^2, \tag{28}$$

which is the Herfindahl index of country size within the first tier.

Proof. Appendix A.6. □

[23] The superscipt N stands for non-cooperative.

The proposition shows that the global abatement rate when the first tier interacts non-cooperatively is proportional to the cooperative global abatement rate, where the proportionality constant is given by an intuitive measure of country size concentration within the first tier. Since the restriction $\phi_i^E = \phi_i^\gamma$ holds roughly for the Tier 1 countries in the calibrated model, I focus on this case for intuition.

The Herfindahl index is a common measure for characterizing the distribution of size within a group. In this case Eq. 28 refers to the distribution of country size within the first tier. As is well known, intuition for the index is easiest to see when countries are the same size. In that case, the Herfindahl index reduces to $1/n$, where n is the number of countries. With one country, the index is 1, indicating the highest possible concentration. With ten countries, the index drops to a tenth.

Eq. 28 shows how the size distribution of countries within the coalition impacts how the absence of within-coalition cooperation affects global abatement. If coalition countries are the same size, Eq. 28 implies that global abatement reduces by $100 \times \frac{n-1}{n}$ percent when moving from the case of full cooperation to Nash. Thus, with five equal-sized participants in the coalition, absence of cooperation within the coalition reduces global abatement by 80 percent, while with two participants the reduction is 50 percent. The finding suggests a further advantage of small climate clubs—apart from their having an easier time negotiating agreement. Small clubs reduce the importance of achieving cooperation in the first tier, which is important if participants in the first tier are too big to punish.

The cost of cooperation The last proposition solves for the minimum Tier 1 penalty needed to induce Tier 1 countries to adopt the cooperative policy.

Proposition 5. *Let $\theta_2 = 2$. Countries differ in size but scale symmetrically ($\phi_i \equiv \phi_i^Q = \phi_i^E = \phi_i^\gamma$ and $\theta_{1,i} = \theta_1$ all $i \in \Omega$). The minimum Tier 1 penalty needed to support the cooperative policy—Eq. 18—is*

$$\omega_{1,i}^C = \Gamma(\phi_\Omega - \phi_i)^2, \text{ for } i \in \Omega.$$

where

$$\Gamma = \frac{\gamma^2\sigma^2}{2\theta_1}\left(1 + \alpha\frac{1-\phi_\Omega}{\phi_\Omega}\right)^2 > 0.$$

Proof. Appendix A.7. □

The minimum penalty to induce compliance with the cooperative policy depends on country size. Smaller Tier 1 countries, which internalize a smaller portion of global climate damages, require a higher penalty. The relationship between country size and the minimum penalty is nonlinear, with the penalty increasing with the square of the distance between the size of the country and the size of the first tier. If we are constrained to apply the same penalty (as a fraction of GDP) to all Tier 1 countries then

$$\bar{\omega}^C \equiv \max_{i\in\Omega} \omega_{1,i}^C$$

is the minimum penalty to ensure compliance from all Tier 1 countries.

4 Quantitative model

This section presents the calibration and numerical approach.

4.1 Calibration

To calibrate the full model from Section 2, I depart from Nordhaus (2015) in two ways. First, I revise the set of world regions. Second, I update the data.

Nordhaus (2015) divides the world into 15 regions, including 7 aggregate regions, such as "Latin America," "Eurasia," and "Southeast Asia." These regions, by assumption, behave in a cooperative manner when adopting climate policy. In order to model incentives for actual governments, I focus on national incentives for large countries, while keeping the EU as a unified region (as is done in most of the literature). To keep the model computationally tractable, I divide the world into the following economies: the EU, plus the ten biggest national economies outside the EU by GDP, plus the rest of the world (ROW).[24] Emission response in the ROW takes into account the domestic marginal climate damage for each country, but these countries do not have the option to join the first tier. The eleven modeled economies comprise 75 percent of global CO2 emissions, and the largest economy in the ROW region is Mexico.

To update the data, I revise Nordhaus (2015)'s (mostly) 2011 calibration using 2019 data (the last full year before COVID). A key determinant of Tier 1 policy ambition, as shown in Proposition 1, is the ratio of Tier 2 to Tier 1 energy intensity. If the first tier entails the US, EU, and UK, this ratio increased about 30 percent between 2011 and 2019. Not accounting for this large transition in the global economy over this decade would reduce the quantitative realism of the results. The eleven modelled economies plus ROW are separately calibrated to reflect differences in GDP, CO2 emissions, domestic climate damages, domestic abatement costs, and tariff-impact parameters. The country-level parameters are reported in Table 1.

For GDP and CO2 emissions, I use 2019 World Bank data. For domestic climate damages, I employ the "three-model" estimate of regional climate damages from Table B-2 of the Online Appendix to Nordhaus (2015). These estimates average the regional SCC calculations for the RICE, FUND, and PAGE models. For countries not in Nordhaus's list of regions, I downscale climate damages by assuming that damages within region are proportional to GDP. I further assume that regional SCCs are a constant fraction of the global SCC (γ) which I vary separately. For abatement costs, I follow the calibration in Nordhaus (2015). The abatement cost scale parameter $\theta_{1,i}$ varies across regions, while the exponent is constant—$\theta_2 = 2$. On average, countries with higher emissions intensity have higher abatement costs. Since the regions here differ from those in Nordhaus (2015), I adjust the abatement costs in Nordhaus (2015) by multiplying all $\theta_{1,i}$ parameters by a constant scale factor to ensure a 25 USD per ton CO2 carbon tax generates an 18 percent reduction in global CO2 emissions, which is the calibration target in Nordhaus (2015).

[24] An almost identical specification of regions is used in Farrokhi and Lashkaripour (2020).

	USA	EU	China	Japan	UK	India	Brazil	Can.	Russia	Korea	Aus.
ϕ_j^Q:	24.4	17.9	16.3	5.8	3.3	3.2	2.1	2	1.9	1.9	1.6
γ_j/γ:	10.6	11.7	11	2.4	2.1	11.7	2.9	1.0	3.5	3.5	3.0
σ_j:	0.061	0.047	0.204	0.058	0.033	0.237	0.063	0.091	0.274	0.101	0.076
$\theta_{1,j}$:	0.0297	0.0210	0.0545	0.0275	0.0210	0.0328	0.0065	0.0383	0.0519	0.0275	0.0297
α_j:	0.733	0.611	0.581	0.41	0.611	0.585	0.682	0.685	0.654	0.641	0.641
β_j:	0.621	0.555	0.618	0.427	0.555	0.975	0.758	0.646	0.617	0.605	0.605

Table 1: Heterogeneity parameters: ϕ_j^Q is GDP as a fraction of global GDP in percent; γ_j/γ is climate damages as a fraction of global damages in percent; σ_j is emissions intensity in tons carbon per thousand USD; $\theta_{1,j}$ is the unitless abatement cost scale parameter; α_j and β_j are the tariff-impact parameters.

For tariff impact functions maintain the values from Nordhaus (2015) for the economies that coincide with regions in his model. For the other economies, I assume they are the same as for the corresponding Nordhaus region in which they are contained. Thus, for the UK, I assume they are the same as for the EU, while for South Korea and Australia, I assume they are the same as for the South East Asian region in Nordhaus (2015).

To quantify trade flows between model regions, I use 2019 UNCTAD data[25]. I combine total international merchandise trade and total trade in services. Appendix A.8 provides bilateral trade matrices for each category separately. To calibrate the relationship between aggregate coalition GDP and the maximum feasible match rate, I assume $b = 1$. This means that a match rate of one is only feasible if the coalition includes the whole world. The plausibility of this calibration assumption is evaluated in Section 5.3 where I derive the implied minimum tariffs needed to support the equilibrium outcome for the calibrated model. I evaluate the sensitivity of results to the value of b in Section 5.4.

Table 2 reports values for the remaining parameters that do not vary across countries. Units are explained in the figure caption. I assume the global SCC is \$55 per ton CO2, the middle estimate from the Interagency Working Group report (EPA 2021). Converting the SCC to units of thousands of USD per ton carbon gives the value of γ reported in the table. I consider the sensitivity of results to the global SCC in Section 5.4.

4.2 Numerical approach

To study the cooperative policy for a given match rate, I numerically solve for the harmonized Tier 1 carbon price that maximizes coalition surplus. The objective function takes into account the optimal response behavior of all non-coalition countries, including countries within the ROW region for which optimal emissions response is separately modeled. Faced with minimum carbon price target $\hat{\tau}_2$—and a sufficient trade threat to make it in the country's interest to avoid the penalty—each Tier 2 country

[25]The data is available here: https://unctadstat.unctad.org/EN/BulkDownload.html

Q	E	σ	γ	b
87.7	9.5	0.108	0.202	1.0

Table 2: Other model parameters that are not country-specific. Q is global GDP in trillions of USD per year. E is global emissions in GtC. σ is aggregate emissions intensity in tons carbon per thousand USD. γ is the global SCC in trillions of USD per GtC (or 1000's of USD per ton carbon). b is the unitless proportionality constant in Eq. 12.

i implements domestic carbon price

$$\tau_i = \max(\gamma_i, \hat{\tau}_2), \tag{29}$$

where the domestic SCC γ_i is $i's$ unilateral best response absent a binding conditional trade threat.

To study the noncooperative equilibrium in which coalition countries take the match rate as given and play Nash with each other, I employ an iterative algorithm. The object of iteration is a vector of carbon prices for each Tier 1 country. The iteration step has each Tier 1 country take the carbon prices of the other Tier 1 countries as given—specified by the last iteration vector—then solve for its unilateral best response. The vector of best responses for all Tier 1 countries gives the next iteration vector. Carbon prices for the Tier 2 countries are pinned down by this procedure since they are a simple function of the Tier 1 carbon prices. I start the algorithm at an arbitrary initial vector of carbon prices for each Tier 1 country, and I find that the algorithm consistently converges to the same fixed point starting from a wide range of initial conditions. In each case, the problem is implemented in MATLAB with the Knitro solver.

To further validate the numerical approach, I analytically solve a restrictive version of the problem, then show that the numerical algorithm replicates the analytical result under the restrictive assumptions. Under the assumption that countries outside the coalition do not abate without external coercion, Appendix A.9 extends the analytical solutions (both upper and lower bounds) for the case in which countries differ in terms of emissions intensity and abatement costs in addition to differing in terms of climate damages, emissions, and GDP. The formula for the cooperative policy becomes

$$\tau_i^{COOP} = \gamma_C \frac{\sigma}{\theta_2} \frac{\sum_j \phi_j^E a_j + (1-\phi_C^E) a_R \alpha^b}{\sum_{i=1}^n \phi_i^Q \theta_{1,i} a_i^{\theta_2}},$$

where

$$a_i = \left(\frac{\sigma_i}{\theta_{1,i}\theta_2}\right)^b$$

and

$$b = \frac{1}{\theta_2 - 1}.$$

In addition, the corresponding non-cooperative policy solves the following system of equations in the same number of unknowns:

$$\tau_i = \gamma_i \left[1 + \alpha^b \frac{1-\phi_C^E}{\phi_C^E} \left(\frac{\sigma_R}{\sigma_i} \frac{\theta_{1,i}}{\theta_{1,R}}\right)^b \left(\frac{\tau_i}{\tau^{AVG}}\right)^{1-b}\right], \quad \text{for } i \in \Omega,$$

where (as before) $\tau^{AVG} = \frac{1}{\phi_C^E} \sum_{i \in \Omega} \phi_i^E \tau_i$.

5 Quantitative results

This section presents the numerical results. For the baseline calibration, I study Tier 1 stability, the range of achievable policies, and the tariffs needed to support them. Next, relaxing the baseline assumptions, I examine the sensitivity of outcomes to key assumptions.

5.1 Coalition stability

To study coalition stability, I evaluate the internal and external stability conditions—Equations 15 and 16—under all possible permutations of the eleven economies in the quantitative model. The main result is that the stable first tier is unique for the baseline calibration, and it consists of the US, EU, and UK. In equilibrium, the stable first tier achieves 55 percent of the efficient level of global abatement. A unique stable coalition contrasts with analyses of coalition stability for single-tier climate clubs. For example, Hagen and Schneider (2021) find many possible stable coalitions. Section 5.4 shows that while outcomes with multiple equilibria are possible, they are less common.

To develop intuition for the country-level incentives that underlie the stability results, Figure 3 compares payoffs for each economy depending on whether it joins a given Tier 1 coalition or not. The horizontal axis depicts a series of three candidate coalitions, increasing in size from left to right. The left-most coalition entails the Nash equilibrium in which all countries play their unilateral best response. This is described as consisting either of the US or the EU so the meaning of adding another country to the candidate coalition is well defined. In each case, the solid black line depicts payoffs for a country if it joins the indicated coalition (assuming it is not already in) or stays in (if it is already in). The dashed-grey line depicts payoffs if it stays out (if it is not already in) or leaves (if it is already in). The third coalition, the stable first tier, adds the UK to the EU-US coalition. Panels are sorted in order of descending incentive to join the stable first tier.

The first row shows that the EU has a strong incentive to join the stable first tier, while the US and UK both have a positive but marginal incentive to join. The US incentive to join is much higher when the UK is out. In that case, the US benefits about 9 billion USD per year from participating in the coalition. The difference stems from the assumption that the TCC with a positive match rate only kicks in when there is more than one participant in the coalition. Thus, the EU alone does not implement a TCC, but the EU plus the UK does. Given the superior trading prowess of the US, it is probably unreasonable to expect a tiered climate club to form without strong leadership from the US.[26] If this is true, then the actual incentive for the US to join the stable coalition is closer to 9 billion USD per year.

The US benefits less than the EU mainly because its economy is more energy intensive, so a given carbon price is more costly. In addition, because the US comprises a larger fraction of global CO2 emissions, the US joining the coalition undermines the incentive for other coalition countries to abate

[26]This conclusion is reached in Farrokhi and Lashkaripour (2020) who, using a multiregion trade model, show that US participation in a climate club is critical to achieve substantial abatment outcomes.

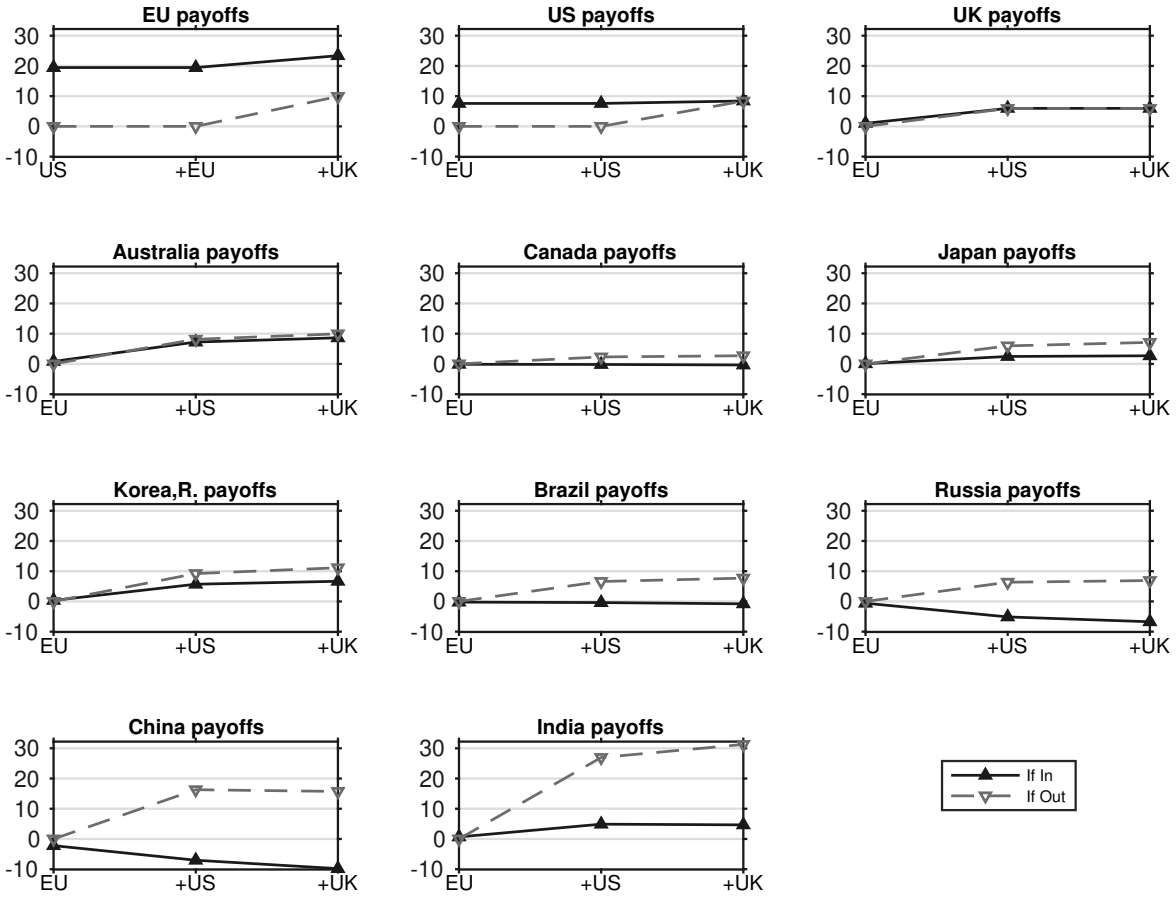


Figure 3: Payoffs relative to Nash payoff (in billions of USD per year). Horizontal axis depicts alternative (Tier 1) coalitions. Left coalition is Nash. Middle coalition is US-EU. Right coalition is US-EU-UK. Solid-black line indicates payoffs if country stays in or joins the indicated coalition; dashed-grey line indicates payoffs if it stays out or leaves. Economies sorted in order of descending net benefit to join.

carbon at a high level. The latter effect will turn out to be a major explanation for China's strong preference to stay out.

Among the eight depicted countries that prefer to stay out of the stable coalition, Australia is closest to wanting to join. For each country, the incentive to stay out increases as the size of the existing coalition increases. This feature of the problem underlies the tendency for the model to give rise to a unique stable first tier (or at most two, at least for the parameter range considered in the sensitivity analysis below). The last row of panels in Figure 3 shows that both India and China are starkly better off staying out of the first tier. For both countries, the payoff if it stays out is roughly 25 billion USD per year higher than if it joins. India still benefits about 5 billion USD per year if it joins, while China is strictly worse off. Despite this difference, China still benefits substantially from the agreement provided it stays in the second tier.

The figure also shows the extent to which different countries benefit from the stable agreement, whether they are inside the first tier or not. Of all countries, India benefits the most, gaining about 30

billion USD per year relative to the Nash equilibrium. The EU is next, gaining about 22 billion USD per year, followed by China, which gains 16 billion USD per year. Next, the US, Australia, Brazil, Korea, and Russia all benefit a similar amount at around 10 billion USD per year. Importantly, while the agreement has the Tier 1 countries adopt a coercive stance toward Tier 2 countries, all countries are strictly better off.[27] This leaves room for diplomatic efforts to convince Tier 2 countries to support a TCC agreement.

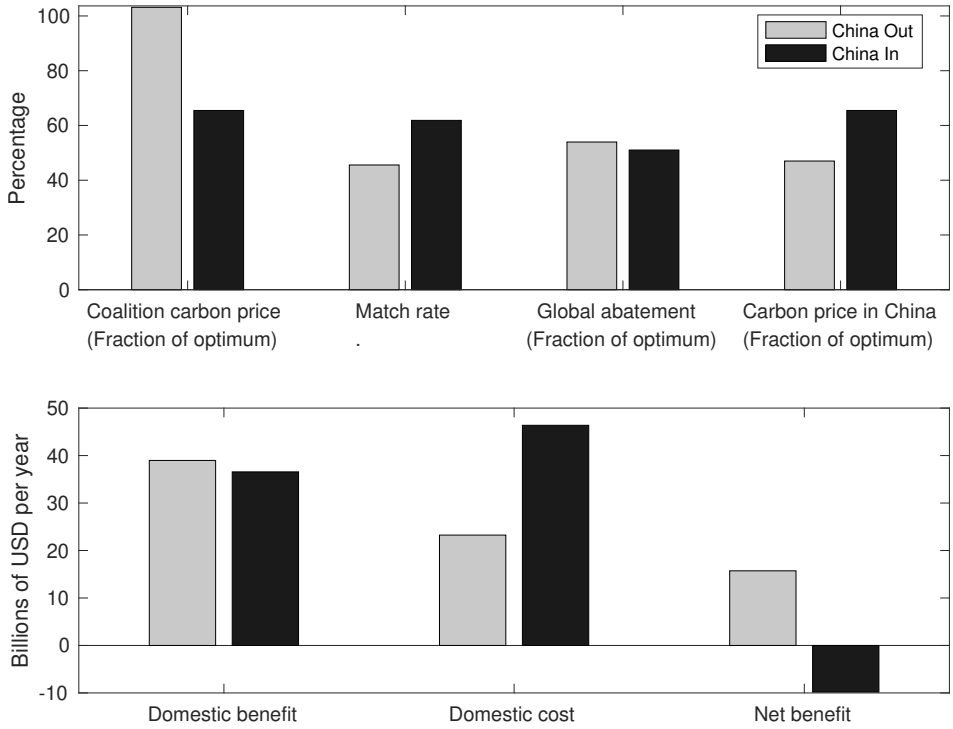


Figure 4: Grey bars indicate outcomes under the stable coalition with China out (EU+US+UK), while black bars indicate outcomes with China in (EU+US+UK+China). The top panel compares equilibrium outcomes under each coalition. The bottom panel decomposes the corresponding payoffs to China.

Given China's out-sized importance as the world's largest emitter, Figure 4 analyzes why China is better off staying out of the stable first tier. The top panel shows how China's choice to join or not would impact equilibrium outcomes, including the resulting carbon price in the first tier, the match rate, and the total amount of global abatement. The bottom panel shows the corresponding impact on benefits and costs for China.

The top panel shows that while the optimal carbon price without China is approximately equal to the globally efficient carbon price, it drops by almost 40 percent if China joins. The reason for the large effect can be seen from the formula for the cooperative carbon price in Eq. 18. In particular, since China comprises 30 percent of global CO2 emissions, it provides a large incentive for the Tier 1

[27]This fact hinges on the adopted calibration of damages, and would change, for example, if some countries benefit from climate change.

countries to price carbon at a high level when it remains outside the first tier. But if China joins, the relative size ratio $\frac{1-\phi_C^E}{\phi_C^E}$ drops from $\frac{1-\phi_C^E}{\phi_C^E} = \frac{1-0.230}{0.230} = 3.35$ to $\frac{1-\phi_{C'}^E}{\phi_{C'}^E} = \frac{1-0.542}{0.542} = 0.85$. Consequently, the amplification factor is bigger by a factor of 4.0 when China is out. In addition, China has higher emissions intensity, so it optimally prefers a lower carbon price and hence would drag down the average carbon price chosen by the coalition.

Apart from affecting the carbon price, China would also increase the match rate if it were to join. Nevertheless, because China's size in GDP terms (as a fraction of the global total) is roughly half its size in CO2 terms, its impact on the match rate is less important. The combined effect of China joining is to marginally reduce global abatement. The bottom panel shows the net impact of these changes on China's economy. By joining the coalition, China's domestic benefit falls slightly, while its domestic cost rises sharply. On net, it goes from gaining 16 billion USD per year to losing 10 billion per year.

5.2 Range of supportable policies

In the analysis above, the match rate for the US-EU-UK coalition is 45.6 percent and Tier 1 policy is cooperative. Next, I explore how policy and outcomes change if the first tier stays the same but the match rate and degree of Tier 1 cooperation vary.

Figure 5 plots the range of supportable carbon prices (left panel) and the the corresponding range of global abatement levels (right panel) as the match rate varies from zero to one and as the degree of cooperation varies from the cooperative upper bound to the non-cooperative lower bound. For each panel, the black line shows the outcome with cooperative policy, and the grey line shows the outcome non-cooperative policy. The shaded region in between shows the range of policies that could be supported for intermediate values of the Tier 1 penalty. Values are plotted as a percent of the global optimum, and the vertical dashed line shows the equilibrium match rate.

When the first tier acts cooperatively, global abatement substantially overshoots the efficient level if the match rate were increased to one. Even with the equilibrium match rate of 45.6 percent, the Tier 1 carbon price already slightly exceeds the globally efficient carbon price, though global abatement is only 55 percent of the efficient level since non-coalition countries abate less. If the match rate equals 67 percent, global abatement exactly equals the efficient level. In this case, the outcome remains inefficient since carbon prices differ between the first and second tiers. With a match rate of 100 percent, global abatement is almost twice the globally efficient level.

The overshooting discussion in Section 3.1 provides intuition for the results. Proposition 1 shows (in a more restrictive setting) that two key driver's of overshooting are the extent to which Tier 1 climate damages are a bigger fraction of the global total than coalition emissions and the extent to which emissions intensity outside the coalition exceeds emissions intensity within it. The corresponding ratios for the US-EU-UK coalition are $\frac{\phi_C^\gamma}{\phi_C^E} = 1.1$ and $\frac{\sigma_R}{\sigma_C} = 2.8$. Thus, while both channels induce overshooting, the main effect comes from the relatively low emissions intensity of the first tier.

For low match rates, the relationship between the match rate and the coalition carbon price is

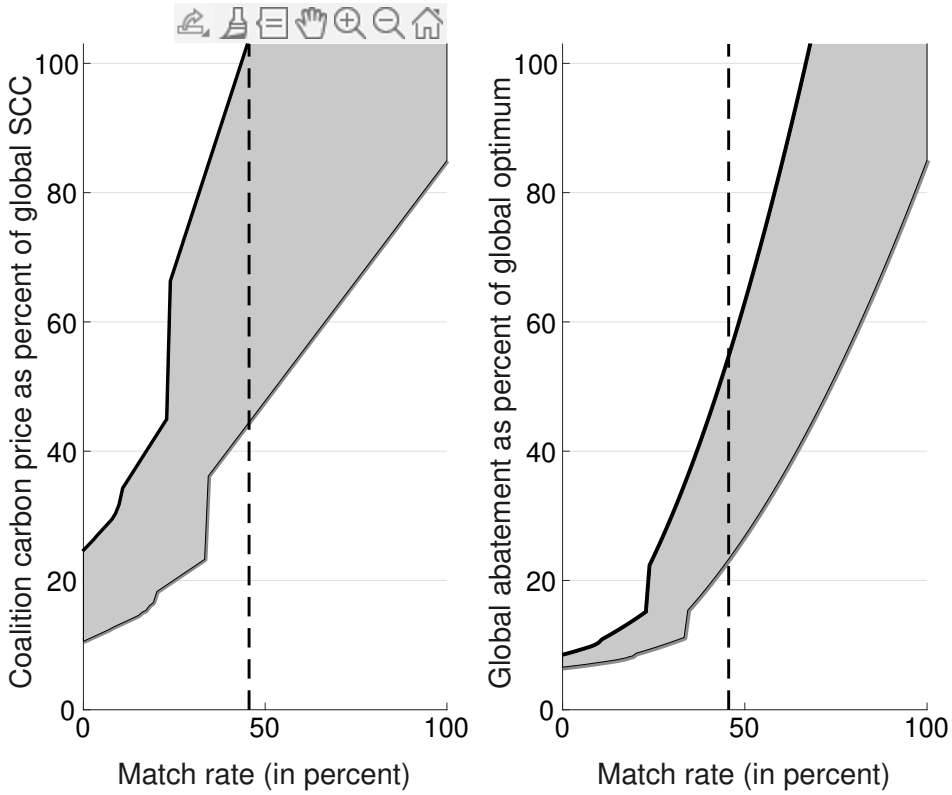


Figure 5: The left panel plots the range of supportable average Tier 1 carbon prices with a US-EU-UK coalition as a function of α. The right panel plots the corresponding range of global abatement levels. In each panel, the grey line shows the non-cooperative policy, while the black line shows the cooperative policy. The gray region shows the range of policies that could hypothetically be supported for each value of α. The dashed line indicates the equilibrium match rate.

jagged. A low match rate reduces the minimum carbon price target for Tier 2 countries, which increases the fraction of Tier 2 countries for which the domestic SCC exceeds the minimum target. When this happens, the linkage between coalition abatement and non-coalition abatement is nonbinding, so the amplification effect does not operate for that portion of non-coalition emissions. The effect is most pronounced for China and India. Because the two countries have roughly the same damages (as a fraction of global damages) in the model, the match rate threshold at which the minimum carbon price target begins to bind is around 23 percent for both. Because China and India together comprise 38 percent of global CO2 emissions,[28] this transition substantially increases the coalition incentive to price carbon at a higher level, which explains the sharp jump in the cooperative carbon price at a match rate of 23 percent. For the non-cooperative case (grey line) the corresponding jump occur at a higher match rate because the non-cooperative carbon price is lower.

When the match rate is sufficiently high, both the cooperative and non-cooperative carbon price curves are linear functions of the match rate. Linearity is a consequence of the quadratic abatement cost assumption in the quantitative model. In both cases, the curves are steep due to the small size of

[28]Based on 2019 emissions data.

the US-EU-UK coalition, which comprises only 23 percent of global CO2 emissions. A small coalition means the match rate applies to a larger non-coalition region.

The non-cooperative policy provides an important lower bound on how much could be achieved if the first tier countries are reluctant to threaten each other with tariffs. Figure 5 shows that while abatement is substantially lower in the non-cooperative case, the agreement could still achieve 85 percent of the globally efficient level of abatement without threatening punishment against Tier 1 countries, provided the match rate is one.

It is perhaps surprising that the non-cooperative policy generates as much abatement as it does. The reason non-coalition abatement is as large a fraction of coalition as it is stems from the small size of the coalition. Proposition 4 showed (under more restrictive assumptions) that global abatement in the non-cooperative case is proportional to global abatement in the cooperative case, with the Herfindahl index of country size within the coalition serving as the proportionality constant. For a US-EU-UK coalition, the index equals 49 percent. Proposition 4 would then imply that for a given match rate the ratio of non-cooperative abatement to cooperative abatement should be 49 percent. If we compute the same ratio for the quantitative model—restricting attention to match rates above 35 percent or so to avoid the jagged regions of the parameter space—we find that the ratio is roughly constant around 43 percent. The divergence from the theoretical prediction of 49 percent stems from the impact of heterogeneity in abatement costs.

Finally, Figure 5 also shows how a single-tier climate club would perform if it were implemented by the US, EU, and UK together, with other countries choosing not to participate. In this case, the cooperative carbon price for the club would be around 25 percent of the global SCC, while the resulting amount of global abatement would be only 10 percent of the globally efficient level. Global abatement is low because a single-tier club cannot take advantage of cheap abatement opportunities outside the club. In contrast, by requiring non-club countries to share part of the burden, the TCC leads to global abatement that is higher by more than a factor of five.

5.3 Tariffs

In this section, I back out the implied minimum tariffs needed to support different policies. The minimum tariff thresholds provide a check on the plausibility of the match rate model parameter b, which determines the equilibrium match rate α. Since bilateral trade flows into the first tier are a key input to this analysis, I also plot export flows from each of the model countries into the US, the EU, and the UK. The results are in Figure 9, which to save space is relegated to Appendix A.10.

To back out the implied minimum tariffs, I use the tariff-impact model described in Section 2.1 along with bilateral trade data summarized in Figure 9. In each case, the same uniform *ad valorem* tariff is applied on imports into each of the punishing countries. The first case I consider is the baseline equilibrium of the calibrated model. This is the case in which the US, EU, and UK impose a match rate of 45.6 percent and act cooperatively. The results are shown in Figure 6.

The left panel of Figure 6 shows how big uniform tariffs need to be against the Tier 1 countries

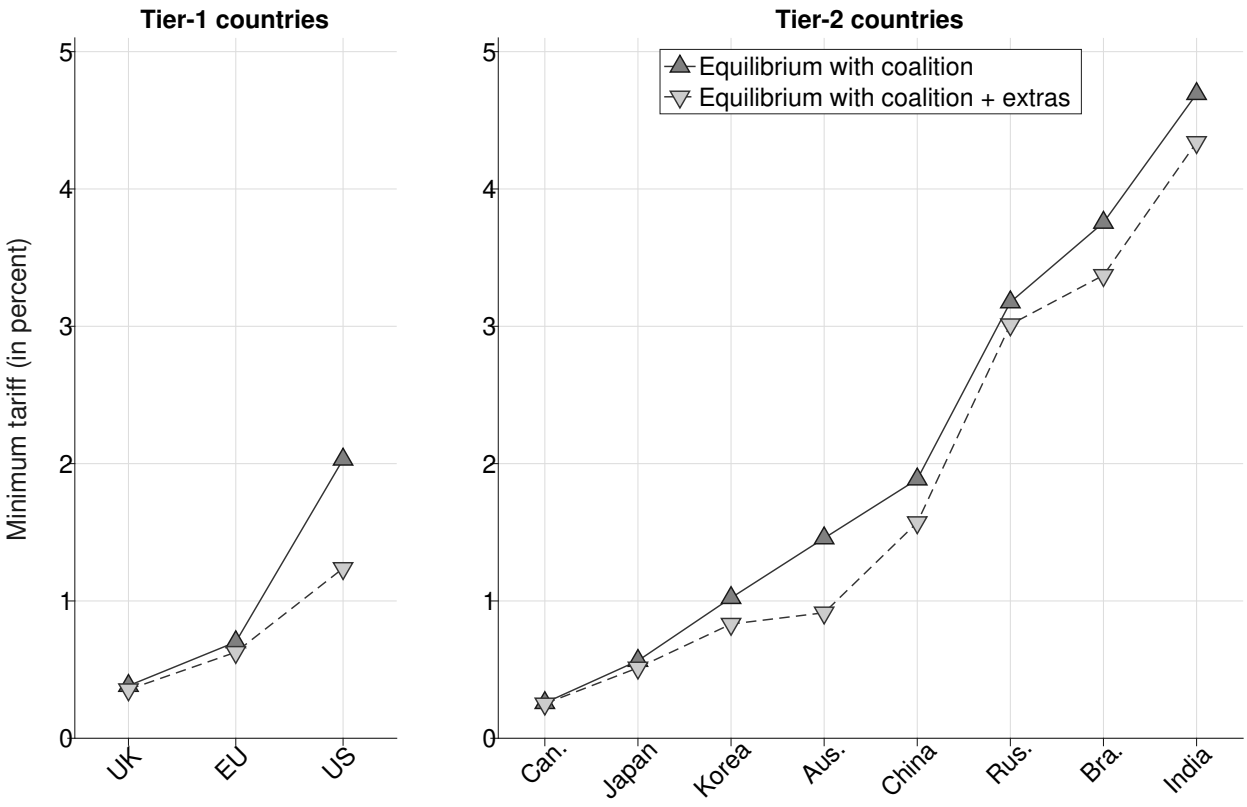


Figure 6: Minimum tariff from participating members of stable Tier 1 coalition to support the equilibrium ($\alpha = 0.46$). Dark-grey triangles assume penalties imposed by US, EU, and UK. Light-grey (upside-down) triangles Japan, UK, and Australia also impose tariffs.

to achieve the cooperative outcome. The right panel shows how big tariffs need to be against the Tier 2 countries. In each case, there are two lines, which correspond to two different sets of countries imposing the tariffs. In the first case (dark-grey, upward-pointed triangles) punishments are imposed by the stable first tier—the US, EU, and UK. In the second case (light-grey, downward-pointed triangles) punishments are imposed by the stable first tier plus Japan, Canada, and Australia, which I refer to as "extras". The reason for considering the impact of adding the extras, is that one might consider including other high income countries to help in meting our punishments. Since I want the club to remain stable, I assume that the extras participate in meeting out the punishments but are still held to the abatement obligation that applies to other Tier 2 countries. Since tariffs are not actually imposed in equilibrium, including these countries in the set of punishing countries does not appreciably change their incentive to go along with the agreement.

Figure 6 shows that the equilibrium outcome can be supported with relatively low tariffs on the 11 high-GDP economies in the model. This finding provides evidence that the calibration for the endogenous match rate model is plausible since the equilibrium match rate of 45.6 percent for the US-EU-UK coalition can be supported with relatively low tariffs. Among the Tier 1 countries, the necessary tariff is largest for the US, equalling 2.0 percent if penalties are meted out by the EU and UK alone. For all Tier 2 countries, it is below 5 percent, and for all Tier 2 countries except Russia,

Brazil, and India, it is below 2.0 percent. Among all countries, the impact of expanding the punishing-country set to include Japan, Canada, and Australia is largest for the US and Australia. The figure also shows that adding the extras to the set of countries imposing tariffs has only a modest impact on the minimum tariff for most countries, though it could have an important effect for the US and Australia.

In addition to computing minimum tariffs to support the baseline equilibrium, I also back out the minimum tariffs to support two more-ambitious alternatives. The first alternative has the stable first tier behave cooperatively with a match rate of 67 percent—the minimum match rate that achieves the efficient level of global abatement. The second has the first tier behave non-cooperatively, while increasing the match rate to 100 percent. The results for these more ambitious cases are shown in Figure 10, though again to save space, I present this figure in Appendix A.10. The analysis presented in Figure 10 suggests two main findings. First, the minimum tariffs to support each of these more ambitious cases is roughly the same for the Tier 2 countries, though much higher for the Tier 1 countries, since when the Tier 1 countries interact non-cooperatively, the minimum tariff against each other is zero. Second, for the Tier 2 countries, the minimum tariffs are higher than in the equilibrium case (Figure 6) by roughly a factor of four.

5.4 Sensitivity analysis

Next, I consider how Tier 1 stability (Figure 7) and equilibrium abatement (Figure 8) change when the parameters b and γ vary from the baseline calibration in which $b = 1.0$ and $\gamma = \$55$ per ton C02. I focus on these parameters because both have a significant effect on the results and both are highly uncertain.

Figure 7 shows how the stable Tier 1 set varies as b and γ vary. For each tile in the figure, I solve for the set of stable coalitions for the b and γ values that align with the mid-point of the tile. The tile color and the presence or absence of a dot indicate the set of stable coalitions. Dark-grey indicates the absence of a stable coalition, so all countries play Nash. Mid-grey (without a dot) indicates the US-EU coalition is uniquely stable. Light-grey (without a dot) indicates the US-EU-UK coalition is uniquely stable. A dot on top of a colored tile indicates that a second coalition is stable, and in each case, the second coalition consists of the EU, UK, and Australia. For the parameters in the figure, there is no instance with more than two stable coalitions, and the stable coalition is unique in 88 percent of cases.

While it is possible to have more than one stable coalition, one might expect the coalition with larger countries to be more likely. In instances in which the US-EU-UK and the EU-UK-Australia are both stable, Australia is effectively keeping the US out when it chooses to join. By doing so, the amount of global abatement falls since the amount of trading prowess (hence the equilibrium match rate) is substantially lower without the US. In this case, Australia would be better off staying out and letting the alternate coalition form, but it doesn't consider this possibility because the considered version of coalition stability only allows for unilateral deviations. In instances in which the US-EU and the EU-UK-Australia are both stable, the same logic applies to the UK and Australia together.

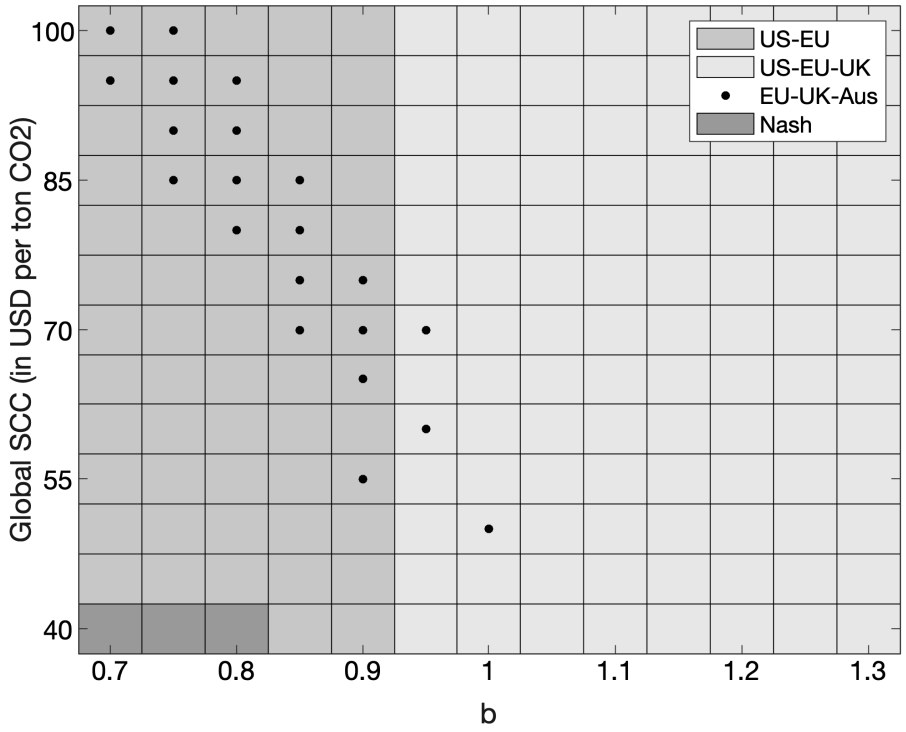


Figure 7: Map of stable Tier 1 coalitions. Black dots indicate cases in which two coalitions are stable: the first indicated by the shaded color and the second consisting of the EU, UK, and Australia. The dark-grey shade ("Nash") indicates cases in which no coalition is stable. The other stable coalitions are indicated by tile color, as explained in the legend.

A further reason to expect the coalition with bigger countries is because the match rate model in Eq. 12 likely understates the trading prowess of large countries. Modifying the match rate model to better reflect the strength of large countries would increase the tendency toward a unique equilibrium with large countries.

Next, in Figure 8, I show how the abatement varies across the parameter space. I assume that whenever there is more than one stable coalition, the larger one in terms of combined GDP prevails. Thus, I assume that the US-EU and US-EU-UK coalitions prevail in instances in which the EU-UK-Australia coalition is also stable.[29] Since the globally-efficient abatement level varies with the global SCC, I plot aggregate abatement as a fraction of the efficient level.

While the underlying coalition for each parameter combination is not labeled in Figure 8, it can be inferred by comparing the figure with Figure 7, which spans the same parameter set. Aggregate abatement jumps sharply when moving across regions in which the stable coalition changes. The three dark squares in the lower left corner correspond to the case in which the Nash equilibrium prevails. In this case, aggregate abatement is around 10 percent. For the remainder of the parameter space, the

[29]The assumption is not important since it only reflects the summary information I choose to display in the figure and this could be easily changed.

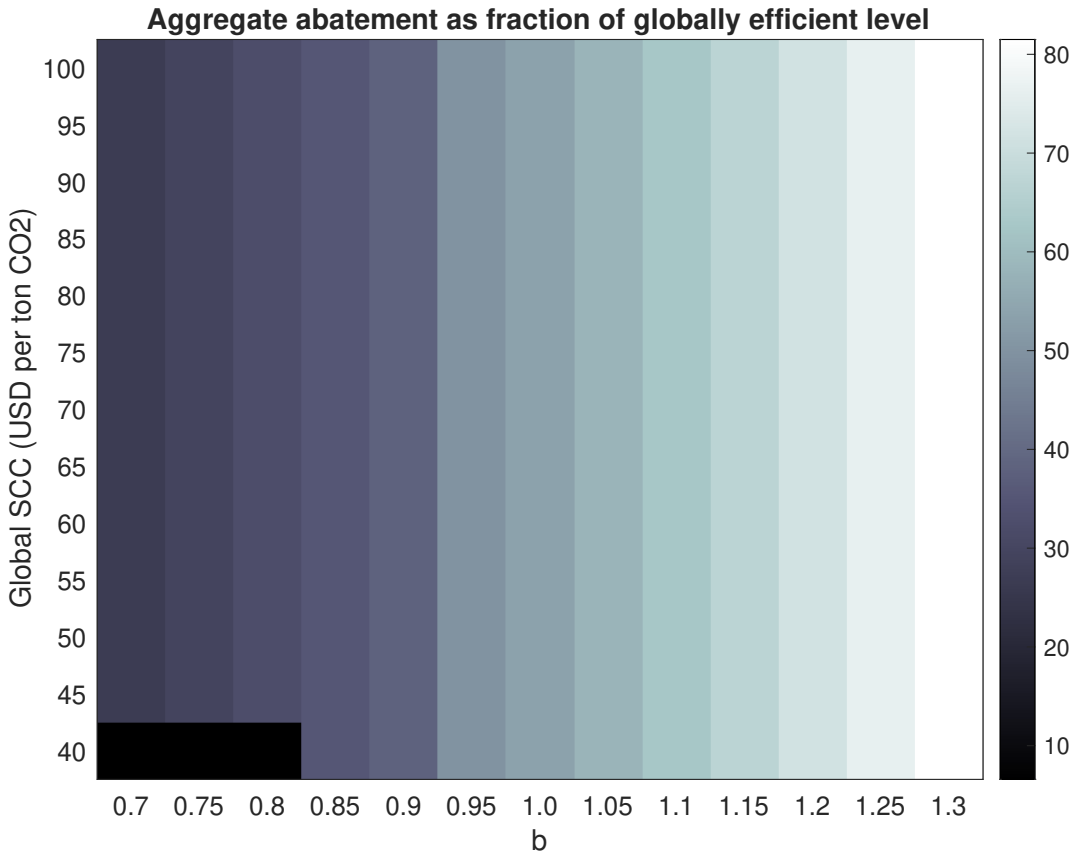


Figure 8: Heatmap shows global abatement under equilibrium TCC as a fraction of the global optimum. Horizontal axis varies b. Vertical axis varies γ. Abatement reflects optimal coalition policy for the largest stable coalition.

US-EU coalition obtains when $b < 0.95$, and the US-EU-UK coalition obtains when $b \geq 0.95$. In both case Holding fixed b and γ, adding the UK to the coalition increases aggregate abatement by about 11 percent.

Another result apparent in the figure is that aggregate abatement as a fraction of the efficient level does not change with the global SCC (γ), holding fixed b. In particular, the heatmap color doesn't change in the vertical direction for regions in which the coalition is fixed. Higher γ (i.e., higher global climate damages) naturally increases the efficient global abatement rate, but aggregate abatement increases in proportion, so the ratio is constant. This result stems from the assumption of quadratic abatement costs in the quantitative model. In particular, quadratic abatement costs imply a linear relationship between the carbon price and the abatement rate (see Section 2.1), while an increase in the global SCC implies a proportionate increase in both the coalition carbon price and the efficient carbon price.

6 Conclusion

Negotiations for a top-down global climate agreement have reached an impasse, with nations embracing the voluntary framework established by the 2015 Paris Agreement. This shift towards a non-coercive approach in international climate policy means that countries undertake abatement measures only to the extent that they align with national interests. Given that each nation accounts for only a small portion of global climate damages, collective abatement efforts fall far short of what is required for global efficiency. This paper introduces a novel approach for achieving coordinated global climate action. A relatively small coalition of economically powerful economies (the first tier) takes the lead. Rather than focus on reducing its own emissions, however, the coalition uses its combined leverage in trade negotiations to induce the rest of the world (the second tier) to match its carbon price at a less than one-for-one rate. The linkage enhances abatement incentives for the first tier, while taking advantage of the cheapest abatement opportunities in the second tier. In the calibrated model, a unique stable coalition emerges. The US, EU, and UK lead an effort that achieves over 50 percent of the efficient level of global abatement. Fairness considerations are outside the model, yet the equilibrium outcome aligns with intuitive notions of fairness—the economies with the greatest historical responsibility for carbon emissions contribute the most.

References

Al Khourdajie, Alaa, and Michael Finus. 2020. "Measures to enhance the effectiveness of international climate agreements: The case of border carbon adjustments." *European Economic Review*, 124: 103405.

Barrage, Lint, and William D Nordhaus. 2023. "Policies, Projections, and the Social Cost of Carbon: Results from the DICE-2023 Model." National Bureau of Economic Research.

Barrett, Scott. 1994. "Self-enforcing international environmental agreements." *Oxford economic papers*, 46(Supplement_1): 878–894.

Barrett, Scott. 1997. "The strategy of trade sanctions in international environmental agreements." *Resource and Energy Economics*, 19(4): 345–361.

Barrett, Scott. 2003. *Environment and statecraft: The strategy of environmental treaty-making: The strategy of environmental treaty-making.* OUP Oxford.

Böhringer, Christoph, and Thomas Rutherford. 2017. "Paris after Trump: An inconvenient insight."

Böhringer, Christoph, Carolyn Fischer, Knut Einar Rosendahl, and Thomas J Foxon. 2022. "Potential impacts and challenges of border carbon adjustments." *Nature Climate Change*, 12(1): 22–29.

Böhringer, Christoph, Edward J Balistreri, and Thomas F Rutherford. 2012. "The role of border carbon adjustment in unilateral climate policy: Overview of an Energy Modeling Forum study (EMF 29)." *Energy Economics*, 34: S97–S110.

Böhringer, Christoph, Jared C Carbone, and Thomas F Rutherford. 2016. "The strategic value of carbon tariffs." *American Economic Journal: Economic Policy*, 8(1): 28–51.

Branger, Frédéric, and Philippe Quirion. 2014. "Would border carbon adjustments prevent carbon leakage and heavy industry competitiveness losses? Insights from a meta-analysis of recent economic studies." *Ecological Economics*, 99: 29–39.

Carraro, Carlo. 2003. *The endogenous formation of economic coalitions.* Edward Elgar Publishing.

Carraro, Carlo, and Domenico Siniscalco. 1993. "Strategies for the international protection of the environment." *Journal of public Economics*, 52(3): 309–328.

Carraro, Carlo, and Domenico Siniscalco. 1998. "International institutions and environmental policy: international environmental agreements: incentives and political economy." *European economic review*, 42(3-5): 561–572.

Carraro, Carlo, Domenico Siniscalco, et al. 1995. "R&D cooperation and the stability of international environmental agreements." CEPR Discussion Papers.

d'Aspremont, Claude, Alexis Jacquemin, Jean Jaskold Gabszewicz, and John A Weymark. 1983. "On the stability of collusive price leadership." *Canadian Journal of economics*, 17–25.

Dubey, Amitrajeet, Michael Keen, and E. Marianna Taxell. 2014. "The impact of border carbon adjustments on carbon leakage: a general equilibrium analysis." *The Economic Journal*, 124(579): F165–F190.

Falkner, Robert. 2016. "A minilateral solution for global climate change? On bargaining efficiency, club benefits, and international legitimacy." *Perspectives on Politics*, 14(1): 87–101.

Farrokhi, Farid, and Ahmad Lashkaripour. 2020. "Trade, Firm-Delocation, and Optimal Climate Policy." Mimeo.

Felder, Stefan, and Thomas F Rutherford. 1993. "Unilateral CO2 reductions and carbon leakage: the consequences of international trade in oil and basic materials." *Journal of Environmental Economics and management*, 25(2): 162–176.

Fischer, Carolyn, and Alan K Fox. 2012. "Comparing policies to combat emissions leakage: Border carbon adjustments versus rebates." *Journal of Environmental Economics and management*, 64(2): 199–216.

Folmer, Henk, Pierre V Mouche, and Shannon Ragland. 1993. "Interconnected games and international environmental problems." *Environmental and Resource Economics*, 3: 313–335.

Golosov, Mikhail, John Hassler, Per Krusell, and Aleh Tsyvinski. 2014. "Optimal taxes on fossil fuel in general equilibrium." *Econometrica*, 82(1): 41–88.

Group, Interagency Working, et al. 2021. "Technical support document: social cost of carbon, methane, and nitrous oxide interim estimates under executive order 13990." Tech. rep., White House. URL https://www. whitehouse. gov/wp-content/uploads ....

Gwatipedza, Johnson, and Edward B Barbier. 2014. "Environmental regulation of a global pollution externality in a bilateral trade framework: The case of global warming, China and the US." *Economics*, 8(1).

Hagen, Achim, and Jan Schneider. 2021. "Trade sanctions and the stability of climate coalitions." *Journal of Environmental Economics and Management*, 109: 102504.

Helm, Carsten, and Robert C Schmidt. 2015. "Climate cooperation with technology investments and border carbon adjustment." *European Economic Review*, 75: 112–130.

Hoel, Michael, and Kerstin Schneider. 1997. "Incentives to participate in an international environmental agreement." *Environmental and Resource economics*, 9: 153–170.

Iverson, Terrence. 2022. "Advancing Global Carbon Abatement with a Two-Tier Climate Club." *CESifo Working Paper No. 9831*.

Kotchen, Matthew J. 2018. "Which social cost of carbon? A theoretical perspective." *Journal of the Association of Environmental and Resource Economists*, 5(3): 673–694.

Lessmann, Kai, Robert Marschinski, and Ottmar Edenhofer. 2009. "The effects of tariffs on coalition formation in a dynamic global warming game." *Economic Modelling*, 26(3): 641–649.

Marrouch, Walid, Amrita Ray Chaudhuri, et al. 2016. "International environmental agreements: doomed to fail or destined to succeed? A review of the literature." *International Review of Environmental and Resource Economics*, 9(3–4): 245–319.

Naim, Moises. 2009. "Minilateralism." *Foreign policy*, , (173): 136.

Nordhaus, William. 2015. "Climate clubs: Overcoming free-riding in international climate policy." *American Economic Review*, 105(4): 1339–70.

Nordhaus, William D. 2008. "Six: The Economics Of Participation." *A Question of Balance: Weighing the Options on Global Warming Policies*, 116–122. Yale University Press.

Nordhaus, William D. 2017. "Revisiting the social cost of carbon." *Proceedings of the National Academy of Sciences*, 114(7): 1518–1523.

Nordhaus, William D. 2021. "Climate Club Futures: On the Effectiveness of Future Climate Clubs."

Ossa, Ralph. 2014. "Trade wars and trade talks with data." *American Economic Review*, 104(12): 4104–4146.

Parry, Ian, Simon Black, and James Roaf. 2021. "Proposal for an international carbon price floor among large emitters."

Petrakis, Emmanuel, and Anastasios Xepapadeas. 1996. "Environmental consciousness and moral hazard in international agreements to protect the environment." *Journal of Public Economics*, 60(1): 95–110.

Rubio, Santiago J, and Alistair Ulph. 2007. "An infinite-horizon model of dynamic membership of international environmental agreements." *Journal of Environmental Economics and Management*, 54(3): 296–310.

Tagliapietra, Simone, and Guntram B Wolff. 2021. "Form a climate club: United States, European Union and China."

Traeger, Christian P. 2023. "Ace—analytic climate economy." *American Economic Journal: Economic Policy*, 15(3): 372–406.

Weyant, John P. 1999. "The costs of the Kyoto Protocol: a multi-model evaluation." *Energy Journal*.

A Appendix

A.1 The abatement rate as a function of the carbon price

Given carbon price τ_i, country i will abate until the marginal cost of abatement (denominated in emission units) equals the carbon price. To convert abatement from the unitless fraction μ to units of emission reduction, define

$$\hat{\mu}_i = \mu_i E_i,$$

so that $\hat{\mu}_i$ is abatement in i denominated in units of emissions reduced. Abatement costs rewritten as a function of abatement measured in units of emission reduction are:

$$\psi(\hat{\mu}_i) = Q_i \theta_1 \left(\frac{\hat{\mu}_i}{E_i} \right)^{\theta_2},$$

where Q_i is output in region i.

Marginal abatement costs are

$$\frac{\partial \psi_i}{\partial \hat{\mu}_i} = \theta_2 Q_i \theta_1 \left(\frac{\hat{\mu}_i}{E_i} \right)^{\theta_2 - 1} \frac{1}{E_i} \tag{30}$$

$$= \theta_1 \theta_2 \frac{Q_i}{E_i} \left(\frac{\hat{\mu}_i}{E_i} \right)^{\theta_2 - 1} \tag{31}$$

$$= \frac{\theta_1 \theta_2}{\sigma_i} \mu_i^{\theta_2 - 1}, \tag{32}$$

where $\sigma_i \equiv \frac{E_i}{Q_i}$ is the emissions intensity of output (carbon emissions over GDP) and $\mu_i = \frac{\hat{\mu}_i}{E_i}$.

Next, set marginal abatement costs equal to the carbon price:

$$\frac{\theta_1 \theta_2}{\sigma_i} \mu_i^{\theta_2 - 1} = \tau_i. \tag{33}$$

Solving for μ_i gives:

$$\mu_i = \left[\frac{\sigma_i}{\theta_1 \theta_2} \right]^{\frac{1}{\theta_2 - 1}} \tau_i^{\frac{1}{\theta_2 - 1}} \equiv G_i(\tau_i). \tag{34}$$

I often write

$$\mu_i = a_i (tau_i)^b, \tag{35}$$

where

$$a_i = \left[\frac{\sigma_i}{\theta_1 \theta_2} \right]^{\frac{1}{\theta_2 - 1}} \tag{36}$$

and

$$b = \frac{1}{\theta_2 - 1}. \tag{37}$$

If $\theta_2 = 2$, then $b = 1$, so the abatement rate is linear in the carbon price:

$$\mu_i = \left[\frac{\sigma_i}{\theta_1 \theta_2} \right] \tau_i.$$

Importantly, the function $G_i(\cdot)$ is independent of region size as long as abatement within the region is allocated efficiently across subsidiary jurisdictions.

A.2 Alternative statement of the optimal policy problem

To capture intermediate cases between the cooperative and noncooperative scenarios, we can alternatively assume that the first tier takes α and ω_1 as given, then chooses carbon price targets to maximize the Tier 1 surplus subject to the constraint that no country has an incentive to deviate from the agreement. It solves:

$$\max_{\{\hat{\tau}_i \geq 0\}_{i\in\Omega}, \{\hat{\tau}_j \geq 0\}_{j\notin\Omega}} \sum_{i\in\Omega} \Pi_i(\hat{\tau}_i, \{\hat{\tau}_j\}_{j\neq i}; \hat{\tau}_i, \omega_1) \tag{$\boldsymbol{P3}$}$$

subject to

$$\Pi_i(\hat{\tau}_i, \{\hat{\tau}_j\}_{j\neq i}; \hat{\tau}_i, \omega_1) \geq \max_{\tau_i \geq 0} \Pi_i(\tau_i, \{\hat{\tau}_j\}_{j\neq i}; \hat{\tau}_i, \omega_1), \text{ for all } i \in \Omega,$$

$$\Pi_j(\hat{\tau}_j, \{\hat{\tau}_k\}_{k\neq j}; \hat{\tau}_j, \omega_2) \geq \max_{\tau_j \geq 0} \Pi_j(\tau_j, \{\hat{\tau}_k\}_{k\neq j}; \hat{\tau}_j, \omega_2), \text{ for all } \notin \Omega,$$

and

$$\hat{\tau}_j \geq \alpha \tau^{AVG}, \text{ for all } j \notin \Omega,$$

where $\tau^{AVG} = \sum_{i\in\Omega} \hat{\phi}_i^E \hat{\tau}_i$ and the payoff function $\Pi_i(\cdot)$ is defined in Eq. 10. The first constraint ensures no Tier 1 country can benefit by violating the agreement, the second constraint ensures no Tier 2 country can benefit by violating the agreement, and the last constraint ensures the Tier 2 targets oblige the match rate.

Under this interpretation, we would treat ω_1 as something that the first tier takes as being politically constrained. Given the value for this penalty, the first tier then solves for the target carbon prices that maximize coalition surplus subject to the constraint that no countries want to deviate, including the Tier 1 countries, each of whom faces a conditional trade penalty equal to fraction ω_1 of their GDP.

The non-cooperative policy problem is the special case of this problem when $\omega_1 = 0$. The cooperative policy problem is the special case of this problem when ω_1 is set at the minimum level needed to get all Tier 1 countries to comply with the cooperative carbon price.

In a prior version of the paper, I used this specification throughout, but I decided to focus on the cooperative and non-cooperative cases to make the paper easier to read.

A.3 Proof of Proposition 1

The solution to Problem ***P1*** has the first tier adopt a harmonized carbon price, which I denote τ^C. Suppose instead that the solution to the cooperative problem involved at least one $i \in \Omega$ with a different target carbon price than the other countries. Then by shifting one unit of abatement from a higher marginal abatement cost country to a lower marginal abatement cost country (through appropriate adjustments in carbon prices) then one would lower Tier 1 costs without changing Tier 1 benefits. This contradicts the supposition that a non-harmonized carbon prices were optimal.

Also, since $\theta_2 = 2$, while emissions intensity is σ_Ω for $i \in \Omega$ and σ_R for $j \notin \Omega$, we have $\mu_i = \frac{\sigma_\Omega}{2\theta_1}\tau_i$ for $i \in \Omega$ and $\mu_j = \frac{\sigma_R}{2\theta_1}\tau_j$ for $j \notin \Omega$. Recalling also the proposition assumption that Tier 2 countries don't abate without coercion, the non-cooperative policy problem can be written as follows:

$$\max_{\{\tau^C \geq 0\}_{i\in\Omega}} \sum_{i\in\Omega} \left[\gamma_i E \mu(\{\tau^C; \alpha) - C_i(\tau^C)\right] \tag{38}$$

where

$$\mu(\{\tau^C;\alpha) = \phi_\Omega^E \frac{\sigma_\Omega}{2\theta_1}\tau^C + (1-\phi_\Omega^E)\frac{\sigma_R}{2\theta_1}\alpha\tau^C.$$

The first-order condition gives

$$\gamma_\Omega E\big[\phi_\Omega^E \frac{\sigma_\Omega}{2\theta_1} + (1-\phi_\Omega^E)\frac{\sigma_R}{2\theta_1}\alpha\big] = \sum_{i\in\Omega} C_i'(\tau^C)$$

From Eq. 5, $C_i(\tau^C) = Q_i\theta_1\big(\frac{\sigma_\Omega}{2\theta_1}\big)^2(\tau^C)^2$, so

$$C_i'(\tau_i) = \frac{1}{2\theta_1}Q_i\sigma_\Omega^2\tau^C$$

Combining and simplifying gives

$$\gamma_\Omega E\big[\phi_\Omega^E\sigma_\Omega + (1-\phi_\Omega^E)\sigma_R\alpha\big] = Q_\Omega\sigma_\Omega^2\tau^C.$$

Simplifying further gives

$$\tau^C = \gamma_\Omega\big[1 + \frac{1-\phi_\Omega^E}{\phi_\Omega^E}\frac{\sigma_R}{\sigma_\Omega}\alpha\big]$$

To complete the proof, I show that

$$\frac{\mathrm{d}a_G}{\mathrm{d}a_C} = 1 + \frac{1-\phi_C^E}{\phi_C^E}\frac{\sigma_R}{\sigma_C}\alpha.$$

Derivative interpretation Let μ_Ω and μ_R denote abatement in percent for the coalition and non-coalition regions, and let a_Ω and a_R denote the corresponding abatement in units of emission reduction. By definition,

$$a_\Omega = \mu_\Omega\phi_\Omega^E E,$$

where E is global carbon emissions. Similarly,

$$a_R = \mu_R(1-\phi_\Omega^E)E.$$

We also have, for arbitrary τ, that

$$\mu_\Omega(\tau) = \frac{\sigma_\Omega}{2\theta_1}\tau$$

and

$$\mu_R(\tau) = \frac{\sigma_R}{2\theta_1}\tau.$$

The TCC policy requires $\tau^R = \alpha\tau^C$, so

$$\begin{aligned}\mu^R &= \frac{\sigma_R}{2\theta_1}\alpha\tau^C \\ &= \alpha\frac{\sigma^R}{\sigma^\Omega}\frac{\sigma^\Omega}{2\theta_1}\tau^C \\ &= \alpha\frac{\sigma^R}{\sigma^\Omega}\mu^\Omega\end{aligned}$$

Thus,

$$\begin{aligned} a_R &= (1-\phi_\Omega^E)E\mu_R \\ &= (1-\phi_\Omega^E)E\alpha\frac{\sigma_R}{\sigma_\Omega}\mu_\Omega \\ &= \frac{1-\phi_\Omega^E}{\phi_\Omega^E}\mu_\Omega\phi_\Omega^E E\alpha\frac{\sigma_R}{\sigma_\Omega} \\ &= \frac{1-\phi_\Omega^E}{\phi_\Omega^E}\alpha\frac{\sigma_R}{\sigma_\Omega}a_\Omega \end{aligned}$$

It follows that global abatement in emission units is

$$\begin{aligned} a_G &= a_\Omega + a_R \\ &= \left[1+\alpha\frac{\sigma_R}{\sigma_\Omega}\frac{1-\phi_\Omega^E}{\phi_\Omega^E}\right]a_\Omega. \end{aligned}$$

Hence,

$$\frac{\mathrm{d}a_G}{\mathrm{d}a_\Omega} = 1+\alpha\frac{\sigma_R}{\sigma_\Omega}\frac{1-\phi_\Omega^E}{\phi_\Omega^E}.$$

A.4 Proof of Proposition 2

Proposition 2 is a special case of a more general proposition, which I state and prove here. In the more general proposition, the ROW is broken into n sub-regions, where ϕ_j denotes the size of sub-region j, and $0 \leq \alpha_j \leq 1$ is the match rate in j. Given coalition carbon price τ_C, the carbon price in ROW sub-region j is

$$\tau_j = \alpha_j \tau_C. \tag{39}$$

The model coincides with Nordhaus (2008) when $\alpha_j = 0$ for all j. As in Nordhaus (2008), countries differ in size but are otherwise homogeneous, including abatement opportunities that scale proportionally with size (see Appendix A.1). To condense notation, I let $i = 0$ index the coalition, while $i = 1, \ldots, n$ indexes the sub-regions in the ROW. I also define $\alpha_0 = 1$, $\boldsymbol{\alpha} = (\alpha_0, \alpha_1, \ldots, \alpha_n)$, and $\boldsymbol{\phi} = (\phi_0, \phi_1, \ldots, \phi_n)$, where $\sum_{i=0}^n \phi_i = 1$.

Proposition 6. *If a coalition comprising fraction ϕ_0 of the global economy implements climate policy with a harmonized carbon price and the n sub-regions in the ROW match the coalition carbon price at rates indicated by $\boldsymbol{\alpha}$, then global abatement costs can be expressed as the product of a cost penalty $P(\boldsymbol{\phi}, \boldsymbol{\alpha})$ and the cost-minimizing global abatement cost function $\Psi^*(\mu) = Q\theta_1\mu^{\theta_2}$:*

$$\Psi(\mu; \boldsymbol{\phi}, \boldsymbol{\alpha}) = P(\boldsymbol{\phi}, \boldsymbol{\alpha}) \times \Psi^*(\mu).$$

The cost penalty is given by

$$P(\boldsymbol{\phi}, \boldsymbol{\alpha}) = \frac{\phi_0 + \sum_{i=1}^n \phi_i \alpha_i^{\theta_2/(\theta_2-1)}}{\left(\phi_0 + \sum_{j=1}^n \phi_j \alpha_j^{1/(\theta_2-1)}\right)^{\theta_2}}. \tag{40}$$

For the special case with a single non-coalition region ($n = 1$):

$$P(\phi_0, \alpha) = \frac{\phi_0 + (1-\phi_0)\alpha^{\theta_2/(\theta_2-1)}}{\left[\phi_0 + (1-\phi_0)\alpha^{1/(\theta_2-1)}\right]^{\theta_2}}. \tag{41}$$

Proof. Recall that index $j = 0$ denotes the coalition region. By assumption,

$$\tau_j = \alpha_j \tau_0, \quad \text{for j=0,...,n.}$$

It follows that the global abatement rate is

$$\begin{aligned}\mu &= \phi_0 G(\tau_0) + \phi_1 G(\alpha_1 \tau_0) + \ldots + \phi_n G(\alpha_n \tau_0) \\ &= G(\tau_0) \sum_{i=0}^{n} \phi_i \alpha_i^{1/(\theta_2-1)}.\end{aligned}$$

Thus,

$$G(\tau_0) = \Gamma \mu,$$

where

$$\Gamma = \frac{1}{\sum_{i=0}^{n} \phi_i \alpha_i^{1/(\theta_2-1)}}.$$

Global abatement costs are

$$\Psi = \sum_{j=0}^{n} \Psi_j \tag{42}$$

$$= \sum_{j=0}^{n} \phi_j Q \theta_1 G(\alpha_j \tau_0)^{\theta_2} \tag{43}$$

$$= Q\theta_1 G(\tau_0)^{\theta_2} \sum_{j=0}^{n} \phi_j \alpha_j^{\frac{\theta_2}{\theta_2-1}} \tag{44}$$

$$= Q\theta_1 \mu^{\theta_2} \cdot \Gamma^{\theta_2} \cdot \sum_{j=0}^{n} \phi_j \alpha_j^{\frac{\theta_2}{\theta_2-1}} \tag{45}$$

$$= \frac{\sum_{i=0}^{n} \phi_i \alpha_i^{\theta_2/(\theta_2-1)}}{\left(\sum_{j=0}^{n} \phi_j \alpha_j^{1/(\theta_2-1)}\right)^{\theta_2}} \cdot Q\theta_1 \mu^{\theta_2} \tag{46}$$

Thus, global abatement costs increase by the multiplicative penalty

$$P(\boldsymbol{\phi}, \boldsymbol{\alpha}) = \frac{\sum_{i=0}^{n} \phi_i \alpha_i^{\theta_2/(\theta_2-1)}}{\left(\sum_{j=0}^{n} \phi_j \alpha_j^{1/(\theta_2-1)}\right)^{\theta_2}}.$$

Suppose $n = 1$ and $\alpha_1 \equiv \alpha$. Then

$$\sum_{i=0}^{n} \phi_i \alpha_i^{\theta_2/(\theta_2-1)} = \phi_0 \alpha_0^{\theta_2/(\theta_2-1)} + \phi_1 \alpha_1^{\theta_2/(\theta_2-1)} \tag{47}$$

$$= \phi + (1-\phi)\alpha_1^{\theta_2/(\theta_2-1)}. \tag{48}$$

Equation 41 follows.

If $\alpha = 0$,

$$P(\phi, 0) = \frac{\phi + (1-\phi)0}{(\phi + (1-\phi)0)^{\theta_2}} = \frac{\phi}{\phi^{\theta_2}} = \phi^{1-\theta_2},$$

as in Nordhaus (2008).

□

First derivative of the penalty function (one non-coalition region) Define $\gamma = \frac{1}{\theta_2 - 1}$ and assume $\theta_2 > 1$ (so $\gamma > 0$). Then the penalty function in Equation 41 can be written

$$P(\phi, \alpha) = \frac{f(\phi, \alpha)}{g(\phi, \alpha)},$$

where $f(\phi, \alpha) = \phi + (1 - \phi)\alpha^{\theta_2 \gamma} > 0$ and $g(\phi, \alpha) = [\phi + (1 - \phi)\alpha^{\gamma}]^{\theta_2} > 0$.

Letting subscripts on functions denote partial derivatives, we have

$$f_\alpha = (1 - \phi)\theta_2 \gamma \alpha^{\theta_2 \gamma - 1}$$

and

$$g_\alpha = \theta_2 [\phi + (1 - \phi)\alpha^{\gamma}]^{\theta_2 - 1} \gamma (1 - \phi)\alpha^{\gamma - 1}.$$

I want to show

$$P_\alpha = \frac{f_\alpha g - g_\alpha f}{g^2} < 0.$$

This is true if and only if

$$f_\alpha g - g_\alpha f < 0,$$

if and only if

$$(1 - \phi)\theta_2 \gamma \alpha^{\theta_2 \gamma - 1} [\phi + (1 - \phi)\alpha^{\gamma}]^{\theta_2} < \theta_2 [\phi + (1 - \phi)\alpha^{\gamma}]^{\theta_2 - 1} \gamma (1 - \phi)\alpha^{\gamma - 1} [\phi + (1 - \phi)\alpha^{\theta_2 \gamma}]. \quad (49)$$

Substituting shows that $\theta_2 \gamma - 1 = \gamma$. We use this to substitute for the exponent on α, then divide both sides by the common positive factor $(1 - \phi)\theta_2 \gamma \alpha^{\gamma} [\phi + (1 - \phi)\alpha^{\gamma}]^{\theta_2 - 1}$. Thus, (49) holds if and only if

$$[\phi + (1 - \phi)\alpha^{\gamma}] < \alpha^{-1} [\phi + (1 - \phi)\alpha^{\theta_2 \gamma}].$$

if and only if

$$\alpha\phi + (1 - \phi)\alpha^{\gamma + 1} < \phi + (1 - \phi)\alpha^{\theta_2 \gamma}.$$

But $\gamma + 1 = \theta_2 \gamma$, so this is true if and only if

$$\alpha\phi < \phi,$$

which is true as long as $\alpha < 1$ as assumed.

A.5 Proof of Proposition 3

By assumption, $\theta_2 = 2$, while emissions intensity is σ_Ω for $i \in \Omega$ and σ_R for $j \notin \Omega$. Thus, $\mu_i = \frac{\sigma_\Omega}{2\theta_1} \tau_i$ for $i \in \Omega$ and $\mu_j = \frac{\sigma_R}{2\theta_1} \tau_j$ for $j \notin \Omega$. Using this together with the assumption that Tier 2 countries don't abate without coercion, the non-cooperative policy problem can be written as follows:

$$\max_{\tau_i \geq 0} \left[\gamma_i E\mu(\tau_i, \{\hat{\tau}_j\}_{j \in \Omega, j \neq i}; \alpha) - C_i(\tau_i) \right], \text{ for each } i \in \Omega,$$

where

$$\mu(\tau_i, \{\hat{\tau}_j\}_{j \in \Omega, j \neq i}; \alpha) = \phi_i^E \frac{\sigma_\Omega}{2\theta_1} \tau_i + \sum_{j \in \Omega, j \neq i} \phi_j^E \frac{\sigma_\Omega}{2\theta_1} \hat{\tau}_j + \sum_{k \notin \Omega} \phi_k^E \frac{\sigma_R}{2\theta_1} \left[\alpha [\hat{\phi}_i^E \tau_i + \sum_{j \in \Omega, j \neq i} \hat{\phi}_j^E \hat{\tau}_j \right]$$

The first-order condition for an interior solution is

$$\gamma_i E\left[\phi_i^E \frac{\sigma_\Omega}{2\theta_1} + \sum_{k\notin\Omega} \phi_k^E \frac{\sigma_R}{2\theta_1}\alpha\hat{\phi}_i^E\right] = C_i'(\tau_i)$$

From Eq. 5, $C_i(\tau_i) = Q_i\theta_1\left(\frac{\sigma_\Omega}{2\theta_1}\right)^2(\tau_i)^2$, so

$$C_i'(\tau_i) = \frac{1}{2\theta_1}Q_i\sigma_\Omega^2\tau_i$$

Combining and simplifying gives

$$\gamma_i E\left[\phi_i^E\sigma_\Omega + \alpha\sigma_R\hat{\phi}_i^E\sum_{k\notin\Omega}\phi_k^E\right] = Q_i\sigma_\Omega^2\tau_i,$$

which simplifies to

$$\tau_i = \gamma_i\left[1 + \alpha\frac{1-\phi_C^E}{\phi_C^E}\frac{\sigma_R}{\sigma_C}\right].$$

Proof of the derivative interpretation is provided in the proof of Proposition 1. The result follows.

A.6 Proof of Proposition 4

Global abatement in the non-cooperative case is given by

$$\begin{aligned}\mu^N(\alpha) &= \sum_{i\in\Omega}\phi_i^E\mu_i + (1-\phi_\Omega^E)\mu_R \\ &= \sum_{i\in\Omega}\phi_i^E a_C\tau_i^N(\alpha) + (1-\phi_\Omega^E)a_R\alpha\tau^{AVG},\end{aligned}$$

where

$$\begin{aligned}\tau^{AVG} &= \sum_{i\in\Omega}\hat{\phi}_i^E\tau_i^N \\ &= \frac{1}{\phi_\Omega^E}\sum_{i\in\Omega}\phi_i^E\gamma_i\left[1+\alpha\frac{1-\phi_\Omega^E}{\phi_\Omega^E}\frac{\sigma_R}{\sigma_\Omega}\right] \\ &= \frac{\gamma}{\phi_\Omega^E}\left[1+\alpha\frac{1-\phi_\Omega^E}{\phi_\Omega^E}\frac{\sigma_R}{\sigma_\Omega}\right]\sum_{i\in\Omega}\phi_i^E\phi_i^\gamma,\end{aligned}$$

where $\phi_i^\gamma = \frac{\gamma_i}{\gamma}$. Combining gives

$$\begin{aligned}\mu^N(\alpha) &= \sum_{i\in\Omega}\phi_i^E a_\Omega\gamma_i\left[1+\alpha\frac{1-\phi_\Omega^E}{\phi_\Omega^E}\frac{\sigma_R}{\sigma_\Omega}\right] + (1-\phi_\Omega^E)a_R\alpha\frac{\gamma}{\phi_\Omega^E}\left[1+\alpha\frac{1-\phi_\Omega^E}{\phi_\Omega^E}\frac{\sigma_R}{\sigma_\Omega}\right]\sum_{i\in\Omega}\phi_i^E\phi_i^\gamma \\ &= a_\Omega\gamma\left[1+\alpha\frac{1-\phi_\Omega^E}{\phi_\Omega^E}\frac{\sigma_R}{\sigma_\Omega}\right]\left[1+\alpha\frac{1-\phi_\Omega^E}{\phi_\Omega^E}\frac{a_R}{a_\Omega}\right]\sum_{i\in\Omega}\phi_i^E\phi_i^\gamma\end{aligned}$$

Next, do the same for $\mu^C(\alpha)$

$$\begin{aligned}\mu^C(\alpha) &= \phi_\Omega^E\mu_C + (1-\phi_\Omega^E)\mu_R \\ &= \phi_\Omega^E a_\Omega\gamma_\Omega\left[1+\alpha\frac{1-\phi_\Omega^E}{\phi_\Omega^E}\frac{\sigma_R}{\sigma_\Omega}\right] + (1-\phi_\Omega^E)a_R\alpha\gamma_\Omega\left[1+\alpha\frac{1-\phi_\Omega^E}{\phi_\Omega^E}\frac{\sigma_R}{\sigma_\Omega}\right] \\ &= a_\Omega\gamma_\Omega\phi_\Omega^E\left[1+\alpha\frac{1-\phi_\Omega^E}{\phi_\Omega^E}\frac{\sigma_R}{\sigma_\Omega}\right][1+\frac{1-\phi_\Omega^E}{\phi_\Omega^E}\frac{a_R}{a_\Omega}\alpha]\end{aligned}$$

Dividing gives

$$\frac{\mu^N(\alpha)}{\mu^C(\alpha)} = \sum_{i\in\Omega} \hat{\phi}_i^E \hat{\phi}_i{}^{\gamma},$$

where $\hat{\phi}_i^E \equiv \frac{E_i}{E_\Omega}$ is i's share of Tier 1 emissions and $\hat{\phi}_i^E \equiv \frac{\gamma_i}{\Omega}$ is i's share of Tier 1 climate damages.

If each Tier 1 country i has the same share of global damages as of global emissions (i.e., $\phi_i^\gamma = \phi_i^E$, then

$$\sum_{i\in\Omega} \hat{\phi}_i^E \hat{\phi}_i{}^{\gamma} = \sum_{i\in\Omega} \left(\hat{\phi}_i^E\right)^2,$$

which can be interpreted as the Herfindahl index of country size within the first tier. Size can be interpreted to mean either carbon emissions or marginal climate damages.

A.7 Proof of Proposition 5

The assumptions are stated in the proposition. In the quadratic case, we have

$$G(\tau) = \frac{\sigma}{\theta_1\theta_2}\tau,$$

so $a = \frac{\sigma}{\theta_1\theta_2}$ and $b = 1$. Also,

$$\mu(\tau_i, \{\tau_j\}_{j\neq i}; \alpha) = \phi_i \frac{\sigma}{\theta_1\theta_2}\tau_i + \sum_{j\neq i} \phi_j \frac{\sigma}{\theta_1\theta_2}\tau_j + (1-\phi)\frac{\sigma}{\theta_1\theta_2}\alpha\hat{\phi}_i\tau_i + (1-\phi)\frac{\sigma}{\theta_1\theta_2}\alpha\sum_{j\neq i}\hat{\phi}_j\tau_j.$$

Let $\hat{\boldsymbol{\phi}} \equiv (\hat{\phi}_1, \ldots, \hat{\phi}_n)$ denote the size distribution of countries within the coalition, where $\sum_i \hat{\phi}_i = 1$.

I refer to the maximizers of the optimal coalition policy problem as $\{\tau_i^*(\alpha)\}_{i\in\Omega}$.

I develop the proof as a sequence of Lemmas.

Lemma 1. *The objective function in the OCP problem can be rewritten as the sum of n functions, each of which depend on one τ_i only:*

$$\sum_{i=1}^{n} \Pi_i(\tau_i, \{\tau_j\}_{j\neq i}; \alpha) = \sum_{i=1}^{n} \Omega(\tau_i; \alpha).$$

For each $i = 1, \ldots, n$ (independent of size) $\Omega(\tau_i; \alpha)$ is a strictly concave quadratic function that attains a maximum when

$$\tau_i = \gamma\phi[1 + \frac{1-\phi}{\phi}\alpha]. \tag{50}$$

Proof. The objective function for the OCP problem is

$$\sum_{i=1}^{n} \Pi_i(\tau_i, \{\tau_j\}_{j\neq i}; (\tau_i, \omega_1, \alpha, \omega_2))$$
$$= \sum_{i=1}^{n}\left[\phi_i\gamma E[\phi_i\frac{\sigma}{\theta_1\theta_2}\tau_i + \sum_{j\neq i}\phi_j\frac{\sigma}{\theta_1\theta_2}\tau_j + (1-\phi)\frac{\sigma}{\theta_1\theta_2}\alpha\hat{\phi}_i\tau_i + (1-\phi)\frac{\sigma}{\theta_1\theta_2}\alpha\sum_{j\neq i}\hat{\phi}_j\tau_j] - Q_i\theta_1\left(\frac{\sigma}{\theta_1\theta_2}\right)^2(\tau_i)^2\right]$$

Pick $i = k$ and combine all terms from the summation that depend on τ_k; also, let $C_k(\tau_k) = Q_k\theta_1\left(\frac{\sigma}{\theta_1\theta_2}\right)^2(\tau_k)^2$. This gives

$$\begin{aligned}
\Omega_k(\tau_k;\alpha) &= \gamma E\frac{\sigma}{\theta_1\theta_2}\big[\phi_k^2\tau_k + \sum_{i\neq k}\phi_i\phi_k\tau_k + \phi_k\alpha(1-\phi)\hat{\phi}_k\tau_k + \sum_{i\neq k}\phi_i\alpha(1-\phi)\hat{\phi}_k\tau_k\big] - C_k(\tau_k)\\
&= \gamma E\frac{\sigma}{\theta_1\theta_2}\big[\phi_k\tau_k\sum_i\phi_i + \alpha(1-\phi)\hat{\phi}_k\tau_k\sum_i\phi_i\big] - C_k(\tau_k)\\
&= \gamma E\frac{\sigma}{\theta_1\theta_2}\big[\phi_k\phi + \alpha(1-\phi)\hat{\phi}_k\phi\big]\tau_k - C_k(\tau_k)\\
&= \gamma E\frac{\sigma}{\theta_1\theta_2}\big[\phi_k\phi + \alpha(1-\phi)\phi_k\big]\tau_k - C_k(\tau_k)\\
&= \phi\gamma E\frac{\sigma}{\theta_1\theta_2}\phi_k\big[1+\frac{1-\phi}{\phi}\alpha\big]\tau_k - C_k(\tau_k).
\end{aligned}$$

It follows that

$$\sum_{i=1}^n\Pi_i(\tau_i,\{\tau_j\}_{j\neq i};(\tau_i,\omega_1,\alpha,\omega_2)) = \sum_{k=1}^n\left[\phi\gamma E\frac{\sigma}{\theta_1\theta_2}\phi_k\big[1+\frac{1-\phi}{\phi}\alpha\big]\tau_k - \phi_k Q\theta_1\left(\frac{\sigma}{\theta_1\theta_2}\right)^2(\tau_k)^2\right].$$

Taking the first-order condition with respect to τ_i in the OCP optimization problem gives

$$\frac{\partial}{\partial\tau_i}\Omega_i(\tau_i;\alpha) = \gamma\big(\phi_i^2 + \phi_i\phi_{-i} + \alpha(1-\phi)\phi_i\big) - Q_i 2\theta_1\frac{\sigma}{\theta_1\theta_2}\tau_i,$$

if and only if,

$$\tau_i = \gamma\phi[1+\frac{1-\phi}{\phi}\alpha].$$

Moreover,

$$\frac{\partial^2}{\partial\tau_i^2}\Omega_i(\tau_i;\alpha) = -Q_i 2\theta_1\frac{\sigma}{\theta_1\theta_2} < 0,$$

so the Ω_i functions are quadratic and strictly concave. □

An immediate consequence of Lemma 1 is that the optimal cooperative policy entails a harmonized carbon price in which all countries price carbon at the rate in (50).

Lemma 2. *Provided the penalty term is zero, the payoff for country i when the rest of the coalition chooses $\{\tau_j\}_{j\neq i}$—$\Pi_i(\tau_i,\{\tau_j\}_{j\neq i};\omega_1=0)$—is a strictly-concave, quadratic function that attains a maximum at*

$$\tau_i^* = \phi_i\gamma[1+\frac{1-\phi}{\phi}\alpha] = \hat{\phi}_i\tau^C(\alpha) < \tau^C(\alpha). \tag{51}$$

The maximizer τ_i^* is independent of $\{\tau_j\}_{j\neq i}$.

Proof. Modifying Equation 10 for the quadratic case and setting $\omega_1 = 0$,

$$\Pi_i(\tau_i,\{\tau_j\}_{j\neq i}) = \phi_i\gamma E\left[\phi_i\frac{\sigma}{\theta_1\theta_2}\tau_i + \sum_{j\neq i}\phi_j\frac{\sigma}{\theta_1\theta_2}\tau_j + (1-\phi)\frac{\sigma}{\theta_1\theta_2}\alpha\hat{\phi}_i\tau_i + (1-\phi)\frac{\sigma}{\theta_1\theta_2}\alpha\sum_{j\neq i}\hat{\phi}_j\tau_j\right] \tag{52}$$

$$- \phi_i Q\theta_1\left(\frac{\sigma}{\theta_1\theta_2}\right)^2(\tau_i)^2. \tag{53}$$

Taking the first-order condition with respect to τ_i gives

$$\phi_i\gamma E\frac{\sigma}{\theta_1\theta_2}\left[\phi_i + (1-\phi)\alpha\hat{\phi}_i\right] - 2\phi_i Q\theta_1\left(\frac{\sigma}{\theta_1\theta_2}\right)^2(\tau_i) = 0.$$

If and only if,

$$\gamma\sigma\left[\phi_i + (1-\phi)\alpha\hat{\phi}_i\right] = 2\theta_1\frac{\sigma}{\theta_1\theta_2}(\tau_i).$$

Simplifying gives

$$\tau_i = \phi_i\gamma\left[1 + \frac{1-\phi}{\phi}\alpha\right].$$

□

Lemma 2 implies that individual payoffs are decreasing in the interval $[\tau_i^*, \tau_C^*]$, while Lemma 1 implies that coalition payoffs are increasing over the same interval. Because of this, the coalition would not want to pick a target for country i outside the interval $[\tau_i^*, \tau_C^*]$. If it picked $\tau_C < \tau_i^*$, the participation constraint would be slack and it could increase coalition surplus with $\tau_i = \tau_i^*$. Alternately, if $\tau_i > \tau_C^*$, it could increase coalition surplus by instead choosing $\tau_i = \tau_C^*$ and the participation constraint would still hold.

By similar logic, it is also clear that, given $\omega_1 \geq 0$, the $\hat{\tau}_i \in [\tau_i^*, \tau_C^*)$ that solves the OCP problem must be one at which the participation constraint s. If it didn't, then it would be possible to pick $\hat{\tau}_i + \epsilon$ for $\epsilon > 0$ where for ϵ small enough the participation constraint would still hold and coalition surplus would be strictly bigger, since coalition surplus is strictly increasing in τ_i within the interval.

Since the participation constraint must bind for each i, the $\hat{\tau}_i$ that solves the OCP problem must (for each i) solve

$$\Pi_i(\hat{\tau}_i, \{\hat{\tau}_j\}_{j\neq i}; (\hat{\tau}, \omega_1, \alpha, \omega_2)) = \max_{\tau_i \geq 0} \Pi_i(\tau_i, \{\hat{\tau}_j\}_{j\neq i}; (\hat{\tau}, \omega_1, \alpha, \omega_2)), \text{ for } i = 1, \ldots, n.$$

Moreover, from the geometry of the problem, it is clear that given $\omega_1 > 0$, the $\hat{\tau}_i$ that solves the OCP problem for each i will entail $\hat{\tau} > \tau_i^*$. It solves

$$\phi_i\gamma E[\phi_i\frac{\sigma}{\theta_1\theta_2}\hat{\tau}_i + \sum_{j\neq i}\phi_j\frac{\sigma}{\theta_1\theta_2}\hat{\tau}_j + (1-\phi)\frac{\sigma}{\theta_1\theta_2}\alpha\hat{\phi}_i\hat{\tau}_i + (1-\phi)\frac{\sigma}{\theta_1\theta_2}\alpha\sum_{j\neq i}\hat{\phi}_j\hat{\tau}_j] - Q_i\theta_1\left(\frac{\sigma}{\theta_1\theta_2}\right)^2(\hat{\tau}_i)^2$$
$$= \phi_i\gamma E[\phi_i\frac{\sigma}{\theta_1\theta_2}\tau_i^* + \sum_{j\neq i}\phi_j\frac{\sigma}{\theta_1\theta_2}\hat{\tau}_j + (1-\phi)\frac{\sigma}{\theta_1\theta_2}\alpha\hat{\phi}_i\tau_i^* + (1-\phi)\frac{\sigma}{\theta_1\theta_2}\alpha\sum_{j\neq i}\hat{\phi}_j\hat{\tau}_j] - Q_i\theta_1\left(\frac{\sigma}{\theta_1\theta_2}\right)^2(\tau_i^*)^2 - \omega_1 Q_i.$$

Cancelling like terms gives

$$\phi_i\gamma E[\phi_i\frac{\sigma}{\theta_1\theta_2}\hat{\tau}_i + (1-\phi)\frac{\sigma}{\theta_1\theta_2}\alpha\hat{\phi}_i\hat{\tau}_i] - Q_i\theta_1\left(\frac{\sigma}{\theta_1\theta_2}\right)^2(\hat{\tau}_i)^2$$
$$= \phi_i\gamma E[\phi_i\frac{\sigma}{\theta_1\theta_2}\tau_i^* + (1-\phi)\frac{\sigma}{\theta_1\theta_2}\alpha\hat{\phi}_i\tau_i^*] - Q_i\theta_1\left(\frac{\sigma}{\theta_1\theta_2}\right)^2(\tau_i^*)^2 - \omega_1 Q_i.$$

Further combining similar terms gives

$$\phi_i\gamma E[\phi_i\frac{\sigma}{\theta_1\theta_2} + (1-\phi)\frac{\sigma}{\theta_1\theta_2}\alpha\hat{\phi}_i](\hat{\tau}_i - \tau_i^*)$$
$$= Q_i\theta_1\left(\frac{\sigma}{\theta_1\theta_2}\right)^2[(\hat{\tau}_i)^2 - (\tau_i^*)^2] - \omega_1 Q_i.$$

Dividing through by $Q_i\frac{\sigma}{\theta_1\theta_2}$ gives

$$\gamma\sigma[\phi_i + (1-\phi)\alpha\hat{\phi}_i](\hat{\tau}_i - \tau_i^*)$$
$$= \theta_1\frac{\sigma}{\theta_1\theta_2}[(\hat{\tau}_i)^2 - (\tau_i^*)^2] - \omega_1\frac{\theta_1\theta_2}{\sigma}.$$

Simplifying gives

$$\phi_i\gamma[1+\frac{1-\phi}{\phi_i}\alpha](\hat{\tau}_i-\tau_i^*)=\frac{1}{\theta_2}[(\hat{\tau}_i)^2-(\tau_i^*)^2]-\omega_1\frac{\theta_1\theta_2}{\sigma^2}.$$

If and only if

$$\tau_i^*(\hat{\tau}_i-\tau_i^*)=\frac{1}{\theta_2}[(\hat{\tau}_i)^2-(\tau_i^*)^2]-\omega_1\frac{\theta_1\theta_2}{\sigma^2}. \tag{54}$$

To simplify the quadratic equation, I define $x\equiv\hat{\tau}_i-\tau_i^*$. It follows that

$$(\hat{\tau}_i)^2-(\tau_i^*)^2=(\hat{\tau}_i-\tau_i^*)(\hat{\tau}_i+\tau_i^*) \tag{55}$$

$$=x(x+2\tau_i^*) \tag{56}$$

$$=x^2+2\tau_i^*x. \tag{57}$$

Substituting into Eq. 54 gives

$$\tau_i^*x=\frac{1}{\theta_2}(x^2+2\tau_i^*x)-\omega_1\frac{\theta_1\theta_2}{\sigma^2}.$$

Simplifying gives

$$x^2=\frac{4}{\sigma^2}\theta_1\omega_1.$$

Since we know $\hat{\tau}_i>\tau_i^*$, the answer is the positive square root. Hence,

$$\hat{\tau}_i=\tau_i^*+\frac{2}{\sigma}\sqrt{\theta_1\omega_1}.$$

It follows that the penalty needed to support the cooperative outcome solves

$$\tau_i^C(\alpha)=\tau_i^N(\alpha)+\frac{2}{\sigma}\sqrt{\theta_1\omega_1^C}.$$

If and only if,

$$\frac{\phi_C^E}{\phi_i^E}\tau_i^N(\alpha)=\tau_i^N(\alpha)+\frac{2}{\sigma}\sqrt{\theta_1\omega_1^C}.$$

If and only if,

$$\sqrt{\theta_1\omega_1^C}=\frac{\sigma}{2}\left(\frac{\phi_C^E}{\phi_i^E}-1\right)\tau_i^N(\alpha).$$

If and only if,

$$\omega_{1,i}^C=\frac{1}{\theta_1}\left[\frac{\sigma}{2}\left(\frac{\phi_C^E}{\phi_i^E}-1\right)\tau_i^N(\alpha)\right]^2.$$

If we substitute for $\tau_i^N(\alpha)$, this becomes

$$\omega_{1,i}^C=\frac{1}{\theta_1}\left[\frac{\sigma}{2}(\phi_C^E-\phi_i^E)\gamma\left(1+\frac{1-\phi_C^E}{\phi_C^E}\alpha\right)\right]^2.$$

A.8 Bilateral trade data

USA	EU	China	Japan	UK	India	Brazil	Canada	Russia	KoreaR	Australia
0	461.6	472.5	147	64.1	59.9	32.1	326.6	23.2	79.9	11
289.1	0	472.5	89.4	197.6	49.5	34	24.4	168.3	57.6	10.1
123.8	252.7	0	171.8	23.9	18	80	28.2	61.2	173.6	121.3
81.4	81.1	169.3	0	8.1	5.4	8	11.8	14.3	29.6	45.5
130.4	698.8	128.9	25.9	0	19.5	6	29.6	23.6	10.2	14.5
36.6	45	75.6	12.5	6.7	0	3.1	3.9	7.1	16.5	10.7
36.6	36	37.5	4.9	2.6	4.7	0	2.6	4	5.2	1
230.1	51.5	56.5	12.4	7	4	4.1	0	1.4	7.2	1.8
8.1	97.7	53.4	8	3.7	3.3	1.8	0.6	0	8.2	0.6
62.1	51.6	107.2	47.6	4.2	5.6	4.3	5.7	14.6	0	20.6
26.2	35.1	56.9	15.5	5.2	3.4	0.6	2	0.3	8.6	0

Table 3: Bilateral trade data for merchandise goods in billions of USD per year (UNCTAD 2019). The columns indicate the exporting country. The rows indicate the importing country in the same order as the column labels.

USA	EU	China	Japan	UK	India	Brazil	Canada	Russia	KoreaR	Australia
NA	NA	19.9	36.3	64.7	29.6	6.6	38.9	1.7	10.9	8.7
255.5	NA	37	18.2	195.2	20.3	8.7	15.5	14	8.5	7.3
NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA
65.7	NA	12.2	NA	13.8	2	0.4	2.3	1.1	7.4	2.4
53.2	NA	2.9	6.9	NA	8.3	0.6	4	1	0.9	3.3
NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA
NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA
68.7	NA	2.5	2.5	6.8	2	0.3	NA	0.7	0.4	1.1
4	NA	3.7	0.6	5.2	0.4	0	0.3	NA	1.1	0.1
31.3	NA	17.1	10.2	NA	NA	NA	NA	NA	NA	NA
12.7	NA	2.3	3.2	7.1	1.9	NA	1.7	0.1	0.6	NA

Table 4: Bilateral data for trade in services in billions of USD per year (UNCTAD 2019). The columns indicate the exporting country. The rows indicate the importing country in the same order as the column labels.

A.9 Extension of results to validate numerical model

In this section, I extend the Nash carbon price, $\tau_i^N(\alpha)$, and the cooperative carbon price, $\tau_i^{COOP}(\alpha)$, to allow for "full" heterogeneity across Tier 1 countries.

Cooperative policy

Lemma 3. *Suppose Tier 1 countries differ in terms of ϕ_i^E, ϕ_i^Q, γ_i, and $\theta_{1,i}$. The ROW has a common carbon intensity σ_R and a common $\theta_{1,R}$, and as before it consists of a continuum of infinitesimal countries that don't abate in the absence of external coercion. Then the cooperative policy is (for all i in the coalition)*

$$\tau_i^{COOP} = \gamma_C \frac{\sigma}{\theta_2} \frac{\sum_j \phi_j^E a_j + (1-\phi_C^E) a_R \alpha^b}{\sum_{i=1}^n \phi_i^Q \theta_{1,i} a_i^{\theta_2}},$$

where

$$a_i = \left(\frac{\sigma_i}{\theta_{1,i}\theta_2} \right)^b$$

and

$$b = \frac{1}{\theta_2 - 1}.$$

Proof. Given local carbon price τ, the abatement rate in country or region i is

$$\mu_i = G_i(\tau) \equiv a_i \tau^b,$$

where

$$a_i = \left(\frac{\sigma_i}{\theta_{1,i}\theta_2} \right)^b$$

and

$$b = \frac{1}{\theta_2 - 1}.$$

Define $h(\tau) = \sum_{i=1}^n \Pi_i(\tau, \tau)$. When it does not create confusion, I suppress the dependence of functions on the background policy $(\tau, \omega_1, \alpha, \omega_2)$.

Since the penalty is never incurred if all countries move in sync,

$$\begin{aligned}
\Pi_i(\tau,\tau) &= B_i(\tau,\tau) - C_i(\tau) \\
&= \gamma_i E\left[\sum_j \phi_j^E G_j(\tau) + (1-\phi_C^E) G_R(\alpha\tau) \right] - \phi_i^Q Q \theta_{1,i} G_i(\tau)^{\theta_2} \\
&= \gamma_i E\left[\sum_j \phi_j^E a_j \tau^b + (1-\phi_C^E) a_R (\alpha\tau)^b \right] - \phi_i^Q Q \theta_{1,i} a_i^{\theta_2} (\tau)^{b\theta_2} \\
&= \gamma_i E \tau^b \left[\sum_j \phi_j^E a_j + (1-\phi_C^E) a_R \alpha^b \right] - \phi_i^Q Q \theta_{1,i} a_i^{\theta_2} (\tau)^{b\theta_2} \\
&= \gamma_i E \tau^b \Theta_1 - \phi_i^Q Q \theta_{1,i} a_i^{\theta_2} (\tau)^{b\theta_2},
\end{aligned}$$

where $\Theta_1 \equiv \sum_j \phi_j^E a_j + (1-\phi_C^E) a_R \alpha^b$.

Thus,

$$\begin{aligned} h(\tau) &= \sum_{i=1}^{n} \Pi_i(\tau, \tau) \\ &= \sum_{i=1}^{n} \left[\gamma_i E \tau^b \Theta_1 - \phi_i^Q Q \theta_{1,i} a_i^{\theta_2} (\tau)^{b\theta_2}\right] \\ &= \gamma_C E \Theta_1 \tau^b - Q \tau^{b\theta_2} \sum_{i=1}^{n} \phi_i^Q \theta_{1,i} a_i^{\theta_2} \\ &= \gamma_C E \Theta_1 \tau^b - Q \tau^{b\theta_2} \Theta_2, \end{aligned}$$

where

$$\Theta_2 \equiv \sum_{i=1}^{n} \phi_i^Q \theta_{1,i} a_i^{\theta_2}.$$

Taking the derivative,

$$h'(\tau) = \gamma_C \Theta_1 E b \tau^{b-1} - Q \Theta_2 b \theta_2 \tau^{b\theta_2 - 1}$$

and (using the fact that $b\theta_2 - 1 = b$)

$$h''(\tau) = \gamma_C \Theta_1 E b (b-1) \tau^{b-2} - Q \Theta_2 b \theta_2 b \tau^{b-1}$$

Since $b = 1/(\theta_2 - 1) \in (0, 1]$ given the assumption $\theta_2 \geq 2$, it is easy to see that $h''(\tau) < 0$: If $\theta_2 = 2$, then the first term is zero and the second term strictly negative, while if $\theta_2 > 2$ then both terms are strictly negative.

Since $h(\tau)$ is strictly concave for all $\tau \geq 0$, it attains a global maximum when $h'(\tau) = 0$. This implies

$$\gamma_C \Theta_1 E b \tau^{b-1} = Q \Theta_2 b \theta_2 \tau^b.$$

Define the τ at which this occurs as τ^{COOP}, since it is the τ at which the cooperative optimum is achieved. It solves

$$\tau^{COOP} = \gamma_C \frac{\sigma}{\theta_2} \frac{\Theta_1}{\Theta_2} \tag{58}$$

$$= \gamma_C \frac{\sigma}{\theta_2} \frac{\sum_j \phi_j^E a_j + (1 - \phi_C^E) a_R \alpha^b}{\sum_{i=1}^{n} \phi_i^Q \theta_{1,i} a_i^{\theta_2}}. \tag{59}$$

By strict concavity of $h(\cdot)$, it follows that the $h(\cdot)$ is strictly increasing to the left of τ^{COOP}, so the result follows. □

Nash policy The Tier-1 participation constraint for country i is

$$B_i(\hat{\tau}_i, \{\hat{\tau}_j\}_{j \neq i}; \alpha) - C_i(\hat{\tau}_i) \geq \max_{\tau_i \geq 0} [B_i(\tau_i, \{\hat{\tau}_j\}_{j \neq i}; \alpha) - C_i(\tau_i)] \tag{60}$$

Since $\tau_i = \hat{\tau}_i$ is a feasible option in the optimization problem on the right-hand side, we must have

$$\max_{\tau_i \geq 0} [B_i(\tau_i, \{\hat{\tau}_j\}_{j \neq i}; \alpha) - C_i(\tau_i)] \geq B_i(\hat{\tau}_i, \{\hat{\tau}_j\}_{j \neq i}; \alpha) - C_i(\hat{\tau}_i)$$

It follows that the constraint must hold with equality for each i.

Next, I show that the objective function in the maximization problem on the right-hand side of (60) is strictly concave, so the solution is unique.

First, it is straightforward to see that $C_i(\tau_i)$ is strictly convex in τ_i provided $\theta_2 > 1$. Thus, a sufficient condition for weak concavity of the objective function is $B_i(\tau_i, \{\hat{\tau}_j\}_{j\neq i}; \alpha)$ weakly concave in τ_i. Since $B_i(\cdot)$ is proportional to $\mu(\cdot)$, it is enough to show that $\mu(\tau_i, \{\hat{\tau}_j\}_{j\neq i}; \alpha)$ is weakly concave in τ_i.

We have

$$\mu(\tau_i, \{\hat{\tau}_j\}_{j\neq i}; \alpha) = \phi_i^E \mu_i + \sum_{j\neq i} \phi_j^E \hat{\mu}_j + (1-\phi_C^E)\mu^R \tag{61}$$

$$= \phi_i^E a\tau_i^b + \sum_{j\neq i} \phi_j^E a\hat{\tau}_j^b + (1-\phi_C^E)a[\alpha\hat{\phi}_i^E\tau_i + \alpha\sum_{j\neq i}\hat{\phi}_j^E\hat{\tau}_j]^b. \tag{62}$$

Thus,

$$\frac{\partial\mu}{\partial\tau_i} = \phi_i^E ab\tau_i^{b-1} + (1-\phi_C^E)ab(\alpha\tau^{AVG})^{b-1}\alpha\hat{\phi}_i^E > 0, \tag{63}$$

and

$$\frac{\partial^2\mu}{\partial^2\tau_i} = \phi_i^E ab(b-1)\tau_i^{b-2} + (1-\phi_C^E)ab(b-1)(\alpha\tau^{AVG})^{b-2}(\alpha\hat{\phi}_i^E)^2.$$

Since $b = \frac{1}{\theta_2-1} > 0$ if $\theta_2 > 1$, while $b-1 = \frac{2-\theta_2}{\theta_2-1} \leq 0$ if $\theta_2 \geq 2$, it follows that a sufficient condition for strict concavity of the objective function is $\theta_2 \geq 2$. I maintain this assumption throughout the paper.

It follows that the unique value of τ_i that satisfies the constraint solves the following fixed point condition for each i:

$$\arg\max_{\tau_i\geq 0}\big[B(\tau_i, \{\hat{\tau}_j\}_{j\neq i}) - C(\tau_i)\big] = \tau_i.$$

Taking the first-order condition of the left side gives:

$$\gamma_i E\frac{\partial\mu}{\partial\tau_i} = \theta_1 a^{\theta_2} b\theta_2(\tau_i)^{b\theta_2-1}Q_i$$

Since $b\theta_2 - 1 = b$, substituting gives

$$\gamma_i E[\phi_i^E ab\tau_i^{b-1} + (1-\phi_C^E)ab(\alpha\tau^{AVG})^{b-1}\alpha\hat{\phi}_i^E] = \theta_1 a^{\theta_2} b\theta_2(\tau_i)^b Q_i \tag{64}$$

Since the percent abatement function coefficient a depends on both parameters, we have

$$a_i = \left(\frac{\sigma_i}{\theta_{1,i}\theta_2}\right)^b.$$

We thus have

$$\gamma_i E[\phi_i^E a_i b\tau_i^{b-1} + \alpha^b(1-\phi_C^E)\hat{\phi}_i^E a_R b(\tau^{AVG})^{b-1}] = \theta_{1,i}a_i^{\theta_2} b\theta_2(\tau_i)^b Q_i$$

Divide through by $a_i bQ\phi_i^E\theta_{1,i}\theta_2\tau_i^{b-1}$ gives

$$\gamma_i\frac{\sigma}{\theta_{1,i}\theta_2}[1 + \alpha^b\frac{1-\phi_C^E}{\phi_C^E}\frac{a_R}{a_i}(\frac{\tau_i}{\tau^{AVG}})^{1-b}] = a_i^{\theta_2-1}\tau_i\frac{\phi_i^Q}{\phi_i^E}$$

iff

$$\gamma_i\frac{\sigma}{\theta_{1,i}\theta_2}[1 + \alpha^b\frac{1-\phi_C^E}{\phi_C^E}\frac{a_R}{a_i}(\frac{\tau_i}{\tau^{AVG}})^{1-b}] = \frac{\sigma_i}{\theta_{1,i}\theta_2}\tau_i\frac{\phi_i^Q}{\phi_i^E}$$

iff

$$\tau_i = \gamma_i \left[1 + \alpha^b \frac{1 - \phi_C^E}{\phi_C^E} \left(\frac{a_R}{a_i} \right) \left(\frac{\tau_i}{\tau^{AVG}} \right)^{1-b} \right], \quad \text{for } i = 1, \ldots, n.$$

iff

$$\tau_i = \gamma_i \left[1 + \alpha^b \frac{1 - \phi_C^E}{\phi_C^E} \left(\frac{\sigma_R}{\sigma_i} \frac{\theta_{1,i}}{\theta_{1,R}} \right)^b \left(\frac{\tau_i}{\tau^{AVG}} \right)^{1-b} \right], \quad \text{for } i = 1, \ldots, n. \tag{65}$$

A.10 Additional analysis related to minimum tariff calculations

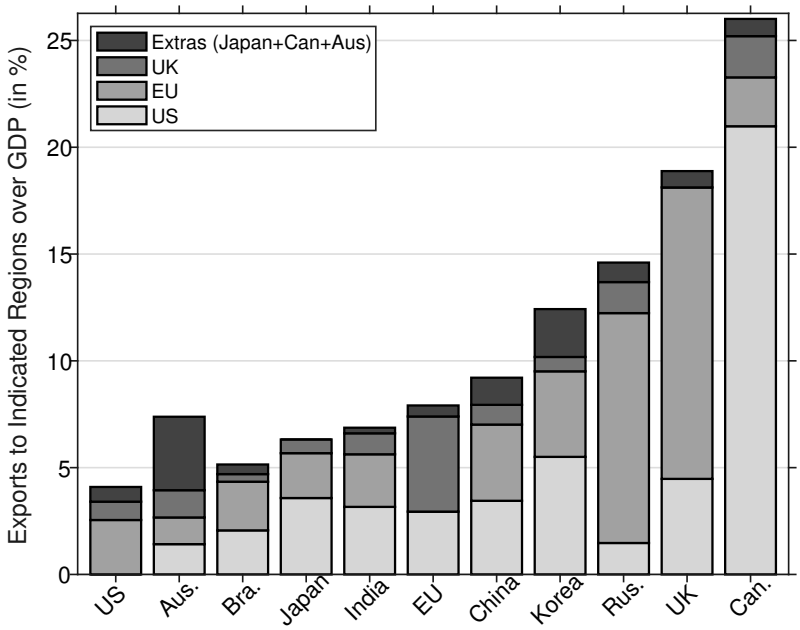


Figure 9: The bar for each country decomposes exports of goods and services to the regions indicated by segment color. Export flows expressed as fraction of export country GDP. Bars sorted low to high in combined exports to the US, EU, and UK.

The bars in Figure 9 decompose exports to each of the color-coded regions in the legend. Trade flows are reported as a fraction of the export country's GDP. Ignoring the "extras" segment at the top of each bar, the combined segments for the US, EU, and UK provide a measure of how much trade leverage the stable coalition has available to penalize other countries. The bars are sorted according to this metric. By this account, the US (at 3.4 percent of GDP) is least exposed to trade threats from the stable coalition. Next is Australia at 3.9 percent, then Brazil (4.7 percent), Japan (6.3 percent), and India (6.6) percent.

In addition to summarizing trade flows to the stable coalition, I also consider how much additional leverage could be gained if the other high-income countries from the 11 model regions—mainly, Japan, Canada, and Australia—join the coalition in meting out punishments. One could imagine adding these countries to the set of punishing countries without changing the abatement obligation they are held to as a Tier 2 country. Since tariffs are not actually imposed in equilibrium appreciably change a country's incentive to go along with the agreement terms. Figure 9 shows that adding Japan, Canada,

and Australia to the set of punishing countries has only a modest impact on the total exports bar for all countries, with the exception of Australia and South Korea.

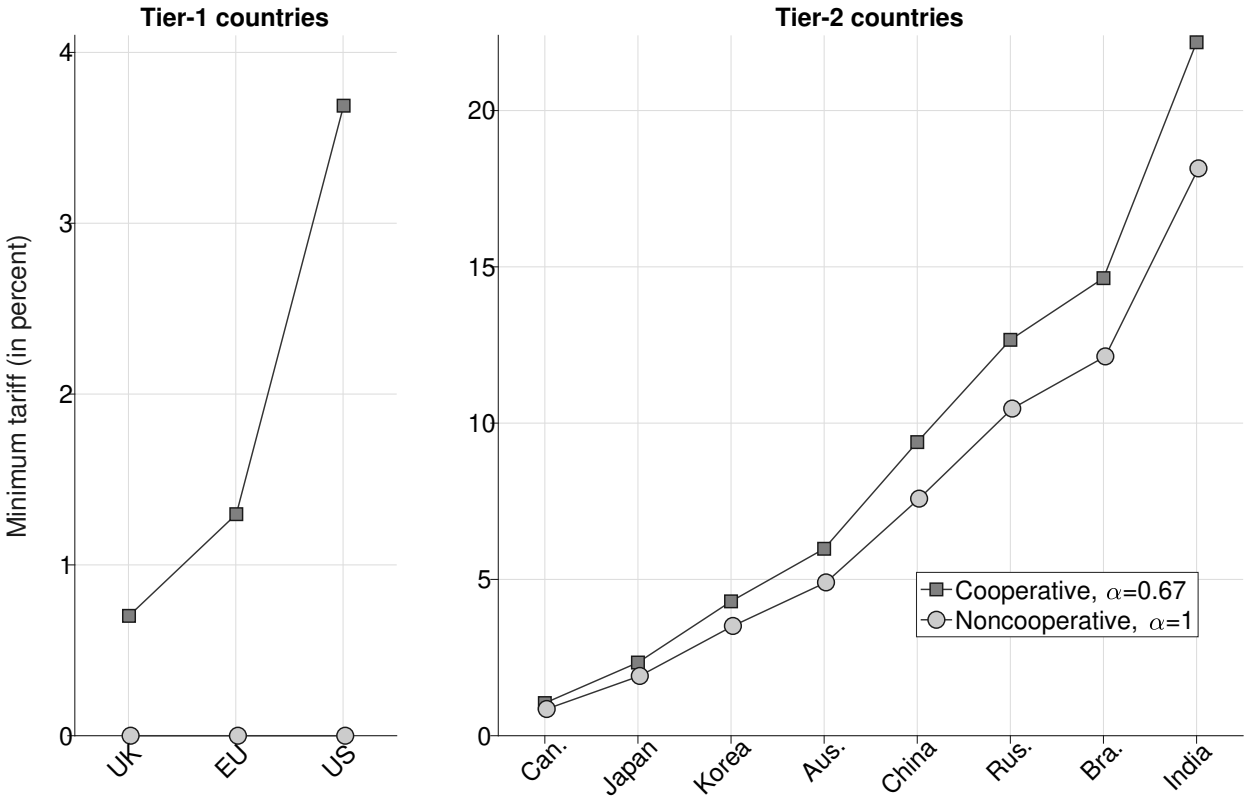


Figure 10: Minimum tariffs to support ambitious outcomes with penalties imposed by US, EU, and UK (no extras). Dark squares assume cooperative outcome with $\alpha = 0.67$, the match rate that generates efficient global abatement. Light circles assume noncooperative case with $\alpha = 1$.

Figure 10 backs out the implied minimum tariffs to support two other policy scenarios. The first policy scenario considers the farthest a US-EU-UK coalition might hope to go without threatening tariffs within the coalition. This non-cooperative case with $\alpha = 1$ achieves 85 percent of the efficient level of global abatement, which is significantly more than the 54 percent level achieved by the equilibrium policy. The second policy considers the cooperative policy with a 67 percent match rate, which by design achieves 100 percent of the globally efficient abatement rate. Interestingly, the tariffs needed to generate the two outcomes are similar in magnitude for all countries.

Comparing Figure 9 and Figure 6, the ordering of countries appears to change in a counter-intuitive manner. For instance, while Japan ranks high in one figure, it is relatively low in another. This apparent discrepancy arises because the estimation of tariff impacts, as presented in Nordhaus (and adopted in this paper), takes into account pre-existing trade distortions that vary considerably across countries. These distortions have a significant impact on the resulting tariff effects, even when controlling for the flow of goods and services between countries.