Tiered Climate Clubs:
Global Abatement Without Global Agreement*

Terrence Iverson†‡

December 31, 2024

Abstract

This paper introduces a novel policy structure to mitigate global carbon emissions without requiring broad multilateral cooperation. Extending Nordhaus's (2015) climate club, 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 stable first-tier coalition consists of the US and EU, which optimally sets a carbon price at 60% of the global Social Cost of Carbon. This policy structure achieves global abatement four and a half times higher than the uncoordinated Nash equilibrium and one-third of the globally efficient level. Quantitative results have been revised due to a corrected calibration error, detailed in Appendix A.1.

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 enable 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 include the following:

1. There are three stable first tier coalitions in the baseline model—the EU-US, the EU-UK, and the EU-Australia. Neverthless, with minimal transfers, the EU-US coalition Pareto dominates the other stable coalitions, making this coalition a natural focal point.

2. If the US and EU adopt a conventional climate club,[3] choosing policy to maximize club surplus, their average carbon price increases by a factor of two compared with the Nash equilibrium (where all countries act independently).[4] Carbon prices outside the club remain at the Nash

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

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

[4] Under Nash, the US and EU price carbon at 11 percent of the globally efficient level. When they use a conventional

level. Global abatement increases from 7 percent of the efficient level under Nash to 9 percent of the efficient level—a 32 percent increase. In contrast, with a TCC, the US-EU first tier sets carbon prices five and a half times higher than under Nash. The second tier adopts a carbon price 43 percent as high, leading to global abatement 32 percent of the efficient level—460 percent higher than the Nash level. Shifting from a conventional climate club to a TCC amplifies the increase in global abatement above the Nash level by a factor of 14.

3. If the match rate were hypothetically increased to 93 percent, a US-EU led TCC would achieve the efficient level of global abatement. In this case, the first tier prices carbon above the efficient level. It is willing to do this because it has lower carbon intensity than the rest of the world, 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",[5] the TCC abates 16 percent of the efficient level—less than half the cooperative outcome, but still 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 partially 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

This 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 (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 substantially more than is accomplished in the Nash equilibrium.

climate club to achieve the cooperative outcome, they price carbon at 22 percent of the globally efficient level.

[5] 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".

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).[6]

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. Specifically, by combining a non-excludable-benefits-producing IEA game with a separate game with club-like payoffs, the approach makes it possible to incorporate targeted punishments into an international climate agreement (Folmer, Mouche and Ragland, 1993). Among the linkage options explored in the literature—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 (Barrett, 1997; Lessmann, Marschinski and Edenhofer, 2009; Helm and Schmidt, 2015; Böhringer, Carbone and Rutherford, 2016; Al Khourdajie and Finus, 2020; Hagen and Schneider, 2021).

A recent and compelling example of this approach is Farrokhi and Lashkaripour (2024) (henceforth FL24), who analytically solve for optimal trade policy in a multi-country, multi-sector, general equilibrium model of trade with a climate externality. By incorporating analytical solutions to the optimal unilateral tariff problem, they numerically study the efficacy of trade-based measures for supporting carbon abatement in a quantitatively plausible setting. FL24 find that while border carbon adjustments are largely ineffective at amplifying global abatement, a climate club in which the US, EU, and China form a committed alliance of core members can support 68 percent of the globally efficient level of abatement.

While the general equilibrium model in FL24 is a clear advance over the partial equilibrium framework adopted in Nordhaus (2015), their analysis relies on a crucial assumption that limits its applicability to the TCC proposal in this paper: FL24 focus on climate clubs in which all countries participate. This restriction is crucial because it allows them to derive analytical formulas for optimal tariff responses based on an envelope theorem result. Without full participation, the equilibrium would not be globally optimum, and the envelope theorem result would not apply. Since the main advantage of a TCC is its potential to exploit coercion from a small group of countries to incentivize partial participation from others, it cannot be viewed as a local perturbation of a global optimum, precluding the use of an envelope theorem type result. One could use their model to study the efficacy of a TCC in a fully numerical way; however, numerical general equilibrium models are notoriously difficult to solve in this context, especially when policy is chosen endogenously, as we do in this paper. Therefore, employing a multi-country, multi-industry quantitative general equilibrium trade model like the one

[6]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).

in FL24 would significantly constrain our ability to explore the wide range of scenarios that form the core results of this paper.

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 paper employs the theoretical framework developed in Nordhaus (2015). The present section outlines the model, defines the Tiered Climate Club agreement structure, establishes the optimal policy problem, and details the adjustments made to update and refine the original model.

2.1 Model

Setup and notation As in Nordhaus (2015), the model is static and employs a partial equilibrium structure with reduced-form functions to capture the main economic forces in the model. There are N heterogeneous "countries" (or regions), and each country acts like a unitary decision maker in matters of climate policy. 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.[7] Global GDP, global

[7]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. (2014) and Traeger (2023).

CO2 emissions, and global marginal damages from CO2 emissions (the Social Cost of Carbon, or SCC) are represented by Q, E, and $\gamma \equiv \sum_i \gamma_i$, respectively. Q denotes global GDP, E denotes global CO2 emissions, and $\gamma \equiv \sum_i \gamma_i$ denotes global marginal damages from CO2 emissions (the Social Cost of Carbon, or SCC). In addition, baseline trade flows are defined by a bilateral trade matrix $\mathbf{X}$.

To denomenate values in terms of their fraction of the global total, I use the variable 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 national/regional abatement cost functions follow Nordhaus (2015):

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

where μ_i is the abatement rate in i and $\theta_{1,i}$ and θ_2 are parameters. Countries set carbon prices, not abatement rates, so I express the cost functions in terms of the carbon price, τ_i. This is done by equating marginal abatement costs in emission units with the carbon price to get the following relationship between the abatement rate and the carbon price:[8]

$$\mu_i = a_i (\tau_i)^b \equiv G_i(\tau_i). \tag{1}$$

where

$$a_i = \left[\frac{\sigma_i}{\theta_1 \theta_2}\right]^{\frac{1}{\theta_2 - 1}} \tag{2}$$

and

$$b = \frac{1}{\theta_2 - 1}. \tag{3}$$

$\sigma_i = \frac{E_i}{Q_i}$ is the emissions intensity of output in country i. Combining gives abatement costs in terms of the carbon price:[9]

$$C_i(\tau_i) = Q_i \theta_1 a_i^{\theta_2} (\tau_i)^{b\theta_2}. \tag{4}$$

Climate benefits The climate benefits from policy depend on the reduction in global emissions, which depends on policy in all countries. To separate the carbon price in i from that in the rest of the world, I use the notation $\tau_{-i} \equiv \{\tau_j\}_{j\neq i}$ to denote climate policies outside i. The global abatement rate can then be written

$$\mu(\tau_i, \tau_{-i}) = \sum_j \phi_j^E G_j(\tau_j), \tag{5}$$

where $G_j(\cdot)$, defined in (1), 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, which depends on policy in all countries:

$$B_i(\tau_i, \tau_{-i}) = \gamma_i E \mu(\tau_i, \tau_{-i}). \tag{6}$$

[8] Details are in Appendix A.2.

[9] Details are in Appendix A.2.

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

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

$$P_{ji}(t_{ji}; \mathbf{X}_{i,j}) = \mathbf{X}_{i,j}(\alpha_{ij} t_{ji} + \beta_{ij} t_{ji}^2), \tag{7}$$

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_{ij} and β_{ij} are the estimated parameters for a uniform tariff in country j on imports from 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

$$\omega_{i,\Omega} = \frac{1}{Q_i} \sum_{j \in \Omega} P_{ji}(t_{ji}; \mathbf{X}_{i,j}). \tag{8}$$

For convenience, I sometimes suppress the dependence of the total penalty on the underlying tariffs. In that case, the conditional trade incentive faced by a country is a pair $(\omega, \underline{\tau})]$, where ωQ_i is the trade penalty on i if it fails to price carbon above $\underline{\tau}$. In the quantitative section, actual bilateral trade flows are used to quantify tariff impacts.

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

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

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, using uniform import tariffs to impose penalties.

Tier-2 incentives Countries in the second tier must price carbon at or above a fixed fraction of the average carbon price adopted in the first tier, or face tariffs. Given observed first tier carbon prices $\{\tau_i\}_{i \in \Omega}$, the minimum carbon price for Tier-2 countries is $\alpha \tau^{AVG}$, where $\alpha \in [0, 1]$ is the specified "match rate" and

$$\tau^{AVG} \equiv \sum_{i \in \Omega} \hat{\phi}_i^E \tau_i \tag{9}$$

[10] This equation differs slightly from the stated "tariff benefit" function in Eq. * of Nordhaus (2015). The quadratic term captures the efficiency loss from a tariff, which is a positive loss for both countries. The linear term captures the terms of trade effect, which is a loss to the country that faces tariffs on its exports (but a gain to the country imposing it). Hence, the penalty to country i that faces tariffs from j is the sum of two positive loss terms.

is the (size-weighted) average Tier-1 carbon price.[11]

The match rate is set by the first tier, which prefers the highest possible match rate, as a higher match rate shifts more abatement responsibility onto the second tier. However, since a higher match rate also imposes higher costs on the second tier, 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." The reduced form relationship captures the most important determinant of general trading leverage, which is the size of the domestic market that another country may hope to gain access to. Specifically, I assume that the (highest achievable) match rate is proportional to the combined GDP of the first tier:

$$\alpha(\Omega) = b\phi_{\Omega}^{Q}, \tag{10}$$

where ϕ_{Ω}^{Q} is the first tier's share of global GDP and $b > 0$ is a calibrated parameter that captures how responsive the (maximum achievable) match rate is to coalition GDP. The calibration of b is discussed in Section 4.1.

While the functional form in Eq. 10 is *ad hoc*, there is a sense in which it can be viewed as conservative. The assumption implies that 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 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. Consequently, the assumption understates the potential role of large countries in a TCC. Since the quantitative results, generated while maintaining the assumption in Eq. 10, finds that the biggest economies are most likely to join the first tier anyway, there is an important sense in which the assumption understates the liklihood that countries like the U.S. and E.U. would choose to join the stable coalition.

Tier 1 incentives Tier 1 countries also use conditional trade incentives with each other to increase cooperation within the first tier. 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 carbon price at or above $\underline{\tau}_i$ or face uniform tariffs on exports to the other Tier 1 countries with total economic value (as a fraction of i's GDP) equal to $\omega_{1,i}$—determined through Eq. 8. When Tier 1 countries set the Tier 1 incentives, a range of behavioral assumptions are possible, as explored below.

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

[11] Recall that $\hat{\phi}_i^E$ is country i's CO2 emissions as a fraction of first-tier emissions.

as a Stackelberg leader vis-a-vis the second tier. The game is solved with backward induction.

2.3.1 Second stage

With the first tier in place, the second stage unfolds as a sequential game. First, tier-1 countries establish the agreement terms, including the match rate, the minimum tier-1 targets, and the penalty threats for tier-1 and tier-2 countries. Next, each tier-1 country chooses its carbon price, considering the conditional trade threat it faces and expectations about how the second tier will respond. Having observed the average tier-1 carbon price, each tier-2 country then chooses its carbon price. Finally, punishments are imposed by participating tier-1 countries.

In solving the second stage with backward induction, the first question is if tier-1 countries will follow through with the indicated punishments. 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). This fact is used to defend the use of trade tariffs as a punishing devise in climate agreements (e.g., Nordhaus, 2015). Reflecting these findings, the model assumes that (absent retaliation) implementing tariffs has no welfare impact on the implementing country. I therefore assume that Tier 1 countries will always follow through with tariff threats as long as the terms of the threat are lenient enough to avoid a trade war. Meanwhile, the procedure for calibrating the match rate for a given tier-1 set ensures that the magnitude of tariffs needed to ensure universal compliance in equilibrium does not trigger retaliation. Hence, in equilibrium, the tier-2 tariff threat is credible.

Working backwards, tier-2 countries price carbon at or above fraction α of the average tier-1 carbon price, or face tariffs. Let

$$\tau_j^* = \arg\max_{\tau_j} \Big[B_j(\tau_j, \tau_{-j}) - C_j(\tau_j) \Big]$$

be tier-2 country j's optimal policy ignoring tariffs, where τ_{-j} is the carbon price of all countries other than j. It is easy to show that the unilateral optimal response equals the domestic Social Cost of Carbon: $\tau_j^* = \gamma_j$. Thus, j's optimal compliance strategy is

$$\hat{\tau}_j = \max(\gamma_j, \alpha\tau^{AVG}), \tag{11}$$

and the corresponding Tier 2 participation constraint is

$$B_j(\hat{\tau}_j, \tau_{-j}) - C_j(\hat{\tau}_j) \geq B_j(\gamma_j, \tau_{-j}) - C_j(\gamma_j) - \omega_{j,\Omega} Q_j \mathbf{1}_{\tau_j^* < \alpha\tau^{AVG}}, \tag{12}$$

where $\omega_{j,\Omega} Q_j$ is the combined economic loss to j from tariffs on exports to Ω.

In the most interesting case, the agreement terms are bind—$\alpha\tau^{AVG} > \tau_j^*$—and the participation constraint is simply

$$B_j(\alpha\tau^{AVG}, \tau_{-j}) - C_j(\alpha\tau^{AVG}) \geq B_j(\gamma_j, \tau_{-j}) - C_j(\gamma_j) - \omega_{j,\Omega} Q_j.$$

To ensure compliance, the penalty must be large enough to incentivize j to adopt a higher carbon price than it would optimally do without coercion.

When setting their own carbon prices, tier-1 countries anticipate how their domestic efforts will amplify abatement in the second tier. Nevertheless, they still have a choice about how ambitiously to design the agreement. The ambition depends on their willingness to use tariff threats with each other.

On the low end, tier-1 countries use tariff threats with the second tier, but avoid using them with each other. The equilibrium in this case has each tier-1 country play its noncooperative best response, internalizing only the portion of global climate damages that falls within its own borders. On the high end, tier-1 countries set tier-1 policy targets to maximize the combined surplus of the first tier. Tier-1 countries internalize climate damages that fall within their collective borders. Supporting this policy requires tariffs.

Non-cooperative policy In the lower-bound scenario, Tier 1 countries interact non-cooperatively. This case coincides with a Tier 1 penalty of zero. Each Tier 1 country $i \in \Omega$ takes the policy targets of the other Tier 1 countries as given—fixed at $\hat{\tau}_j$—then chooses its own carbon price to solve

$$\max_{\tau_i \geq 0} \left[B_i(\tau_i, \{\hat{\tau}_j\}_{j\in\Omega, j\neq i}, \{\hat{\tau}_k\}_{k\notin\Omega}) - C_i(\tau_i) \right], \tag{13}$$

subject to the the tier-2 participation constraints from (12). The tier-2 carbon prices, $\hat{\tau}_k$, depend on the tier-1 carbon prices through (11), and the model components are defined in (1)- (10). Letting n be the number of countries in Ω, the optimization problem in Eq. 13, one for each $i \in \Omega$, defines a system of n first-order conditions in n Tier 1 carbon prices.

To get intuition for how the TCC affects Tier 1 incentives, it is instructive to trace how policy choice in i impacts global abatement. The global abatement rate can be written

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

with $G_i(\cdot)$ defined in Eq. 1.

The expression highlights the two channels through which Tier 1 country i's choice of τ_i impacts global abatement—thus, i's climate benefits. 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 (without tariff threats against the second tier) 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.

Cooperative policy In the upper-bound scenario, Tier 1 countries use trade threats with each other to support the cooperative outcome. Policy is chosen to maximize the combined surplus of the

first tier, so spillovers between Tier 1 countries are internalized. The cooperative case is the standard benchmark in the International Environmental Agreements literature (see Marrouch, Chaudhuri et al. (2016) for a review), and it is the case we will emphasize when studying coalition stability in the quantitative application.

The cooperative policy solves

$$\max_{\{\tau_i \geq 0\}_{i \in \Omega}} \sum_{i \in \Omega} \left[B_i(\tau_i, \{\tau_j\}_{j \in \Omega, j \neq i}, \{\hat{\tau}_k\}_{k \notin \Omega}) - C_i(\tau_i) \right] \tag{14}$$

subject to

$$\hat{\tau}_k = \max(\gamma_k, \alpha \sum_{i \in \Omega} \hat{\phi}_i \tau_i), \quad k \notin \Omega,$$

the model equations (1)-(10), the Tier 2 participation constraints in (12), and the following participation constraint for each $i \in \Omega$:

$$B_i(\tau_i, \tau_{-i}) - C_i(\tau_i) \geq B_i(\gamma_i, \tau_{-i}) - C_i(\gamma_i) - \omega_{i,\Omega \setminus i} Q_i \mathbf{1}_{\tau_i < \alpha \tau^{AVG}}. \tag{15}$$

In the latter participation constraint, the set of penalizing countries for Tier 1 country i excludes i.

Intermediate cases In addition to the cooperative and non-cooperative cases, any carbon price between these bounds can be supported with sufficiently tariffs. 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.3.

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. 10—and on the incentive for Tier 1 countries to price carbon—as explained in Section 3.1.

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.

Provided Tier 1 and Tier 2 participation constraints are included in the optimal policy problem used to identify carbon price targets—as they were for the cooperative and noncooperative Tier 1 problems defined in Section 2.3.1—we can focus on targets that are incentive compatible. Let $\{\tau_j(\Omega)\}_{j\in\Omega}$ be the set of incentive-compatible carbon price targets when the first tier consists of the countries in Ω. The targets could be cooperative, non-cooperative, or somewhere in between. Then the minimum Tier 2 target is $\alpha(\Omega)\tau^{AVG}(\Omega)$, where

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

To study coalition stability, we need to evaluate how a country's decision to join, or not, impacts national payoffs.

Let $\Pi_i^{IN}(\Omega)$ be i's payoff when i is a member of the Tier 1 set Ω. Then

$$\Pi_i^{IN}(\Omega) \equiv B_i(\{\tau_j(\Omega)\}_{j\in\Omega}, \{\hat{\tau}_k\}_{k\notin\Omega},) - C_i(\tau_i(\Omega)), \tag{16}$$

where

$$\hat{\tau}_k = \max(\gamma_k, \alpha\tau^{AVG}(\Omega)). \tag{17}$$

Similarly, the payoff for Tier 2 country j given first tier Ω is

$$\Pi_j^{OUT}(\Omega) \equiv B_i(\{\tau_j(\Omega)\}_{j\in\Omega}, \{\hat{\tau}_k\}_{k\notin\Omega},) - C_j(\hat{\tau}_j) \tag{18}$$

with $\hat{\tau}_k$ defined in (17).

The first tier Ω is stable if it satisfies *internal stability* and *external stability*. Internal stability requires that all Tier 1 countries are better off staying in the first tier than if they unilaterally left. Thus, for all $i \in \Omega$,

$$\Pi_i^{IN}(\Omega) \geq \Pi_i^{OUT}(\Omega \setminus i). \tag{19}$$

The departure of country i from the first tier results in a reduction of the Tier-1 set from Ω to $\Omega \setminus i$, which impacts global abatement through the match rate $\alpha(\Omega)$ and the endogenous Tier 1 carbon price targets, $\{\tau_j(\Omega)\}_{j\in\Omega}$. The coalition satisfies external stability if every $j \notin \Omega$ is better off staying out than unilaterally joining:

$$\Pi_j^{OUT}(\Omega) \geq \Pi_j^{IN}(\Omega \cup j). \tag{20}$$

2.4 Model changes relative to Nordhaus (2015)

The model differs from Nordhaus (2015) in three 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. Finally, I 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. By showing that the stable coalition is unique, I can argue that the obligation to take a lead in setting up a TCC would fall clearly on the shoulders of a specific group of nations.

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 significant portion of the globally efficient level of abatement: 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 use analytical arguments to study the second stage of the agreement formation game. The first tier, and consequently the match rate, are both treated as fixed. The analysis shows the different ways in which a TCC mitigates the difficulties that arise under a small (conventional) climate club, where all abatement is undertaken by a small coalition of countries. The analysis also shows how the degree of cooperation within the first tier of a TCC impacts policy outcomes.

3.1 Mitigating the hurdles of a small climate club

Conventional climate clubs (e.g., Nordhaus (2015)) work well when a large fraction of countries join the club. Small clubs, in contrast, run into the same difficulties that stymie any attempt to abate global carbon emissions at a sub-global scale. Most importantly, small abatement coalitions are hampered by free-rider incentives, excess costs stemming from the inability to exploit cheap abatement opportunities in other countries, and carbon leakage. This section shows how moving from a small (conventional) climate club to a TCC of the same size can go a long way toward addressing these concerns.

3.1.1 The free-rider problem

The free rider problem arises because the jurisdictions adopting climate policy care about the full cost of abatement efforts but only for a small portion of the resulting climate damages. The problem is most severe when countries act independently. In that case, each country only takes into account the portion of damages that falls within its national borders. When a group of countries cooperate, they can mutually benefit by jointly internalizing damages that fall within their collective borders, though they still have no incentive to consider damages that fall outside their collective borders.

To develop intuition for what happens to free rider incentives when the coalition of acting countries adopts a TCC, it is useful to reframe the model from Section 2.1 by changing the choice variable. In the original model, countries choose carbon prices directly, but the model can be equivalently formulated so countries choose abatement denominated in emission units.

Let Ω be the coalition of countries that leads the effort. This could be the club in a conventional climate club or the first tier in a TCC. Either way, the coalition behaves cooperatively. As in the cooperative policy problem in (14), it chooses policy to maximize coalition surplus. Letting a_Ω denote coalition abatement, it solves

$$\max_{a_\Omega \geq 0} \Bigg[\underbrace{\gamma_\Omega \times a_G(a_\Omega)}_{\substack{\text{Coalition} \\ \text{Climate Benefit}}} - \underbrace{C_\Omega(a_\Omega)}_{\substack{\text{Coalition} \\ \text{Abatement Cost}}} \Bigg],$$

where $\gamma_\Omega = \sum_{i \in \Omega} \gamma_i$ is the coalition marginal damage, $a_G(a_\Omega)$ 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 cost function assuming efficient policy within the coalition.

The first-order condition for the coalition problem gives

$$\underbrace{\gamma_\Omega \times \frac{\mathrm{d}a_G}{\mathrm{d}a_\Omega}}_{\text{Marginal benefit}} = \underbrace{C'(a_\Omega)}_{\text{Marginal cost}} . \tag{21}$$

In the standard case—including a conventional climate club[12]—the coalition only controls coalition abatement, so $\frac{\mathrm{d}a_G}{\mathrm{d}a_\Omega} = 1$. Eq. 21 then reduces to the standard result 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 14 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,[13] 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 $\theta_{1,R}$. Abatement costs are quadratic ($\theta_2 = 2$),[14] and Tier 2 countries only abate with external coercion.[15] Then the cooperative policy—defined in 14—imposes the following harmonized carbon price on all $i \in \Omega$:*

$$\tau^C(\alpha) = \gamma_\Omega \times \frac{\mathrm{d}a_G}{\mathrm{d}a_C}, \tag{22}$$

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

$$\frac{\mathrm{d}a_G}{\mathrm{d}a_C} = 1 + \alpha \frac{1-\phi_C^E}{\phi_C^E} \frac{\sigma_R}{\sigma_C} \tag{23}$$

is the amplification factor defined in the text.

Proof. Appendix A.4. □

The amplification factor (Eq. 23) 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

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

[13] Emissions intensity is CO2 emissions over GDP.

[14] 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.

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

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.4). 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:[16]

$$\frac{\tau^C(\alpha)}{\gamma} = \frac{\phi_\Omega^\gamma}{\phi_\Omega^E}\left[\phi_\Omega^E + \alpha\frac{\sigma_R}{\sigma_\Omega}(1-\phi_\Omega^E)\right]. \tag{24}$$

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 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

$$\frac{\tau^C(\alpha)}{\gamma} = \phi_\Omega + \alpha(1-\phi_\Omega), \tag{25}$$

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

[16] The cooperative carbon price defined in Eq. 22 implies

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

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 first tier is instructive since this is the stable first tier we will focus on in most of the quantitative discussion. 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 lower emissions intensity than the rest of the world: $\frac{\sigma_R}{\sigma_C} \approx 1.4$.

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 a given carbon price target and thus prefer a higher target than the average country, all else equal.

Returning to the example of a US-EU first tier, substituting the approximate ratio values into Eq. 24 implies $\tau^C(1) \approx 1.33 \times \gamma$. With a match rate of one, the US and EU would optimally price carbon a third higher than 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 would require less than a 100 percent match rate provided the endogenous coalition has relatively low emissions intensity (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.[17] 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.

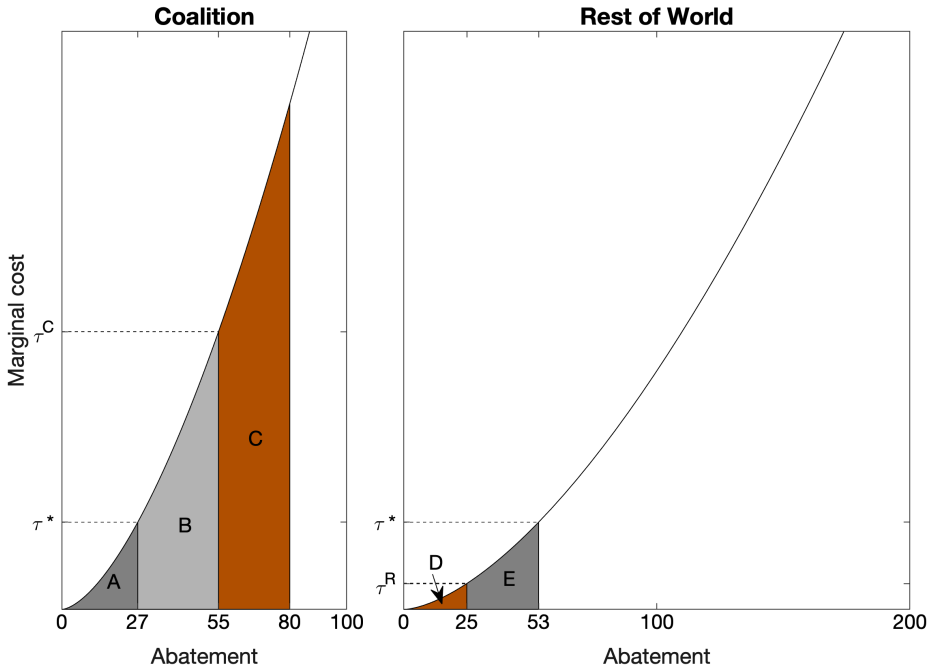

Figure 1: The figure shows marginal abatement costs for a coalition comprising a third of global emissions and for the rest of the world. The curves are generated using the abatement cost function calibrated in Barrage and Nordhaus (2023) (and other recent applications of DICE). Letters indicate colored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

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^*.[18] 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

[17] 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.

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

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 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".[19]

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.5, 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 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

$$P(\phi,\alpha) = \frac{\phi + (1-\phi)\alpha^{\theta_2/(\theta_2-1)}}{\left[\phi + (1-\phi)\alpha^{1/(\theta_2-1)}\right]^{\theta_2}}. \tag{26}$$

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

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

Proof. Appendix A.5. □

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.5). In Figure 2, I use Eq. 46 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

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

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.

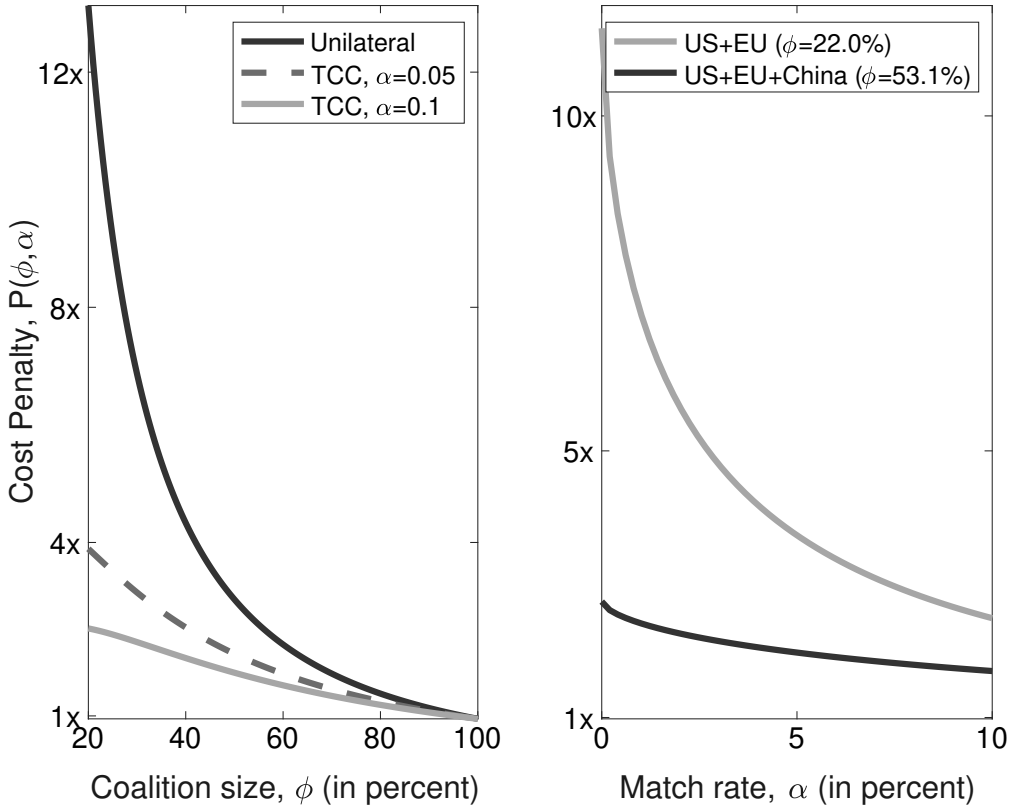

Figure 2: The left panel plots the penalty as a function of ϕ for three different policies. The right panel plots the penalty as a function of α for two values of ϕ: $\phi = 22.0\%$ (US+EU) and $\phi = 53.1\%$ (China+US+EU).

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 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 portion of leakage that arises when firms sell to foreign 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 **??**). 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* **??** *is given by*[20]

$$\tau_i^N(\alpha) = \gamma_i \frac{\mathrm{d}a_G}{\mathrm{d}a_C} = \hat{\phi}_i^{\gamma} \tau^C(\alpha), \qquad i \in \Omega, \tag{27}$$

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

$$\frac{\mathrm{d}a_G}{\mathrm{d}a_C} = 1 + \alpha \frac{1 - \phi_C^E}{\phi_C^E} \frac{\sigma_R}{\sigma_C} \tag{28}$$

and

$$\tau^C(\alpha) = \gamma_\Omega \times \frac{\mathrm{d}a_G}{\mathrm{d}a_C}. \tag{29}$$

Proof. Appendix A.6. □

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

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

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*

$$\mu^N(\alpha) = \tilde{H}\big(\{\hat{\phi}_i^E, \hat{\phi}_i\}_{i\in\Omega}\big)\mu^C(\alpha), \tag{30}$$

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

$$\tilde{H}\big(\{\hat{\phi}_i^E, \hat{\phi}_i\}_{i\in\Omega}\big) = \sum_{i\in\Omega} \hat{\phi}_i^E \hat{\phi}_i{}^{\gamma}. \tag{31}$$

$\tilde{H}(\cdot)$ 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

$$\tilde{H}\big(\{\hat{\phi}_i^E\}_{i\in\Omega}\big) = \sum_{i\in\Omega} \big(\hat{\phi}_i^E\big)^2, \tag{32}$$

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

Proof. Appendix A.7. □

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. 32 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. 32 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. 32 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. 22—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.8. □

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).[21] 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

[21] An almost identical specification of regions is used in Farrokhi and Lashkaripour (2024).

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 and EU, this ratio increased about 12 percent between 2011 and 2019.[22] Not accounting for this 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.

	USA	EU	China	Japan	UK	India	Brazil	Can.	Russia	Korea	Aus.
ϕ_j^Q:	15.4	15.5	17.9	3.9	2.4	7.1	2.4	1.4	3.3	1.6	1.0
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 PPP-adjusted 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 GDP and CO2 emissions, I use 2019 World Bank data. As in Nordhaus (2015), the model employs PPP-adjusted GDP. 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 in sensitivity analysis.

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).

[22]The difference reflects emissions intensity computed using PPP-adjusted GDP. With nominal GDP, the increase in relative emissions intensity over the period is closer to 30 percent.

The tariff impact functions maintain the calibrated parameters 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[23]. I combine total international merchandise trade and total trade in services. Appendix A.9 provides bilateral trade matrices for each category separately.

To calibrate the match rate—parameter b in Eq. 10—I compute the endogenous outcome of the model for a range of b values, holding fixed the rest of the baseline calibration. For each outcome, I back out the implied tariffs needed to ensure tier-2 compliance.[24] Repeating this exercise for a range of b values, I conclude that $b = 1$ is roughly plausible, and I use this value for the baseline calibration. In calibrating b, it is important to note that the "true" value of b would likely depend on a variety of factors outside the model, including international goodwill, the size distribution of Tier 1 countries, and global buy-in to the idea of using trade threats to bolster climate policy. Given these diverse factors, I view the calibration of Eq. 10 as approximate, and I conduct extensive sensitivity analysis around the value of b to see how the outcomes change when this key parameter is adjusted.[25]

Table 2 reports values for the remaining parameters that do not vary across countries. Units are explained in the figure caption. To calibrate the SCC for CO2, I use the latest release of the US EPA's Final Report on the Social Cost of Greenhouse Gases. The middle estimate in this report, which uses a 2 percent discount rate, finds a 193 USD/tCO2 for 2020. I use the somewhat lower estimate of 120 USD/tCO2, derived with a 2.5 percent discount rate. 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 5.5.

Q	E	σ	γ	b
139.2	9.5	0.108	0.440	1.0

Table 2: Other model parameters that are not country-specific. Q is PPP-adjusted 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. 10.

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

[24]An example of backing out the implied tariffs is presented in Figure 6 in Section 5.4 for the value of b in the baseline calibration.

[25]When using Eq. 10 to determine the match rate, I use nominal GDP (not PPP-adjusted GDP) since this reflects the size of the market that other countries export into.

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 i implements domestic carbon price

$$\tau_i = \max(\gamma_i, \hat{\tau}_2), \tag{33}$$

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.10 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, 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 19 and 20—under all possible permutations of the eleven non-passive economies in the quantitative model. For the baseline calibration, there are three stable coalitions: the EU-US, the EU-UK, and the EU-Australia. While the stable coalition is not unique, a strong argument can be made that the EU-US coalition is by far the most compelling.

To see this, Figure 3 presents national payoffs for a subset of countries under each of the three stable coalitions. Payoffs are computed relative to those in the Nash equilibrium. The endogenous match rate for a EU-US first tier is 42 percent, compared with just 21 percent with the EU and UK first tier (which has a somewhat higher match rate than a EU-Australia first tier). As a result, global abatement is roughly three times higher, and all countries except the US overwhelmingly prefer the EU-US coalition to the other two.[26] The US prefers to remain in the second tier under either of the other coalitions, but its benefit from staying out is relatively small. Specifically, the US payoff under the EU-UK first tier is 3.6 billion USD per year higher than in the EU-US first tier, while the combined global gain in moving from the EU-UK coalition to the EU-US coalition is about 320 billion USD per year. Thus, with minimal transfers, the EU-US coalition Pareto dominates the other stable coalitions. For this reason, I focus on the US-EU coalition for the main analysis in the paper.

Importantly, while the TCC agreement has the Tier 1 countries adopt a coercive stance toward Tier 2 countries, all countries are strictly better off. This leaves room for diplomatic efforts to convince Tier 2 countries to support a TCC agreement.

Notably, the EU is a core participant in all three stable coalitions. Moreover, we will see in the robustness section below that when we look across a range of parameter combinations, the set of potential stable first-tier coalitions increases somewhat. in all cases, however, the stable first tier consists of two participants and the EU is one of them.

Three characteristics contribute to the EU having the most to gain by joining the first tier. First, it has relatively high nominal GDP, giving it substantial trading clout that enables it to meaningfully increase the match rate when joining the first tier. It also has relatively high climate damages as a fraction of the global total, which increases its incentive to internalize its own emissions compared to most other countries. While these traits are important, both are shared with the US and China. The key difference between the EU and the US (or China) is the EU's lower emissions intensity (CO2

[26] While the figure only shows a subset of countries, the other countries in the model also benefit most from the EU-US first tier.

emissions over PPP-adjusted GDP). With less carbon emissions per unit of economic activity, a given carbon price is less costly. Emissions intensity in the US is 30 percent higher than in the EU, while that in China is higher by a factor of four.[27] This means that the cost of facing a higher carbon price by moving into the first tier is lower for the EU than for the US or China.

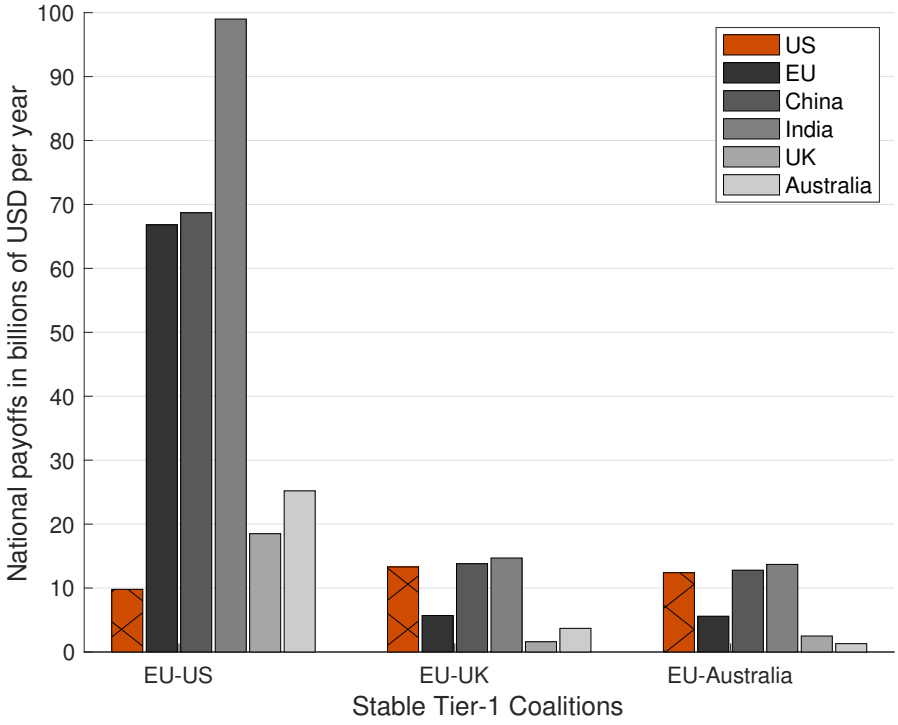

Figure 3: Comparison of payoffs relative to Nash for select nations under the stable tier-1 coalitions.

The set of stable coalitions under a TCC differs from what obtains with a traditional (single-tier) climate club in two notable ways. First, with single-tier climate clubs, small clubs are typically unstable and stable clubs are typically large. In contrast, we find here that the stable coalitions with the TCC have only two countries. In Section 5.5, we study how the set of stable coalitions varies as model parameters change. While it is possible to have other stable coalitions than the three here, there are no stable coalitions with more than two countries.

TCCs give rise to small stable coalitions because the incentive to move from the second tier to the first declines as more countries join the first tier. When the first tier is small, the match rate is low, which means that marginal abatement costs are very low in the second tier. By joining the first tier in that situation, the effect on global emissions from increasing the match rate is more pronounced due to the low marginal abatement costs in the second tier. Moreover, when the first tier is small, the second tier is large, which itself makes the amplification effect large for the existing first tier. However, when additional countries join the first tier, they reduce the size of the second tier, which lowers the amplification effect. Both effects imply declining benefits of moving into the first tier as the set of

[27] In Appendix A.11, I show how payoffs for different countries vary depending on whether or not they choose to enter or stay out of the EU-US coalition.

tier-1 countries increases.

The tendency for small clubs suggests a major advantage of a TCC. While large coalitions have more trading clout, enabling them to achieve more global abatement, they also require a much larger degree of international coordination to achieve. Large clubs are consequently much harder to get started.

The second difference between the set of stable coalitions under a TCC and a single-tier climate club has to do with the number of stable coalitions. For a conventional climate club, there are typically many stable coalitions (Hagen and Schneider, 2021, e.g.,). In contrast, here we find just three, and two of the three are marginal and unlikely to obtain in practice. The sensitivity analysis in Section 5.5 shows that the number of stable coalitions remains small across a wide range of the parameter space. The tendency for a small number of stable coalitions is useful since it gives clearer guidance about which countries should be expected to lead the establishment of a TCC.

5.2 China's interest

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 EU-US first tier. The top panel shows how China's choice to join or not would impact equilibrium outcomes, including the match rate, the first tier carbon price, and the total amount of global abatement. The bottom panel shows the corresponding impact on benefits and costs for China.

At first pass, there might seem to be an inconsistency between the impact of Chinese participation on the match rate and its impact on the first tier coalition's carbon price. By joining the first tier, China increases the match rate for the EU-US coalition from 42 percent to 59 percent. According to the formula for the cooperative carbon price in Eq. 22, this change should heighten the amplification effect, increasing the first tier carbon price. But the change in the first their carbon price when China joins goes the opposite direction, dropping from 60 percent of the efficient level when China is out to 51 percent when it is in. The reason can be seen from the same formula—Eq. 22. In particular, since China comprises 30 percent of global CO2 emissions, when it is part of the second tier, it provides a large incentive for the Tier 1 countries to price carbon at a high level. However, if it joins the first tier, 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 $\frac{1-\phi_{C'}^E}{\phi_{C'}^E} = \frac{1-0.542}{0.542} = 0.85$. Because China's size in nominal 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, so the net effect is to reduce the first tier carbon price.

The bottom panel shows the net impact of these changes on China's economy. By joining the coalition, China's domestic climate benefit goes up, but not by near as much as the increase in domestic abatement costs. On net, it goes from gaining almost 70 billion USD per year to losing 10 billion.

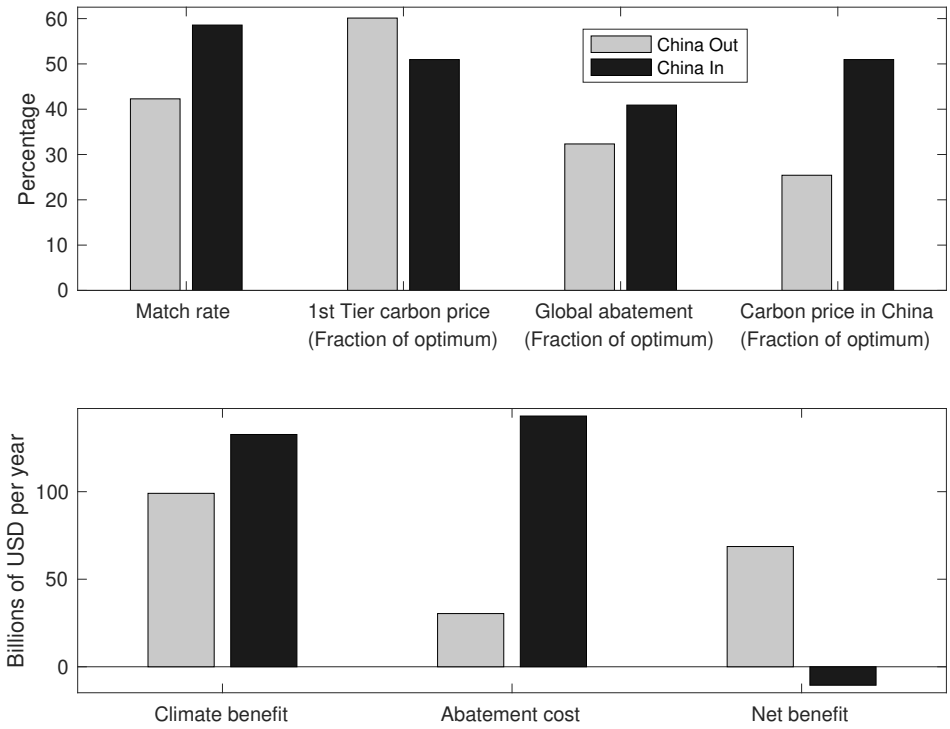

Figure 4: Grey bars indicate outcomes under the stable coalition with China out (EU+US), while black bars indicate outcomes with China in (EU+US+China). The top panel compares equilibrium outcomes under each coalition. The bottom panel decomposes the corresponding payoffs to China.

5.3 Range of supportable policies

In the analysis above, the endogenous match rate for the US-EU first-tier coalition is 42.3 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 tier-1 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 under non-cooperative policy. The shaded region in between shows the range of policies that could be supported for intermediate degrees of cooperation within the first tier. 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, the equilibrium match rate (coincident with the dashed vertical line) implies a first tier carbon price 60 percent of the globally efficient level and global abatement 33 percent of the efficient level. To appreciate the accomplishment of a TCC led by the EU and US, it is important to compare it with the corresponding outcome without external coercion—thus, in a world that resembles the status quo arrangement under the Paris Agreement. In the Nash equilibrium—equivalently, in the noncooperative TCC with a match rate of zero—global abatement is just 7 percent of the efficient level. Thus, the EU-US TCC leads to a nearly five-fold increase in global

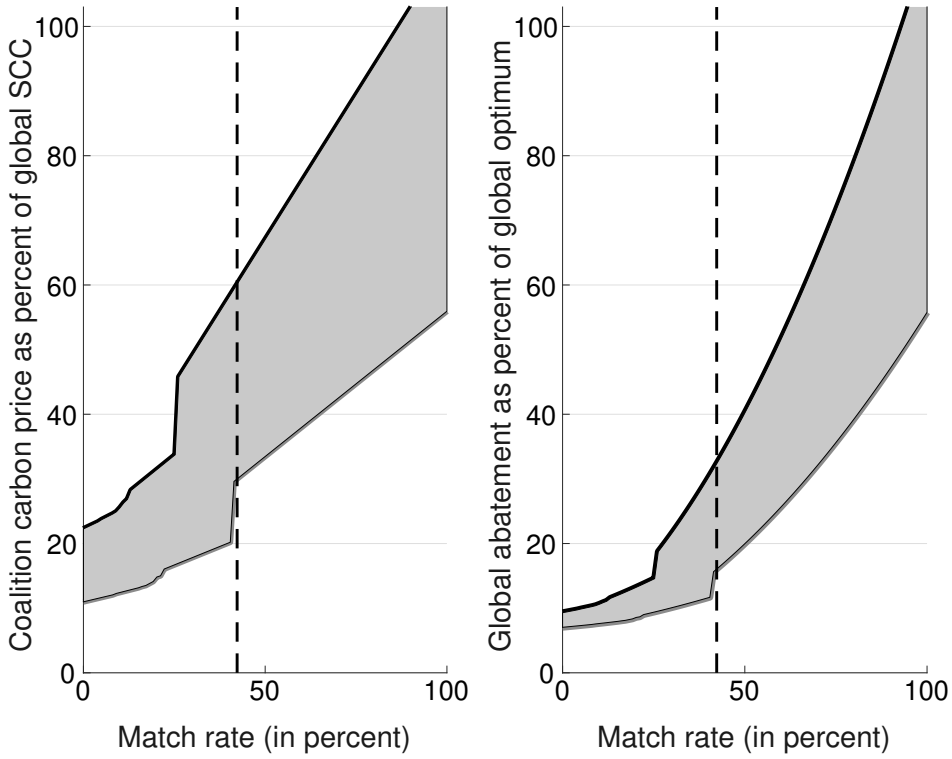

Figure 5: The left panel plots the range of supportable average Tier 1 carbon prices with a US-EU coalition as a function of α. The right panel plots the corresponding range of global abatement levels. In each panel, the grey line shows the non-cooperative policy, while the black line shows the cooperative 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.

abatement.

While the calibrated match rate is 42 percent under an EU-US first tier, it is interesting to see how close we might come to achieving global efficiency under a EU-US led TCC. If the match rate increases to 87 percent, the cooperative first tier would set its carbon price at exactly the global SCC. In this case, global abatement remains below the efficient level because the second tier does somewhat less. However, if the match rate increases to 93 percent, then the policy achieves the efficient level of global abatement.[28] In this case, the climate outcome coincides with the global optimum, though the overall outcome remains inefficient since carbon prices differ somewhat across regions, implying inefficiently high aggregate abatement costs.

Recall from the overshooting discussion in Section 3.1 (in a more restrictive setting) that the 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 coalition are $\frac{\phi_C^\gamma}{\phi_C^E} = 1.1$ and $\frac{\sigma_R}{\sigma_C} = 1.6$. Thus, while both channels induce overshooting, the main effect comes from

[28]The degree of overshooting is substantially lower in this version of the paper due to the corrected calibration error described in Appendix A.1.

the relatively low emissions intensity of the first tier.

Inspecting Figure 5 shows that the relationship between the match rate and the coalition carbon price is jagged when the match rate is low. 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—both large nations with a relatively high fraction of global climate damages. 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 30 percent for both. Because China and India together comprise 38 percent of global CO2 emissions,[29] this transition substantially increases the coalition incentive to price carbon at a higher level, which explains the sharp jump in the cooperative carbon price in the viscinity of a match rate of 25 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 than the cooperative carbon price.

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 the US-EU coalition, which comprises only 22 percent of global CO2 emissions. A small coalition means the match rate applies to a larger non-coalition region. The slope of the carbon price line is higher after the threshold is passed under which the minimum carbon price target on China and India is binding. This reflects the fact that only when their participation constraint binds does marginal tier-1 policy impact marginal policy in China and India.

The non-cooperative policy provides an important lower bound on how much could be achieved if the first tier countries are unable 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 56 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 coalition, the index equals 54 percent. Proposition 4 would then imply that for a given match rate the ratio of non-cooperative abatement to cooperative abatement should be 54 percent. If we compute the same ratio for the quantitative model—restricting attention to match rates above 30 percent or so to avoid the jagged regions of the parameter space—we find that the ratio is roughly

[29] Based on 2019 emissions data.

constant around 50 percent. The slight divergence from the theoretical prediction stems from the impact of heterogeneity in abatement costs, which is not allowed under the assumptions of Proposition 4.

Finally, Figure 5 also shows how a standard (single-tier) climate club would perform if it were implemented by the US and EU together, with other countries choosing not to participate. This coincides with the cooperative case (black line) when the match rate is zero. In this case, the cooperative carbon price for the first tier is 22 percent of the global SCC, while the resulting amount of global abatement is just 9 percent of the globally efficient level. The small traditional climate club increases global abatement roughly 30 percent above what is achieved in the fully non-cooperative Nash equilibrium (where the gray line intersects the vertical axis). Global abatement remains low because a single-tier club only internalizes climate damages within its collective borders, and it cannot take advantage of cheap abatement opportunities outside the club. In contrast, by requiring tier-2 countries to share part of the burden, the TCC leads to global abatement that is 460 percent above Nash.

5.4 Tariffs

The maximum match rate that a first-tier coalition can implement without provoking trade retaliation is determined by the reduced-form match-rate function in Equation 10. With the calibrated value of b, a EU-US first-tier can implement a 42% match rate, whereas a EU-Australia coalition is limited to a 19% match rate. To evaluate the plausibility of this calibration, I derive the implied tier-1 tariffs needed to secure participation from each major economy in the cooperative equilibrium of the model. This analysis is conducted both for tier-2 countries, assuming full participation of tier-1 countries in imposing penalties, and for tier-1 countries with penalties levied by participating tier-1 countries only.

For tier-2 country i, the minimum tariff to ensure compliance is the minimum tier-1 tariff at which i's optimal deviation from the terms of the TCC does not justify the economic cost of the tariff. Suppose i faces minimum carbon price target $\underline{\tau}$ and the rest of the world complies with the TCC terms, implying carbon prices τ_{-i}. Then the minimum tariff to induce compliance from i is the $\hat{t}$ at which the following participation constraint binds:

$$\sum_{j \in \Omega} P_{ji}(\hat{t}) \geq \max_{\tau_i} \Big[B_i(\tau_i, \tau_{-i}) - C_i(\tau_i) \Big] - \Big[B_i(\underline{\tau}, \tau_{-i}) - C_i(\underline{\tau}) \Big]. \tag{34}$$

$P_{ji}(\hat{t})$, defined in Eq. 7, is the net income loss in i when j applies tariff $\hat{t}$ on imports from i, so the left side is the total net income loss when all tier-1 countries impose $\hat{t}$ on imports from i. The right side is the increase in net benefits, ignoring tariff costs, if i optimally deviates from the minimum carbon price target.

Given this relationship, Figure 6 backs out the minimum tariffs to induce compliance for each country in the cooperative equilibrium. The left panel shows minimum tariffs for tier-1 countries, and the right panel shows minimum tariffs for tier-2 countries. The dark line (upward-pointed triangles) shows the baseline case with punishments imposed by the EU and US first tier. The lighter dashed line

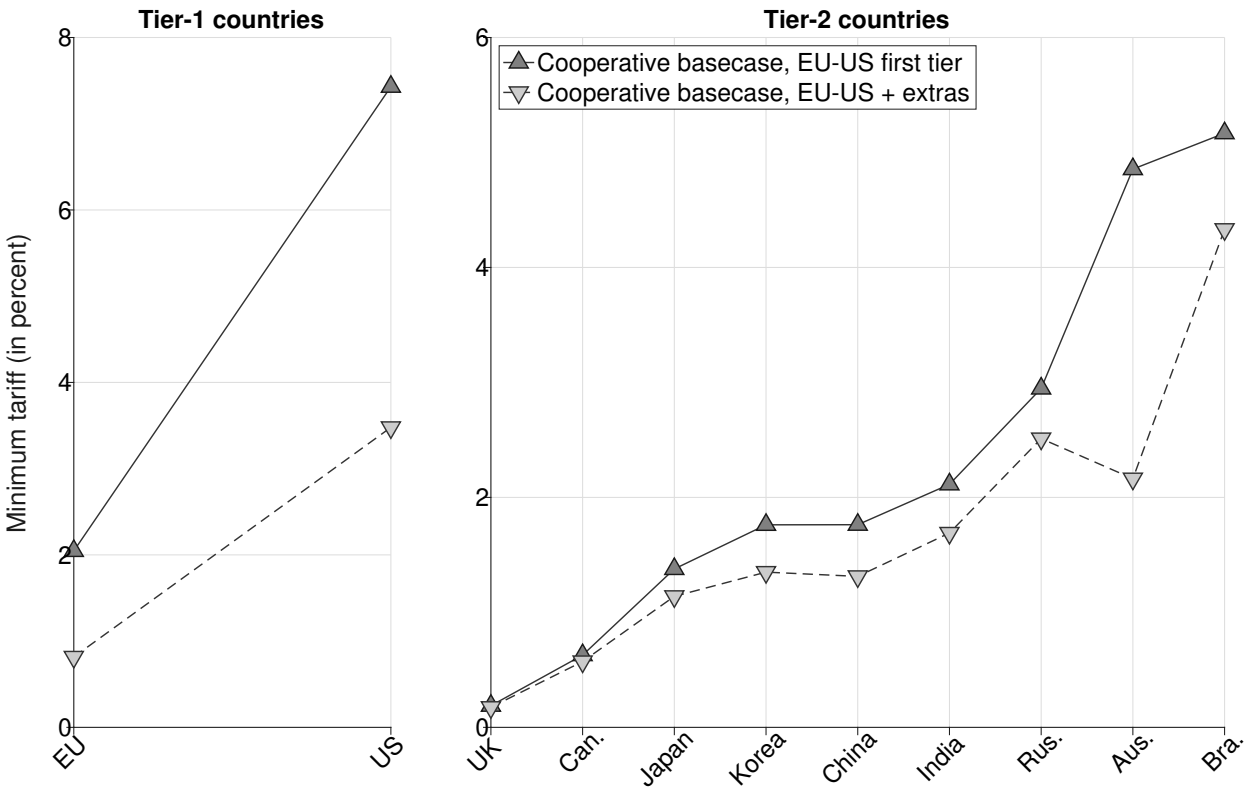

Figure 6: Minimum tariffs to sustain the cooperative equilibrium ($\alpha = 0.42$). Dark gray triangles represent penalties imposed by the EU and US. Light gray (inverted) triangles indicate scenarios where Japan, the UK, Canada, and Australia also impose tariffs.

(downward-pointed triangles) presents the alternative case in which punishments are imposed by the EU and US plus the other high-income countries from the 11 model regions—mainly, Japan, the UK, Canada, and Australia. The case with these additional "extras" shows how the minimum tariffs could be reduced if more countries are brought into the fold of administering punishments. The extras are held to the same abatement obligation that apply to other Tier-2 countries, so including these countries in the set of punishing countries does not change their incentive to comply with the agreement.

The minimum tariffs vary widely across countries. For example, to induce compliance from the UK and Canada, tariffs well below 1 percent are sufficient. This is mainly due to the high degree of trade connectivity between these countries and the US and EU. In contrast, Australia and Brazil, countries with far lower trade connectivity with the US and EU, require tariffs closer to 5 percent. Given the central role of trade connectivity in determining the implied tariffs, I plot export flows from each of the model countries into the US and the EU (and into the US and EU plus extras) in Figure 9. To save space, this figure is relegated to Appendix A.12.

Despite the variation across tier-2 countries, the implied tariffs are plausibly low, suggesting that the calibrated value of b is at least roughly plausible. Nevertheless, it is also clear that the "true" value of b is hard to know with precision. For example, the match rate that a given first tier could get away with without triggering a trade war would ultimately depend on a variety of diplomatic factors,

such as the degree of global buy-in to the idea of using trade tariffs in a coercive way to support global efforts to reduce carbon emissions. in practice, a given coalition would likely begin with a relatively low value for the match rate, then slowly ratchet it up over time with an eye to seeing how high a value would be diplomatically feasible. Given the inevitable imprecision around our calibration of b, I use the basecase calibrated model as a central reference, while also, studying the impact of alternative values of b on the main results (Section 5.3).

A further takeaway from Figure 6 is that the greatest challenge to achieving the cooperative equilibrium is likely to come in the need to secure US compliance. For the US to be willing to follow through with the cooperative outcome, the EU would need to threaten import tariffs against the US exceeding 7%. The threshold drops below 4% if the punishing coalition expands to include Japan, the UK, Canada, and Australia. The challenge of incentivizing US compliance with the cooperative policy suggests that it might be necessary to implement the policy with a less than fully cooperative outcome within the first tier. Indeed, as discussed in Section 5.3, a range of possible outcomes are possible depending on how willing the tier-1 countries are to use tariff threats with each other. At the bottom of this range, the noncooperative outcome depicted in Figure 5 corresponds to the case in which the tier-1 countries do not use tariff threats with each other. In that case, While global abatement drops significantly without tier-1 cooperation, it is still more than twice as high achieved in the Nash equilibrium without a TCC and more than 70% higher than under a traditional EU-US climate club. Moreoever, intermediate values of global abatement could be achieved with tariffs threats above zero but still below the level needed to support the fully cooperative outcome. Including other high income countries in the punishment regime would also help substantially in pushing the US further in the direction of the cooperative outcome.

5.5 Sensitivity analysis

The discussion in Section 5.1 showed that the stable first tier is not unique. In this section, I consider how Tier-1 stability changes when the match rate parameter b and the global Social Cost of Carbon γ vary from the baseline calibration. While the set of potential stable first tiers increases somewhat relative to the baseline calibration, the broad conclusions do not change.

The calibrated match rate ($b = 1$) assumes that a given coalition can impose a match rate on the rest of the world equal to the coalition's fraction of global GDP. Here, I consider a range of b values between 0.7 and 1.0. At the lower end of the range, a EU-US first tier, which makes up 42 percent of global GDP, can only impose a match rate of 29 percent on the rest of the world. In addition, the SCC γ varies between 100 USD/tCO2 and 150 USD/tCO2. The results are summarized in Figure 7.

The left axis (blue bars) shows global abatement for each first tier (computed for the baseline calibration). For comparison, global abatement achieved in the Nash equilibrium is indicated by the dashed black line. The right axis (black line with solid dots) shows how frequently each first tier appears as a stable coalition when looking across the range of parameters.

Across the parameter space, five stable first tier coalitions occur. Of these, the EU-US, EU-UK,

and EU-Australia are most common, each arising in roughly 70 percent of simulations. In addition, the EU-India is stable in 45 percent of simulations, and the EU-Korea is stable in 20 percent of simulations. As we saw before, there is a large difference between the global abatement achieved by the EU-US first tier and that achieved by the other stable coalitions. The EU-UK, EU-Australia, and EU-Korea each increase global abatement about 50 percent above what is achieved in the Nash equilibrium. The EU-India does somewhat better, increasing abatement 90 percent above Nash.[30] In contrast, the EU-US coalition increases global abatement 460 percent above Nash.

Despite the expanded set of stable coalitions, the broad conclusion from Section 5.1 remains unchanged. The EU-US first tier generates a far more attractive outcome for the world than any other stable first tier. While the US would slightly prefer one of the other stable coalitions in which it remains in the second tier, the increase in global surplus generated in moving to the EU-US-led TCC vastly outweighs inconvenience to the US, suggesting that modest transfers could achieve a Pareto improvement for all countries.

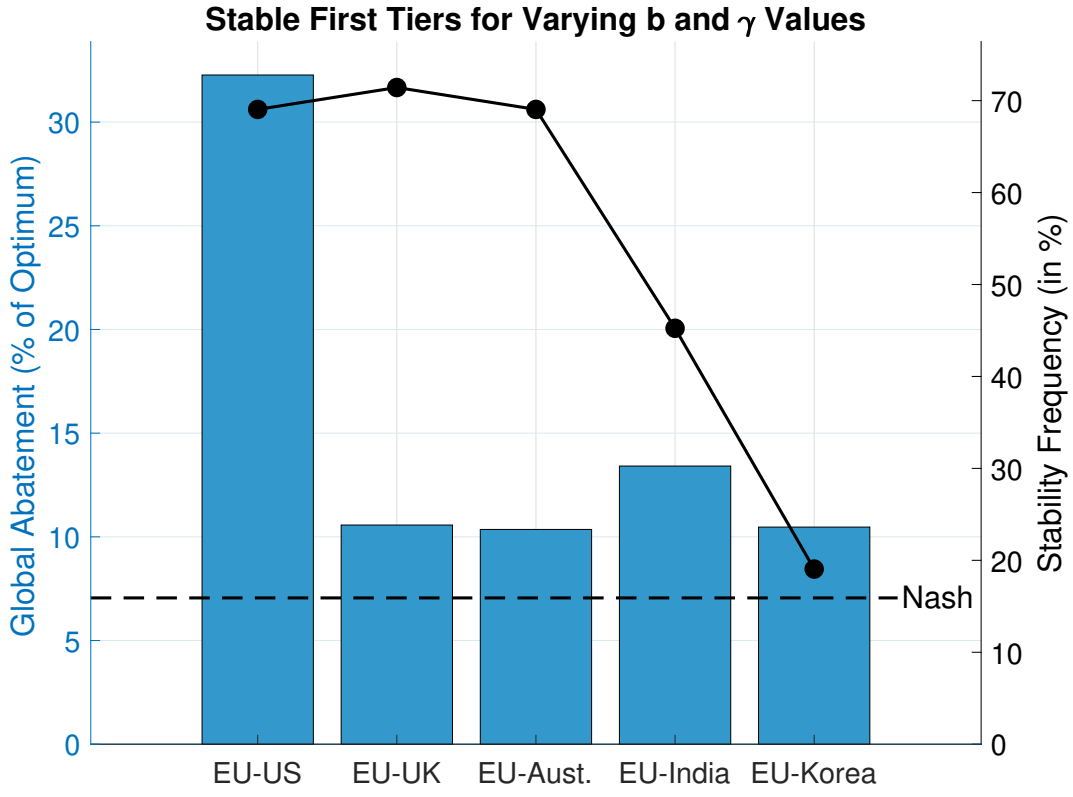

Figure 7: Alternative stable Tier 1 coalitions based on simulations across a range of parameter values described in the text. Each stable first tier includes the EU plus one other country. Left axis (blue bars) show global abatement; right axis (black line) shows occurence frequency for each stable first-tier coalition across simulations.

[30]The EU-India coalition achieves more abatement both because India has higher GDP, implying a higher match rate, and because it has higher climate damages and lower emissions intensity, both of which increase its incentive to price carbon at a higher level when participating in the first tier.

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

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. 2024. "Can trade policy mitigate climate change." *Unpublished Working Paper*.

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.

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.

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.

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 Explanation for change in quantitative results

Earlier versions of this paper used nominal GDP when calibrating the model, which is inconsistent with the implementation in Nordhaus (2015). The current version correctly uses PPP-adjusted GDP instead. The change impacts both the coalition stability results and the amount of global abatement achieved by a given coalition. When nominal GDP is used, a first tier that includes the EU and US is much less carbon intensive than the rest of the world. When PPP-adjusted GDP is used instead, this effect is attenuated since less developed countries have higher effective GDP, lowering their effective carbon intensity. The net effect is to significantly reduce the optimal carbon price of the first tier (as a fraction of the global optimum). The mechanism for this effect is explained in the explanation of Proposition 1.

A.2 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

$$\frac{\partial \psi_i}{\partial \hat{\mu}_i} = \theta_2 Q_i \theta_1 \left(\frac{\hat{\mu}_i}{E_i}\right)^{\theta_2 - 1} \frac{1}{E_i} \tag{35}$$

$$= \theta_1 \theta_2 \frac{Q_i}{E_i} \left(\frac{\hat{\mu}_i}{E_i}\right)^{\theta_2 - 1} \tag{36}$$

$$= \frac{\theta_1 \theta_2}{\sigma_i} \mu_i^{\theta_2 - 1}, \tag{37}$$

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:

$$\frac{\theta_1 \theta_2}{\sigma_i} \mu_i^{\theta_2 - 1} = \tau_i. \tag{38}$$

Solving for μ_i gives:

$$\mu_i = \left[\frac{\sigma_i}{\theta_1 \theta_2}\right]^{\frac{1}{\theta_2 - 1}} \tau_i^{\frac{1}{\theta_2 - 1}} \equiv G_i(\tau_i). \tag{39}$$

I often write

$$\mu_i = a_i (tau_i)^b, \tag{40}$$

where

$$a_i = \left[\frac{\sigma_i}{\theta_1 \theta_2}\right]^{\frac{1}{\theta_2 - 1}} \tag{41}$$

and

$$b = \frac{1}{\theta_2 - 1}. \tag{42}$$

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.3 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. **??**. 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.4 Proof of Proposition 1

The solution to Problem 14 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:

$$\max_{\{\tau^C \geq 0\}_{i\in\Omega}} \sum_{i\in\Omega} \left[\gamma_i E\mu(\{\tau^C;\alpha) - C_i(\tau^C)\right] \tag{43}$$

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

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

From Eq. 4, $C_i(\tau^C) = Q_i\theta_1\left(\frac{\sigma_\Omega}{2\theta_1}\right)^2(\tau^C)^2$, so

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

Combining and simplifying gives

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

Simplifying further gives

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

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.5 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

$$\tau_j = \alpha_j\tau_C. \tag{44}$$

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.2). 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

$$P(\boldsymbol{\phi}, \boldsymbol{\alpha}) = \frac{\phi_0 + \sum_{i=1}^{n} \phi_i \alpha_i^{\theta_2/(\theta_2-1)}}{\left(\phi_0 + \sum_{j=1}^{n} \phi_j \alpha_j^{1/(\theta_2-1)}\right)^{\theta_2}}. \tag{45}$$

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

$$P(\phi_0, \alpha) = \frac{\phi_0 + (1-\phi_0)\alpha^{\theta_2/(\theta_2-1)}}{\left[\phi_0 + (1-\phi_0)\alpha^{1/(\theta_2-1)}\right]^{\theta_2}}. \tag{46}$$

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

$$\Psi = \sum_{j=0}^{n} \Psi_j \tag{47}$$

$$= \sum_{j=0}^{n} \phi_j Q\theta_1 G(\alpha_j\tau_0)^{\theta_2} \tag{48}$$

$$= Q\theta_1 G(\tau_0)^{\theta_2} \sum_{j=0}^{n} \phi_j \alpha_j^{\frac{\theta_2}{\theta_2-1}} \tag{49}$$

$$= Q\theta_1 \mu^{\theta_2} \cdot \Gamma^{\theta_2} \cdot \sum_{j=0}^{n} \phi_j \alpha_j^{\frac{\theta_2}{\theta_2-1}} \tag{50}$$

$$= \frac{\sum_{i=0}^{n} \phi_i \alpha_i^{\theta_2/(\theta_2-1)}}{\left(\sum_{j=0}^{n} \phi_j \alpha_j^{1/(\theta_2-1)}\right)^{\theta_2}} \cdot Q\theta_1\mu^{\theta_2} \tag{51}$$

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

$$\sum_{i=0}^{n} \phi_i \alpha_i^{\theta_2/(\theta_2-1)} = \phi_0 \alpha_0^{\theta_2/(\theta_2-1)} + \phi_1 \alpha_1^{\theta_2/(\theta_2-1)} \tag{52}$$

$$= \phi + (1-\phi)\alpha_1^{\theta_2/(\theta_2-1)}. \tag{53}$$

Equation 46 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 46 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

$$(1-\phi)\theta_2 \gamma \alpha^{\theta_2 \gamma - 1} [\phi + (1-\phi)\alpha^{\gamma}]^{\theta_2} < \theta_2 [\phi + (1-\phi)\alpha^{\gamma}]^{\theta_2 - 1} \gamma (1-\phi)\alpha^{\gamma - 1} [\phi + (1-\phi)\alpha^{\theta_2 \gamma}]. \tag{54}$$

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, (54) 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.6 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. 4, $C_i(\tau_i) = Q_i\theta_1\left(\frac{\sigma_\Omega}{2\theta_1}\right)^2(\tau_i)^2$, 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.7 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.8 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

$$\tau_i = \gamma\phi[1 + \frac{1-\phi}{\phi}\alpha]. \tag{55}$$

Proof. The objective function for the OCP problem is

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

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

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

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 (55).

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*

$$\tau_i^* = \phi_i\gamma[1 + \frac{1-\phi}{\phi}\alpha] = \hat{\phi}_i\tau^C(\alpha) < \tau^C(\alpha). \tag{56}$$

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

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

$$\Pi_i(\tau_i, \{\tau_j\}_{j\neq i}) = \phi_i\gamma E\left[\phi_i\frac{\sigma}{\theta_1\theta_2}\tau_i + \sum_{j\neq i}\phi_j\frac{\sigma}{\theta_1\theta_2}\tau_j + (1-\phi)\frac{\sigma}{\theta_1\theta_2}\alpha\hat{\phi}_i\tau_i + (1-\phi)\frac{\sigma}{\theta_1\theta_2}\alpha\sum_{j\neq i}\hat{\phi}_j\tau_j\right] \tag{57}$$

$$- \phi_i Q\theta_1\left(\frac{\sigma}{\theta_1\theta_2}\right)^2(\tau_i)^2. \tag{58}$$

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

$$\tau_i^*(\hat{\tau}_i - \tau_i^*) = \frac{1}{\theta_2}[(\hat{\tau}_i)^2 - (\tau_i^*)^2] - \omega_1\frac{\theta_1\theta_2}{\sigma^2}. \tag{59}$$

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

$$(\hat{\tau}_i)^2 - (\tau_i^*)^2 = (\hat{\tau}_i - \tau_i^*)(\hat{\tau}_i + \tau_i^*) \tag{60}$$
$$= x(x + 2\tau_i^*) \tag{61}$$
$$= x^2 + 2\tau_i^* x. \tag{62}$$

Substituting into Eq. 59 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.9 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.10 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,

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

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

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

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

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

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

$$\tau^{COOP} = \gamma_C \frac{\sigma}{\theta_2} \frac{\Theta_1}{\Theta_2} \tag{63}$$

$$= \gamma_C \frac{\sigma}{\theta_2} \frac{\sum_j \phi_j^E a_j + (1-\phi_C^E) a_R \alpha^b}{\sum_{i=1}^n \phi_i^Q \theta_{1,i} a_i^{\theta_2}}. \tag{64}$$

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

$$B_i(\hat{\tau}_i, \{\hat{\tau}_j\}_{j\neq i}; \alpha) - C_i(\hat{\tau}_i) \geq \max_{\tau_i \geq 0} [B_i(\tau_i, \{\hat{\tau}_j\}_{j\neq i}; \alpha) - C_i(\tau_i)] \tag{65}$$

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 (65) 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

$$\mu(\tau_i, \{\hat{\tau}_j\}_{j\neq i}; \alpha) = \phi_i^E \mu_i + \sum_{j\neq i} \phi_j^E \hat{\mu}_j + (1-\phi_C^E)\mu^R \tag{66}$$

$$= \phi_i^E a \tau_i^b + \sum_{j\neq i} \phi_j^E a \hat{\tau}_j^b + (1-\phi_C^E) a [\alpha \hat{\phi}_i^E \tau_i + \alpha \sum_{j\neq i} \hat{\phi}_j^E \hat{\tau}_j]^b. \tag{67}$$

Thus,

$$\frac{\partial \mu}{\partial \tau_i} = \phi_i^E ab \tau_i^{b-1} + (1-\phi_C^E) ab (\alpha \tau^{AVG})^{b-1} \alpha \hat{\phi}_i^E > 0, \tag{68}$$

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

$$\gamma_i E[\phi_i^E ab\tau_i^{b-1} + (1-\phi_C^E)ab(\alpha\tau^{AVG})^{b-1}\alpha\hat{\phi}_i^E] = \theta_1 a^{\theta_2} b\theta_2 (\tau_i)^b Q_i \tag{69}$$

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

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

iff

$$\tau_i = \gamma_i \left[1 + \alpha^b \frac{1-\phi_C^E}{\phi_C^E}\left(\frac{\sigma_R}{\sigma_i}\frac{\theta_{1,i}}{\theta_{1,R}}\right)^b (\frac{\tau_i}{\tau^{AVG}})^{1-b}\right], \quad \text{for } i = 1, \ldots, n. \tag{70}$$

A.11 National payoffs in stable coalition

A.12 Additional analysis related to minimum tariff calculations

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. Analysis based on 2019 UNCTAD data. Ignoring the "extras" segment at the top of each bar, the combined segments for the US and EU 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 2.5 percent of GDP) is least exposed to trade threats from the stable coalition (in this case just the EU). Next is Australia at 2.7 percent, then the EU (2.9 percent), and Brazil (4.3 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, the UK, 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, the UK, Canada, and Australia has the biggest effect on the US, Australia, and the EU, which are also the countries least connected to the stable first tier.

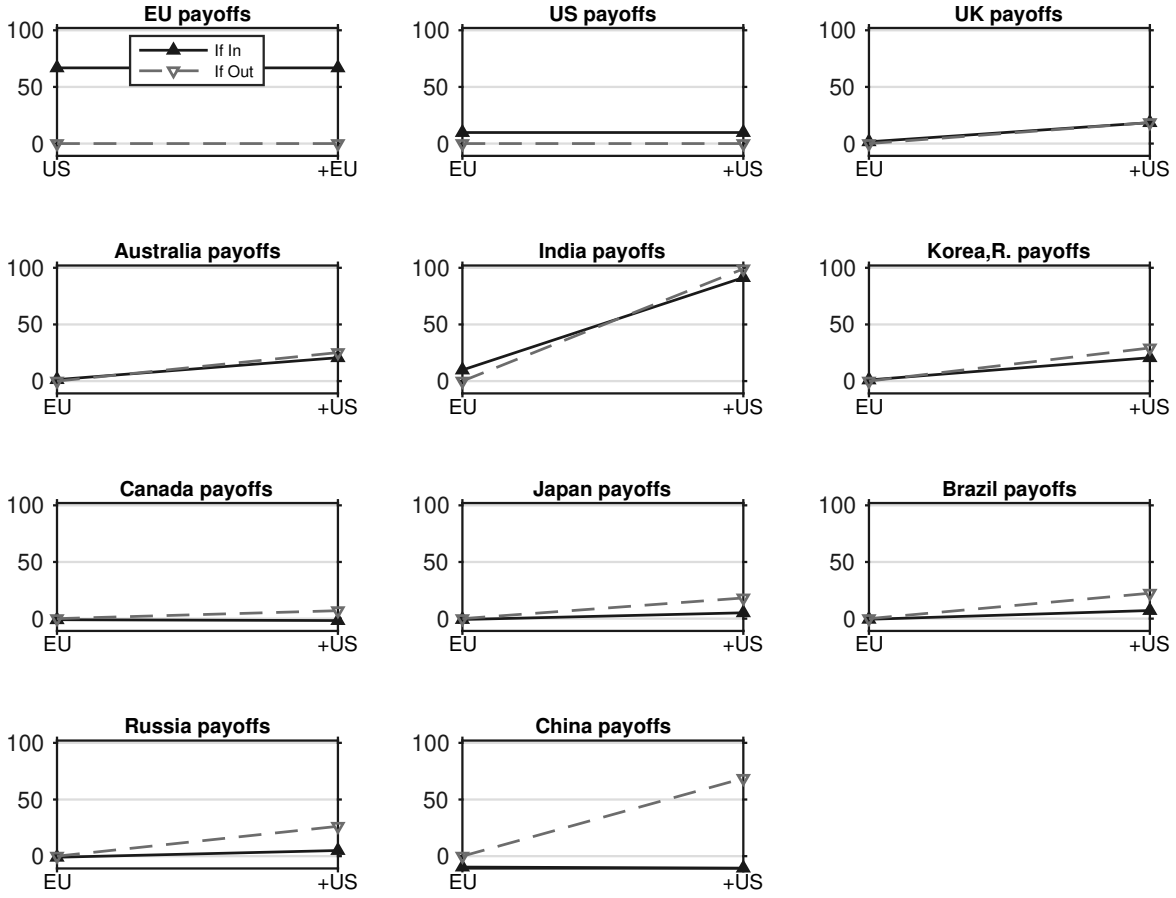

Figure 8: Payoffs relative to Nash payoff (in billions of USD per year). Horizontal axis depicts alternative coalitions. Left coalition is Nash. Right coalition is EU-US. Solid-black line indicates payoffs if country stays in or joins the indicated coalition; dashed-grey line indicates payoffs if it stays out or leaves. Economies sorted in order of descending net benefit to join.

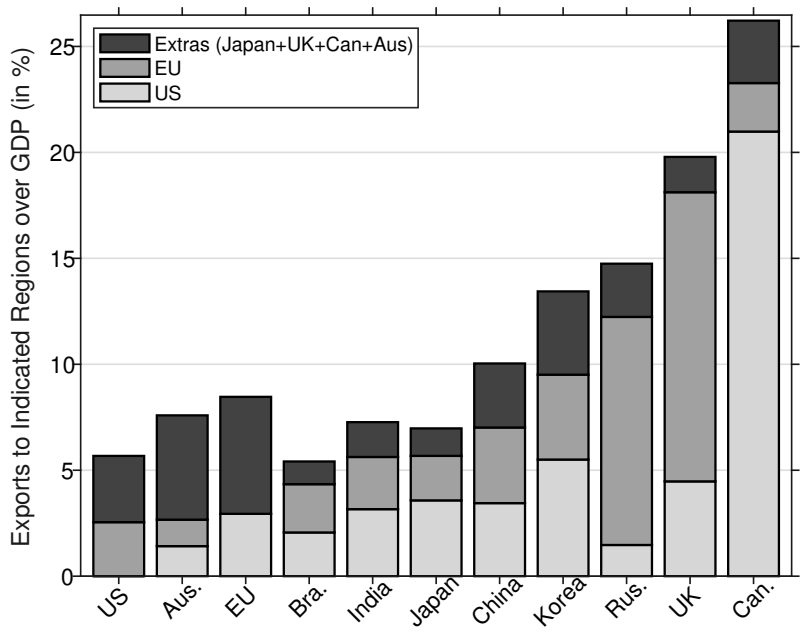

Figure 9: The bar for each country decomposes exports of goods and services to the regions indicated by segment color. Export flows expressed as fraction of export country GDP. Bars sorted low to high in combined exports to the US and EU.