Iverson, Terrence

Working Paper

Advancing Global Carbon Abatement with a Two-Tier Climate Club

CESifo Working Paper, No. 9831

Provided in Cooperation with:
Ifo Institute – Leibniz Institute for Economic Research at the University of Munich

Suggested Citation: Iverson, Terrence (2022) : Advancing Global Carbon Abatement with a Two-Tier Climate Club, CESifo Working Paper, No. 9831, Center for Economic Studies and ifo Institute (CESifo), Munich

This Version is available at:
https://hdl.handle.net/10419/263761

CESifo WORKING PAPERS

9831
2022

July 2022

Advancing Global Carbon Abatement with a Two-Tier Climate Club

Terrence Iverson

CESifo

Impressum:

CESifo Working Papers
ISSN 2364-1428 (electronic version)
Publisher and distributor: Munich Society for the Promotion of Economic Research - CESifo GmbH
The international platform of Ludwigs-Maximilians University's Center for Economic Studies and the ifo Institute
Poschingerstr. 5, 81679 Munich, Germany
Telephone +49 (0)89 2180-2740, Telefax +49 (0)89 2180-17845, email office@cesifo.de
Editor: Clemens Fuest
https://www.cesifo.org/en/wp
An electronic version of the paper may be downloaded
· from the SSRN website: www.SSRN.com
· from the RePEc website: www.RePEc.org
· from the CESifo website: https://www.cesifo.org/en/wp

CESifo *Working Paper No. 9831*

Advancing Global Carbon Abatement with a Two-Tier Climate Club

Abstract

A two-tier climate club exploits the comparative advantage of large countries to mete out punishments through trade, while taking their capacity to resist punishment as a constraint. Countries outside the coalition price carbon at a fixed fraction of the average carbon price adopted within the coalition, or face tariffs. Coalition countries abate more since doing so induces matching abatement elsewhere. If the rate at which noncoalition countries match coalition abatement goes to one, equilibrium abatement approximates the globally efficient outcome even though the coalition only internalizes damages within its borders. Even with a low match rate, the arrangement drastically reduces aggregate abatement costs. In contrast to a single-tier climate club in which many stable coalitions are possible, the stable coalition in the calibrated model is unique and consists of the US and the EU. Global abatement achieved by the stable agreement is about 40 percent of the efficient level.

JEL-Codes: Q540, Q560, Q580, F180, F530, H230, H410.

Keywords: international environmental agreement, climate club, trade sanctions, retaliation, incomplete participation costs, country-size heterogeneity.

Terrence Iverson
Department of Economics
Colorado State University
1771 Campus Delivery
USA – Fort Collins CO, 80523
terry.iverson@colostate.edu

July 4, 2022
The author thanks Ed Barbier, Jo Burgess, Jared Carbone, Achim Hagan, Larry Karp, David Kelly, Hiroaki Sakamoto, Robert Schmidt, Christian Traeger and participants at the Front Range Energy Economics Workshop and SURED 2022 for helpful conversations and feedback. The author is grateful for funding from the SoGES Resident Fellows Program at Colorado State University.

1 Introduction

When a country reduces its carbon emissions, most of the benefit accrues to other countries. Self-interest therefore pushes it to abate too little. The problem goes away if countries are forced to move in sync (Cramton et al. 2016). But since we lack a global government with the capacity to enforce such an arrangement, participants in a climate agreement need a mechanism for punishing free riders. A growing literature demonstrates the promise of trade penalties (Barrett 1997, Lessman et al. 2009, Helm and Schmidt 2015, Bohringer et al. 2016, Khourdajie and Finus 2020, Hagan et al. 2021). In a version of the idea, Nordhaus (2015) suggests a "climate club" in which countries pay dues in the form of carbon abatement in exchange for benefits in the form of avoided trade tariffs. The arrangement can support substantial abatement in a stable coalition with only modest tariffs on non-participants.

While compelling, at least three barriers hinder progress in developing an effective climate club. First, the potential for retaliatory tariffs by nonmembers raises the risk of a destructive trade war. Hagan et al. (2021) show that retaliation destabilizes small climate clubs since a critical mass of participants is needed to discourage retaliation. In addition, Bohringer and Rutherford (2017) show that big countries like the United States will be very hard to punish. Second, we lack clear guidance on how to get started. With a single-tier climate club, there are many possible stable coalitions (Nordhaus 2008, Hagan et al. 2021), and it is unclear who should go first. Third, since many stable coalitions include only a subset of countries, there is likely to be a substantial portion of the world outside the agreement, at least initially. Incomplete participation drastically increases global abatement costs, as discussed below.

To overcome these barriers, a two-tier climate club (TCC) exploits heterogeneity in country size, a feature of the problem mostly ignored in prior analyses of climate clubs. Big countries, like the US and the EU,[1] are vastly more powerful in trade negotiations than smaller countries (Broda et al. 2008, Ossa 2014). As a result, they are both better situated to mete out punishments using trade penalties and also harder to punish. In addition, big countries internalize a larger fraction of global climate damages and hence have more incentive to pursue ambitious climate policy as part of a coalition.

To exploit these differences, a two-tier climate club is led by a coalition of relatively large countries that together possess a high degree of leverage in trade negotiations.[2] The coalition uses conditional trade incentives to enforce a minimum carbon price that is set distinctly for countries inside the coalition—the tier-1 countries—and for those outside the coalition—the tier-2 countries. While tier-2 countries are not deliberate participants in the agreement, they still face conditional trade incentives that require them to price carbon at a minimum level or incur tariffs on exports to the coalition region. The minimum carbon price for tier-2 countries is a fixed fraction of the (size-weighted) average carbon price adopted within the coalition. I refer to this fixed fraction as the match rate. Tier-1 countries take the match rate as given when choosing their own policy. The linkage increases the incentive for tier-1 countries to abate since doing so increases abatement in the rest of the world—abatement that is effectively free from the perspective of the coalition. Finally, the climate club defines conditional trade incentives within the coalition that can be used to increase the degree of cooperation among tier-1 countries.

The proposed agreement surmounts the noted barriers in the following ways. First, it reduces the risk of a trade war by concentrating trading prowess within the coalition and differentiating trade threats across different types of countries to better reflect the capacity of the coalition to induce abatement in different countries. Also, because bigger countries are harder to punish, the increased incentive to contribute induced by the matching abatement structure reduces the extent to which tier-1 countries need to threaten each other to achieve a given level of abatement. Second, in contrast to a standard climate club for which many stable coalitions exist, I find in the calibrated model that the stable coalition is unique. The finding is explained below. Third, the two-tier structure sharply reduces the cost penalty for incomplete participation. To show this, I extend the analysis of incomplete participation costs in Nordhaus (2008) assuming non-coalition countries match a coalition carbon price at a fixed rate. The analysis reduces to that in Nordhaus (2008) when the match rate is zero, and it mimics a two-tier climate club when the match rate is positive. A US-EU coalition with a zero match increases global abatement costs by a factor of 10.5 relative to the efficient policy, but increasing the

[1] I loosely refer to the EU as a "country" assuming it has the capacity to choose climate policy that is in the collective best interest of its members.

[2] In the paper, the coalition is endogenous, but I find that the countries with the largest trading prowess have the largest incentive to join.

match to just 10 cents on the dollar decreases the cost penalty by more than a factor of four.[3]

The paper studies the impact of a TCC in a static model with abatement costs, climate damages, and heterogeneous countries. The results are divided into two parts. Section 4 analytically studies the optimal coalition problem when the set of coalition countries is fixed, and Section 5 numerically studies coalition stability.

In the analytical section, the "tier-1 penalty" (equivalently, the degree of issue linkage among tier-1 countries) and the match rate are both treated as politically determined parameters that reflect the extent to which a given coalition can induce abatement from tier-1 and tier-2 countries, respectively. I derive an analytical expression for the optimal coalition policy, which maximizes coalition surplus subject to participation constraints for all countries. Holding fixed the tier-1 penalty, the optimal policy increases in a term I call the amplification factor, which captures the extent to which matching abatement in the non-coalition region induces tier-1 countries to price carbon at a higher level than they otherwise would. The amplification factor captures the fact that the coalition gets matching abatement in the rest of the world, so the more emissions there are in the rest of the world relative to the coalition, the higher the coalition prices carbon.

I also provide a useful limiting property. If the coalition behaves cooperatively and the match rate is raised to one, then global abatement achieved by the policy is (at least) as high as the globally efficient level provided the portion of climate damages that accrue within the coalition are as large as (larger than) the portion of global CO2 emissions that arise within the coalition. The result obtains even though the coalition only internalizes damages within its borders. The intuition is that a coalition that only internalizes (say) 20 percent of global damages also recognizes that its abatement induces matching (thus, effectively "free") abatement in 80 percent of the world, which roughly makes up for the fact that the coalition only comprises (roughly) 20 percent of global damages. It follows that a two-tier climate club led by a modest number of large countries has the potential to replicate an efficient global agreement in an incentive-compatible way.

A further analytical result—Proposition 5—derives an expression for the global abatement rate when tier-1 countries play Nash with each other (i.e., the tier-1 penalty is zero). If the match rate is zero, the expression reduces to the known result (with quadratic abatement costs) that global abatement in the Nash equilibrium equals the efficient abatement rate times the Herfindahl index of country size for the world (Nordhaus 2015). But if the match rate increases to one, the global abatement rate becomes the efficient abatement rate times the (typically much larger) Herfindahl index of country size within the coalition. The effect is large if the set of coalition countries is small.

The second set of results study coalition stability. To study stability, it is necessary to take a stand on how changes in the set of coalition countries impacts both the degree of cooperation among tier-1 countries (equivalently, the tier-1 penalty) and the match rate. While the analytical section studies the range of supportable policies as the tier-1 penalty increases above zero, the section on coalition stability focuses on the simple case, consistent with most of the IEA literature since Barrett (1994), in which the coalition behaves cooperatively. In particular, the coalition implements the cooperative policy as long as it includes two or more countries; otherwise, each country plays Nash.

To model how changes in coalition size impact the match rate, I assume that the match rate is proportional to the fraction of global GDP inside the coalition. While rough, I use GDP because it is a reasonably good indicator of trading prowess (Ossa 2014). A coalition is stable if it is both internally stable—no one inside the coalition would be better off out—and externally stable— no one outside the coalition would be better off in. The quantitative model allows countries to differ in terms of trading clout, GDP, CO2 emissions (hence also emissions intensity), climate damages, and abatement costs.

Under these assumptions, the stable coalition is unique, and it consists of the US and the EU. Both countries are substantially better off joining the agreement than they would be under Nash: the EU gains over 20 billion USD per year, while the US gains about 10 billion USD per year. Under the stable agreement, tier-2 countries match the coalition carbon price at (roughly) 40 percent. Playing the tier-2 role in a US-EU led agreement, China benefits over 15 billion USD per year. In contrast, if China were to join the agreement as a tier-1 country, it would be worse off than under Nash. Global abatement achieved by the agreement is over 40 percent of the globally efficient level.

While the 40 percent match rate is intended to reflect the capacity of a US-EU coalition to induce noncoalition countries to match the tier-1 carbon price, it is also interesting to consider how much abatement a EU-US coalition would achieve if the match rate were increased to one. An option I

[3]See Figure 3 in Section 2.2.

discuss for increasing the match rate is to "promote" some economically powerful tier-2 countries into the role of helping to threaten import tariffs against noncompliant countries without changing the abatement requirements that these countries face as tier-2 countries (see Section 6). With a match rate of one, the cooperative carbon price that a US-EU-led coalition would optimally choose substantially overshoots the global Social Cost of Carbon (SCC). The overshooting occurs because a US-EU coalition accounts for a greater portion of global climate damages and a smaller portion of global CO2 emissions, controlling for size, than does the noncoalition region. In addition, the US and EU have lower emissions intensity than the rest of the world, which lowers the relative domestic cost of a given carbon price. A consequence of the high willingness of the US-EU coalition to price carbon at a relatively high level is that this coalition could achieve over 70 percent of the efficient global abatement rate even if the US and EU behave noncooperatively with each other. This strong result stems in part from the fact that a US-EU coalition has a high degree of concentration (Herfindahl index) of "country" size within the coalition, since there are only two economies of roughly equal size. As a result the noncooperative coalition outcome is a large fraction of the cooperative coalition outcome (see Proposition 5). The finding shows that the stable coalition can go a long way toward achieving the globally efficient outcome without needing to threaten trade penalties against the largest—and hence hardest to punish—economies.

While a number of 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 et al. 2014; Tagliapietra et al. 2021) to my knowledge this is the first paper to show that leadership by major economies follows strictly from self interest. My analysis also shows that while the US and EU strictly benefit from leading the process, China does not. Finally, I provide an agreement structure under which a US-EU coalition is stable and could form the basis for a scalable climate agreement with the potential to achieve something close to the globally efficient outcome.

Section 2 provides motivation, including the noted extension of Nordhaus's (2008) study of incomplete participation costs. Section 3 presents the model and defines the proposed agreement structure. Section 4 presents the analytical results. Section 5 studies coalition stability in a quantitative numerical model. Section 6 considers related policy issues, and Section 7 concludes.

2 Motivation

2.1 Big countries are qualitatively different

While much of the literature on International Environmental Agreements starts from the premise that countries are symmetric (e.g., Barrett 1994, Nordhaus 2015) the assumption misses an important feature of the real-world problem. Figure 1 shows the sense in which the US, the EU, and China are qualitatively different from other countries/economies. The top panels show 2019 World Bank data for GDP and Total Imports for the top-eight "countries" with the US, EU and China in black. The bottom panels show 2018 CO2 emissions (World Bank Development Indicators) along with climate damages constructed as the average of the three regional climate damage estimates used in Nordhaus (2015).[4] With the exception of climate damages for India, the US, EU and China comprise a markedly higher fraction of the aggregate world quantity for all four indicators.

More incentive to contribute The domestic SCC in panel four approximates each country's incentive to contribute to the global public good. Absent external pressure, a country acting in its own self interest would optimally price carbon at the domestic SCC (Kotchen 2018). The panel shows that the EU, India, China, and the US all have a markedly higher incentive to contribute than other countries.

A domestic SCC around 10 percent means these countries internalize only 10 percent of global climate damages, much less than would a global planner. While less than ideal, it is still a far more promising starting point for initiating climate action then with other countries. Indeed, of the UN member states outside the EU, about 90 percent of them have a domestic SCC that is less than one percent of the global SCC (author's calculations). For these countries, there is almost no incentive to contribute to the global public good without external pressure.

[4] See table B-2 in the online appendix for Nordhaus (2015).

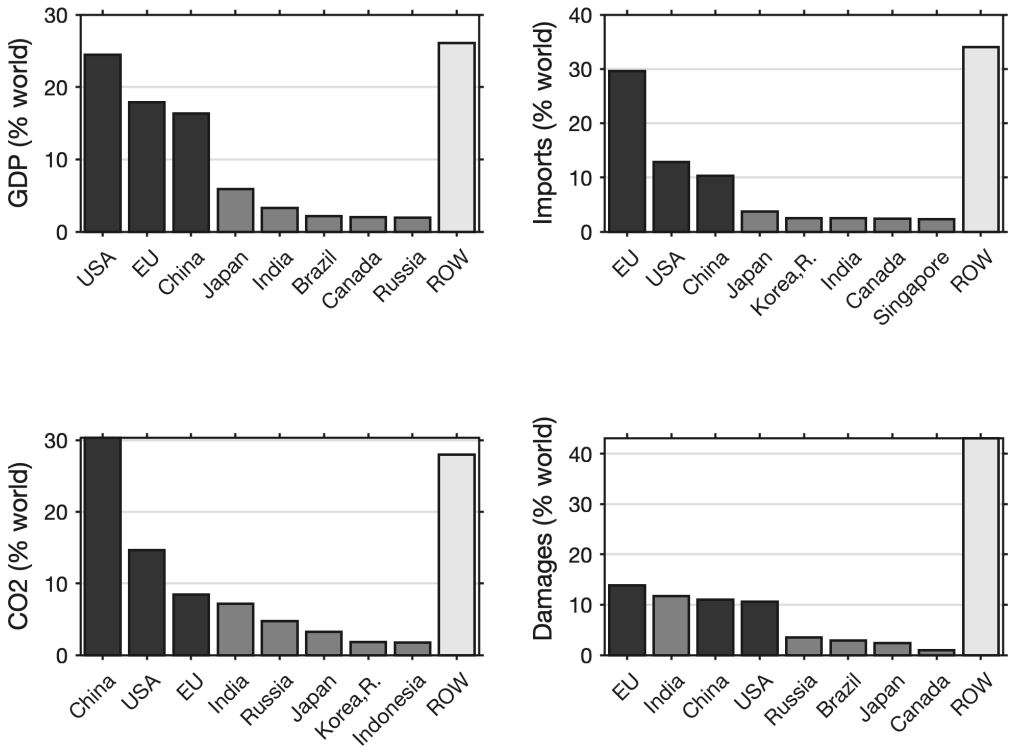

Figure 1: The panels display four economic variables for the top eight countries in the world plus the rest of the world (ROW). For each variable, quantities are represented as a percent of the global total. The USA, EU, and China are shown in black, and the next five countries are shown in grey. GDP, Imports and CO2 are shown for 2019 based on World Bank data; climate damages in panel four are based on the three-model average used in Nordhaus (2015).

Greater bargaining power in trade negotiations To see how differences in country size translate into differences in trade-negotiation leverage, I employ the reduced form model calibrated in Nordhaus (2015). For each bilateral country/region pair in RICE, he uses the multi-country, multi-industry, general equilibrium trade war model from Ossa (2014) to compute the optimal uniform import tariff absent retaliation and to estimate the parameters of a reduced form tariff benefit function.

I combine Nordhaus's quantification of tariff costs with a simple gravity model of trade without trade frictions. Simulating trade flows with the simple model shows how relative size impacts the relative consequences of a trade war absent the "noise" of geographic effects.[5] To simulate payoffs in a trade war, I assume that each country plays the unilateral best response according to Nordhaus's simulations of the Ossa (2014) model. I then study the ratio of losses (as a fraction of GDP) that country i incurs relative to losses in the coalition (as a fraction of GDP). I assume the parameters of the reduced form tariff benefit function for the coalition is an average of the values for the US, the EU, and China, while the volume of trade between the coalition and other countries is scaled up or down as indicated by the reduced form trade model as the coalition size varies. The results are shown in Figure 2. Further details are presented in Appendix A.1.

Figure 2 shows that the US, the EU and China are indeed qualitatively different from other countries in terms of trade leverage. Even for the most powerful of the remaining countries, a trade war with the EU, US and China together would be disastrous, costing the country over 20 times more per unit of GDP than it would cost the coalition.[6] It follows that it would be relatively easy for the coalition to threaten these countries with penalty tariffs without having to worry about the risk of an ensuing trade war.

In contrast, China, the EU, and the US are all much harder to punish, though there is substantial

[5] In the quantitative section, I use actual trade flows between countries to quantify the magnitude of outcomes in a more realistic way.

[6] Nordhaus (2015) notes that India and Japan have relatively weak tariff benefits for their size due to preexisting tariffs included in the Ossa 2014 model.

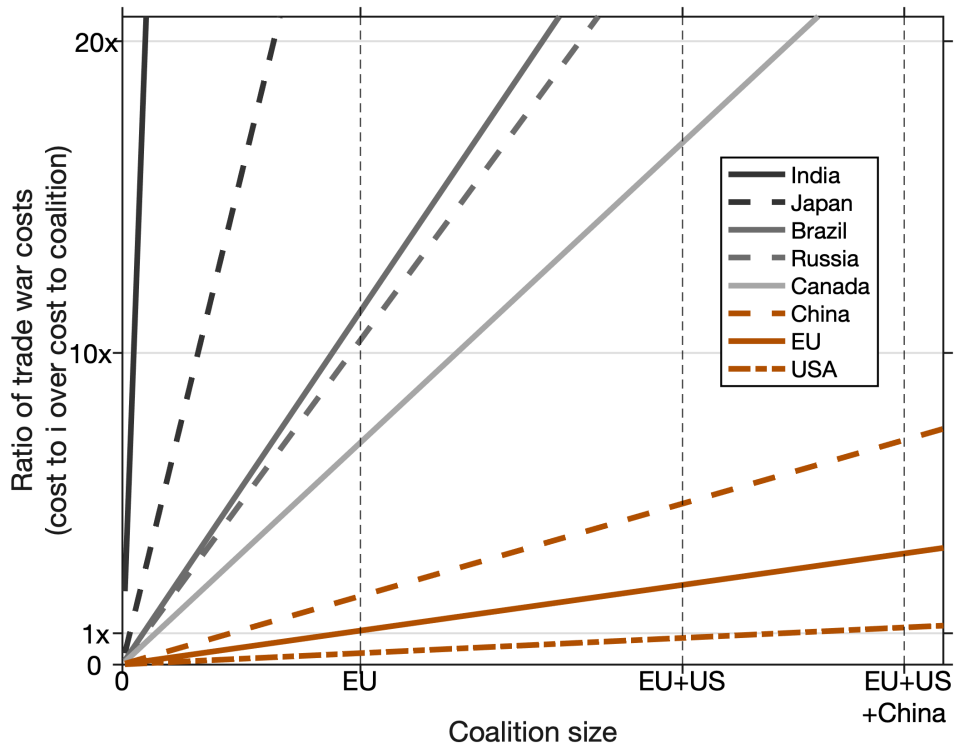

Figure 2: Costs of a trade war to indicated country relative to the trade war cost to the coalition.

variation within this group. The US is by far the hardest to punish, followed by the EU, and then China. The contrast stems in part from the large trade deficit in the US and the large trade surplus in China. Böhringer and Rutherford (2017) use a global CGE to demonstrate the considerable difficulty of punishing the US for climate negligence.

2.2 A modest match drastically lowers aggregate cost

As further motivation for the two-tier climate club described in the next section, I extend Nordhaus's (2008) study of incomplete participation to show that the very high cost of sub-global abatement is drastically reduced if coalition abatement is paired with a modest carbon price in the rest of the world.

Nordhaus (2008) shows that at any given level of global abatement, the associated global abatement costs increase by the multiplicative factor $\phi^{1-\theta_2}$ if abatement is done entirely by a coalition that comprises fraction ϕ of the world economy, where θ_2 is the curvature parameter in the abatement cost function. In the most recent DICE model (Nordhaus 2016) it is calibrated by assuming $\theta_2 = 2.6$. With this calibration, a coalition that comprises 20 percent of the global economy would face abatement costs that are higher than the efficient level by a factor of $(0.2)^{1-2.6} = 13.1$.

In contrast, if the rest of world shares just a small portion of the burden, thus exploiting the cheapest available abatement opportunities, the cost penalty from incomplete participation drops sharply. To allow for the possibility that countries in the rest of the world differ in their willingness to match the coalition price, I assume 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

$$\tau_j = \alpha_j \tau_C. \tag{1}$$

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 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 $\sum_{i=0}^{n} \phi_i = 1$.

As in Nordhaus (2008), I quantify how the global cost of achieving a given global abatement rate μ increases relative to the cost-minimizing level when abatement is led by a sub-global coalition.

Proposition 1. *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

$$P(\boldsymbol{\phi}, \boldsymbol{\alpha}) = \frac{\phi_0 + \sum_{i=1}^{n} \phi_i \alpha_i^{\theta_2/(\theta_2-1)}}{\left(\phi_0 + \sum_{j=1}^{n} \phi_j \alpha_j^{1/(\theta_2-1)}\right)^{\theta_2}}. \tag{2}$$

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

$$P(\phi_0, \alpha) = \frac{\phi_0 + (1-\phi_0)\alpha^{\theta_2/(\theta_2-1)}}{\left[\phi_0 + (1-\phi_0)\alpha^{1/(\theta_2-1)}\right]^{\theta_2}}. \tag{3}$$

Proof. Appendix A.3. □

The proposition shows that for any given level of global abatement, costs increase above the globally efficient cost by the amount indicated in the penalty function. The cost penalty reduces to that in Nordhaus (2008) when $\alpha = \mathbf{0}$:

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

When $n = 1$ (penalty given by Equation 3) the penalty is strictly decreasing in both ϕ and α (Appendix A.4).

Figure 3 uses Eq. 3 to quantify how much the cost penalty from incomplete participation decreases when coalition policy is paired with a matching carbon price in the ROW. I restrict attention to the special case of a single non-coalition region that matches the coalition carbon price at rate α. The left panel plots $P(\phi, \alpha)$ as a function of ϕ for three values of the match rate α. The solid-gray line shows the penalty when $\alpha = 0$, which coincides with the Nordhaus (2008) penalty function. In this case, global abatement costs rise very rapidly as ϕ gets small, reaching 13.1 times the efficient cost when $\phi = 0.2$. In contrast, for even small values of α, the penalty declines markedly. For example, if $\phi = 0.2$, the penalty decreases by a factor of five (from 13.1 to 2.5) when the match rate increases from 0 to 10 percent.

The right panel of Figure 3 plots $P(\phi, \alpha)$ as a function of α for two empirically interesting values of ϕ. The first ϕ value coincides with a coalition between the the United States and the European Union (excluding the UK). Using 2018 carbon dioxide emission data from Climatewatch.org (World Resources Institute 2020), the coalition comprises 23.0 percent of global emissions. The second ϕ value adds China, which increases the coalition size to 53.3 percent of global emissions.

Without China, the abatement cost penalty is immense, though it also declines very rapidly in α. Without a matching carbon price ($\alpha = 0$) abatement costs are 10.5 times higher for the US+EU coalition than if the burden were efficiently shared across all countries. When $\alpha = 0.05$, the penalty decreases by roughly a factor of three, and when $\alpha = 0.1$, it decreases by more than a factor of four.

Adding China has a huge effect on the abatement cost penalty, especially for α small. When $\alpha = 0$, adding China to the coalition reduces the penalty from 10.5 to 2.7. With the bigger coalition, the cost penalty is less sensitive to α, in part because the non-coalition region is a smaller fraction of the world economy. Nevertheless, adding a matching carbon price outside a China-US-EU coalition would still matter a great deal. Increasing α from zero to 5 percent reduces the penalty by more than a quarter (from 2.7 to 2.0), and increasing α from zero to 10 percent reduces the penalty by almost 40 percent (from 2.7 to 1.7).

A possible takeaway from the right panel of Figure 3 is that including China in a global climate agreement is critical since doing so drastically reduces the cost penalty from incomplete participation. Indeed, this fact may partly explain the emphasize in recent policy debates on including China in a potential agreement. Nevertheless, the figure also suggests an alternative interpretation. Mainly, the impact on the cost penalty from adding China is roughly the same as the impact on the cost penalty if the match rate under a US-EU coalition is increased from zero to ten percent. This fact is

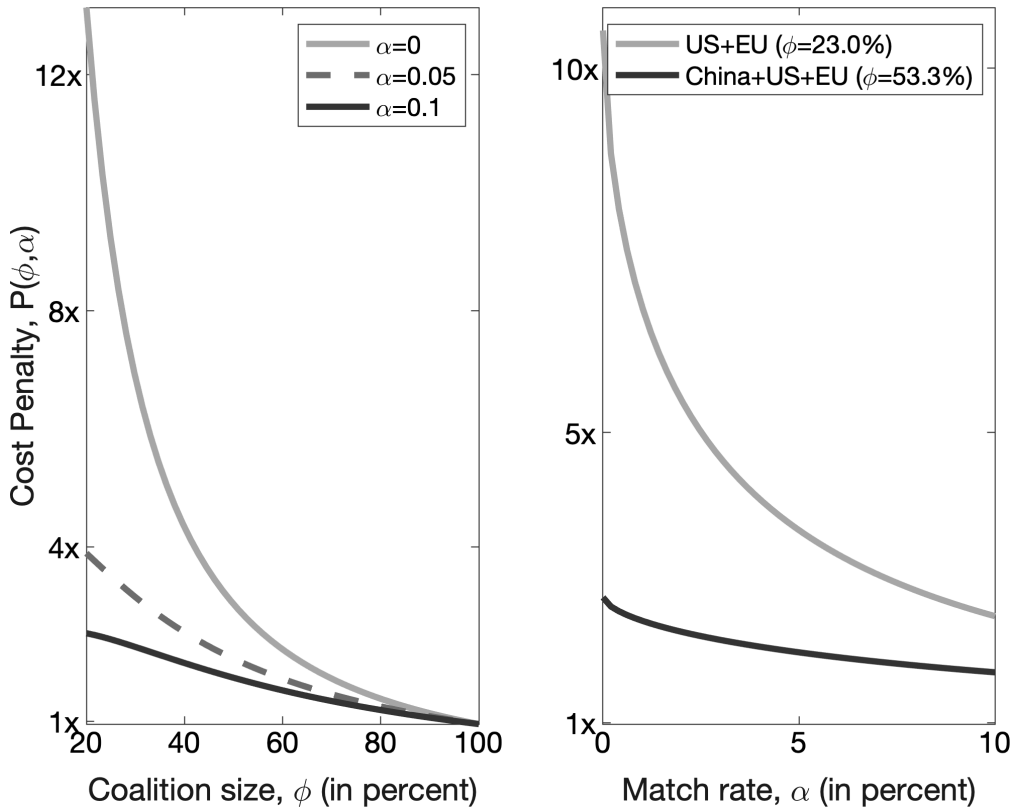

Figure 3: The left panel plots the cost penalty as a function of ϕ for three values of α. The right panel plots the cost penalty as a function of α for two values of ϕ: $\phi = 23.0\%$ (US+EU) and $\phi = 53.3\%$ (China+US+EU).

worth keeping in mind when interpreting the results in Section 5 where we find that the unique stable coalition with a TCC consists of the US and EU with China out.

In the policy considered in this section, the "match rate" α is placed on the carbon price ($\tau^R = \alpha\tau_C$) though one could alternatively place it on the abatement rate (i.e., requiring $\mu^R = \alpha\mu^C$). A reason one might prefer a policy that matches price is because it achieves more abatement for a given match rate α than would a policy that matches the abatment rate. Specifically, as shown in Appendix A.2, a match rate α applied to the carbon price is equivalent to a match rate $\alpha^{1/(\theta_2-1)}$ applied to the abatement rate. Since $\alpha < 1$, $\alpha^{1/(\theta_2-1)} > \alpha$ as long as $\theta_2 > 2$. Given $\theta_2 = 2.6$, a match rate of $\alpha = 5\%$ applied to the coalition carbon price is equivalent to a requirement that the non-coalition region abate 15 percent of the abatement level in the coalition (since $0.05^{1/(\theta_2-1)} = 15.4\%$). If $\theta_2 = 2$, there is no difference between putting the match rate on the carbon price or on the abatement rate.

Intuition To explain the dramatic results in Figure 3, I use Figure 4 to show why global abatement costs scale so quickly and why a matching carbon price is so effective at reducing cost. The left panel shows marginal abatement costs for the coalition, and the right panel shows marginal abatement costs for the rest of the world. In each case, the abatement cost functions are calibrated with $\theta_2 = 2.6$. Without abatement, coalition emissions are normalized to 100 units and rest-of-world emissions are twice as much, so the coalition comprises a third of global emissions.

The coalition's objective is to reduce emissions by 80 units. If the coalition does this on its own, the cost is the area under the coalition's marginal abatement cost curve between 0 and 80, which is the area of $A + B + C$. This cost can be compared to the cost of an efficient global policy at which marginal abatement costs are equalized at the common carbon price τ^*. In this case, the coalition abates 27, the rest of the world abates 53, and the cost is the area of $A + D + E$. When the coalition acts alone, it is forced to engage in very high marginal cost abatement activities (area $B + C$) while leaving unexploited a wide array of very cheap abatement options available elsewhere in the world

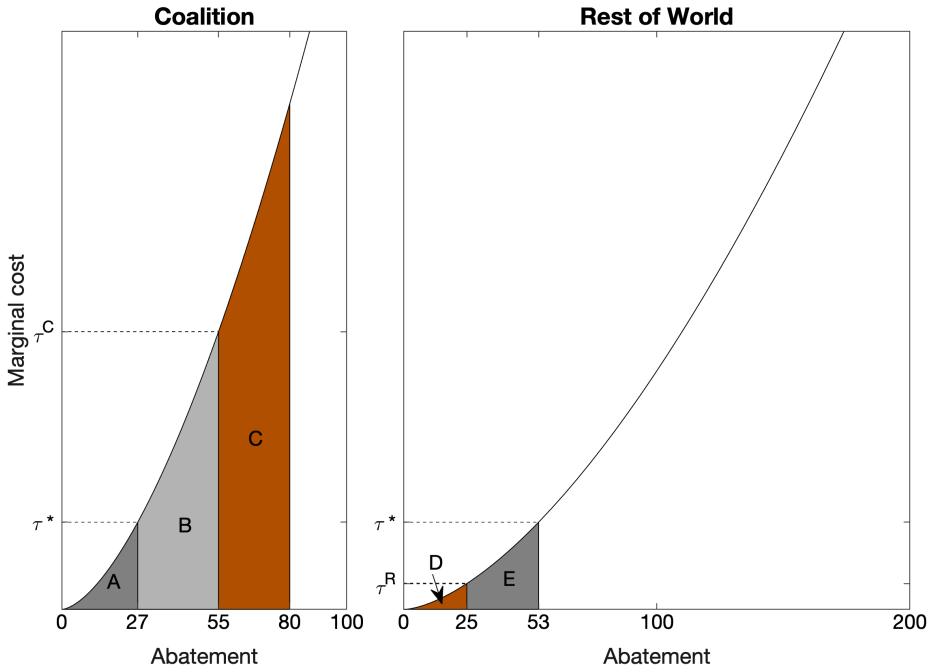

Figure 4: The figure shows marginal abatement costs for a coalition comprising a third of global emissions and for the rest of the world. The curves are generated using the abatement cost function in Nordhaus (2008, 2016) assuming $\theta_2 = 2.6$. Letters indicate colored regions whose area reflects the total cost of moving between the indicated levels of abatement.

(area $D + E$).

Next, consider an alternative arrangement that seeks to benefit from the same action of exchanging high cost abatement for low cost abatement, though with less ambition than the cost-minimizing policy. Under the proposed policy, the coalition imposes carbon price τ^C within its borders, while requiring the rest of the world to impose the much smaller carbon price τ^R. In Figure 4, τ^C induces 55 units of abatement from the coalition, while τ^R induces 25 units of abatement outside. Total abatement is 80 units. Relative to the case in which all abatement is undertaken by the coalition, we exchange costs equal to area C for costs equal to area D. For a given level of ambition, the policy always takes advantage of the most valuable opportunities to swap abatement across regions. This is true because the first units of abatement given up in the coalition are the highest-cost ones, while the first units taken up outside it are the cheapest remaining. The exercise replicates the reallocation of abatement effort achieved by a TCC.

3 Model and agreement structure

This section presents the static model of cross-country cooperation to be used throughout the paper. It defines country-level payoffs, presents the proposed agreement structure, and specifies the conditions for coalition stability.

3.1 Heterogeneity and payoffs

The world is comprised of N countries that vary in size. The size of each country i is characterized in three dimensions: GDP, Q_i, business-as-usual CO2 emissions, E_i, and the domestic Social Cost of Carbon (SCC) γ_i. The domestic SCC is the marginal damage in i of an extra unit of CO2 emissions, which is assumed to be constant.[7] The global SCC is $\gamma = \sum_i \gamma_i$.

[7]Constant marginal damages from atmospheric CO2 is roughly consistent with most of the IAM literature. The assumption holds exactly in Golosov et al. (2014) and Traeger (2018), and it holds approximately in DICE (e.g, Nordhaus 2016) which assumes a convex relationship between atmospheric temperature and economic damages, but a logarithmic relationship between CO2 concentration and atmospheric temperature.

Both aggregate and domestic climate damages depend on global CO2 emissions. Absent policy, global emissions are $E = \sum_i E_i$. The policy problem in each country is to choose a domestic carbon price, $\tau_i \geq 0$, or equivalently, a domestic abatement rate, μ_i. The resulting amount of global abatement, hence the resulting climate damages, depends on the policy choice of all countries.

Let $\tau_{-i} \equiv \{\tau_j\}_{j \neq i}$ denote the carbon price adopted by countries other than i, and let $\mu(\tau_i, \tau_{-i})$ be the global abatement rate. Then the policy benefit accruing to country i is

$$B_i(\tau_i, \tau_{-i}) = \gamma_i E \mu(\tau_i, \tau_{-i}). \tag{4}$$

As in RICE (Nordhaus and Boyer 2003) and C-DICE (Nordhaus 2015), the abatement opportunities in each country are characterized by an abatement cost function in which domestic abatement costs are a power function of the domestic abatement rate times GDP:

$$\Psi_i(\mu_i) = Q_i \theta_1 (\mu_i)^{\theta_2}, \tag{5}$$

where Q_i is GDP in i and θ_1 and θ_2 are scale and shape parameters, respectively. Abatement opportunities scale with country size, and the abatement cost function is consistent with that in the aggregate DICE model (e.g., Nordhaus 2016) if abatement costs are aggregated efficiently across regions.

The abatement rate can be written as a function of i's carbon price, τ_i, by equating τ_i with the marginal cost of an extra unit of abatement (denominated in units of emission reduction, not percentage terms). This implies a power-function relationship between τ_i and the percent abatement rate μ_i (Appendix A.2):

$$\mu_i = a_i (\tau_i)^b \equiv G_i(\tau_i). \tag{6}$$

where

$$a_i = \left[\frac{\sigma_i}{\theta_1 \theta_2} \right]^{\frac{1}{\theta_2 - 1}} \tag{7}$$

and

$$b = \frac{1}{\theta_2 - 1}. \tag{8}$$

Here, $\sigma_i = \frac{E_i}{Q_i}$ is the emissions intensity of output. If σ_i is the same across countries in a region, then $G_i(\cdot)$ is independent of size and the same function can be applied equally to a single country or an aggregate region provided abatement efforts are aggregated efficiently across countries.[8] Combining Equation 5 and Equation 6 gives abatement costs as a function of the carbon price:

$$C_i(\tau_i) = Q_i \theta_1 a_i^{\theta_2} (\tau_i)^{b\theta_2}. \tag{9}$$

A key feature of a climate club is the use of trade tariffs to penalize countries that do not comply with the terms of the agreement. The use of conditional trade incentives to support regulation of transboundary pollutants has been extensively studied, including important examples by Folmer et al. 1993, Barrett 2003, Lessman et al. 2009, Nordhaus 2015, and Böhringer et al. 2016. I define a *conditional trade incentive* as a pair $(\omega, \underline{\tau})$, where ωQ_i is the trade penalty imposed on country i if it fails to oblige the minimum carbon price $\underline{\tau}$. Thus, ω is denominated in units of percent GDP.

The payoff for country i depends on its policy, the policy of other countries, and the conditional trade incentive that it faces:

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

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

3.2 A two-tier climate club

The agreement begins with a set of countries, Ω, committed to establish (and enforce) an arrangement to reduce global carbon emissions. I refer to countries inside the coalition as tier-1 countries and to those outside the coalition as tier-2 countries. A *two-tier climate club* (TCC) extends Nordhaus's

[8] I assume this, for example, when modeling the European Union as a single entity later.

(2015) climate club by differentiating the conditional trade incentives faced by countries in each tier. In either case, trade penalties are carried out by the participating tier-1 countries.

For each tier-1 country i, the policy specifies a minimum carbon price, $\hat{\tau}_i$, that must be obliged to avoid trade tariffs from the other tier-1 countries. The magnitude of the penalty if i fails to meet the minimum obligation is ω_1 denominated in percentage-GDP units for country i.

Tier-2 countries also face a minimum carbon price obligation, though to increase the incentive for tier-1 countries to contribute, the agreement deliberately ties abatement by tier-2 countries with the carbon prices actually chosen by the tier-1 countries. Specifically, the agreement defines a match rate $0 \leq \alpha \leq 1$ with the understanding that the minimum carbon price required by tier-2 countries will be α times the (size-weighted) average carbon price in the coalition. The average coalition carbon price is given by

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

where $\hat{\phi}_i \equiv \frac{\phi_i^E}{\phi_C^E}$ (with $\phi_C^E \equiv \sum_{k \in \Omega} \phi_k^E$) denotes the fraction of the coalition comprised by tier-1 country i and τ_i is the carbon price actually implemented in i. Failure to oblige the minimum carbon price $\alpha\tau^{AVG}$ results in import tariffs on the offending tier-2 country j with combined economic cost $\omega_2 Q_j$, where Q_j is GDP in country j.

The determination of global abatement is modeled as a four-stage game solved with backward induction. In stage one, tier-1 countries establish the agreement, summarized by the vector $\mathbf{P} = (\{\hat{\tau}_i\}_{i \in \Omega}, \omega_1, \alpha, \omega_2)$. In stage two, each tier-1 country i faces conditional trade incentive $(\hat{\tau}_i, \omega_1)$ and chooses τ_i. In stage three, having observed the tier-1 choices in stage 2, each tier-2 country faces conditional trade incentive $(\alpha\tau^{AVG}, \omega_2)$ and chooses τ_j. In stage four, punishments are imposed by the tier-1 countries.

I assume that all countries choose the domestic policy that maximizes the domestic payoff and that the coalition has perfect foresight about future choices when designing the agreement. I thus restrict attention to incentive compatible agreements since the outcome of a non-incentive-compatible agreement could always be emulated with one that is incentive compatible.

A policy is incentive compatible if no country has an incentive to deviate. For tier-2 countries, this means that net payoffs from obliging the agreement are at least as high as under the optimal unilateral deviation ($\tau_j = \gamma_j$):

$$B_j(\alpha\tau^{AVG}, \tau_{-j}) - C_j(\alpha\tau^{AVG}) \geq B_j(\gamma_j, \tau_{-j}) - C_j(\gamma_j) - \omega_2 Q_j \mathbf{1}_{\gamma_j < \alpha\tau^{AVG}}. \tag{12}$$

This will be true as long as ω_2 is big enough to discourage deviations given α (and the level of τ^{AVG} that results in equilibrium). In Eq. 12, the unilateral deviation by country j has no effect on the policy choice of others countries, so τ_{-j} is the same on both sides of the equation. For tier-1 countries, the agreement is incentive compatible provided

$$B_i(\hat{\tau}_i, \tau_{-i}(\hat{\tau}_i)) - C_j(\hat{\tau}_i) \geq B_j(\gamma_i, \tau_{-i}(\gamma_i)) - C_j(\gamma_i) - \omega_1 Q_i \mathbf{1}_{\gamma_i < \hat{\tau}_i}. \tag{13}$$

In this case, the unilateral deviation by tier-1 country i affects τ^{AVG}, so it also affects the carbon price choice by tier-2 countries. For this reason, with a slight abuse of notation, I have written τ_{-i} to be a function of the policy choice in i. Since the policy choice is different on each side of the equation, τ_{-i} is also.

Tier-1 countries benefit when ω_1 increases above zero since this allows the coalition to implement a policy that increases coalition surplus relative to the case in which all tier-1 countries simply pursue their unilateral best response. Thus, if the coalition could choose a value of ω_1, it would always choose a higher value, at least up until the point at which the cooperative outcome for the coalition can be supported. But in reality, the value of ω_1 would be constrained by characteristics of the coalition that are beyond its control. For example, it could depend on the distribution of trading clout among members of the coalition, and it could depend on the history of international relations between these countries. For this reason, I view ω_1 to be a function of the coalition Ω that the current coalition takes as given.

Similarly, the coalition would always prefer a higher match rate α since this increases the amount of abatement done by tier-2 countries, abatement that is effectively free from the perspective of the coalition. Given the participation constraints for tier-2 countries (Eq. 12) the only way in which a higher value of α can be supported is if the coalition threatens higher values of ω_2. But similar to ω_1,

the maximum achievable value of ω_2 is ultimately a political constraint that depends on attributes of the coalition. To account for these considerations, I assume that the coalition takes $\omega_2(\Omega)$ as given and chooses α to be as high as possible, which coincides with the value of α at which the participation constraint in Eq. 12 holds. Equivalently (given perfect foresight) I view the coalition problem as that of taking $\alpha(\Omega)$ as given, while assuming that ω_2 is always set high enough to ensure that Eq. 12 is satisfied.[9]

The model is solved with backward induction. In the final stage, tier-1 countries follow through with the designated trade penalties provided it is incentive compatible to do so. The trade literature finds that modest tariffs on imports tends to benefit the country imposing the tariffs due to a terms of trade effect (Broda et al. 2008). For this reason, it is plausible that tier-1 countries would stand to benefit (at least marginally) from following through with modest tariffs on deviating tier-2 countries as long as the tier-2 countries don't retaliate. For simplicity, I assume the net benefit to tier-1 countries of imposing tariffs on tier-2 countries is always zero, and I assume that they are willing to follow through with the threat when the incentive is zero. Ultimately, $\alpha(\Omega)$ captures the extent to which the coalition can push tier-2 countries using trade threats, while $\omega_1(\Omega)$ captures the extent to which tier-1 countries can push each other. In stage three, tier-2 countries optimally oblige the target $\alpha\tau^{AVG}$, and in stage two the tier-1 countries optimally choose the minimum carbon price specified by the (incentive-compatible) agreement.

Finally, in stage one, the coalition chooses carbon price targets for each tier-1 country to maximize coalition surplus subject to the constraint that no country has an incentive to violate its respective participation constraint. I define the policy that solves this optimization problem as the *Optimal Coalition Policy* (OCP). It takes $\omega_1(\Omega)$ and $\alpha(\Omega)$ as given and solves

$$\max_{\{\hat{\tau}_i \geq 0\}_{i=1}^n} \sum_{i=1}^{n} \Pi_i(\hat{\tau}_i, \{\hat{\tau}_j\}_{j\neq i}; (\hat{\tau}_i, \omega_1(\Omega), \alpha(\Omega))) \tag{$\boldsymbol{P1}$}$$

subject to

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

plus the requirement that ω_2 is set high enough to ensure that the tier-2 participation constraints (Eq. 12) hold. The dependence of the payoff function on the vector $(\hat{\tau}_i, \omega_1(\Omega), \alpha(\Omega))$ indicates the carbon price threshold below which country i will be penalized, and it accounts for how the match rate amplifies the effect of the carbon price in each tier-1 country i through its effect on abatement elsewhere in the world.

I denote the solution of the OCP problem by $\{\tau_i^*(\Omega)\}_{i\in\Omega}$. Since the parameters of the problem depend on the coalition Ω, I write the solutions to be a function of Ω as well.

To see how the proposed mechanism increases i's incentive to contribute, suppose the other tier-1 countries go along with the specified targets while i chooses τ_i (possibly a deviation). Then the global abatement rate is

$$\mu(\tau_i, \{\hat{\tau}_j\}_{j\neq i}; \alpha) = \underbrace{\phi_i^E G_i(\tau_i)}_{\substack{\text{Direct}\\ \text{effect}}} + \sum_{j\neq i} \phi_j^E G_j(\hat{\tau}_j) + (1 - \phi_C^E) G_R\big(\alpha[\underbrace{\hat{\phi}_i^E \tau_i}_{\substack{\text{Indirect}\\ \text{effect}}} + \sum_{j\neq i} \hat{\phi}_j^E \tau_j]\big). \tag{14}$$

The underscored terms highlight the two channels through which i's policy choice impacts global abatement. First, 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, 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 coalition.

It is straightforward to recover the payoff to countries inside and outside the coalition. These payoffs will be used to discuss coalition stability in the next subsection. Let superscript "IN" and superscript "OUT" distinguish payoffs for countries inside and outside the coalition, respectively. For countries inside the coalition ($i \in \Omega$) we have

$$\Pi_i^{IN}(\Omega) = \gamma_i E\mu(\{\tau_j^*(\Omega)\}_{j\in\Omega}, \{\alpha\tau^{AVG}(\Omega)\}_{k\notin\Omega},) - C_i(\tau_i^*(\Omega)), \tag{15}$$

[9] If Eq. 12 were not satisfied, then it would follow that the posited value of $\alpha(\Omega)$ was too high given the politically feasible value of $\omega_2(\Omega)$.

where

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

Similarly, for countries outside the coalition ($j \notin \Omega$) we have

$$\Pi_j^{OUT}(\Omega) = \gamma_j E \mu(\{\tau_k^*(\Omega)\}_{k \in \Omega}, \{\alpha \tau^{AVG}(\Omega)\}_{k \notin \Omega},) - C_i(\alpha \tau^{AVG}(\Omega)). \tag{16}$$

In both cases, I have expressed the global abatement rate to be a function of the equilibrium policy choice in all countries.

3.3 Conditions for coalition stability

In the last subsection, a fixed set of countries Ω formed a TCC with the aim of maximizing coalition surplus subject to the given match rate $\alpha(\Omega)$ and the given tier-1 penalty $\omega_1(\Omega)$. Next, I embed the agreement formation decision within a two-stage endogenous coalition formation game. This step is inline with most of the IEA literature since Barrett (1994). In the first stage, countries decide whether or not to join the agreement. In the second stage, conditional on the set of countries that join, the coalition establishes the agreement terms. I assume the agreement takes the form of a TCC, so the second stage consists of the four-stage subgame described in the last subsection. Thus, the payoffs for countries inside and outside the coalition are given by Equations 15 and 16, respectively.

For the coalition to be stable, two sets of stability conditions must hold. *Internal stability* requires that every country that joins is better off staying in than it would have been if it hadn't joined. In particular, for every $i \in \Omega$, we must have

$$\Pi_i^{IN}(\Omega) \geq \Pi_i^{OUT}(\Omega \setminus i), \tag{17}$$

where $\Omega \setminus i$ denotes the set of countries in the set Ω less i. $\Pi_i^{IN}(\Omega)$ and Π_i^{OUT} is defined by Equation 15 and Π_i^{OUT} is defined by Equation 16.

In addition, *external stability* requires that all countries outside the coalition are better off having stayed out. Thus, for every $j \notin \Omega$, we require

$$\Pi_j^{OUT}(\Omega) \geq \Pi_j^{IN}(\Omega \cup j). \tag{18}$$

For the stability conditions to be well defined, it is necessary to take a stand on the relationship between any given set of countries Ω and the two key constraints facing the coalition when forming a TCC: the match rate $\alpha(\Omega)$ and the tier-1 penalty $\omega_1(\Omega)$. Before doing this, Section 4 takes the coalition set as given and analytically studies the optimal design of a TCC (the OCT problem) for the feasible range of values for α and ω_1.[10] Next, Section 5 specifies functional form assumptions for $\alpha(\Omega)$ and $\omega_1(\Omega)$, then studies coalition stability.

4 Analytical results

This section assumes the set of coalition countries (Ω) is given, then solves the OCP problem for the feasible range of values for α and ω_1. Since α and ω_1 are taken in this section to be independent of Ω, I express the maximizers of the OCP problem as $\{\tau_i^*(\Omega, \alpha, \omega_1)\}_{i \in \Omega}$. The propositions characterize these maximizers and the resulting amount of global abatement for different values of α and ω_1.

I begin with notation. While the coalition takes the set Ω as given, the set is generically defined with n countries, each characterized by the tuple (Q_i, E_i, γ_i). For convenience, I define $\phi_C^E = \frac{1}{E} \sum_{i \in \Omega} E_i$ as the fraction of global CO2 emissions in the coalition and $\gamma_C = \sum_{i \in \Omega} \gamma_i$ as the coalition SCC.

The analytical results make use of one further assumption that I state explicitly.

Assumption 1. *Every tier-2 country j behaves as if $\gamma_j = 0$.*

The assumption simplifies the response behavior of tier-2 countries since their optimal unilateral response is to abate zero without external coercion. While the assumption is not exactly true, it is a

[10] In principle, $\omega_1 \geq 0$ is unbounded above, though there is no need to consider values higher than that needed to support the cooperative outcome since the coalition would never choose to exceed this level of policy.

rough approximation if the coalition consists of the biggest countries. Also, according to my analysis the domestic SCC is below one percent of the global SCC for approximately 90 percent of UN countries.

While Assumption 1 implies that individual tier-2 countries do not internalize a meaningful fraction of global climate damages, the set of all tier-2 countries in aggregate do incur climate damages equal to $\gamma_R = \gamma - \gamma_C$. The subscript "R" refers to the non-coalition countries in aggregate as the rest of the world (ROW). The assumption is consistent with a model in which the ROW consists of a continuum of infinitesimal countries, none of which internalizes a meaningful portion of global damages, even though in aggregate the region adds up to a meaningful portion of global damages. I further assume that the emissions intensity in each tier-2 country equals the average emissions intensity outside the coalition, which is $\sigma_R = \frac{E - \sum_i E_i}{Q - \sum_i Q_i}$.

The remainder of the section characterizes solutions to the OCP problem for a range of values for α and ω_1. Propositions 2 and 3 characterize policies for the "general" case in which $\theta_2 \geq 2$, while Propositions 4 and 5 extend these results for the special case in which abatement costs are quadratic ($\theta_2 = 2$).

The restriction to abatement cost functions with $\theta_2 \geq 2$ is made to ensure that the individual payoff function (absent the conditional trade penalty) is strictly concave in τ_i. $\theta_2 > 1$ would be enough to ensure strict convexity of the cost function (hence, strict concavity of the negative of the cost function) but $\theta_2 \geq 2$ ensures weak concavity of the benefit function $B_i(\tau_i, \tau_{-i})$ in τ_i. See Appendix A.5 for details.[11]

Proposition 2 begins with the least ambitious case: coalition countries use trade threats to induce cooperation from tier-2 countries, but they do not threaten trade penalties against each other (i.e., $\omega_1 = 0$). This is an important special case given the noted difficulty of penalizing large countries (see Figure 2).

Proposition 2. *Let $\theta_2 \geq 2$ and $\sigma_i = \sigma$ for all countries. If $\omega_1 = 0$, then the solutions to **P1** solve the following system of equations:*[12]

$$\tau_i^N(\alpha) \equiv \tau_i^M(0, \alpha) = \gamma_i \left[1 + \frac{1 - \phi_C^E}{\phi_C^E} \alpha^{1/(\theta_2 - 1)} \left(\frac{\tau_i^*}{\tau^{AVG}} \right)^{1 - \frac{1}{\theta_2 - 1}} \right], \text{ for } i = 1, \ldots, n, \tag{19}$$

where $\tau^{AVG} = \sum_{i=1}^n \hat{\phi}_i^E \tau_i^*$ If $\alpha = 0$, then $\tau_i^N(0) = \gamma_i$, which is the standard Nash equilibrium response. $\tau_i^N(\alpha)$ is strictly increasing in α. If we further assume that climate damages scale proportionally with energy use, $\frac{\gamma_i}{\gamma} = \phi_i^E$ for all i, then

$$\tau_i^N(\alpha) = \gamma_i A(\phi_C^E, \alpha), \text{ for } i = 1, \ldots, n,. \tag{20}$$

where

$$A(\phi_C^E, \alpha) \equiv \left[1 + \frac{1 - \phi_C^E}{\phi_C^E} \alpha^{\frac{1}{\theta_2 - 1}} \right]. \tag{21}$$

Proof. Appendix A.5. □

Without issue linkage between tier-1 countries, the OCP increases in the domestic SCC (γ_i) and in the match rate (α). A higher domestic SCC means the country internalizes a greater portion of global climate damages, which increases its incentive to contribute to the global public good. A higher match rate means that every unit of abatement by country i leads to more matching abatement outside the coalition, abatement that is effectively free from the perspective of country i.

The intuition for the OCP formula is easiest to see in the special case in which damages scale with energy use (Eq. 20). In this case, the OCP carbon price for country i equals its domestic SCC times $A(\phi_C^E, \alpha)$, which I refer to as the "amplification factor". The amplification factor captures the extent to which a given tier-1 country is induced to contribute more to the global public good than it otherwise would due to the fact that higher coalition abatement causes non-coalition countries to abate more (the cost of which falls outside the coalition).

[11] The assumption is a weak sufficient condition and could be relaxed somewhat, though at some inconvenience. Nordhaus assumes $\theta_2 \geq 2$ in all applications of the DICE model that I am familiar with, so the assumption is in step with much of the literature.

[12] I loosely refer to the OCP carbon price without issue linkage among big countries as "Nash" since the big countries are playing the Nash equilibrium with each other, though not with the rest of the world (unless $\alpha = 0$). Hence, the superscript "N" stands for Nash.

When $\alpha = 0$, then $A(\phi_C^E, 0) = 1$ (for $\phi_C^E > 0$) and the OCP reduces to the standard Nash result that countries price carbon at the domestic SCC (Kotchen 2018). When $\alpha > 0$, the second term in the amplification factor is the product of two components. The first component, $\alpha^{\frac{1}{\theta_2 - 1}}$, captures the extent to which a carbon price within the coalition induces more abatement outside the coalition. When $\theta_2 = 2$, the exponent is simply one, so higher α has a one-for-one effect on the term. If $\theta_2 > 2$, then the exponent is less than one, which makes a given α have a bigger effect since $\alpha < 1$. The other term, $\frac{1-\phi_C^E}{\phi_C^E}$, captures the fact that the amount of non-coalition abatement induced by a higher coalition carbon price is greater when the relative size of the non-coalition region is bigger. If $1 - \phi_C^E > \phi_C^E$ then there is a more than one-for-one impact of coalition policy on non-coalition abatement, while if $1 - \phi_C^E < \phi_C^E$ the opposite is true.

If we relax the assumption that climate damages scale with country size, then the OCP is given by the system of equations in (19). If abatement costs are quadratic, then the exponent on $\frac{\tau_i^*}{\tau^{AVG}}$ is zero, and the interdependent system again reduces to the set of independent equations in (20). If $\theta_2 \neq 2$, then each of the equations in (19) depends on τ^{AVG}, which itself depends on $\tau_i^N(\alpha)$ for each i. In this case, the determinants of $\tau_i^N(\alpha)$ are similar to those described for the special case above, but there is a further term that pushes countries with relatively high γ_i to have a somewhat higher (lower) $\tau_i^N(\alpha)$ provided $\theta_2 > 2$ (< 2).

The next proposition shows the upper bound on what could be achieved if tier-1 countries use trade threats against each other as a means to increase cooperation amongst themselves.

Proposition 3. *Let $\theta_2 \geq 2$ and $\sigma_i = \sigma$ for all countries. The agreement that maximizes coalition surplus (the "cooperative policy") imposes the following harmonized carbon price on each coalition country i:*

$$\tau^C(\alpha) = \gamma_C A(\phi_C^E, \alpha), \tag{22}$$

where $\gamma_C \equiv \sum_{i=1}^n \gamma_i$ and $A(\phi_C^E, \alpha)$ is defined in Eq. 21. This policy can be supported by threatening each tier-1 country with a sufficiently high though finite penalty ω_1.

Proof. Appendix A.6. □

The proposition shows the range of policies that can be supported if the tier-1 penalty (ω_1) increases above zero. If the penalty is sufficiently large, then the cooperative policy can be supported, which is as ambitious as the coalition would want to achieve. Comparing Equations 20 and 22 (for the special case in which climate damages scale exactly with country size) it is easy to see that

$$\tau_i^N(\alpha) = \frac{\gamma_i}{\gamma_C} \tau^C(\alpha). \tag{23}$$

Since $\frac{\gamma_i}{\gamma_C}$ is the fraction of coalition climate damages that accrue to country i, Eq. 23 says that the OCP without issue linkage among big countries ($\omega_1 = 0$) is $\frac{\gamma_i}{\gamma_C}$ times the (harmonized) OCP if the coalition acts cooperatively. This is similar to the standard result we would expect in a pure Nash equilibrium (with $\alpha = 0$) where jurisdictions internalize the portion of damages that fall within their borders. Here we see that an equivalent relationship holds for any value of α between zero and one.

The proposition also reveals how far a two-tier climate club could be expected to go. In the best case scenario, the coalition implements the cooperative policy and the match rate is one. Substituting $\alpha = 1$ into Eq. 22 implies

$$\begin{aligned} \tau_C(1) &= \left[1 + \frac{1 - \phi_C^E}{\phi_C^E}\right] \gamma_C \\ &= \frac{1}{\phi_C^E} \gamma_C \\ &= \left[\frac{\gamma_C/\gamma}{\phi_C^E}\right] \gamma. \end{aligned}$$

If climate damages scale with country size ($\gamma_C/\gamma = \phi_C^E$) then $\tau_C(1) = \gamma$, the global SCC. Since the match rate is one, the policy would result in the globally efficient carbon price harmonized across all countries and hence would achieve the globally-efficient outcome. More generally, the policy would achieve at least the globally efficient level of abatement provided $\gamma_C/\gamma \geq \phi_C^E$. Given the importance of this finding, I state it as a corollary.

Corollary 1. *Suppose* $\frac{\gamma_C}{\gamma} \geq \phi_C^E$*, so the portion of global climate damages that fall inside the coalition is at least as high as the fraction of global energy use in the coalition. Then if the match rate* (α) *equals one and if* ω_1 *is set high enough to achieve the cooperative policy, then the global abatement rate will be at least as high as the globally efficient rate.*

Importantly, the result obtains even though the coalition only internalizes the portion of global climate damages that fall within its borders. It occurs because the amplification factor is just big enough to induce the coalition to price carbon at the globally efficient level. For example, if we consider the special case in which damages scale with size ($\frac{\gamma_i}{\gamma} = \phi_i^E$) then the amplification factor is $1/\phi_C^E$, which exactly offsets the fact that the coalition in this case only internalizes fraction ϕ_C^E of global emissions.

So far, the results have focused on the somewhat general case in which $\theta_2 \geq 2$. With this assumption, we solve explicitly for the range of policies that can be supported by a TCC if we increase the tier-1 penalty above zero, but I am not able to show analytically precisely how the OCP depends on the tier-1 penalty. This can be done for the special case in which abatement costs are quadratic ($\theta_2 = 2$). The model with quadratic abatement costs is highly tractable, and I use it to extend the results in multiple ways. Proposition 4 presents an expression for the OCP carbon price as a function of α and ω_1, and Proposition 5 presents an expression for the global abatement rate when $\omega_1 = 0$.

Proposition 4. *Let* $\theta_2 = 2$*. Tier-1 countries differ in size but climate damages scale with size* ($\gamma_i = \phi_i^E \gamma$ *all* i*) and emissions intensity is constant* ($\sigma_i = \sigma$ *all* i*). If* $\omega_1 = 0$ *then*

$$\tau_i^N(\alpha) \equiv \tau_i^M(0, \alpha) = \gamma_i \left[1 + \frac{1 - \phi_C^E}{\phi_C^E}\alpha\right], \text{ for } i = 1, \ldots, n. \tag{24}$$

If $\omega_1 \geq 0$*, the OCP for tier-1 country* i *is:*

$$\tau_i^M(\omega_1, \alpha) = \tau_i^N(\alpha) + \frac{2}{\sigma}\sqrt{\theta_1 \omega_1}. \tag{25}$$

The cooperative policy is

$$\tau_i^C(\alpha) = \gamma_C \left[1 + \frac{1 - \phi_C^E}{\phi_C^E}\alpha\right],$$

and the penalty needed to support the cooperative policy for tier-1 country i *is*

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

Proof. Appendix A.7. □

For the case with quadratic abatement costs, the OCP without issue linkage among big countries ($\omega_1 = 0$) is $\frac{\gamma_i}{\gamma}$ times the cooperative carbon price. Eq. 25 shows that as ω_1 increases above zero, the OCP increases monotonically, though with diminishing returns since it depends on the square root of ω_1. The tier-1 penalty needed to achieve the cooperative outcome differs across tier-1 countries of different size: smaller countries need a larger penalty to induce participation since they internalize a smaller portion of global climate damages and thus benefit less from policy.

While the OCP solution shows the carbon price for each tier-1 countries, the corresponding amount of global abatement depends on the carbon price in all countries through Eq. 14. Despite the complexity of the mapping from carbon prices to global abatement, we can derive a tractable (and intuitive) formula for global abatement under quadratic abatement costs for the case in which $\omega_1 = 0$.

Proposition 5. *Let* $\theta = 2$ *and* $\omega_1 = 0$*. The global abatement rate achieved under the OCP policy (Eq. 24) is*

$$\mu(\alpha)\,|_{\omega_1 = 0} = H(\hat{\boldsymbol{\phi}})\mu^C(\alpha), \tag{26}$$

where $H(\hat{\boldsymbol{\phi}}) = \sum_{i=1}^n \hat{\phi}_i^2$ *is the Herfindahl index of country size within the coalition and* $\mu^C(\alpha)$ *is global abatement when the coalition behaves cooperatively. At the lower bound of* $\alpha = 0$*,*

$$\mu(0)\,|_{\omega_1 = 0} = H(\boldsymbol{\phi})\mu^*,$$

where $H(\boldsymbol{\phi})$ is the Herfindahl index of country size for all countries and μ^* is the globally optimal abatement rate. At the upper bound of α,

$$\mu(1)\,|_{\omega_1=0} = H(\hat{\boldsymbol{\phi}})\mu^*.$$

Proof. Appendix A.8. □

The Herfindahl index of country size within the coalition, $H(\hat{\boldsymbol{\phi}})$, measures the degree of concentration within the coalition. If the coalition consisted of a single country, the index would equal one, while if it consisted of n equal-sized countries, it would equal $1/n$.[13] If $\alpha = 0$, then the result in Eq. 26 is equivalent to the previously noted result that global abatement in the Nash equilibrium of a static abatement game with quadratic abatement costs equals the Herfindahl index of country size times the globally efficient abatement rate (Nordhaus 2015). What is interesting here is that the global abatement rate increases in α and equals $H(\hat{\boldsymbol{\phi}})\mu^*$ if the match rate is one. By definition, $H(\boldsymbol{\phi}) = \phi^2 H(\hat{\boldsymbol{\phi}})$, so the effect is large for modest ϕ. For example, if $\phi = 0.5$, then $\phi^2 = 0.25$, so increasing α from 0 to 1 would increase global abatement by a factor of four.

5 Quantitative evaluation

Next, I use a numerical model to study coalition stability and to characterize the range of outcomes that could be achieved by a TCC in an incentive-compatible way.

5.1 Calibration and numerical approach

The quantitative model includes twelve regions: the EU, plus the ten biggest national economies outside the EU (by GDP),[14] plus a rest-of-world (ROW) region that is assumed to consist of a continuum of small countries that do not internalize climate damages within their borders.

Heterogeniety across countries varies separately in terms of GDP, CO2 emissions, domestic climate damages, and domestic abatement costs. For GDP and emissions, I use 2018 World Bank data. For domestic climate damages, γ_i, 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 the baseline calibration, I assume the global SCC is \$51 per ton CO2 as in the middle estimate from the (interim) Interagency Working Group report (EPA 2021).

Finally, for abatement costs, I follow the C-DICE calibration in Nordhaus (2015). This calibration allows the scale parameter $\theta_{1,i}$ to vary across countries, while assuming $\theta_2 = 2$. 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 the calibration target in Nordhaus (2015) is satisfied. The adjustment ensures that a 25 USD per ton CO2 carbon tax generates an 18 percent reduction in global CO2 emissions.

	USA	EU	China	Japan	UK	India	Brazil	Canada
ϕ_i^Q:	24.5	17.9	16.3	5.9	3.3	3.3	2.1	2.0
γ_i/γ:	10.6	11.3	11.0	2.4	2.5	2.5	2.9	1.0
ϕ_i^E:	14.6	8.4	30.3	3.3	1.1	7.2	1.3	1.7
$\theta_{1,i}$:	0.030	0.021	0.055	0.028	0.021	0.033	0.007	0.039

Table 1: Dimensions of heterogeneity shown for the eight largest economies by GDP. All values expressed in percent except $\theta_{1,i}$.

Table 5.1 shows the key dimensions of heterogeneity for the eight largest economies, sorted in order of declining GDP. The first row shows GDP as a fraction of global GDP (ϕ_i^Q); the second row shows

[13] It is well known that $1/n$ is the lower bound on the Herfindahl index when there are n countries.
[14] The smallest national economy is Australia.

climate damages as a fraction of global damages (γ_i/γ); the third row shows carbon emissions as a fraction of global emissions (ϕ_i^E); and the last row shows the economy-specific abatement cost scale parameter, $\theta_{1,i}$.

When solving for the optimal coalition policy, I maintain the approximation (from the analytical section) that countries outside the coalition do not abate without external coercion. The assumption simplifies the calculation of OCP policies, though it somewhat overstates the carbon price for low values of the match rate. Nevertheless, the extent to which the optimal policy is overstated is modest and has no effect when the match rate is above the level that is easily passed for the calibrated model when either the EU or the US is included in the coalition.

To solve the problem while allowing for "full" heterogeneity across countries, Appendix A.9 extends the analytical solutions (both the 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 (Appendix A.9):

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

where

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

and

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

In addition, the formula for the noncooperative policy becomes (Appendix A.9):

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

Despite making the approximation that the coalition does not expect other countries to abate without external coercion, when I compute actual abatement for non-coalition countries (as in the right panel of Figure 7) I relax the assumption and assume that tier-2 countries faced with minimum carbon price $\hat{\tau}_2$ would implement domestic carbon price

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

where the domestic SCC γ_i is $i's$ unilateral best response. This assumption attenuates the effect of a low match rate on aggregate abatement when it is too low to trigger an increase in abatement for larger tier-2 countries, since these countries would already have incentive to abate since the Nash level exceeds the small requirement from the coalition.

5.2 Coalition stability

To study coalition stability, it is necessary to take a stand on how changes in the coalition set Ω affect the degree of cooperation within the coalition and the maximum match rate that it can enforce. I follow Barrett (1994), and much of the IEA literature since, in assuming that the coalition always behaves cooperatively. If no countries join (or the coalition size is one) then all countries play Nash, otherwise the coalition chooses policy to maximize its joint surplus.

To model $\alpha(\Omega)$, it is natural to assume that α rises in the combined trading clout of the coalition since the penalty used to enforce the match rate stems from tariffs on imports into the coalition region. Ideally, we would have a micro-founded bargaining model to predict the outcome of a trade negotiation between a given coalition and countries outside the coalition. Instead, to keep the analysis straightforward, I simply assume that the maximum feasible value of α is proportional to the combined GDP of the coalition region:

$$\alpha(\Omega) = b\phi_C^Q,$$

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

The assumption likely understates the relative trading prowess of large countries. Under the assumption, a group of countries with combined GDP that adds up to the GDP of the US would have

the same bargaining power in a trade negotiation as the US. But it seems more likely that the bargaining power of the single large country would be larger. Thus, the assumption might be viewed as conservatively understating the potential role of large countries in a two-tier climate club. In the baseline calibration, I assume $b = 1$. This means that the match rate equals one only if the coalition includes all countries. This is potentially also an understatement. If so, it shows the potential for a TCC to achieve substantial carbon abatement even under relatively conservative assumptions.

To study coalition stability, I consider all possible permutations of the eleven economies in the quantitative model. As described in Section 3.3, a candidate coalition is stable if it is both internally stable and externally stable, thus if it satisfies Eq. 17 for all economies inside the coalition and Eq. 18 for all economies outside the coalition.

The main result is that the set of stable coalitions is a singleton for the baseline calibration. It consists of the US and the EU—thus, roughly, the Annex I countries from the Kyoto Protocol. The finding that the stable coalition is unique contrasts sharply with analyses of coalition stability for single-tier climate clubs. For example, both Nordhaus (2015) and Hagen et al. (2021) find many possible stable coalitions. The finding of a unique stable coalition is not a general result, since I am able to find areas of the parameter space in which there are two stable coalitions. Nevertheless, the case of many possible coalitions does appear to be ruled out by the structure imposed with a TCC.

To develop intuition for the result, Figure 5 compares payoffs for the eight largest economies depending on whether the country joins a given coalition or not. The horizontal axis depicts a series of three candidate coalitions, increasing in size from left to right. The left-most coalition entails the Nash equilibrium in which all countries play their unilateral best response. This is described as consisting either of the US or the EU so the meaning of adding another country to the candidate coalition is well defined. In each case, the solid black line depicts payoffs for a country if it joins the indicated coalition (assuming it is not already in) or stay in (if it is already in) while the dashed-grey line depicts payoffs if it either stays out (if it is not already in) or leaves (if it is already in). The third coalition adds the UK to the EU-US coalition. The candidate coalition with the EU, US, and UK is included to show how payoffs change as the coalition size grows. The UK is included rather than another country because it would be next in line to join, in the sense that it has the smallest disincentive keeping it out. Reflecting this fact, the eight panels in Figure 5 are sorted in order of descending incentive to join the stable coalition, and the UK is third.

The first two panels show that both the EU and US have a strong incentive to join the stable coalition, though the incentive is more than twice as high for the EU. The US receives a net benefit from joining the stable coalition of about 10 billion USD per year. The graph also shows that the US incentive to join would decline if the coalition added other countries. The large benefit for the US to join stems from the assumption that the alternative coalition would involve all countries playing Nash. But if the US could count on the EU to set up a TCC on its own (perhaps together with the UK) then the US benefits only modestly from joining this coalition compared to staying out. Nevertheless, given the unparalleled trading prowess of the US, it is probably reasonable to assume that a two-tier climate club would only be possible with strong leadership from the US. If this is true, then the strong incentive for the US to join found in the model is reasonable.

The reason the US benefits less from participation than the EU is because its economy is more energy intensive, which makes the higher carbon price more costly. In addition, because the US comprises a larger fraction of global CO2 emissions, the US joining undermines the incentive for other coalition countries to abate carbon at a high level. The latter effect will turn out to be a major explanation for China's strong preference to stay out. China's incentives are discussed below in the context of Figure 6.

Among the six depicted countries that prefer to stay out of the stable coalition, the UK is closest to wanting to join. Indeed, the difference between the UK's payoff if it stays out of a US-EU coalition and its pay if it joins the coalition is barely perceptible on the graph. Japan, Canada, and Brazil each have successively stronger incentives to stay out. Moreover, for each of these countries, the incentive to stay out increases as the size of the existing coalition increases. This feature of the problem underlies the tendency for the model to give rise to a unique stable coalition (or at least very few stable coalitions).

The last two panels show that both India and China are starkly better off staying out of the agreement. For both countries, the payoff if it stays out of the stable coalition is roughly 20 billion USD per year higher than if it were to join. One major difference is that India would still benefit on the order of 10 billion USD per year if it were to join the stable coalition, assuming the next best

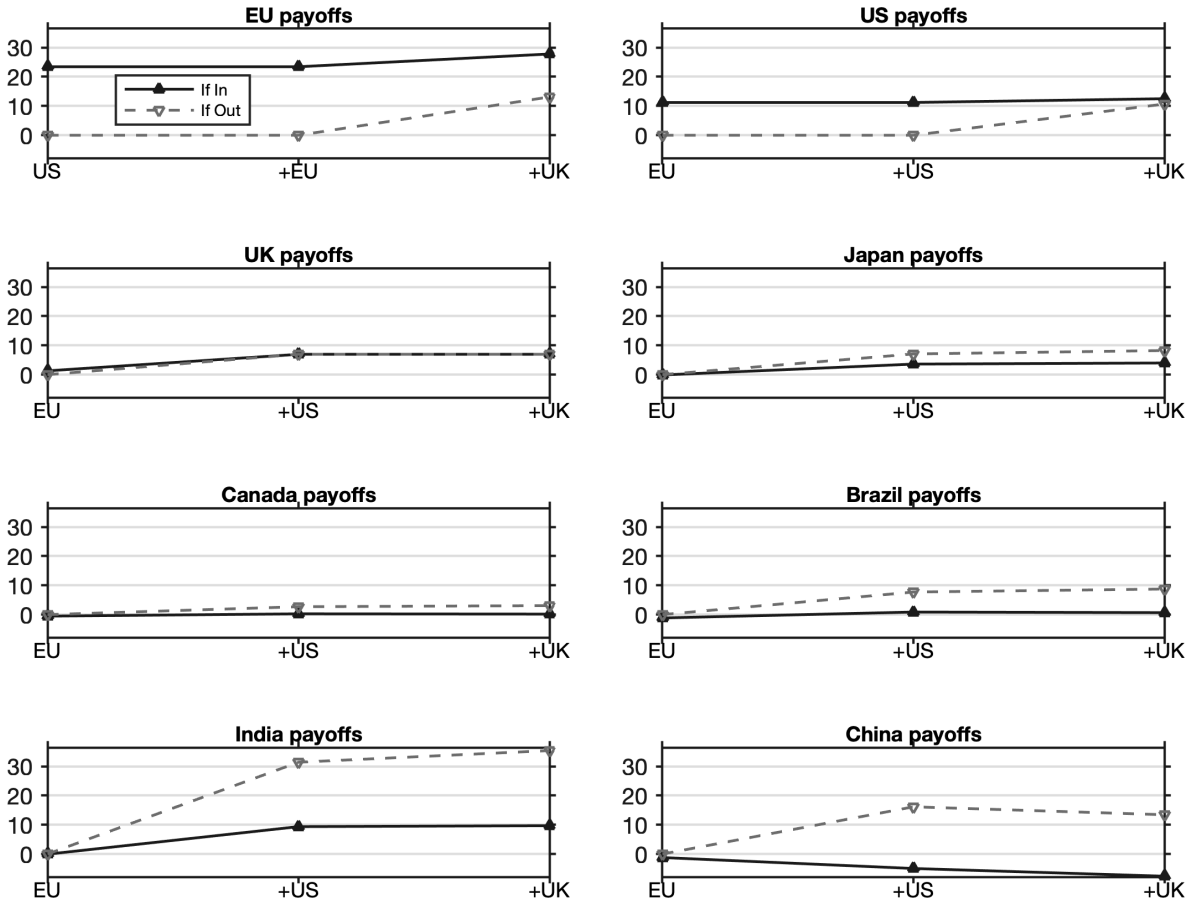

Figure 5: Payoffs relative to Nash payoff (in billions of USD per year) for top eight economies by GDP. Horizontal axis shows three possible coalitions. The left-most coalition is the Nash equilibrium, the second entails the EU and US, and the third adds the UK. The solid-black line indicates payoffs if the country stays in or joins the indicated coalition, while the dashed-grey line indicates payoffs if the country stays out or leaves. Countries sorted in order of descending net benefit to join.

alternative were Nash. In contrast, China is strictly worse off joining the stable coalition than it is under Nash. Despite this difference, China still benefits substantially from the agreement if it stays in the second tier. Relative to Nash, China gains roughly 17 billion USD per year from the stable agreement.

Given its crucial role as the world's largest CO2 emitter, it is important to understand the root of China's incentive to stay out. Figure 6 compares equilibrium outcomes (top panel) and the corresponding benefits and costs to China (bottom panel) if China either stays out of the stable coalition (grey bars) or joins the stable coalition (black bars).

The top panel shows that if China were to join an EU-US coalition, a consequence is that the cooperative carbon price implemented by the coalition would fall. The reason for this effect can be seen in the formula for the cooperative carbon price in Eq. 22. In particular, since China comprises 30 percent of global CO2 emissions, it provides a large incentive for the US and EU to price carbon at a high level when it remains outside the coalition through the mechanism described by the amplification factor. But if China joins the coalition, the relative size of the non-coalition region (i.e., the ratio $\frac{1-\phi_C^E}{\phi_C^E}$) drops sharply. As a result, the effect of China joining is to decrease the cooperative carbon price, thus also the carbon price implemented in the US and EU.

The bottom panel shows the net impact of these changes on China's economy. By joining the stable coalition, China would increase its domestic climate benefit by three and a half billion dollars per year, but to accomplish this, China's domestic abatement cost would increase by 25 billion USD per year. The net effect is that while China stands to benefit from a TCC in which it remains a tier-2 member,

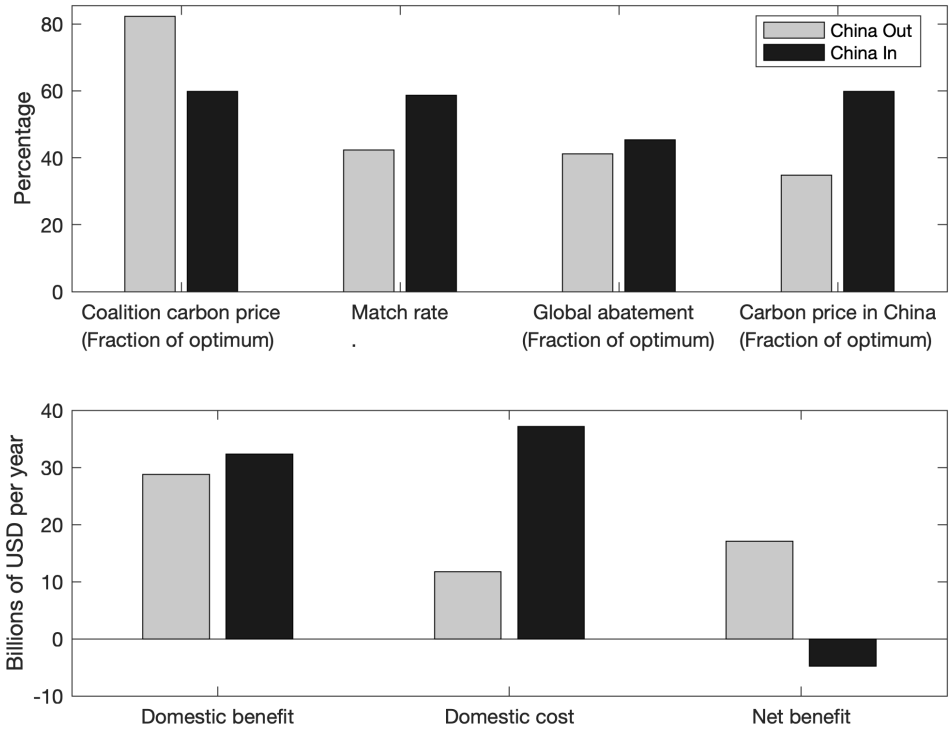

Figure 6: Grey bars indicate outcomes under the stable coalition with China out (EU+US), while black bars indicate outcomes with China in (EU+US+China). The top panel compares equilibrium outcomes under each coalition. The bottom panel compares and decomposes the corresponding payoffs to China.

it would be worse of it were a leading member of the coalition.

Importantly, China still benefits substantially from the existence of the stable coalition—to the tune of 17 billion USD per year. Thus, while China would not want to join, it would also have substantial incentive to support a two-tier climate agreement led by the EU and the US.

It is worth noting that the model generates the result that self-interested behavior should induce the EU and US to lead global climate action without taking into account the fairness issues that have been at the heart of many international policy discussions. Reflecting these issues, cumulative CO2 emissions between 1750 and 2020 were 78 billion metric tons of CO2 for the UK and 16 billion metric tons for Brazil. Since Brazil's population is roughly three times bigger than the UK's, it follows that the historical emissions burden per current living person is a factor of fifteen higher for the UK than it is for Brazil. These differences suggest a large moral impetus for European and North American countries to lead global climate policy. Taking such considerations into account would strongly reinforce the findings here.

5.3 Range of supportable policies

While the analysis in the last subsection takes a concrete stand on the match rate that a given coalition could achieve, there remains considerable uncertainty about what match rate would actually obtain. Moreover, an attractive feature of the proposed agreement structure is the potential for the same agreement to support rising ambition over time—a possibility discussed in the next section. To allow for this possibility, the analysis in this section solves for the range of outcomes that could be supported when varying the match rate from zero to one and varying the degree of cooperation within the coalition from the noncooperative lower bound to the cooperative upper bound.

The left panel of Figure 7 plots the range of supportable carbon prices, and the right panel plots the corresponding range of global abatement levels. In each panel, the grey line shows the outcome in the noncooperative case without issue linkage among tier-1 countries ($\omega_1 = 0$), while the black line shows the cooperative outcome. The gray region in the middle depicts the range of policies that could be supported under varying degrees of issue linkage within the coalition. Values are plotted as a percent

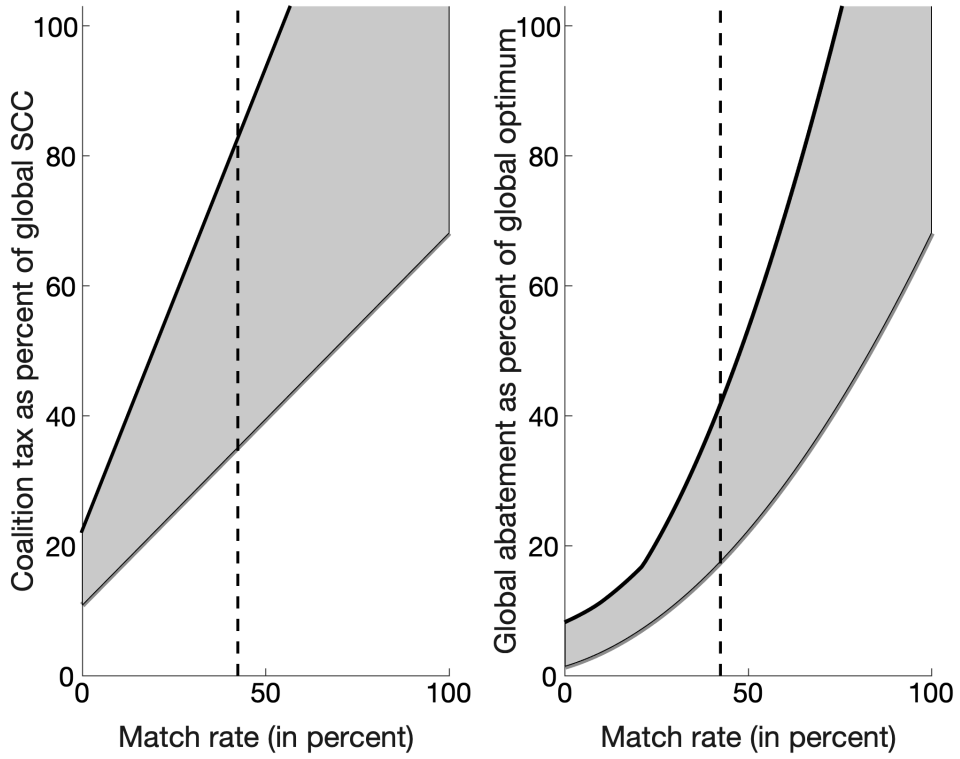

Figure 7: The left panel plots the range of supportable carbon prices with an EU-US coalition as a function of α. The right panel plots the corresponding range of global abatement levels. In each panel, the grey line shows the Optimal Coalition Policy without issue linkage among tier-1 countries ($\omega_1 = 0$), while the black line shows the cooperative policy. The gray region shows the range of policies that can be supported for each value of α. The dashed line indicates the endogenous match rate in the calibrated model with a EU-US coalition.

of the global optimum. The dashed vertical line shows the endogenous match rate under an EU-US coalition in the calibrated model.

The first clear finding is that an EU-US coalition acting cooperatively would substantially overshoot the global SCC if the match rate were set near one. The reason for this "overshooting" result stems from the fact that the coalition as a region has both higher climate damages and lower emissions intensity, controlling for size, than the rest of the world. In addition to the effects described in the analytical section, allowing for heterogeneity in emissions intensity and abatement costs in the current setting amplifies the latter effect.

A second finding is that both optimal policy and aggregate abatement are highly sensitive to the match rate. This high sensitivity is a consequence of the small size of the coalition, which comprises just over 20 percent of global CO2 emissions. Having a small coalition increases the amplification effect that underlies the optimal coalition carbon price, and it also amplifies how much non-coalition abatement is brought on line as the match rate increases.

The next important finding is that even though the amount of abatement achieved by the agreement is substantially higher when the coalition acts cooperatively, the agreement could still achieve a large amount of abatement without issue linkage between the US and EU. Indeed, even when the coalition behaves noncooperatively, with each country choosing its own unilateral best response (within the two-tier matching structure of the TCC) the agreement still achieves 68 percent of the globally efficient level of abatement when the match rate is one.

The finding that global abatement in the noncooperative case is a large fraction of the efficient abatement case when the match rate if high is especially interesting since—as noted in the introduction—big countries like the US will be especially difficult to punish (Böhringer and Rutherford 2017). It suggests that a TCC led by the EU and US could go a long way toward achieving the goals of an effective global climate agreement without that substantial trade threats be imposed on the economies

for which this would be most difficult.

The intuition for the latter finding is closely related to the main result in Proposition 5. There, we found that global abatement in the noncooperative case is proportional to global abatement in the cooperative case, where the proportionality constant is the Herfindahl index of country size within the coalition. Thus, the gap between the cooperative and noncooperative outcomes depends on the degree of concentration within the coalition. In the case of an EU-US coalition, there are two economies of roughly equal size, so the Herfindahl index of country size within the coalition is just under a half. This means that global abatement when $\alpha = 1$ should be just under half what global abatement would be under the cooperative policy when $\alpha = 1$. Because policy overshoots the efficient level in the cooperative case, half of this quantity amounts to almost 70 percent of the efficient rate.

6 Discussion

The results show that a two-tier climate club offers a promising structure through which a small group of economically-powerful economies could stand in for the missing global authority at the root of the climate problem. This section discusses a variety of related practical considerations.

All boats rise Given the risk that tier-2 countries might see the arrangement as coercive, it is important to emphasize the benefits that accrue to countries outside the coalition. The payoffs in Figure 5 show that all tier-2 countries benefit substantially from the stable agreement. Of the eight economies shown in the figure, India benefits the most, gaining roughly 30 billion USD per year when playing its tier-2 role in the agreement. This benefit is about 50 percent greater than the benefit that accrues to the EU, even though the EU stands to gain the most from global carbon abatement. The difference arises because the EU carries a larger share of the abatement burden as a tier-1 participant. Relative to India, China benefits over half as much, while the UK, Japan and Brazil all benefit nearly a third as much.

As a consequence of the arrangement, tier-2 countries are presented with a situation in which the coalition stands in for the missing global authority. While tier-2 countries are required to abate carbon emissions, the carbon abatement they do is matched equally or more by all other countries. Countries benefit from this exchange on average in so far as the global benefit from carbon abatement exceeds the cost. Stated differently, the arrangement will typically pass a domestic cost-benefit test because the free-rider incentive has been removed (Cramton et al. 2016).[15]

Protection against carbon leakage Carbon leakage is a major political obstacle to adopting significant climate policy. While the model does not allow for the possibility that economic activity could move across national borders in response to differences in energy prices, it is straightforward to see that a two-tier climate club would go a long way toward eliminating this concern. The requirement that countries outside the coalition impose a matching carbon price creates a global price floor on fossil energy that reduces the potential for leakage to arise; in doing so, it reduces leakage concerns for the tier-1 countries. In addition, for tier-2 countries, the possibility of carbon leakage is eliminated entirely since they are always at the bottom end of global carbon prices.

The latter effect suggests a powerful reason why China might strongly prefer a TCC over other climate agreement structures, especially if it is allowed to play the role of a tier-2 country. As the manufacturing hub of the global economy, China has substantial grounds to worry that unilateral climate action could cause mobile manufacturing firms to relocate operations to other countries. Meanwhile, being the source of roughly 30 percent of global CO2 emissions—more than two-and-a-half times the next biggest emitter—China faces rising pressure to address its own emissions. A TCC would provide a structure in which it could do this without risking an important source of competitive advantage.

Getting started Reflecting on how a stable climate club might come about, Nordhaus (2015) notes:

[15] It is possible that some countries near the Arctic could benefit from higher global temperatures (e.g., Burke et al. 2015) in which case they could be worse off by the existence of an agreement, but I do not consider this possibility here. More generally, it is possible that some tier-2 countries could be somewhat worse off by the agreement if the match rate is high and their share of global damages is low.

> An important question is, how would a top-down Climate Club get started? There are no clear answers to these questions. International organizations evolve in unpredictable ways. Sometimes, it takes repeated failures before a successful model is developed. The histories of the gold and dollar standards, cholera conventions, the WTO, the European Union, and the Internet all emphasize the unpredictability in the development of international regimes (for some histories, see Cooper et al. 1989). The destination of a Climate Club is clear, but there are many roads that will get there.

In contrast to a single-tier climate club in which many stable coalitions exist and there are many possible options for getting started, a two-tier climate club creates a clear incentive for the economies with the strongest trading prowess and the most to gain from a strong global climate agreement to take the initiative in getting things starts. The finding that the coalition should consist of the EU and US is especially convenient since it aligns with the intuitive fairness considerations that have played an important role in global policy discussions but are outside the model.[16] It is also clear that reducing the number of negotiating partners to two would drastically reduce the complexity of developing and managing a global climate agreement.[17]

To get the agreement started, it would be helpful to keep the match rate low initially. Keeping the stakes low would make it easier to develop buy-in while also setting up the monitoring infrastructure necessary for a climate club to work. Since the key mechanism at the heart of the arrangement hinges on carbon prices being observable, it would be critical to develop a trustworthy international authority to monitor carbon pricing around the world. Probably the easiest way to do this would be to require that countries impose a carbon tax at the point of entry of fossil fuels into the economy (e.g., at the port or at the well). The monitoring apparatus would then need to confirm that the fossil fuel price throughout the economy accurately reflects the tax, ensuring black markets are avoided. In practice, it would be important to allow other forms of policy as well, such as emissions trading, but I do not delve into such details here.

In addition to developing the infrastructure for reliable monitoring, it would also be important to develop an international norm for using trade-based penalties to incentivize carbon pricing in other countries. The approach would entail a significant departure from climate agreements, and it would be important to diplomatically involve all countries, while communicating the benefits to be created and shared through the agreement.

Only after these preliminary steps are accomplished would it make sense to increase the match rate, gradually ratcheting up ambition in the direction of an efficient global agreement.

Multiple tiers and promotion In principle, there is no reason the number of climate-club tiers must stop at two. For example, it would probably make sense to exclude countries with very low CO2 emissions per capita. If the agreement were to exclude countries with per capita emissions below that of India (1.8 metric tons per capita in 2018) it would exclude 85 of the 193 UN countries, while only missing about 5 percent of global CO2 emissions (author's calculations). In this case, there would be two groups of "tier-2" countries: one with a positive match rate and one with a match rate of zero.

More generally, there could be multiple groups of tier-2 countries, each with a different match rate. This flexibility could accommodate ethical intuitions that countries with higher per capita emissions or greater cumulative emissions should carry a greater share of the burden of current abatement. The incentive effect on tier-1 countries would, in this case, depend on a weighted average of the matching abatement across the different tier-2 groups.

An alternative extension that would help increase the match rate without changing the coalition would be to "promote" specific tier-2 countries into a broader role in which they share the burden of meting out punishments (through tariffs on imports into their own country) without changing the abatement obligation that they are held to as a tier-2 country. Countries with a high degree of trading prowess but less incentive to join the stable coalition—such as China, the UK, and Canada—would be important early candidates for this type of promotion, though in the long run, the coalition would

[16] The expectation on the part of low and middle income countries that richer economies who created the problem should take the lead in solving it led to the negotiated structure of the Kyoto Protocol in which abatement requirements fell entirely on Annex I countries, which consisted of the high-income economies that roughly aligned with the EU-US coalition.

[17] While the negotiation would be led by the US and EU, it would still be important to get buy in and explicit support from as many countries from the second tier as possible.

ideally promote as many countries as possible. Since tariffs are not actually imposed in equilibrium provided the combined threat of punishment is sufficiently high, this type of promotion would not appreciably change a country's incentive to go along with the agreement terms.

Comparison with border carbon adjustments The proposed approach is related to the idea of using Border Carbon Adjustments (BCAs) as a means to induce countries outside an abating coalition to impose climate policy within their own borders (Böhringer et al. 2016; Helm and Schmidt 2016). An important advantage of the BCA approach is that it would likely be WTO compliant (Monjon et al. 2011). In contrast, a climate club would require changes to existing WTO rules—what Nordhaus (2015) refers to as "climate ammendments". Nevertheless, there are at least four reasons why a BCA approach would be less effective than a TCC.

First, as Nordhaus (2015) emphasizes, the penalty effect of a BCA is substantially less than that created by a uniform tariff. Thus, if the goal is to create an adequate incentive to induce the target country to impose domestic climate policy, there is more room for a given coalition to induce a bigger response if it uses uniform tariffs instead of tariffs that only apply to the carbon content of imports.

Second, climate policy triggered by a BCA would likely be more restrictive than the economy-wide carbon price imagined in the TCC proposed here. In particular, to oblige the requirements of a BCA policy, a country would only need to change the energy inputs to the export sector, which would typically comprise only a modest fraction of the carbon emissions generated by the entire economy. This means that a large portion of the cheap abatement opportunities available in this country would be missed, thus reducing the efficiency of global abatement achieved by the policy. In contrast, the terms of a TCC explicitly require that a harmonized carbon price be applied across all sectors of the target economy.

Third, as shown in the paper, a TCC has the potential to scale to a fully efficient global agreement. In contrast, the BCA approach would have a lower ceiling on how far it could go. This is true largely because of points two and three above, though a related consideration is that the policy response in non-coalition countries is less predictable under a BCA and thus the inducement for coalition countries to abate more because of the noncoalition match is less clear.

Finally, a major concern with BCAs stems from the complexity of measuring the carbon content of imports, especially those with extended supply chains (Afionis et al. 2017). The simpler requirement with a TCC to tax fossil fuels at the point of entry into the economy avoids the need for complex accounting.

7 Conclusion

The paper proposes a novel structure for a climate agreement that requires non-coalition countries to match coalition abatement at a less than one-for-one rate or face tariffs from the coalition. The arrangement increases the incentive for coalition countries to abate, while reducing the penalty for incomplete participation, a daunting feature of sub-global abatement. In the calibrated model, there is a unique stable coalition that consists of the US and EU, and the coalition achieves over 40 percent of the efficient level of global abatement. In addition, even without threatening punishment against the US or the EU, thus allowing the coalition countries to interact *noncooperatively*, the coalition policy achieves almost 70 percent the efficient global abatement rate if the match rate is increased to one. The proposed structure uses the trading prowess of large economies—economies like the US that may be too big to punish (Böhringer and Rutherford 2017)—to induce matching abatement elsewhere, and in so doing, creates sufficient incentive to induce these economies to establish and enforce a climate agreement that approaches the globally efficient outcome on the basis of self interest alone.

References

[1] Stavros Afionis, Marco Sakai, Kate Scott, John Barrett, and Andy Gouldson. Consumption-based carbon accounting: Does it have a future? *Wiley Interdisciplinary Reviews: Climate Change*, 8(1):e438, 2017.

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

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

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

[5] Christoph Böhringer and Thomas Rutherford. Paris after Trump: An inconvenient insight. 2017.

[6] Christian Broda, Nuno Limao, and David E Weinstein. Optimal tariffs and market power: the evidence. *American Economic Review*, 98(5):2032–65, 2008.

[7] Marshall Burke, Solomon M Hsiang, and Edward Miguel. Global non-linear effect of temperature on economic production. *Nature*, 527(7577):235–239, 2015.

[8] Peter Cramton, David JC MacKay, Axel Ockenfels, and Steven Stoft. *Global carbon pricing: the path to climate cooperation*. The MIT Press, 2017.

[9] Christian Gollier, Jean Tirole, et al. Negotiating effective institutions against climate change. 2015.

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

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

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

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

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

[15] Fabian Kesicki and Paul Ekins. Marginal abatement cost curves: a call for caution. *Climate Policy*, 12(2):219–236, 2012.

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

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

[18] Stéphanie Monjon and Philippe Quirion. A border adjustment for the eu ets: Reconciling wto rules and capacity to tackle carbon leakage. *Climate Policy*, 11(5):1212–1225, 2011.

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

[20] William D. Nordhaus. *Six: The Economics Of Participation*, pages 116–122. Yale University Press, 2008.

[21] William D Nordhaus. Climate club futures: On the effectiveness of future climate clubs. 2021.

[22] William D Nordhaus and Joseph Boyer. *Warming the world: economic models of global warming*. MIT press, 2003.

[23] Ian Parry. Proposal for an international carbon price floor among large emitters. *IMF Climate Notes*, 2021(001), 2021.

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

[25] Christian P Traeger. Ace–analytic climate economy (with temperature and uncertainty). 2018.

[26] Martin L Weitzman et al. How a minimum carbon price commitment might help to internalize the global warming externality. Technical report, National Bureau of Economic Research, 2016.

A Appendix

A.1 Asymmetric trade costs

If country j imposes uniform tariff τ_{ij} on country i, then Nordhaus (2015) assumes that the net income gain to j is

$$\Delta Y_{ij} = X_{i,j}(\alpha_{ij}\tau_{ij} - \beta_{ij}\tau_{ij}^2),$$

where α_{ij} and βij are estimated parameters based on the net benefit of tariffs in the Ossa model between country i and country j. Nordhaus employs separate α and β parameters for each bilateral region pairing. Si

Nordhaus (2015) uses the multi-country, multi-industry, general equilibrium trade war model developed in Ossa (2014) to quantify the impact of tariffs on the net income of different countries. For each bilateral country/region pair in C-DICE, Nordhaus uses the Ossa model to compute the economic impact of a range of uniform import tariffs. The simulation results are used to estimate the parameters of a reduced form tariff benefit function that takes the following form:

$$\Delta Y_{ij} = X_{i,j}(\alpha_{ij}\tau_{ij} - \beta_{ij}\tau_{ij}^2). \tag{27}$$

ΔY_{ij} is the net income gain to j from levying uniform tariff rate τ_{ij} on imports from i, $X_{i,j}$ is imports from i into j, and α_j and βj are the estimated parameters for country j based on the Ossa model simulations.[18] The linear term in (27) captures the terms of trade effect, which entails a gain to the country imposing the tariff and a commensurate loss to the country on which the tariff is levied. The quadratic term captures the simultaneous efficiency loss.

To simulate how the volume of trade varies with country size, the current section employs a simple gravity model of trade without trade frictions. Exports from i to j are given by

$$X_{i,j} = \frac{cY_iY_j}{Y_w},$$

where Y_i (Y_j) is output in i (j), Y_w is world output, and c is a proportionality constant. In this simple world, trade with other countries depends only on the other countries relative size, so geography doesn't matter. In addition, trade between all countries is balanced. I employ this simple model in this section to show how relative size impacts the relative consequences of a trade war absent the "noise" of geographic effects. In the quantitative section, I use actual trade flows between countries to quantify the magnitude of outcomes in a more realistic way.

Given the structure of tariff payoffs, each country benefits from imposing tariffs that are not "too high" provided the other country does not retaliate. Without retaliation, the optimal tariff for country k is

$$\tau_k^* = \frac{\alpha_k}{2\beta_k}. \tag{28}$$

To simulate payoffs in a trade war, I assume that a trade war entails each country playing the unilateral best response in (28). Thus, if country i engages in a trade war with country j then the net loss to country i, as a fraction of GDP, is

$$\frac{L_i(\tau_i^*, \tau_j^*)}{Y_i} = -c\phi_j[\tau_i^*(\alpha_i - \beta_i\tau_i^*) - \alpha_j\tau_j^*]. \tag{29}$$

To see how the consequences of a trade war differ across countries of different sizes, I study the ratio of losses when country i faces coalition "c". The corresponding ratio is

$$\Lambda_i = \frac{L_i(\tau_i^*, \tau_c^*)/Y_i}{L_c(\tau_c^*, \tau_i^*)/Y_c} = \frac{\phi_c}{\phi_i}\left(\frac{\tau_i^*(\alpha_i - \beta_i\tau_i^*) - \alpha_c\tau_c^*}{\tau_c^*(\alpha_c - \beta_c\tau_c^*) - \alpha_i\tau_i^*}\right). \tag{30}$$

In Nordhaus's (2015) calibration derived from the Ossa (2014) model, the optimal tariff is similar across countries as are the α and β parameters. This is not exactly true, but it is useful to note that if

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

these parameters (and the optimal tariff) were the same across countries, then the ratio in (30) would reduce to

$$\Lambda_i \approx \frac{\phi_c}{\phi_i}.$$

In this rough approximation, the relative impact of a trade war equals the inverse of i's size relative to the coalition. This shows that a large coalition would yield substantial clout in being able to threaten trade tariffs against non-coalition countries without needing to worry excessively about the risk of retaliation.

Figure 2 uses the calibrated α and β values from Nordhaus (2015) to compute Λ_i for eight of the biggest countries by GDP. The α and β parameters for the coalition are taken to be an average of each parameter across the US, the EU and China. With this calibration fixed, the x-axis considers the effect of increasing coalition size from zero to a maximum size that corresponds with the size of a combined coalition with all three large countries. Because the optimal tariff of each country is independent of the size of the coalition, the relative penalty is linear in coalition size.

The figure shows that the tier-1 countries are indeed qualitatively different from the tier-2 countries, though the US and China, for example, differ substantially in their capacity to be punished.

A.2 Details of a carbon price policy

The assumptions are the same as in Nordhaus (2008). Countries differ in size but are otherwise homogeneous, including abatement opportunities that scale proportionally with size. Abatement costs in each country i are a power function of the abatement rate expressed as a fraction of own country GDP:

$$\Psi_i(\mu_i) = Q_i\theta_1(\mu_i)^{\theta_2}, \tag{31}$$

where Q_i is GDP in i, μ_i is percent abatement in i, and θ_1 and θ_2 are parameters. The abatement cost function in (**??**) is consistent with that in all recent versions of the DICE model, including Nordhaus (2016). To aggregate abatement costs across countries in an efficient way, marginal abatement costs are equalized by setting the abatement rate (or carbon price) at the same value in each country. Abatement costs aggregated in this way retain the same functional form as in (**??**) with Q_i reflecting GDP for the aggregate region.[19]

A key feature of the abatement cost function for the results in the paper is the degree of convexity, which is governed by the parameter θ_2. For the analytical results, I assume $\theta_2 \geq 2$. Moreover, as a baseline, I follow the most recent DICE calibration (Nordhaus 2016) in assuming $\theta_2 = 2.6$. This calibration implies a substantial amount of cheap abatement opportunities with marginal costs rising rapidly as the abatement rate increases. Proposition 3* considers the case of quadratic abatement costs, and I consider the robustness of key quantitative results to alternative values of θ_2 in the appendix.

Since the paper focuses on policy in the form of a carbon price in each country, I rewrite abatement costs to be a function of country i's carbon price, τ_i. Equating the marginal cost of an extra unit of abatement (denominated in units of emission reduction) with the carbon price implies a power-function relationship between τ_i and the percent abatement rate μ_i:[20]

$$\mu_i = a(\tau_i)^b \equiv G(\tau_i). \tag{33}$$

where

$$a = \left[\frac{\sigma}{\theta_1\theta_2}\right]^{\frac{1}{\theta_2-1}} \tag{34}$$

and

$$b = \frac{1}{\theta_2 - 1}, \tag{35}$$

[19] Aggregate abatement costs are

$$\sum_i \left[Q_i\theta_1\mu^{\theta_2}\right] = \left[\sum_i Q_i\right]\theta_1\mu^{\theta_2} = Q\theta_1\mu^{\theta_2}, \tag{32}$$

where Q is aggregate GDP and μ is the common abatement rate in each country, thus also the aggregate abatement rate.

[20] Details are in Appendix *.

where σ is the carbon intensity of output. Because $G(\cdot)$ is independent of size, it applies equally to a single country or an aggregate region provided abatement efforts are aggregated efficiently across countries. Combining Equation 32 and Equation 33 implies abatement costs

$$C_i(\tau_i) = Q_i \theta_1 a^{\theta_2} (\tau_i)^{b\theta_2}. \tag{36}$$

In addition to having abatement opportunities scale with size, each country i comprises both fraction ϕ_i of global output and fraction ϕ_i of global energy use. It follows that

$$\frac{E_i}{Q_i} = \frac{E}{Q} \equiv \sigma, \tag{37}$$

where Q is global output and E is global energy use. Thus, the carbon intensity of output is the same in all countries.

Given carbon price τ_i, the economy abates 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 region 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

$$\frac{\partial \psi_i}{\partial \hat{\mu}_i} = \theta_2 Q_i \theta_1 \left(\frac{\hat{\mu}_i}{E_i} \right)^{\theta_2 - 1} \frac{1}{E_i} \tag{38}$$

$$= \theta_1 \theta_2 \frac{Q_i}{E_i} \left(\frac{\hat{\mu}_i}{E_i} \right)^{\theta_2 - 1} \tag{39}$$

$$= \frac{\theta_1 \theta_2}{\sigma} \mu_i^{\theta_2 - 1}, \tag{40}$$

where $\sigma = \frac{E}{Q}$ is emissions intensity of output and $\mu_i = \frac{\hat{\mu}_i}{E_i}$.

Next, set this equal to the carbon price:

$$\frac{\theta_1 \theta_2}{\sigma} \mu_i^{\theta_2 - 1} = \tau_i, \tag{41}$$

then solve for μ_i as a function of the carbon price:

$$\mu_i = \left[\frac{\sigma}{\theta_1 \theta_2} \right]^{\frac{1}{\theta_2 - 1}} \tau_i^{\frac{1}{\theta_2 - 1}} \equiv G(\tau_i). \tag{42}$$

This is Equation 33. The function defines a mapping from the carbon price to the abatement rate that is independent of region size, so it can be applied to any region to get local abatement as a function of the local carbon price.

Suppose the coalition enacts carbon price τ^C and the rest of the world matches with $\tau^R = \alpha \tau^C$. From (42),

$$\mu^R = \left[\frac{\sigma}{\theta_1 \theta_2} \right]^{\frac{1}{\theta_2 - 1}} (\alpha \tau^C)^{\frac{1}{\theta_2 - 1}} \tag{43}$$

$$= \alpha^{\frac{1}{\theta_2 - 1}} \left[\frac{\sigma}{\theta_1 \theta_2} \right]^{\frac{1}{\theta_2 - 1}} (\tau^C)^{\frac{1}{\theta_2 - 1}} \tag{44}$$

$$= \alpha^{\frac{1}{\theta_2 - 1}} \mu^C, \tag{45}$$

so a requirement to price carbon at rate $\alpha\tau^C$ is equivalent to a commitment abate at rate $\mu^R = \alpha^{\frac{1}{\theta_2-1}}\mu^C$.

Global abatement in percent becomes

$$\mu = \phi\mu^C + (1-\phi)\mu^R \tag{46}$$
$$= \phi\mu^C + (1-\phi)\alpha^{\frac{1}{\theta_2-1}}\mu^C \tag{47}$$
$$= [\phi + (1-\phi)\alpha^{\frac{1}{\theta_2-1}}]\mu^C \tag{48}$$

Defining $\Gamma \equiv [\phi + (1-\phi)\alpha^{\frac{1}{\theta_2-1}}]^{-1}$,

$$\mu^C = \Gamma\mu,$$

and

$$\mu^R = \alpha^{\frac{1}{\theta_2-1}}\Gamma\mu.$$

A.3 Proof of Proposition 1

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

$$\Psi = \sum_{j=0}^{n}\Psi_j \tag{49}$$
$$= \sum_{j=0}^{n}\phi_j Q\theta_1 G(\alpha_j\tau_0)^{\theta_2} \tag{50}$$
$$= Q\theta_1 G(\tau_0)^{\theta_2}\sum_{j=0}^{n}\phi_j\alpha_j^{\frac{\theta_2}{\theta_2-1}} \tag{51}$$
$$= Q\theta_1\mu^{\theta_2}\cdot\Gamma^{\theta_2}\cdot\sum_{j=0}^{n}\phi_j\alpha_j^{\frac{\theta_2}{\theta_2-1}} \tag{52}$$
$$= \frac{\sum_{i=0}^{n}\phi_i\alpha_i^{\theta_2/(\theta_2-1)}}{\left(\sum_{j=0}^{n}\phi_j\alpha_j^{1/(\theta_2-1)}\right)^{\theta_2}}\cdot Q\theta_1\mu^{\theta_2} \tag{53}$$

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

$$\sum_{i=0}^{n} \phi_i \alpha_i^{\theta_2/(\theta_2-1)} = \phi_0 \alpha_0^{\theta_2/(\theta_2-1)} + \phi_1 \alpha_1^{\theta_2/(\theta_2-1)} \tag{54}$$

$$= \phi + (1-\phi)\alpha_1^{\theta_2/(\theta_2-1)}. \tag{55}$$

Equation 3 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).

A.4 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 3 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

$$(1-\phi)\theta_2 \gamma \alpha^{\theta_2 \gamma - 1} [\phi + (1-\phi)\alpha^{\gamma}]^{\theta_2} < \theta_2 [\phi + (1-\phi)\alpha^{\gamma}]^{\theta_2 - 1} \gamma (1-\phi)\alpha^{\gamma - 1} [\phi + (1-\phi)\alpha^{\theta_2 \gamma}]. \tag{56}$$

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, (56) holds if and only if

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

if and only if

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

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

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

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

A.5 Proof of Proposition 2

Assume $\omega = 0$ and suppose $\{\hat{\tau}_i\}_{i=1}^n$ denotes the minimum carbon price required by the agreement for each coalition country i. Then the tier-1 participation constraint in $\boldsymbol{P1}$ for country i is

$$B_i(\hat{\tau}_i, \{\hat{\tau}_j\}_{j \neq i}; \alpha) - C_i(\hat{\tau}_i) \geq \max_{\tau_i \geq 0} [B_i(\tau_i, \{\hat{\tau}_j\}_{j \neq i}; \alpha) - C_i(\tau_i)] \tag{57}$$

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 (57) 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

$$\mu(\tau_i, \{\hat{\tau}_j\}_{j\neq i}; \alpha) = \phi_i^E \mu_i + \sum_{j\neq i} \phi_j^E \hat{\mu}_j + (1-\phi_C^E)\mu^R \tag{58}$$

$$= \phi_i^E a\tau_i^b + \sum_{j\neq i} \phi_j^E a\hat{\tau}_j^b + (1-\phi_C^E)a[\alpha\hat{\phi}_i^E\tau_i + \alpha\sum_{j\neq i}\hat{\phi}_j^E\hat{\tau}_j]^b. \tag{59}$$

Thus,

$$\frac{\partial\mu}{\partial\tau_i} = \phi_i^E ab\tau_i^{b-1} + (1-\phi_C^E)ab(\alpha\tau^{AVG})^{b-1}\alpha\hat{\phi}_i^E > 0, \tag{60}$$

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

$$\gamma_i E[\phi_i^E ab\tau_i^{b-1} + (1-\phi_C^E)ab(\alpha\tau^{AVG})^{b-1}\alpha\hat{\phi}_i^E] = \theta_1 a^{\theta_2} b\theta_2(\tau_i)^b Q_i \tag{61}$$

Dividing through by $abQ_i\phi_i^E\theta_1\theta_2(\tau_i)^{b-1}$ gives the following equation in τ_i :

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

This implies a series of n equations in n unknowns. If γ_i is the same for all countries, then $\tau^{AVG} = \tau_i$ for all i. Alternatively, if $\theta_2 = 2$ then $b-1 = 0$. In either case,

$$\tau_i = \gamma_i\left[1 + \frac{1-\phi_C^E}{\phi_C^E}\alpha^b\right] = \frac{\gamma_i}{\gamma_C}\tau^{COOP}.$$

A.6 Proof of Proposition 3

The cooperative policy solves

$$\max_{\{\tau_i\geq 0\}_{i=1}^n}\sum_{i=1}^n \Pi_i(\tau_i, \{\tau_j\}_{j\neq i}; \tau_i, \alpha).$$

Since countries move in sync, the tier-1 penalty is never paid, so the problem becomes:

$$\max_{\{\tau_i\geq 0\}_{i=1}^n}\sum_{i=1}^n [\gamma_i E\mu(\tau_i, \{\tau_j\}_{j\neq i}; \alpha) - C_i(\tau_i)].$$

The objective function is strictly concave in each τ_i since Π_i was shown to be strictly concave in τ_i in Appendix A.5, and since the sum of strictly concave functions is strictly concave. I denote the solution to the cooperative problem by $\{\tau_i^C\}_{i=1}^n$.

The first-order condition gives

$$E\frac{\partial \mu}{\partial \tau_i}\sum_i \gamma_i = C_i'(\tau_i^C),$$

where $\frac{\partial \mu}{\partial \tau_i}$ is defined in Eq. 60. Since $b\theta_2 - 1 = b$, substituting gives

$$\gamma_C E[\phi_i^E ab\tau_i^{b-1} + (1-\phi_C^E)ab(\alpha\tau^{AVG})^{b-1}\alpha\hat{\phi}_i^E] = \theta_1 a^{\theta_2} b\theta_2 (\tau_i)^b Q_i, \tag{62}$$

where $\gamma_C \equiv \sum_i \gamma_i$. Dividing through by $abQ_i\phi_i^E\theta_1\theta_2(\tau_i)^{b-1}$ gives:

$$\tau_i^C = \gamma_C\left[1 + \frac{1-\phi_C^E}{\phi_C^E}\alpha^b\left(\frac{\tau_i}{\tau^{AVG}}\right)^{1-b}\right].$$

Since the condition is the same for each i, we must have $\tau^{AVG} = \tau_i$ for each i. This implies

$$\tau_i^C = \gamma_C\left[1 + \frac{1-\phi_C^E}{\phi_C^E}\alpha^b\right],$$

which would amount to a harmonized carbon price across the coalition countries.

Next, I consider what is needed to support the cooperative policy under a TCC. Suppose all countries except i implement the cooperative policy τ_j^C, and suppose there were no penalty to country i for deviating. Then i would choose τ_i to maximize

$$\gamma_i E\mu(\tau_i, \{\tau_j^C\}_{j\neq i}; \alpha) - C_i(\tau_i).$$

Taking the first-order condition gives

$$E\frac{\partial \mu}{\partial \tau_i}\gamma_i = C_i'(\tau_i^{BRWP}),$$

where the superscipt *BRW* stands for "Best Response Without Penalty". As before, $\frac{\partial \mu}{\partial \tau_i}$ is defined in Eq. 60. This is the same condition as in Eq. 62 with γ_C replaced with γ_i. Hence analogous algebra implies

$$\tau_i^{BRW} = \gamma_i\left[1 + \frac{1-\phi_C^E}{\phi_C^E}\alpha^b\left(\frac{\tau_i^{BRW}}{\tau^{AVG}}\right)^{1-b}\right].$$

Previously, we had $\frac{\tau_i}{\tau^{AVG}} = 1$, but here we will instead have $\frac{\tau_i}{\tau^{AVG}} < 1$ since $\gamma_i < \gamma_C$. This fact reinforces the incentive to deviate downwards, and it follows that $\tau_i < \tau_i^C$. Since the individual payoff function for i is strictly concave in τ_i, the only way the coalition can keep country i from deviating away from the cooperative policy is if it imposes a penalty $\omega_1 > 0$ if i chooses $\tau_i < \tau_i^C$.

The minimum penalty needed to get i to go along with the cooperative policy can be found by forcing the participation constraint for i to equality when all other countries choose τ_j^C. This implies

$$\omega_{1,i} = [B_i(\tau_i^{BRW}, \{\tau_j^C\}_{j\neq i}; \alpha) - C_i(\tau_i^{BRW})] - [B_i(\tau_i^C, \{\tau_j^C\}_{j\neq i}; \alpha) - C_i(\tau_i^C)].$$

It follows from the derivation above that the individual payoff function $B_i(\tau_i, \{\tau_j^C\}_{j\neq i}; \alpha) - C_i(\tau_i)$ is strictly concave and attains a max at $\tau_i = \tau_i^{BRW}$. Moreover, since $\tau_i^{BRW} < \tau_i^C$, it follows that $\omega_{1,i} > 0$. Taking the max over all i gives

$$\bar{\omega}_1 = \max(\omega_{1,i}, \ldots, \omega_{n,i}),$$

which is the minimum uniform penalty which, if applied to all i, would support the cooperative policy. ω_1 is finite by construction, and it is positive since all $\omega_{1,i}$ are positive.

A.7 Proof of Proposition 4

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 OCP problem as $\{\tau_i^M(\alpha)\}_{i=1}^n$.
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

$$\tau_i = \gamma\phi[1 + \frac{1-\phi}{\phi}\alpha]. \tag{63}$$

Proof. The objective function for the OCP problem is

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

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

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

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 (63).

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*

$$\tau_i^* = \phi_i\gamma[1 + \frac{1-\phi}{\phi}\alpha] = \hat{\phi}_i\tau^C(\alpha) < \tau^C(\alpha). \tag{64}$$

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

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

$$\Pi_i(\tau_i, \{\tau_j\}_{j\neq i}) = \phi_i\gamma E\left[\phi_i\frac{\sigma}{\theta_1\theta_2}\tau_i + \sum_{j\neq i}\phi_j\frac{\sigma}{\theta_1\theta_2}\tau_j + (1-\phi)\frac{\sigma}{\theta_1\theta_2}\alpha\hat{\phi}_i\tau_i + (1-\phi)\frac{\sigma}{\theta_1\theta_2}\alpha\sum_{j\neq i}\hat{\phi}_j\tau_j\right] \tag{65}$$

$$- \phi_i Q\theta_1\left(\frac{\sigma}{\theta_1\theta_2}\right)^2(\tau_i)^2. \tag{66}$$

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

$$\tau_i^*(\hat{\tau}_i - \tau_i^*) = \frac{1}{\theta_2}[(\hat{\tau}_i)^2 - (\tau_i^*)^2] - \omega_1\frac{\theta_1\theta_2}{\sigma^2}. \tag{67}$$

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

$$(\hat{\tau}_i)^2 - (\tau_i^*)^2 = (\hat{\tau}_i - \tau_i^*)(\hat{\tau}_i + \tau_i^*) \tag{68}$$
$$= x(x + 2\tau_i^*) \tag{69}$$
$$= x^2 + 2\tau_i^* x. \tag{70}$$

Substituting into Eq. 67 gives

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

Simplifying gives

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

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

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

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

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

If and only if,

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

If and only if,

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

If and only if,

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

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

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

A.8 Proof of Proposition 5

Without issue linkage in the coalition, Eq. 64 implies that

$$\tau^{AVG} = \sum_i \hat{\phi}_i\tau_i = \sum_i \hat{\phi}_i\phi_i\gamma\left[1 + \frac{1-\phi}{\phi}\alpha\right] = \phi\gamma\left[1 + \frac{1-\phi}{\phi}\alpha\right]\sum_i \hat{\phi}_i^2 = H(\hat{\boldsymbol{\phi}})\tau^C(\alpha),$$

where $H(\hat{\boldsymbol{\phi}}) = \sum_i \hat{\phi}_i^2$ is the Herfindahl index of country size within the coalition. Since abatement is linear in the tax, the resulting amount of global abatement is

$$\begin{aligned}\mu(\alpha;\omega_1 = 0) &= \phi a\tau^{AVG} + (1-\phi)a\alpha\tau^{AVG} \\ &= [\phi + (1-\phi)\alpha]H(\hat{\boldsymbol{\phi}})a\tau^C(\alpha).\end{aligned}$$

Comparing this to global abatement under the cooperative policy,

$$\begin{aligned}\mu^C(\alpha) &= \phi a\tau^C(\alpha) + (1-\phi)a\alpha\tau^C(\alpha) \\ &= [\phi + (1-\phi)\alpha]a\tau^C(\alpha),\end{aligned}$$

implies that

$$\mu(\alpha;\omega_1 = 0) = H(\hat{\boldsymbol{\phi}})\mu^C(\alpha).$$

If $\alpha = 0$, then $\mu^C(0) = a\phi^2\gamma$, so

$$\mu(0;\omega_1 = 0) = H(\hat{\boldsymbol{\phi}})\mu^C(0) = \phi^2 H(\hat{\boldsymbol{\phi}})a\gamma = H(\phi)\mu^*.$$

If $\alpha = 1$, then $\mu^C(1) = a\gamma = \mu^*$, so

$$\mu(1;\omega_1 = 0) = H(\hat{\boldsymbol{\phi}})\mu^*.$$

A.9 Extension of results for quantitative 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.

Nash policy Next, I extend the derivation in Appendix A.5 for the case in which tier-1 countries are fully heterogeneous with ϕ_i^E, ϕ_i^Q, γ_i, and $\theta_{1,i}$ all distinct.

I begin with Eq. 61, but I extend it to allow for different σ_i and $\theta_{1,i}$ in each i. 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 $a_i b Q \phi_i^E \theta_{1,i} \theta_2 \tau_i^{b-1}$ 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

$$\tau_i = \gamma_i \left[1+\alpha^b \frac{1-\phi_C^E}{\phi_C^E}\left(\frac{\sigma_R}{\sigma_i}\frac{\theta_{1,i}}{\theta_{1,R}}\right)^b(\frac{\tau_i}{\tau^{AVG}})^{1-b}\right], \quad \text{for } i=1,\ldots,n. \tag{71}$$

Cooperative policy

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

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

where

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

and

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

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

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

where

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

and

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

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

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

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

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

Thus,

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

where

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

Taking the derivative,

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

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

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

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

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

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

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

$$\tau^{COOP} = \gamma_C \frac{\sigma}{\theta_2} \frac{\Theta_1}{\Theta_2} \tag{72}$$

$$= \gamma_C \frac{\sigma}{\theta_2} \frac{\sum_j \phi_j^E a_j + (1-\phi_C^E) a_R \alpha^b}{\sum_{i=1}^{n} \phi_i^Q \theta_{1,i} a_i^{\theta_2}}. \tag{73}$$

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. □

Next, suppose all countries have the same emissions intensity of output ($\sigma_i = \sigma$ for all i) and the same abatement cost function scale parameter ($\theta_{1,i} = \theta_1$ for all i). It follows that $\phi_i^E = \phi_i^Q$ for all i. Thus,

$$\begin{aligned} \tau^{COOP} &= \gamma_C \frac{\sigma}{\theta_1 \theta_2} a^{1-\theta_2} \frac{\sum_j \phi_j^E + (1-\phi_C^E)\alpha^b}{\sum_{i=1}^{n} \phi_i^Q} \\ &= \gamma_C \frac{\phi_C^E + (1-\phi_C^E)\alpha^b}{\phi_C^Q} \\ &= \gamma_C \left(1 + \frac{1-\phi_C^E}{\phi_C^E} \alpha^b\right) \end{aligned}$$