Stable IEAs when countries choose emission and abatement separately

Effrosyni Diamantoudi, Concordia University
Eftichios S Sartzetakis, University of Macedonia*
Stefania Strantza, Thompson Rivers University

October 2022 (Preliminary version)

Abstract

The paper contributes to the literature that examines the formation of international environmental agreements as a two-stage game, where in the first countries decide whether to sign or not the agreement and in the second they choose either emissions or abatement. When the second stage is modelled as a Nash equilibrium only small coalitions are stable, while when is modelled as a Stackelberg equilibrium larger coalitions become stable but attain little welfare improvement, a result coined as the paradox of cooperation. In the present paper we allow countries to choose both emissions and abatement separately and we assume Stackelberg leadership in the second stage. Using general functional forms, we confirm the paradox of cooperation, avoiding breaking positivity constraints and we offer clear intuition. When benefits from emissions and damages from global net emissions are substantial, coalition's leadership advantage cannot overcome free-riding incentives if abatement costs are high, restricting the size of stable coalitions. Large coalition become stable only when the cost of abatement decreases substantially reducing free-riding incentives. However, the benefits they attain can also be reached by countries acting independently.

Keywords: International Environmental Agreements, Choice over abatement and emissions, Stable coalition size

JEL: D6, Q5, C7

*Corresponding author: E.S. Sartzetakis, Department of Economics, University of Macedonia, 156 Egnatia Str., 54636. Email: esartz@uom.edu.gr

1 Introduction

Rapidly developing global environmental crises, including climate change, require internationally coordinated action. Responding to this need, a large and expanding body of literature examining the formation of international environmental agreements has developed. A significant part of this literature uses game theoretic tools. The main body of this literature arrives at the pessimistic result that a stable agreement will either be ratified by a small number of countries, or if larger coalitions form they achieve very little welfare improvements, a result coined by Barret (1994) as the paradox of cooperation. The present paper contributes to this literature by modelling countries' choice of emissions and abatement separately, which facilitates the provision of a clear intuition of the above results using general functional forms.

The literature examines the formation of international environmental agreements as a two-stage game, originally developed in D'Aspremont et al. (1983): In the first stage countries, assumed symmetric, decide whether or not to participate in the agreement and in the second they take an action (either emissions or abatement) with signatories internalizing the effect of their actions on all other coalition members while nonsignatories maximize their own welfare. The first stage is a non-cooperative Nash equilibrium while the second stage is modelled as either a Nash or a Stackelberg equilibrium. Under the assumption that signatories and nonsignatories choose their emissions or abatement levels simultaneously, only small coalitions are stable.[1] Under the assumption that the coalition acts as a Stackelberg leader, larger stable coalition become possible. Using abatement as the only choice variable, Barrett (1994) finds that, depending on the cost of abatement, any large coalition, including the grand coalition, is stable. However, large coalitions require extensive abatement effort, which, if the model is expressed in terms of emissions, imply negative emissions as shown in Diamantoudi and Sartzetakis (2006). If emissions are constrained to be positive, only relatively small coalition sizes are stable, except at the corner as shown by Ulph and Rubio (2006). Moreover, Barrett (1994) pointed to the paradox of cooperation, which states that stable coalitions achieve very little especially when the

[1] The simultaneous moves game has been examined by Carraro and Siniscalco (1993), Finus and Rundshagen (2001) and Rubio and Casino (2001).

potential benefits from cooperation are large. The literature has discussed this result extensively providing interesting insights. The framework of our analysis using general functional forms allows us to provide the following interpretation to the paradox of cooperation: Coalitions aiming at reducing a public bad (net emissions) even when they have a leadership advantage over free-riders cannot become immune to deviations and thus their membership is limited, when individual benefits (from production and consumption) are valued high relative to damages from aggregate activity (net emissions) and technology solutions (abatement) are very expensive relative to benefits. Improvements in technology that substantially reduce the cost of abatement allow for large stable coalitions since they reduce free-riding incentives, that is, make participation to the coalition effortless since choices of members and free-riders differ very little.

This literature, up until recently, uses specific functional forms or directly resort to simulations. Diamantoudi and Sartzetakis (2015) using general functional forms and emissions as the only choice variable determine the size of the stable coalition in terms of the slopes of marginal benefits and marginal damages. More recently Finus et al. (2021) also showed, using general functional forms and abatement as the only choice variable, that assuming the coalition acts as a Stackelberg leader in the second stage yields at least as large stable coalitions as the simultaneous moves game. They also provide very interesting intuition emphasizing the role of the best response functions. However, the emphasis of the first paper is on farsighted behavior and of the second on adaptation, which is consider as an additional choice variable.

In the present paper we allow countries to choose both emissions and abatement separately and we assume Stackelberg leadership in the second stage. Allowing countries to choose emissions and abatement separately, yields larger stable coalitions without violating the emissions positivity constraint; however, these large coalitions offer very little welfare improvement. Using general functional forms we provide clear intuition by identifying the two effects that determine the size of the stable coalition: On the one hand, each coalition member internalizes the damages its own net emissions impose on all other members, which leads them to choose lower emissions and higher abatement relative to the outsiders. On the other hand, assuming that the

coalition acts as a leader, its members take into account the reaction of non-members to their choices and thus, they choose higher emissions and lower abatement. We show that both these opposing effects depend on the coalition size: when the coalition is small, each signatory internalizes damages on a small number of countries, while exploits its leadership advantage on a larger number of nonsignatories, but as the coalition size increases the damage internalization effect is augmented while the leadership effect diminishes. There is a critical coalition size at which the two effects offset each other and thus, emissions and abatement and therefore welfare of signatories and nonsignatories become equal, that is, they return to their non-cooperative levels.

We solidly associate this critical size to the smallest size of a stable coalition[2] and thus, we are able to identify its determinants: Under the assumptions of concave benefits and convex damages and abatement costs, nonsignatories respond to an increase in the coalitions net emissions by decreasing their own emissions and increasing their abatement. An improvement in abatement technology that reduces the slope of the marginal cost of abatement makes nonsignatories' reaction functions more responsive, allowing signatories to further exploit their leadership position. Thus, the two effects identified above become equal at a higher coalition size, yielding a larger stable coalition. However, since stable coalitions are closely associated to the critical size that yields the non-cooperative outcome, any stable coalition, regardless of its size, yields minimal welfare improvements. It is also clear that if decisions in the second stage are made simultaneously, only very small coalitions can be stable since there is no leadership effect. Therefore, in this two-stage game framework and under the assumption of concave-convex functions, it is the leadership effect that facilitates larger stable coalitions; furthermore, the leadership effect becomes stronger as abatement costs decrease and damages are substantial. However, reductions in net emissions attained by the coalition are obtained even when all countries act independently. Thus, the paradox of cooperation simply describes situations in which the free-riding incentives that exists when there are substantial benefits from

[2]Using general functonal forms we cannot exclude the posibility that there is another, larger stable coalition, although we are not able to find specifi functional forms that yield more than one stable coalition.

emissions and damages from net emissions, are reduced or even eliminated by the improvements in abatement technology.

To the best of our knowledge, McGinty (2020) is the first to explain the paradox of cooperation by identifying the damage internalization and the leadership effects. However, it does so using specific functional forms in a model with abatement as the only choice variable and without mentioning positivity constraints. The paper further considers transfer schemes relaxing the assumption of symmetric countries. Finus et al. (2021), extend the analysis using general functional forms also in a model using abatement as the only choice variable. This paper's focus is in examining the role of adaptation, which is considered as a separate choice variable. Since both these papers set up the model with abatement as the only choice variable, their analysis takes into account only benefits from global abatement and cost of abatement. In contrast, by separating the choice of emissions from abatement, apart from creating a more realistic framework, it allows us to provide clear intuition at all steps of the analysis. We are able to demonstrate that the result coined as paradox of cooperation is not paradoxical at all: assuming net emissions inflict large damages relative to benefits from emissions, if abatement costs are high the only way to decrease net emissions is by decreasing production and consumption, which naturally poses a limit on the size of the coalition; if abatement costs are low, countries engage extensively in abatement, eliminating most, or even all, emissions and enjoy the benefits from production and consumption which though can be attained by countries acting independently.

The above literature builds on a framework that highlights the existence of strong free-riding incentives which are moderated only when forming coalitions is not relevant, in the sense that it does not provide any welfare benefits over non-cooperation. However, this pessimistic result depends on a number of assumptions: Karp and Simon (2013), using an abatement model and dropping the assumption of strictly convex abatement costs, show using specific examples that larger coalitions are stable and yield significant welfare benefits. Furthermore, if R&D investments that reduce abatement costs exhibit increasing returns to scale, Barret (2006) shows that even the grand coalition is stable. In general, technology agreements and R&D cooperation are considered as club goods whose attractiveness may outweigh the incentive to

free-ride,[3] the same as linking trade to environmental agreements. Optimistic results are also obtained if we assume that countries are farsighted, that is, when a country contemplating joining or leaving the coalition takes into account all other countries' participation decisions.[4] More recently there is an interesting debate as to whether introducing adaptation as a choice could lead to larger stable coalitions.[5]

The rest of the paper is organized as follows. Section 2 lays out the model. Section 3 presents the results using general functional forms. Section 4 provides the specific example of quadratic benefit, damage and abatement cost functions and derives complete analytical solutions that enhance the intuition of the previous Section's results. It also provides some simulation results to highlight some important comparative statics. The last Section concludes the paper.

2 The Model

We assume that there exist n symmetric countries, $N = \{1, \ldots, n\}$. Production and consumption activities in each country i, $i \in N$, yield benefits but they also generate emissions, $e_i > 0$, of a global pollutant. Aggregate emissions of the global pollutant, $E = \sum_{i=1}^{n} e_i$, generate damages in each country. Each country, responding to the adverse effects of emissions, could engage in mitigation consisting of activities that reduce emissions either by reducing production and consumption, or by engaging in abatement $x_i \geqslant 0$. Country i assumes the complete cost of its mitigation effort while the benefits are spread globally: while country i's emissions create a negative externality, its abatement efforts generate a positive externality. Therefore, countries participate in a decision game with two choice variables: emission and abatement.

Country i's social welfare, W_i, is defined as the total benefits country i receives from emitting, that is, from production and consumption activities,

[3] See for example Hoel and de Zeeuw (2010) and Goeschl and Perino (2012).

[4] Diamantoudi and Sartzetakis (2015) and (2018), de Zeeuw (2008), and Osmani and Tol (2009). However, Benchekroun and Chaudhuri (2012), using a farsighted stability concept, find that eco-innovations can reduce the stability of IEAs.

[5] Breton and Sbragia (2017), Benchekroun et al. (2017), Bayramoglu et al. (2018), Rubio (2018) and Finus et al. (2021).

$B_i(e_i)$, minus the environmental damages, $D_i(NE)$, suffered from the aggregate global net emissions, $NE = E - X$, defined as the difference between global emissions E and global abatement $X = \sum_{i=1}^{n} x_i$. To complete the definition of country i's net benefits, we subtract the cost of abatement, $C_i(x_i)$. Since countries are assumed to be identical we henceforth drop the subscripts and the individual country's welfare function is,

$$W = B(e_i) - D(NE) - C(x_i).$$

We assume that $B(e_i)$ is strictly concave, that is, $B(0) = 0$, $B' \geq 0$ and $B'' < 0$. We further assume that environmental damages $D(NE)$ are strictly convex in net emissions, that is, $D(0) = 0$, $D'(NE) \geq 0$ and $D''(NE) > 0$. Finally, we assume that the cost of abatement is strictly convex, that is $C(0) = 0$, $C' \geq 0$ and $C'' > 0$.

3 General results

Ratification of the IEA is depicted by the formation of a coalition. In particular, a set of countries $S \subset N$ sign an agreement and $N \backslash S$ do not. Let the size of coalition be denoted by $|S| = s$, the coalition's aggregate net emissions by $NE_s = E_s - X_s$ while each member's net emissions are $ne_s = e_s - x_s$, such that $NE_s = sne_s$. In a similar manner, each nonsignatory's net emissions are $ne_{ns} = e_{ns} - x_{ns}$, giving rise to a total emission level generated by all nonsignatories $NE_{ns} = E_{ns} - X_{ns} = (n-s)ne_{ns}$. The aggregate net emission level is, $NE = NE_s + NE_{ns} = sne_s + (n-s)ne_{ns}$.

Nonsignatories behave non-cooperatively after having observed the choice of signatories. Therefore, their maximization problem gives rise to an indirect welfare function w_{ns} as follows:

$$w_{ns}(e_s, x_s, s) = \max_{e_{ns}, x_{ns}} \left[B(e_{ns}) - D\left[sne_s + (n-s-1)ne_{ns} + ne_{ns} \right] - C_{ns}(x_{ns}) \right].$$

Taking into account that $\frac{\partial NE}{\partial e_{ns}} = 1$ and $\frac{\partial NE}{\partial x_{ns}} = -1$, at the optimum, nonsignatories' emission and abatement satisfy the conditions,

$$B'(e^*_{ns}(e_s, x_s, s, x_{ns})) = D'(NE), \tag{1}$$

$$C'(x^*_{ns}(x_s, e_s, s, e_{ns})) = D'(NE), \tag{2}$$

where, $D'(NE) = \frac{\partial D(NE)}{\partial NE}$ and $NE = s\,(e_s - x_s) + (n-s)(e^*_{ns}(e_s, x_s, x_{ns}) - x^*_{ns}(e_s, x_s, e_{ns}))$. The solution of the above system yields the best response functions $Re = e^*_{ns}(e_s, x_s, s)$ and $Rx = x^*_{ns}(e_s, x_s, s)$.

Expressing the first order conditions (1) and (2) in terms of the coalition's aggregate emissions and abatement, $E_s = se_s$ and $X_s = sx_s$, we can differentiate both with respect to $NE_s = s(e_s - x_s)$ and solve to obtain the slope of nonsignatories' reaction functions around their equilibrium values,

$$Re' = \frac{\partial e^*_{ns}\,(E_s, X_s)}{\partial NE_s} = \frac{-1}{(n-s)\,(1-\psi) - \phi} \tag{3}$$

$$Rx' = \frac{\partial x^*_{ns}\,(E_s, X_s)}{\partial NE_s} = \frac{-\psi}{(n-s)\,(1-\psi) - \phi} \tag{4}$$

where $\phi \equiv \frac{B''(e^*_{ns}(E_s, X_s))}{D''(NE)} < 0$ and $\psi \equiv \frac{B''(e^*_{ns}(E_s, X_s))}{C''(x^*_{ns}(E_s, X_s))} < 0$, given $B'' < 0$, $D'' > 0$, and $C'' > 0$. Since the denominator of both (3) and (4) is positive, the slope of nonsignatories best response functions take values in the following ranges: $\frac{-1}{(n-s)} < Re' \leqslant 0$ and $1 > Rx' \geqslant 0$. Responding to an increase in signatories' net emissions, nonsignatories decrease their emissions and increase their abatement. Combining equations (3) and (4) yields the slope of the best response of nonsignatories' net emissions $RNe' = \frac{-(1-\psi)}{(n-s)(1-\psi)-\phi} < 0$, given that both ϕ and ψ are negative.

Before turning to signatories' choices, we examine the effect that changes in the slope of marginal damages, $D''(NE)$, and marginal abatement cost, $C''(x^*_{ns}\,(E_s, X_s))$ have on the slope of nonsignatories reaction functions, holding the benefit function unchanged. If new discoveries about the effect of net emissions yield a higher $D''(NE)$, the value of ϕ increases yielding a decrease in the denominator of both (3) and (4) and thus, both emissions and abatement reaction functions become more responsive. The overall effect of a change in $D''(NE)$ on the absolute value of RNe' is positive since $\frac{\partial |RNe'|}{\partial \phi} = \frac{1-\psi}{[(n-s)(1-\psi)-\phi]^2} > 0$. If R&D investment in abatement cost lowers $C''(x^*_{ns}\,(E_s, X_s))$, the absolute value of ψ increases and thus, the less responsive the emission reaction function becomes but the more responsive the abatement reaction function becomes. Overall, a decrease in $C''(x^*_{ns}\,(E_s, X_s))$ makes nonsignatories' net emissions reaction function more responsive since $\frac{\partial |RNe'|}{\partial \psi} = \frac{\phi}{[(n-s)(1-\psi)-\phi]^2} < 0$.

Signatories maximize the coalition's welfare, sw_s, taking explicitly into account nonsignatories' behavior. Similarly, the coalition's maximization problem yields an indirect welfare function ω_s as follows:

$$w_s(s) = \frac{1}{s} \max_{e_s, x_s} \left[sB\left(e_s\right) - sD\left[sne_s + (n-s)ne^*_{ns}\right] - C(x_s)\right].$$

The signatories' emissions $e^*_s(s)$ and abatement $x^*_s(s)$ at the equilibrium satisfy the following conditions,

$$B'(e^*_s(s, x_s)) = D'(NE^*)\frac{\partial NE^*}{\partial e_s} \ , \tag{5}$$

$$C'(x^*_s(s, e_s)) = D'(NE^*)\frac{\partial NE^*}{\partial x_s} \ . \tag{6}$$

where, $NE^* = s\left(e^*_s(s, x_s) - x^*_s(s, x_s)\right) + (n-s)\left(e^*_{ns}(e_s, x_s, x_{ns}) - x^*_{ns}(e_s, x_s, e_{ns})\right)$. The derivatives of aggregate net emissions around the equilibrium values $e^*_s(s)$ and $x^*_s(s)$ are: $\frac{\partial NE^*}{\partial e_s} = s + (n-s)\left(\frac{\partial e^*_{ns}(e_s, x_s))}{\partial e_s} - \frac{\partial x^*_{ns}(e_s, x_s))}{\partial e_s}\right)$ and $\frac{\partial NE^*}{\partial x_s} = -s + (n-s)\left(\frac{\partial e^*_{ns}(e_s, x_s))}{\partial x_s} - \frac{\partial x^*_{ns}(e_s, x_s))}{\partial x_s}\right)$. The solution of the above system yields the equilibrium emission and abatement levels for the signatories $e^*_s(s)$ and $x^*_s(s)$.

Proposition 1 extends the result derived in Diamantoudi and Sartzetakis (2015) to the case in which emissions and abatement are treated as separate choice variables. Similar to the case that emissions is the only choice variable, we establish that there exists a critical coalition size, below which signatories emit more, abate less and attain higher welfare than the non-signatories and above which the reverse is true. This critical size is determined by adjusting, to the lower integer, the value of $z^{\min}$, where $z^{\min}$ denotes the intersection of e_s with e_{ns} and of x_s with x_{ns} and thus, of w_s with w_{ns} which lies right at the minimum value of w_s. Denote by e_{nc}, x_{nc} and E_{nc} the individual country's emissions and abatement and the aggregate emissions respectively, in the purely non-cooperative case, where there is no leader and firms compete a la Cournot. To simplify the exposition we use the notation ϕ and ψ defined above. The proof of Proposition 1 is relegated to Appendix I.

Proposition 1 *Consider the indirect welfare functions of signatory and non-signatory countries, $w_s(s)$ and $w_{ns}(e^*_s(s), x^*_s(s), s)$ respectively. Let*

$$z^{\min} = \frac{n\left(1-\psi\right) - \phi}{1 - \phi - \psi}$$

then,
1. $e_s^*(s) \gtreqless e_{ns}^*(s) \Leftrightarrow s \gtreqless z^{\min}$ *and* $x_s^*(s) \gtreqless x_{ns}^*(s) \Leftrightarrow s \gtreqless z^{\min}$,
2. *if* $s = z^{\min}$ *then* $e_s^*(s) = e_{ns}^*(s) = e_{nc}$ *and* $x_s^*(s) = x_{ns}^*(s) = x_{nc}$,
3. $w_s(s)$ *increases (decreases) in* s *if* $s > z^{\min}$ $(s < z^{\min})$,
4. $z^{\min} = \arg\min_{s \in \Re \cap [0,n]} w_s(s)$,
5. $w_s(s) \gtreqless w_{ns}(e_s^*(s), x_s^*(s), s) \Leftrightarrow s \gtreqless z^{\min}$.

To discuss the intuition of the properties of countries' choice variables presented in Proposition 1, recall that what differentiates the behavior of coalition members from free-riders is that they internalize the damages their net emissions inflict on all other members of the coalition and they exercise their leadership power over free-riders. These two effects move in opposing directions and thus, when they offset each other, coalition's members choices are the same as those of the non-members. Formally, nonsignatories in choosing e_{ns}^* and x_{ns}^* according to (1) and (2), they set $\frac{\partial NE}{\partial e_{ns}} = 1$ and $\frac{\partial NE}{\partial x_{ns}} = -1$, while the signatories, in choosing e_n^* and x_n^* according to (5) and (6), set $\frac{\partial NE}{\partial e_s} = s + (n-s)\left(\frac{\partial e_{ns}^*(e_s,x_s))}{\partial e_s} - \frac{\partial x_{ns}^*(e_s,x_s))}{\partial e_s}\right)$ and $\frac{\partial NE}{\partial x_s} = -s + (n-s)\left(\frac{\partial e_{ns}^*(e_s,x_s))}{\partial x_s} - \frac{\partial x_{ns}^*(e_s,x_s))}{\partial x_s}\right)$. Given that we assume completely homogeneous countries, signatories and nonsignatories' first order conditions differ only in the response of NE to their choice variables. The slope of the reaction function of all $(n-s)$ nonsignatories' net emissions to changes in one of the signatories' emissions, using (3) and (4) and noting that here we differentiate with respect to e_s and not E_s, is $\frac{\partial NE_{ns}}{\partial e_s} = (n-s)\frac{\partial (e_{ns}^*(e_s,x_s) - x_{ns}^*(e_s,x_s))}{\partial e_s} = \frac{-s(n-s)(1-\psi)}{(n-s)(1-\psi)-\phi}$. Therefore, the difference between signatories and nonsignatories optimization conditions, is $(s-1) - \frac{s(n-s)(1-\psi)}{(n-s)(1-\psi)-\phi}$.

A marginal increase (decrease) in one of the signatories' emissions (abatement) increases marginal damages to all other coalition members and thus, the term $(s-1)$ captures the internalization of marginal damages effect: the increase in each coalition's member marginal damage coming indirectly from internalizing the marginal damages all other coalition members $(s-1)$ suffer from a marginal increase (decrease) in that member's emissions (abatement). In the same time, nonsignatories respond to the increase (decrease) in one of the signatories' emissions (abatement) by decreasing their net emissions on aggregate by $\frac{s(n-s)(1-\psi)}{(n-s)(1-\psi)-\phi}$, that is, this term captures the leadership effect: the decrease in each coalition's member marginal damage coming indirectly

through the response of all nonsignatories' net emissions to a marginal increase (decrease) in that member's emissions (abatement). Setting the difference equal to zero and solving yields the critical value $z^{\min} = \frac{n(1-\psi)-\phi}{1-\phi-\psi}$. Note that for $z^{\min} > x$, where x is a positive number, $(n-x) > \frac{-(x-1)\phi}{1-\psi}$ which is true for $x = 1$, since $n > 1$, $\phi < 0$ and $\psi < 0$. As ϕ decreases an/or ψ increases, that is as D'' increases and/or C'' decreases relative to B'', the higher is the value of $z^{\min}$.

Coalition members, in choosing collectively their net emissions, take into account the damages that each member's net emissions impose on all members, and in the same time they exploit their leadership advantage. The size of these two effects on coalition members' choice of net emissions depends on the size of the coalition. When the coalition has few members, the first effect is smaller since lower aggregate damages are internalized, while the second effect is larger, since the coalition makes gains out of a larger number of followers. At the critical size of the coalition $z^{\min}$ these two effects offset each other. For coalition sizes higher than $z^{\min}$ the reverse is true and thus, the welfare of nonsignatories exceeds that of signatories.

The above results are established in the literature (McGinty, 2019 and Finus et al., 2021) in the case of a single choice variable. Although technically the approach is similar, modelling abatement separately from emissions allows us to take into account both dimensions of the problem, that is, individual benefits and global damages from emissions and global benefits and individual costs from abatement. Given the importance of the critical size $z^{\min}$ in determining the size of the stable coalition, as we will establish in what follows, we explore its properties and compare its value to that of the single choice variable models. Denoting by $z^{\min}_{x=0}$ the critical value of the size of the coalition in the case that emissions is the only choice variable and utilizing the corresponding value derived in Diamantoudi and Sartzetakis (2015), $z^{\min}_{x=0} = \frac{n-\phi}{1-\phi}$, where here $\phi = \frac{B''(e_{nc})}{D''(E_{nc})}$, since when $x = 0$, $E_{nc} = NE_{nc}$, the following Proposition summarizes the comparison.[6]

Proposition 2 *The critical size of the coalition $z^{\min}$, at which emissions and abatement choices of signatories and non-signatories are equal, exceeds the corresponding critical size of the coalition, $z^{\min}_{x=0}$, when emissions is the only*

[6]The proof of Proposition 2 is relegated to Appendix II.

choice variable.
The difference between them is increasing as abatement becomes cheaper:
(i) As the marginal cost of abatement becomes very steep, that is, $C'' \to \infty$*, the critical size of the coalition* $z^{\min}$ *tends to* $z^{\min}_{x=0}$*:* $\lim_{C''\to\infty} z^{\min} \to z^{\min}_{x=0}$.
(ii) As the marginal cost of abatement becomes very flat, that is, $C'' \to 0$*, the critical size of the coalition* $z^{\min}$ *tends to* n*:* $\lim_{C''\to 0} z^{\min} \to n$.

To discuss the intuition of the above results, we compare the slopes of nonsignatories' reaction functions in the case that abatement is and in the case it is not an option. As explained above, the slope of the best response of nonsignatories' net emissions is $RNe' = \frac{-(1-\psi)}{(n-s)(1-\psi)-\phi}$. In the case that emissions is the only choice variable, the slope of nonsignatories' best response function is, $Re'|_{x=0} = \frac{\partial e^*_{ns}(E_s)}{\partial E_s} = \frac{-1}{(n-s)-\phi} < 0$.[7] Therefore, since $|RNe'| > |Re'|_{x=0}|$, nonsignatories respond to an increase in the coalition's net emissions by decreasing their own net emissions faster than in the absence of the abatement option. As a result, the coalition's position as a leader is strengthened, which implies that the leadership effect outweighs the damage control effect for larger coalition sizes relative to the case that abatement is not an option. This effect becomes more prominent as the cost of abatement decreases. That is, the smaller is the cost of abatement, the higher is the absolute value of ψ and thus, the higher is the difference $z^{\min} - z^{\min}_{x=0}$.

The critical coalition size $z^{\min}$, determined in Proposition 1, can be used to broadly define the lowest size of a stable coalition and with the help of the comparison of $z^{\min}$ to $z^{\min}_{x=0}$, established in Proposition 2, we can compare the size of the stable coalition between the case of two choice variables and the case of only one choice variable. As mentioned earlier, the literature examined uses the specific two stage game first developed by D 'Aspremont et al. (1983) which uses the notion of internal and external stability. Formally a coalition of size s^* is,

$$\begin{aligned} \text{internally stable if } w_s(s^*) &\geq w_{ns}(s^*-1) \\ \text{and externally stable if } w_{ns}(s^*) &\geq w_s(s^*+1). \end{aligned}$$

[7] See Diamantoudi and Sartzetakis (2015), p. 543.

As established in Proposition 1, there exists a critical value $z^{\min}$, below which signatories' net emissions and welfare are higher relative to nonsignatories, and above which the reverse is true. Nonsignatories' welfare attains its lowest value in the absence of a coalition, the Cournot Nash level w_{nc}. As small size coalitions start forming, their members' net emissions exceed those of nonsignatories, $Ne_s(s) > Ne_{ns}(s)$, and thus, signatories' welfare level exceeds that of nonsignatories, $w_s(s) > w_{ns}(s)$, with $w_{ns}(s)$ dropping below w_{nc}. As the size of coalitions increases, their members' net emissions decrease relieving the pressure from nonsignatories whose welfare starts increasing while that of signatories decreases. At the critical size $z^{\min}$, net emissions of both signatories and nonsignatories, and thus their welfare level, return back to their Cournot Nash levels. For higher coalition sizes nonsignatories' welfare function is always above that of signatories. Although we are able to determine that $w_s(s)$ is monotonically increasing in s after $z^{\min}$, the same is not possible for $w_{ns}(s)$.

If one member exits the coalition of size s, the welfare of nonsignatories becomes $w_{ns}(s-1)$, which is a function with the same properties as $w_{ns}(s)$ shifted by one. That is, the $w_{ns}(s-1)$ function will cut from below the $w_s(s)$ function at a coalition size higher than $z^{\min}$. Adjusting this critical size to the lower integer, yields the size of a stable coalition, denoted by s_1^*, since below that $w_s(s_1^*) \geq w_{ns}(s_1^* - 1)$ and above that $w_{ns}(s_1^*) \geq w_s(s_1^* + 1)$. For general functional forms it is impossible to either define the value of s_1^* or claim its uniqueness: Depending on the slope of $w_s(s)$ around $z^{\min}$, the size of stable coalition s_1^* could be very close of far away from $z^{\min}$. Given that we cannot determine that $w_{ns}(s)$ is monotonically increasing, the $w_{ns}(s-1)$ curve could intersect from below $w_s(s)$ also at higher coalition sizes. This is the reason that in the literature specific functional forms are used to determine the size of stable coalitions; in the next Section we provide an example. However, we can discuss the effect that changes in the benefit, damage and abatement cost have on the lowest size of a stable coalition, utilizing the results of Proposition 2.

It is evident, from the expressions $z^{\min} = \frac{n(1-\psi)-\phi}{1-\psi-\phi}$ and $z^{\min}_{x=0} = \frac{n-\phi}{1-\phi}$, that in the case abatement is not an option, or an extremely expensive one driving ψ to zero, the only determinant of $z^{\min}_{x=0}$ is the $\phi = \frac{B''(e_{nc})}{D''(E_{nc})}$ ratio. When benefits are far larger than damages, $z^{\min}_{x=0}$ becomes smaller since there is no value

gained from an agreement. When damages become more prominent driving ϕ to zero, $z^{\min}_{x=0}$ tends to n: as damages increase, even the grand coalition becomes an option. However, as damages become very high, large coalitions require that their members reduce their emissions drastically to internalize very large externalities over many countries, which in the case that abatement is not a viable option implies that they drastically decrease production and consumption and, depending on the functional forms of benefits and damages, could require drastic reductions in, or even negative, production and consumption. For example, in the case of quadratic benefit and damage functions, Diamantoudi and Sartzetakis (2006) show that, constraining emissions to be positive, the stable coalition's maximum size is 4. Larger coalitions become stable if one ignores the positivity constraint, which have been ignored in some works using abatement as the only choice variable.

When abatement becomes a viable option, separate from emissions, the ratio $\psi = \frac{B''(e_{nc})}{C''(NE_{nc})}$ plays an important role in determining the value of $z^{\min}$, in addition to ϕ. As noticed already, when the rate of the increase in abatement costs is much higher that the rate of the increase in the benefits from production and consumption, that is as $\psi \to 0$, then $z^{\min} \to z^{\min}_{x=0}$. When abatement is very expensive we have the problems described above, restricting the size of stable coalitions. However, when the cost of abatement decreases such that ψ becomes much higher than ϕ, then larger stable coalitions are stable. That is, for given benefits, the less costly abatement is relative to damages, the higher is the stable coalition.

Although the low cost abatement option can lead to large stable coalitions, the welfare gains that such coalitions attain are very small. This is because $z^{\min}$ yields the lowest possible welfare level for both signatories and nonsignatories and thus, s^*_1, regardless of how much higher than $z^{\min}$ is, yields a welfare level close to w_{nc}. Although this has has been termed in the literature as "green paradox" or "paradox of cooperation" (see for example Barret, 1994 and Finus et al., 2021), in light of the above discussion, it is not paradoxical at all. Countries facing large damages from net emissions and having available low cost abatement, engage extensively in abatement, eliminating most, or even all, emissions and enjoy the benefits from production and consumption. However, under such conditions, engaging in high abatement levels is individually rational, that is, a collective agreement at-

tains very little, if any, welfare improvement. When abatement costs are low relative to damages, the difference between the welfare achieved when all countries act as singletons and the welfare the grand coalition achieves becomes very small. The above discussion is summarized in the following Corollary.

Corollary 1 *When emissions and abatement are modeled separately, larger coalitions become stable when the rate of the increase in abatement cost is much less that the rate of the increase in damages. However, large coalitions become stable only when they can offer very little improvement to global welfare.*

When abatement is very cheap countries can increase their production and consumption, since they can eliminate all generated emissions costessly: all countries could join such a coalition that requires no effort and attains the highest possible welfare, which though they can also achieve acting independently.

Before providing a specific example assuming quadratic functions, a note is in order regarding the case that the coalition has no leadership power. In such a case, members of any size coalition choose lower emissions relative to nonsignatories, since they internalize the externalities on all coalition's members and cannot exercise any power over outsiders. Therefore, starting from a welfare level w_{nc} for all countries, as a coalition of two countries forms the welfare of signatories is lower than that of nonsignatories and their difference increases as the size of the coalition increases. Given that in this case both w_s and w_{ns} are monotonically increasing in s, only one stable coalition exists which lies to the right of $z^{\min} = 1$. In the case of quadratic functions the literature, using a single choice variable, has shown that the maximum size of a stable coalition is two. Since there is no leadership effect, nothing changes in the case we model abatement separately from emissions. Therefore, large coalitions can be stable only if the coalition acts as a leader and abatement costs are not very high relative to damages from emissions.

4 Specific example: quadratic functions

Following the literature, we assume quadratic benefits, $B_i(e_i) = b(ae_i - \frac{1}{2}e_i^2)$, with $a > 0$ and $b > 0$ and quadratic damages from aggregate net emissions, $D_i(NE) = \frac{1}{2}c(NE)^2$, with $c > 0$. We further assume that country i's cost of abatement is quadratic, $C_i(x_i) = \frac{1}{2}dx_i^2$, with $d > 0$. Therefore, country i's social welfare is,

$$W_i = b(ae_i - \frac{1}{2}e_i^2) - \frac{1}{2}c(NE)^2 - \frac{1}{2}dx_i^2. \tag{7}$$

Before examining the maximum size of stable, self-enforcing coalitions, we present the two benchmark cases: pure non-cooperation and full cooperation.

In the pure non-cooperative case, we assume that, in the first stage, all countries act individually and no coalition is formed. In the second stage, each country i chooses e_i and x_i in order to maximize its own welfare W_i, given in (7), taking the other countries' emission and abatement levels as given. The first order conditions, $\frac{\partial W_i}{\partial e_i} = 0 \implies ba - be_i = cNE$, and $\frac{\partial W_i}{\partial x_i} = 0 \implies cNE = dx_i$, yield country i's emission and abatement as functions of the rest of the countries' net emissions, $NE_{-i} = \sum_{j \neq i}(e_j - x_j)$,

$$e_i(NE_{-i}) = \frac{(1+\delta)\,a}{1+\delta+\gamma\delta} - \frac{\gamma\delta NE_{-i}}{1+\delta+\gamma\delta}\ , \tag{8}$$

$$x_i(NE_{-i}) = \frac{a}{1+\delta+\gamma\delta} + \frac{NE_{-i}}{1+\delta+\gamma\delta}\ , \tag{9}$$

respectively, where $\gamma = \frac{c}{b}$ and $\delta = \frac{d}{c}$. The parameter γ is the ratio between the constant slope of marginal damages from net emissions and the constant slope of marginal benefits from emissions and δ is the ratio between the constant slope of marginal abatement cost and the constant slope of marginal damages from net emissions. We use a different notation from the one used in the previous Section, to avoid inappropriately confusing this particular example, in which third order derivatives are assumed zero, to the general case examined before. Notice that in this particular example, $B''(e_i) = -b$, $D''(NE) = c$, and $C''(x_i) = d$ at any level of emissions and abatement and thus, we can write $\phi = \frac{-1}{\gamma}$ and $\psi = \frac{-1}{\gamma\delta}$. However, ϕ and ψ take different values depending on the functional forms used and in general are not independent of emissions and abatement.

Country i's best reaction to an increase in the rest of the countries' net emissions is to decrease its own net emissions both by decreasing its emissions, the slope of the emission reaction function is $\frac{\partial e_i(NE_{-i})}{\partial NE_{-i}} = -\frac{\gamma\delta}{1+\delta+\gamma\delta}$, and by increasing its abatement effort, the slope of the abatement reaction function is $\frac{\partial x_i(NE_{-i})}{\partial NE_{-i}} = \frac{1}{1+\delta+\gamma\delta}$. It is interesting to note that the speed of reaction of emission exceeds that of abatement if $\gamma\delta > 1 \Longrightarrow d > b$, that is, when the slope of the marginal cost of abatement exceeds the slope of the marginal benefits from emission: If benefits from emissions are relatively higher, country i adjusts mostly by increasing abatement, while when abatement costs are relatively higher, it adjusts by primarily reducing emission.

Since all countries are symmetric, at the equilibrium they all choose the same level of emissions, denoted by e_{nc}, and abatement, denoted by x_{nc}. The system of reaction functions (8) and (9) yields the non-cooperative level of emission and abatement, $e_{nc} = \frac{(n+\delta)a}{\delta+(1+\gamma\delta)n}$ and $x_{nc} = \frac{na}{\delta+(1+\gamma\delta)n}$. Therefore, each country's net emissions Ne_{nc} are, $Ne_{nc} = e_{nc} - x_{nc} = \frac{\delta a}{\delta+(1+\gamma\delta)n}$.[8] Aggregate emission and abatement levels under the non-cooperative case are, $E_{nc} = ne_{nc} = n\frac{a(n+\delta)}{\delta+(1+\gamma\delta)n}$ and $X_{nc} = nx_{nc} = n\frac{an}{\delta+(1+\gamma\delta)n}$, respectively and therefore net emissions are, $NE_{nc} = E_{nc} - X_{nc} = n\frac{a\delta}{\delta+(1+\gamma\delta)n}$.

In the case of full cooperation, we assume that in the first stage the grand coalition is formed and is stable. In the second stage emission and abatement decisions are taken collectively, that is, countries choose e_i and x_i so as to maximize aggregate welfare, $\sum_{i=1}^{n} W_i$, where W_i is given in (7).

We use the notation defined above and we denote equilibrium values of emission and abatement levels by a subscript c, $e_c = \frac{a(n^2+\delta)}{n^2+\delta+n^2\gamma\delta}$, $x_c = \frac{an^2}{n^2+\delta+n^2\gamma\delta}$. Thus, each country's net emissions Ne_c are, $Ne_c = e_c - x_c = \frac{\delta a}{n^2+\delta+n^2\gamma\delta}$.[9] Aggregate emission and abatement levels under the full cooperation case are, $E_c = ne_c = n\frac{a(n^2+\delta)}{n^2+\delta+n^2\gamma\delta}$ and $X_c = nx_c = n\frac{an^2}{n^2+\delta+n^2\gamma\delta}$, respectively and therefore net emissions are, $NE_c = E_c - X_c = n\frac{a\delta}{n^2+\delta+n^2\gamma\delta}$.

Country i's net emissions are lower in the full cooperation case, i.e. $e_c - x_c < e_{nc} - x_{nc}$. In the full cooperation case, each country emits less ($e_c < e_{nc}$)

[8] In the absence of net environmental damages, $c = 0$, emissions take their highest value, $e_{nc|c=0} = a$, and abatement effort approaches zero, $x_{nc|c=0} = 0$. Also, if abatement becomes costless, $d = 0$, countries choose the highest level of emissions, $e_{nc|d=0} = a$, since they can costlessly abate the total amount of emissions, $x_{nc|d=0} = a$.

[9] In the two extreme cases of $c = 0$ and $d = 0$, emissions and abatement take the same values as in the case of non-cooperation.

and abates more ($x_c > x_{nc}$) relative to the non-cooperative case. That is, aggregate net emissions are lower when all countries cooperate. It can also be shown that aggregate welfare is higher under full cooperation.

4.1 Coalition formation

We now move to examine the stable size of the coalition. Using the notation established in Section 3, aggregate net emissions are, $NE = E - X = s\,(e_s - x_s) + (n - s)\,(e_{ns} - x_{ns})$. Substituting total net emissions of the rest of the countries, $\sum_{j \neq i}(e_j - x_j) = NE_s + (n - s - 1)\,(e_{ns} - x_{ns})$, where $NE_s = s\,(e_s - x_s)$, into the reaction functions (8) and (9), we calculate nonsignatories' reaction functions, $Re = e_{ns}(e_s, x_s)$ and $Rx = x_{ns}(e_s, x_s)$. For brevity, we only report nonsignatories net emissions' reaction function,[10]

$$RNe = Ne_{ns}(NE_s) = \frac{\delta a}{\delta + (1 + \gamma\delta)\,(n - s)} - \frac{(1 + \gamma\delta)\,NE_s}{\delta + (1 + \gamma\delta)\,(n - s)}. \quad (10)$$

It is clear that when all countries act as singletons, $s = 0$, the above collapses to the Nash equilibrium, $Ne_{ns}|_{s=0} = Ne_{nc}$, where Ne_{nc} is defined by subtracting $x_i(NE_{-i})$ (9) from $e_i(NE_{-i})$ given in (8). Furthermore, in the absence of damages, $c = 0$, $RNe|_{c=0} = a$, while when abatement is costless, $d = 0$, $RNe|_{d=0} = 0$. Finally, if abatement was not an option, which is the same as if it was extremely expensive, that is, $d \to \infty$, then $x_{ns} = x_s \to 0$ and $RNe|_{d\to\infty} \to \frac{a}{1+\gamma(n-s)} - \frac{\gamma}{1+\gamma(n-s)} s e_s$, which is exactly the same reaction function as the one reported in Diamantoudi and Sartzetakis (2006).

We can compare the slopes of the net emission reaction functions with and without the option of abatement to verify and extent the discussion in Section 3. It is clear that when technological advancements reduce the cost of abatement substantially, non-signatory countries adjust their net emissions, responding to a change in the coalition's emissions, faster relative to when abatement is not an option. Denote the slope of the reaction function in (10) by $RNe' = \frac{\partial Ne_{ns}(NE_s)}{\partial NE_s} = -\frac{(1+\gamma\delta)}{\delta+(1+\gamma\delta)(n-s)}$ and the slope in the absence of abatement option by $RNe'|_{x_s=0} = \frac{\partial e_{ns}}{\partial E_s} = -\frac{\gamma}{1+\gamma(n-s)}$.[11] Then, $|RNe'| - \left|RNe'|_{x_s=0}\right| = \frac{1}{[1+\gamma(n-s)][\delta+(1+\gamma\delta)(n-s)]} > 0$, which goes to zero as $d \to \infty$ and

[10] The emission and abatement reaction functions will have the same numerators as (8) and (9) respectively and the same denominator as the one in the following expression (10).

[11] See Diamantoudi and Sartzetakis (2006), p. 251.

thus $\delta \to \infty$ and becomes larger the smaller is the slope of the marginal cost of abatement. Thus, nonsignatories adjust faster to the signatories choices as the abatement option becomes cheaper, which implies that the leading coalition can exercise more pressure on singletons when abatement becomes cheaper.

Signatories maximize the coalition's welfare, sW_s, taking explicitly into account nonsignatories' behavior $e_{ns}(e_s, x_s)$ and $x_{ns}(e_s, x_s)$. Given these, aggregate net emissions depend only on signatories' choices, $NE(e_s, x_s) = s(e_s - x_s) + (n-s)(e_{ns}(e_s, x_s) - x_{ns}(e_s, x_s))$. That is, signatories choose e_s and x_s in order to maximize collective welfare,

$$\max_{e_s,\, x_s} \sum W_s = s\left[B_s(e_s) - D_s\left(NE(e_s, x_s)\right) - C_s(x_s)\right].$$

The first-order conditions of the above maximization problem yield signatories emission and abatement effort levels,

$$e_s = a\left(1 - ns\frac{\gamma\delta^2}{\Psi}\right), \tag{11}$$

$$x_s = ans\frac{\delta}{\Psi}, \tag{12}$$

where $\Psi = \Omega^2 + s^2\delta(1+\gamma\delta) > 0$ and $\Omega = \delta + (n-s)(1+\gamma\delta) > 0$. Note that, in the absence of environmental damages, $c = 0$, emissions take their highest value $e_s|_{c=0} = a$, and abatement effort is zero, $x_s|_{c=0} = 0$. Therefore, the signatories' net emissions are,

$$Ne_s = e_s - x_s = a\left(1 - ns\frac{\delta(1+\gamma\delta)}{\Psi}\right). \tag{13}$$

Thus, the coalition's total net emissions are, $NE_s = E_s - X_s = sa\left(1 - ns\frac{\delta(1+\gamma\delta)}{\Psi}\right)$.

Substituting e_s and x_s into the non-signatories' reaction functions, we derive the non-signatories' emission and abatement level,

$$e_{ns} = a\left(1 - n\frac{\gamma\delta\Omega}{\Psi}\right) = e_s + \frac{an\gamma\delta(\delta s - \Omega)}{\Psi}, \tag{14}$$

$$x_{ns} = \frac{an\Omega}{\Psi} = x_s - \frac{an(\delta s - \Omega)}{\Psi}. \tag{15}$$

From (14) and (15), the net emission level of the non-signatories is,

$$Ne_{ns} = e_{ns} - x_{ns} = a\left(1 - n\frac{\Omega(1+\gamma\delta)}{\Psi}\right) = Ne_s + \frac{an(1+\gamma\delta)(\delta s - \Omega)}{\Psi}. \tag{16}$$

The above imply that $e_{ns} \lesseqgtr e_s \Leftrightarrow \delta s \lesseqgtr \Omega$, $x_{ns} \gtreqless x_s \Leftrightarrow \delta s \lesseqgtr \Omega$ and $Ne_{ns} \lesseqgtr Ne_s \Leftrightarrow \delta s \lesseqgtr \Omega$. We will explore this condition further when we compare the welfare of signatories to that of non-signatories. At this point note that since for small coalition sizes $\Omega > \delta s$ regardless of the value of the parameters, when coalition size is small non-signatories emit less and abate more than the signatories. For small coalition sizes, the leadership effect dominates the damage reduction effect.

Nonsignatories' aggregate net emissions are, $NE_{ns} = E_{ns} - X_{ns} = (n - s)\left[Ne_s + \frac{an(1+\gamma\delta)(\delta s-\Omega)}{\Psi}\right]$. Therefore, global net emissions $NE = E - X = (E_{ns} - X_{ns}) + (E_s - X_s)$ are, $NE = E - X = \sum_{i=1}^{n}(e_i - x_i) = an\frac{\delta\Omega}{\Psi}$.

Unlike the non-cooperative and the full cooperation cases, in which the level of emission are strictly positive, $e_{nc} > 0$ and $e_c > 0$, in the coalition formation case we have to restrict the parameters of the model in order to guarantee interior solutions. Therefore, we need to restrict the parameters so that both are positive, given that emissions correspond to production and consumption which cannot be negative. The following Lemma establishes the necessary conditions for interior solutions. The proof is presented to Appendix III.

Lemma 1 *In the case that abatement is not available, that is, $\delta \to \infty$, $e_s > 0$ and $e_{ns} > 0$ if and only if $0 < \gamma < \frac{4}{n(n-4)}$ and $n > 4$. As abatement becomes relatively inexpensive, $e_s > 0$ and $e_{ns} > 0$ hold for higher values of γ, which are increasing as δ decreases.*

The intuition of the above result is clear: as abatement becomes available at low cost, countries both in and outside the coalition are able to decrease their net emissions by engaging in abatement keeping their emissions and thus their production positive, even when damages from emissions are relatively high. This will prove very important for the determination of the maximum size of the stable coalition that follows.

Substituting the equilibrium values of the choice variables from (11), (12), (14) and (15) we derive the indirect welfare function of signatories (w_s) and

non-signatories (w_{ns}),

$$w_s = \frac{ba^2}{2}\left(1 - \frac{\gamma n^2 \delta^2}{\Psi}\right), \tag{17}$$

$$w_{ns} = \frac{ba^2}{2}\left[w_s - \frac{\gamma\delta n^2\left(\delta^2 s^2 - \Omega^2\right)(1+\delta\gamma)}{\Psi^2}\right]. \tag{18}$$

Utilizing the above results, Proposition 3 establishes the properties of these indirect welfare functions, applying the results in Proposition 1 to the specific example of quadratic functions. The proof of Proposition 3 follows the same steps as in the general case and applied for the case of quadratic functions and a single choice variable in Diamantoudi and Sartzetakis (2006).[12,13]

Proposition 3 *We consider the indirect welfare function of signatories and non-signatories,* (w_s) *and* (w_{ns}) *respectively. If we define* $s^{\min} = \frac{\delta+(1+\delta\gamma)n}{1+\delta+\delta\gamma}$, *then,*
(i) $s^{\min} = \arg\min_{s\in\Re\cap[0,n]} w_s(s)$, *that is,* $s^{\min}$ *is the* s *at which* w_s *is minimized,*
(ii) $w_s(s)$ *increases in* s *if* $s > s^{\min}$ *and it decreases in* s *if* $s < s^{\min}$,
(iii) $w_{ns}(s) \lesseqgtr w_s(s)$ *for all* $s \lesseqgtr s^{\min}$.

The above defined properties of the indirect welfare function imply that the indirect welfare function of the non-signatories cuts the indirect welfare function of the signatories from below at its minimum, defined by $s^{\min}$. Note that $s^{\min}$ solves $\Omega = \delta s$, that is, the welfare of the signatories takes its minimum value at the coalition size that equalizes the emission and abatement of signatories and non-signatories.

We know from Section 3 that there are two effects that make coalition members' choices different from those of outsiders: the effect of internalizing damages of all members and the leadership effect which work in opposite directions. At the critical size of coalition $s^{\min}$ these two effects offset each other and thus, signatories and nonsignatories choices are the same as in the absence of a coalition. The damage internalization effect is $(s-1)$ and, in

[12] The proof is available to the interested reader upon request.

[13] Notice that if we substitute the values $\phi = \frac{-1}{\gamma}$ and $\psi = \frac{-1}{\gamma\delta}$, into the definition of $z^{\min}$ defined in Proposition 1, yields the definition of $s^{\min}$.

this particular example, the leadership effect is $\frac{-s(n-s)(1+\gamma\delta)}{(n-s)(1+\gamma\delta)+\delta}$. The sum of these two effects is $\frac{\delta s-\Omega}{(n-s)(1+\gamma\delta)+\delta}$, which becomes zero for $\Omega = \delta s$.

In the case that countries choose only their emission level, Diamantoudi and Sartzetakis (2006) find that the critical coalition size at which $w_{ns}(s) = w_s(s)$, is $s^{\min}_{x=0} = \frac{1+\gamma n}{1+\gamma}$, which is clearly increasing in γ, for $n > 1$. Furthermore, Diamantoudi and Sartzetakis (2006) associate the size of the stable coalition to the integer closer to $s^{\min}_{x=0}$ and after restricting the admissible values of γ in order to have positive emissions, the size of stable coalitions is limited to $s^*_{x=0} \in \{2, 3, 4\}$. In the present case, it is clear that when abatement costs are very low, $\lim_{\delta\to 0} z^{\min} \to n$ which is the specification of the general result presented in Proposition 2 to the quadratic case. The case that abatement cost are very high, $\delta \to \infty$ corresponds to the case that abatement is not an option and we go back to the results obtained in Diamantoudi and Sartzetakis (2006).

The following Corollary compares the case in which emission is the only choice variable with the case we develop in the present paper.

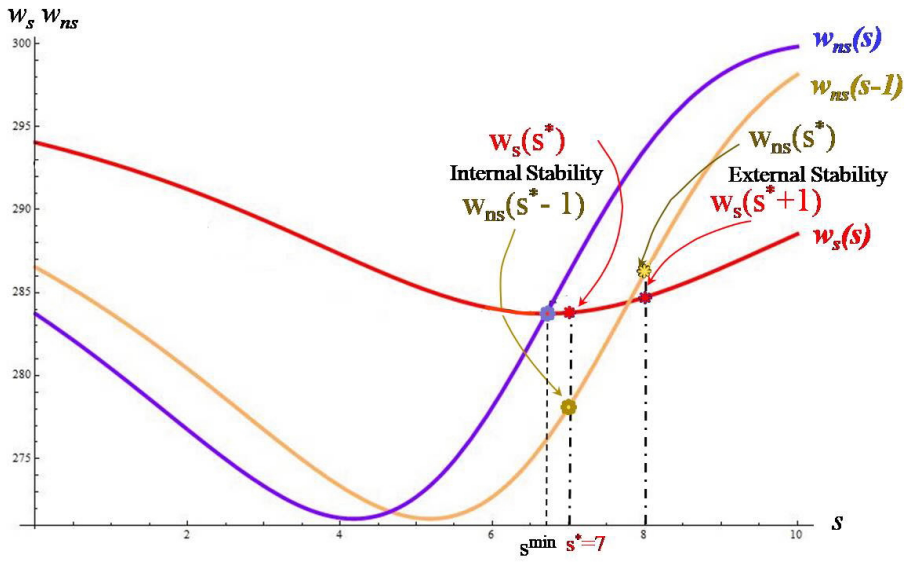
Corollary 2 *Allowing countries to choose abatement separately from emission level increases the size of the stable coalition for any admissible values of benefits damages and cost parameters. This potential enlargement of the coalition size is increasing as the cost of abatement decreases.*

Direct comparison of the two cases reveals that $s^{\min} > s^{\min}_{x=0}$, for $\delta > 0$ and $n > 1$. Furthermore, $\frac{\partial s^{\min}}{\partial \delta} = \frac{1-n}{(1+\delta+\delta\gamma)^2} < 0$, that is, recalling that $\delta = \frac{d}{c}$, as either the cost of abatement decreases or environmental damages increase, the higher is the size of stable coalition. In what follows we illustrate the above results by considering numerical examples.

4.2 Illustration of the results using simulations

To facilitate the comparison to the case that abatement is not a separate choice variable, we choose the same parameter values used in Diamantoudi and Sartzetakis (2006). That is, we assume the following values for the parameters: $n = 10$, $a = 10$, $b = 6$, and $c = 0.39999$, which results in $\gamma = 0.066665$. These values satisfy the restrictions set in *Lemma 1*, since the parameter γ is less than $\gamma < \frac{4}{n(n-4)} < 0.066667$. We also choose a small value

1



1.jpg

Figure 1: Defining the size of a stable IEA

for the parameter δ, such that $c > d$, which means that marginal damage increases faster than marginal abatement cost. We choose $d = 0.239994$, so that given $c = 0.39999$, we have $\delta = 0.6$.

In Figure 1, we plot, using the above defined parameter values, the indirect welfare functions against different coalition sizes s. The red curve depicts $w_s(s)$, the purple curve $w_{ns}(s)$ and the orange curve $w_{ns}(\ s-1)$. Notice that $w_{ns}(s-1)$ is a horizontal shift of $w_{ns}(s)$.

As Figure 1 illustrates, no country wants to exit or join a coalition of seven, that is, a coalition of seven is stable, $s^* = 7$. The internal condition $w_s(s^*) \geq w_{ns}(s^*-1)$ is satisfied since $w_s(7) > w_{ns}(6)$, as the $w_s(s)$ curve is above the $w_{ns}(s-1)$ curve. Moreover, coalition $s^* = 7$ is externally stable i.e. $w_s(s^*+1) \leq w_{ns}(s^*)$ since at $s = s^*+1 = 8$ the $w_{ns}(s-1)$ curve is above the $w_s(s)$ curve. Therefore, the coalition of size $s^* = 7$ is stable.

Similar to the analysis in Diamantoudi and Sartzetakis (2006) the stable coalition size is the higher integer following the size of the coalition for which the welfare of the signatories is at its minimum, $s^{\min}$ and for which $w_s = w_{ns}$. In our example, $s^{\min} = 6.70732$.

The above specified parameter values satisfy the constraints for $e_s > 0$, $e_{ns} > 0$, $e_s - x_s > 0$ and $e_{ns} - x_{ns} > 0$ and also the internal and external

stability conditions. Note that for the same parameter values, the optimal size of the coalition in Diamantoudi and Sartzetakis (2006) is $s^* = 3$. The following Table summarizes the results of the simulations. The first column reports the equilibrium values of emissions, abatement, net emissions and welfare for signatories, the second column for non-signatories and the last column reports the aggregate values.

	$s^* = 7$	$n - s^*$	n
e	9.622	9.665	96.347
x	9.456	8.376	91.322
$e - x$	0.165	1.289	5.025
w	283.8	286.2	2,845.2

Table 1. Coalition formation with abatement for $\delta = 0.6$

These values confirm that signatories emit less and abate more than non-signatories, i.e. $e_s < e_{ns}$ and $x_s > x_{ns}$. Moreover, the net emissions are significantly smaller for the signatories, $e_s - x_s$, than for the non-signatories, $e_{ns} - x_{ns}$. Total net emissions, $\sum_{i=1}^{n}(e_i - x_i)$, include the activities from both signatories and non-signatories.

Total net emissions are slightly smaller relative to the non-cooperative case, $NE_{nc} = 5.4545$, but much larger than the full cooperation case, $NE_c = 0.5736$. The welfare level each of the seven members of the coalition attains is slightly over the one they attain at the non-cooperative case $w_{nc} = 283.736$ and thus the total welfare improvement attained by the coalition is almost negligible and derives from the gains of free-riders.

4.2.1 Comparative statics with respect to abatement cost

In the simulations presented above we used a value for the parameter δ that is below 1, indicating that the abatement cost parameter d is smaller than the emission damage parameter c. In particular we have used the value $\delta = 0.6$. In order to show the effect of changes in the abatement cost which change the ratio δ, we simulate the model for two extreme values of δ, namely, $\delta = 0.0000001$ and $\delta = 1.637$. The first case indicates that abatement cost is negligible relative to environmental damages, while in the second we assume that abatement costs exceed environmental damages. Table 2 presents the

results of the simulations: in the left the case with significantly high abatement cost is presented, which results in $s^* = 5$ and in the right the case with almost zero abatement cost is presented, which yields the grand coalition, $s^* = 10$.

	$\delta = 1.637$			$\delta = 1 * 10^{-7}$		
	$s^* = 5$	$n - s^*$	n	$s^* = 10$	$n - s^*$	n
e	9.079	9.192	91.354	10	–	100
x	8.439	7.406	79.229	10	–	100
$e - x$	0.639	1.785	12.124	$1 * 10^{-8}$	–	$1 * 10^{-7}$
w	244.7	250.7	2,476.0	300.0	–	3,000.0

Table 2. Coalition formation with abatement under different values of δ

The results of the simulations presented in Tables 1 and 2 reveal that the value of the parameter δ is crucial in determining the size of the stable IEA. When $\delta = 0.6$, the size of the stable coalition is $s^* = 7$. When δ takes a very low value, i.e. $\delta = 0.0000001$, the number of signatory countries increases, reaching the grand coalition, $s^* = 10$. When abatement costs are negligible all countries generate the maximum emissions which they completely abate, $\lim_{\delta \to 0} e = \lim_{\delta \to 0} x = a = 10$, to attain the maximum benefits. On the contrary, when δ takes a very high value, i.e. $\delta = 1.637$, the number of signatory countries decreases to $s^* = 5$. Notice though that even when the abatement cost parameter takes the highest value allowed by the model's constraints, presented in Lemma 1, the size of the stable coalition is higher than the case in which countries have only one choice variable.

When δ approaches zero and the the grand coalition emerges, countries' net emissions are the lowest possible achieving the highest welfare. Notice though that as abatement becomes very inexpensive, that is as δ approaches zero, we also have that, $\lim_{\delta \to 0} e_{nc} = \lim_{\delta \to 0} x_{nc} \to a = 10$ and thus, $\lim_{\delta \to 0} w_{nc} \to 300$. That is, countries choices are the same whether they cooperate or not. When the abatement cost becomes negligible, indeed the grand coalition is stable, but it delivers no additional welfare gains relative to the case that countries decide independently. The following Corollary restates the general result presented in Corollary 1.

Corollary 3 *When the coalition acts as a leader and countries choose emissions and abatement separately, larger coalitions can be stable as abatement*

costs decrease, including the grand coalition when costs are negligible. However, very small welfare gains over the non-cooperative equilibrium are attained.

When technology drives abatement cost to zero, it is individually rational for the countries to abate all their emissions and thus they emit the highest possible amounts to attain the highest welfare from production and consumption. As the abatement costs increase, free-riding incentives become significant and the size of stable coalitions decreases, but remains larger compared to the case that abatement is not an option. All stable coalitions cases though yield equilibrium values for signatory and non-signatory countries' welfare that is the same as that they receive acting non-cooperatively. Modelling abatement separately from emissions allows for higher stable coalitions but this is because free-riding incentives are reduced which implies that the larger coalitions yield no welfare gains over the non-cooperative case: The option of abatement enables the coalition to take advantage of its leadership role for larger coalition sizes.

5 Conclusions

The present paper examines the size of stable IEAs employing a non-cooperative leadership framework and assuming that countries choose emission and abatement levels separately. We assume benefits are concave in the country's own emissions; environmental damages are convex in aggregate net emissions and convex in the country's abatement effort. Coalition formation is modelled as a two stage game: in the second stage, countries choose their levels of emission and abatement and in the first stage counties choose whether or not to join a coalition. Within this framework we find using general functional forms that the size of the stable coalition is always larger than in the case in which countries can choose only their emissions level. As the cost of abatement decreases free-riding incentives are reduced since countries can eliminate emissions almost costessly and thus, larger coalitions become stable. However, the same benefits from reduced abatement costs can be reaped by countries acting independently. Regardless of the cost of abatement, coalitions acting as leaders can be stable at larger sizes relative to when

they do not have such advantage, but they achieve this by keeping emissions, abatement and thus welfare to their non-cooperative levels.

6 References

D' Aspremont, C.A., Jacquemin, J. Gabszewicz, J., and Weymark, J.A. (1983). "On the stability of collusive price leadership." *Canadian Journal of Economics*, **16**, 17-25.

Barrett, S. (1994). "Self-enforcing international environmental agreements." *Oxford Economic Papers*, **46**, 878-894.

Barrett, S. (2006) "Climate treaties and breakthrough technologies." *American Economic Review,* **96**, 22–25.

Bayramoglu, B., M. Finus, and J.-F. Jacques (2018). "Climate Agreements in a Mitigation-Adaptation Game." *Journal of Public Economics*, **165**, 101-113.

Benchekroun, H., W. Marrouch, and A. Ray Chaudhuri (2017). "Adaptation Technology and Free-Riding Incentives in International Environmental Agreements." In *Economics of International Environmental Agreements*, edited by M. Ozgur Kayalica, Selim Cagatay, and Hakan Mihci, Chapter 11, Routledge.

Benchekroun H. and Chaudhuri A.R., (2012). "Cleaner technologies and the stability of international environmental agreements." CentER Discussion Paper, 2012-051.

Breton, M. and L. Sbragia (2017). "Adaptation to Climate Change: Commitment and Timing Issues." *Environmental and Resource Economics*, **68**, 975-995.

Carraro, C. and Siniscalco, D. (1993). "Strategies for the international protection of the environment." *Journal of Public Economics*, **52**, 309-328.

de Zeeuw A. (2008) "Dynamic effects on the stability of international environmental agreements." *Journal of Environmental Economics and Management*, **55**(2), 163–174.

Diamantoudi, E. and Sartzetakis, E. (2006). "Stable International Environmental Agreements: An Analytical Approach." *Journal of Public Economic Theory **8**, 247–263.*

Diamantoudi, E. and E. Sartzetakis (2015) "International environmental agreements: coordinated action under foresight." *Economic Theory*, **59**(3), 527-546.

Diamantoudi, E. and E. Sartzetakis (2018) ""International Environmental Agreements-The Role of Foresight." *Environmental & Resource Economics*, **71**(1), 241-257.

Finus, M. and Rundshagen B. (2001), "Endogenous coalition formation global pollution control." *working paper, FEEM,* Nota di Lavoro 43.2001.

Finus, M., F. Furini and A. V. Rohrer (2021) "The efficacy of international environmental agreements when adaptation matters: Nash-Cournot vs Stackelberg leadership." *Journal of Environmental Economics and Management*, **109**, doi.org/10.1016/j.jeem.2021.102461.

Hoel, M. and De Zeeuw A., (2010). "Can a Focus on Breakthrough Technologies Improve the Performance of International Environmental Agreements?" *Environmental and Resource Economics,* **47**(3), 395-406.

Karp, L., Simon, L. (2013) "Participation games and international environmental agreements: a non-parametric model." *Journal of Environmental Economics and Management*, **65** (2), 326–344.

McGinty, M. (2020) "Leadership and Free-Riding: Decomposing and Explaining the Paradox of Cooperation in International Environmental Agreements." *Environmental and Resource Economics*, **77**, 449-474.

Osmani, D. and R. Tol (2009) "Toward farsightedly stable international environmental agreements." *Journal of Public Economic Theory*, **11**, 455–492

Rubio, J. S. and Casino, B. (2001), "International Cooperation in Pollution Control" *mimeo.*

Rubio J.S. (2018). "Self-enforcing international environmental agreements: adaptation and complementarity." FEEM Working Paper 0 29.2018.

Ulph, A & Rubio, S (2006) "Self-enforcing Environmental Agreements Revisited", *Oxford Economic Papers*, **58(2)**, 233-263.

7 Appendix I: Proof of Proposition 1

Although in our model s is a non-negative integer smaller than n, for the ease of exposition and calculations in the following proof we use z to denote a real number taking values from $[0, n]$. At the end we convert back to integer

s.

[**Parts 1-2**] For a coalition size z, we denote the aggregate net emissions level at the equilibrium by $NE^*(z) = s\left(e_s^*(z) - x_s^*(z)\right) + (n-s)\left(e_{ns}^*(z) - x_{ns}^*(z)\right)$, where $e_{ns}^*(z) = e_{ns}^*(e_s^*(z), x_s^*(z), z)$ and $x_{ns}^*(z) = x_{ns}^*(e_s^*(z), x_s^*(z), z)$. Given the assumptions regarding countries' benefit function, $B'' < 0$ we have $e_s^*(z) \gtreqless e_{ns}^*(z) \Leftrightarrow B'\left(e_s^*(z)\right) \lesseqgtr B'\left(e_{ns}^*(z)\right)$. In equilibrium we also have,

$$\begin{aligned} B'(e_s^*(z)) &\equiv D'(NE^*(z)) \left.\frac{\partial NE^*(z)}{\partial e_s}\right|_{e_s=e_s^*(z)} \\ &\text{and} \\ B'(e_{ns}^*(z)) &\equiv D'(NE^*(z)) \left.\frac{\partial NE^*(z)}{\partial e_{ns}}\right|_{e_{ns}=e_{ns}^*(z)}. \end{aligned}$$

where the derivative of aggregate net emissions around the equilibrium value $e_s^*(s)$ is: $\left.\frac{\partial NE^*(z)}{\partial e_s}\right|_{e_s=e_s^*(z)} = z + (n-z)\left.\frac{\partial e_{ns}^*(e_s,x_s))}{\partial e_s}\right|_{e_s=e_s^*(z)} - (n-z)\left.\frac{\partial x_{ns}^*(e_s,x_s))}{\partial e_s}\right|_{e_s=e_s^*(z)}$.

Thus, recalling that $D'(E^*(z)) > 0$ and $\left.\frac{\partial NE^*(z)}{\partial e_{ns}}\right|_{e_{ns}=e_{ns}^*(z)} = 1$, in equilibrium we have,

$$\begin{aligned} e_s^*(z) &\gtreqless e_{ns}^*(z) \Leftrightarrow \\ \left.\frac{\partial NE^*(z)}{\partial e_s}\right|_{e_s=e_s^*(z)} &\lesseqgtr 1. \end{aligned} \tag{A1}$$

Furthermore, given the assumption regarding countries' abatement cost function, $C'' > 0$ we have that $x_s^*(z) \gtreqless x_{ns}^*(z) \Leftrightarrow C'\left(x_s^*(z)\right) \gtreqless C'\left(x_{ns}^*(z)\right)$. In equilibrium we also have,

$$\begin{aligned} C'(x_s^*(z)) &\equiv D'(NE^*(z)) \left.\frac{\partial NE^*(z)}{\partial x_s}\right|_{x_s=x_s^*(z)} \\ &\text{and} \\ C'(e_{ns}^*(z)) &\equiv D'(NE^*(z)) \left.\frac{\partial NE^*(z)}{\partial x_{ns}}\right|_{x_{ns}=x_{ns}^*(z)}. \end{aligned}$$

where the derivative of aggregate net emissions around the equilibrium value $x_s^*(s)$ is: $\left.\frac{\partial NE^*(z)}{\partial x_s}\right|_{x_s=x_s^*(z)} = -z + (n-z)\left.\frac{\partial e_{ns}^*(e_s,x_s))}{\partial x_s}\right|_{x_s=x_s^*(s)} - (n-z)\left.\frac{\partial x_{ns}^*(e_s,x_s))}{\partial x_s}\right|_{x_s=x_s^*(s)}$.
Thus, recalling again that $D'(E^*(z)) > 0$ and $\left.\frac{\partial NE^*(z)}{\partial x_{ns}}\right|_{x_{ns}=x_{ns}^*(z)} = -1$, in

equilibrium we have,

$$\begin{aligned} x_s^*(z) &\gtreqless x_{ns}^*(z) \Leftrightarrow \\ \left.\frac{\partial NE^*(z)}{\partial x_s}\right|_{x_s=x_s^*(z)} &\gtreqless -1. \end{aligned} \tag{A2}$$

The first order condition of the non-signatories with respect to their emissions, yield the identity $B'(e_{ns}^*(e_s)) \equiv D'[z\,(e_s - x_s) + (n-z)(e_{ns}^*(e_s, x_s) - x_{ns}^*(e_s, x_s))]$. Differentiating both sides of this identity with respect to e_s yields,

$$\frac{\partial e_{ns}^*(e_s, x_s)}{\partial e_s} = \frac{zD''(NE(e_s)) - (n-z)D''(NE(e_s))\frac{\partial x_{ns}^*(e_s, x_s)}{\partial e_s}}{B''(e_{ns}^*(e_s)) - (n-z)D''(NE(e_s))}\ . \tag{A3}$$

Similarly, the first order condition of nonsignatories with respect to their abatement, yield the identity $C'(x_{ns}^*(e_s)) \equiv D'[z\,(e_s - x_s) + (n-z)(e_{ns}^*(e_s, x_s) - x_{ns}^*(e_s, x_s))]$. Differentiating both sides of this identity with respect to e_s yields,

$$\frac{\partial x_{ns}^*(e_s, x_s)}{\partial e_s} = \frac{zD''(NE(e_s)) + (n-z)D''(NE(e_s))\frac{\partial e_{ns}^*(e_s, x_s)}{\partial e_s}}{C''(x_{ns}^*(e_s)) + (n-z)D''(NE(e_s))}\ . \tag{A4}$$

The solution of the system of equations (A3) and (A4) yields,

$$\left.\frac{\partial e_{ns}^*(e_s, x_s)}{\partial e_s}\right|_{e_s=e_s^*(z)} = \frac{-z}{(n-z)\,(1-\psi) - \phi}\ , \tag{A5}$$

and

$$\left.\frac{\partial x_{ns}^*(e_s, x_s)}{\partial e_s}\right|_{e_s=e_s^*(z)} = \frac{-z\psi}{(n-z)\,(1-\psi) - \phi}\ . \tag{A6}$$

where $\phi = \frac{B''(e_{ns}^*(E_s, X_s))}{D''(NE)} < 0$ and $\psi = \frac{B''(e_{ns}^*(E_s, X_s))}{C''(x_{ns}^*(E_s, X_s))} < 0$, given $B'' < 0$, $D'' > 0$, and $C'' > 0$. From (A5) and (A6) it is clear that the slope of nonsignatories' emission reaction with respect to a change in signatories' emissions is negative around the equilibrium, $\frac{\partial e_{ns}^*(e_s, x_s)}{\partial e_s} < 0$, while the slope of the abatement reaction function is positive $\frac{\partial x_{ns}^*(e_s, x_s)}{\partial e_s} > 0$.

Substituting (A5) and (A6) into the inequality (A1) yields,

$$z - \frac{z\,(n-z)\,(1-\psi)}{(n-z)\,(1-\psi) - \phi} \lesseqgtr 1.$$

Which reduces to

$$\begin{aligned} e_s^*(z) &\gtreqless e_{ns}^*(z) \Leftrightarrow \\ z &\lesseqgtr \frac{n\left(1-\psi\right)-\phi}{\left(1-\psi\right)-\phi}. \end{aligned} \tag{A7}$$

In a similar manner as above, differentiating with respect to x_s the first order conditions of the non-signatories with respect to their emissions and abatement yields,

$$\frac{\partial e_{ns}^*(e_s,x_s)}{\partial x_s} = \frac{-zD''(NE(e_s)) - (n-z)D''(NE(e_s))\frac{\partial x_{ns}^*(e_s,x_s)}{\partial x e_s}}{B''(e_{ns}^*(e_s)) - (n-z)D''(NE(e_s))} . \tag{A8}$$

$$\frac{\partial x_{ns}^*(e_s,x_s)}{\partial x_s} = \frac{-zD''(NE(e_s)) + (n-z)D''(NE(e_s))\frac{\partial e_{ns}^*(e_s,x_s)}{\partial x_s}}{C''(x_{ns}^*(e_s)) + (n-z)D''(NE(e_s))} . \tag{A9}$$

The solution of the system of equations (A8) and (A9) yields,

$$\left.\frac{\partial e_{ns}^*(e_s,x_s)}{\partial x_s}\right|_{x_s=x_s^*(z)} = \frac{z}{\left(n-z\right)\left(1-\psi\right)-\phi}, \tag{A10}$$

and

$$\left.\frac{\partial x_{ns}^*(e_s,x_s)}{\partial x_s}\right|_{x_s=x_s^*(z)} = \frac{z\psi}{\left(n-z\right)\left(1-\psi\right)-\phi}. \tag{A11}$$

Nonsignatories react to an increase in the signatories' abatement by increasing their emissions, $\frac{\partial e_{ns}^*(e_s,x_s)}{\partial x_s} > 0$, and decreasing their abatement $\frac{\partial x_{ns}^*(e_s,x_s)}{\partial x_s} < 0$.

Substituting (A10) and (A11) into the inequality (A2) yields,

$$-z + \frac{\left(n-z\right)z\left(1-\psi\right)}{\left(n-z\right)\left(1-\psi\right)-\phi} \gtreqless -1.$$

Which reduces to

$$\begin{aligned} x_s^*(z) &\lesseqgtr x_{ns}^*(z) \Leftrightarrow \\ z &\lesseqgtr \frac{n\left(1-\psi\right)-\phi}{\left(1-\psi\right)-\phi}. \end{aligned} \tag{A12}$$

As expected, working either from the comparison of signatories to nonsignatories emissions $e_s^*(z) \gtreqless e_{ns}^*(z)$, or from the comparison of their abatement $x_s^*(z) \gtreqless x_{ns}^*(z)$, we get the same condition presented in (A7) and (A12).

Observe that when $e_s^*(z) = e_{ns}^*(z)$ the non-signatories' first order conditions remains satisfied, i.e.,

$$\begin{aligned} B'(e_{ns}^*(z)) &\equiv D'\left[z\left(e_s - x_s\right) + (n-z)\left(e_{ns}^*(e_s, x_s) - x_{ns}^*(e_s, x_s)\right)\right] \Leftrightarrow \\ B'(e_{ns}^*(z)) &\equiv D'(ne_{ns}^*(z)), \\ C'(x_{ns}^*(e_s)) &\equiv D'\left[z\left(e_s - x_s\right) + (n-z)\left(e_{ns}^*(e_s, x_s) - x_{ns}^*(e_s, x_s)\right)\right] \Leftrightarrow \\ C'(x_{ns}^*(z)) &\equiv D'(ne_{ns}^*(z)). \end{aligned}$$

These are identical to the first order condition of the pure non-cooperative case where countries compete a la Cournot, hence, $e_{ns}^*(z) = e_s^*(z) = e_{nc}$. Note that due to the strict concavity of the benefit function and the strict convexity of both the damage and the abatement cost functions there exists a unique set (e_{nc}, x_{nc}) and, thus, a unique $z^{\min} = \frac{n(1-\psi)-\phi}{(1-\psi)-\phi}$. Reverting the coalition size back to integers yields:

$$\begin{aligned} e_s^*(s) &\gtreqless e_{ns}^*(s) \Leftrightarrow s \lesseqgtr z^{\min}, \\ x_s^*(z) &\lesseqgtr x_{ns}^*(z) \Leftrightarrow s \lesseqgtr z^{\min}. \end{aligned}$$

[Parts 3-4] Since $w_s(e_s^*(z)) \equiv B\left(e_s^*(z)\right) - D\left(NE^*(z)\right) - C\left(x_s^*(z)\right)$ we have

$$\frac{dw_s(z)}{dz} = B'(e_s^*(z))\frac{de_s^*(z)}{dz} - D'(E^*(z))\frac{\partial NE^*(z)}{\partial z} - C'\left(x_s^*(z)\right)\frac{dx_s^*(z)}{dz} \quad \text{(A13)}$$

where, $\frac{\partial NE^*(z)}{\partial z} = e_s^*(z) - x_s^*(z) - (e_{ns}^*(e_s) - x_{ns}^*(z)) + \left(\frac{de_s^*(z)}{dz} - \frac{dx_s^*(z)}{dz}\right) z + (n - z)\left(\frac{de_{ns}^*(z)}{dz} - \frac{dx_{ns}^*(z)}{dz}\right)$. Denoting by $\Delta NE^* = e_s^*(z) - e_{ns}^*(e_s) - (x_s^*(z) - x_{ns}^*(z))$, (A13) can be rewritten as follows:

$$\begin{aligned} \frac{dw_s(z)}{dz} &= \frac{de_s^*(z)}{dz}\left[B'(e_s^*(z)) - zD'(NE^*(z))\right] \\ &\quad - \frac{dx_s^*(z)}{dz}\left[C'(e_s^*(z)) + zD'(NE^*(z))\right] - D'(E^*(z))\Delta NE^* \\ &\quad - D'(E^*(z))(n-z)\left[\frac{de_{ns}^*(z)}{dz} - \frac{dx_{ns}^*(z)}{dz}\right]. \end{aligned} \quad \text{(A14)}$$

From signatories' first order conditions, we know that in equilibrium,

$$\begin{aligned} & B'(e_s^*(z)) - zD'(E^*(z)) \\ \equiv\ & D'(E^*(z))(n-z)\left[\left.\frac{\partial e_{ns}^*(e_s)}{\partial e_s}\right|_{e_s=e_s^*(z)} - \left.\frac{\partial x_{ns}^*(e_s)}{\partial e_s}\right|_{e_s=e_s^*(z)}\right], \\ & C'(e_s^*(z)) + zD'(NE^*(z)) \\ \equiv\ & D'(E^*(z))(n-z)\left[\left.\frac{\partial e_{ns}^*(e_s)}{\partial x_s}\right|_{x_s=x_s^*(z)} - \left.\frac{\partial x_{ns}^*(e_s)}{\partial x_s}\right|_{x_s=x_s^*(z)}\right]. \end{aligned}$$

Furthermore, (A5) and (A6) yield $\left.\frac{\partial e_{ns}^*(e_s)}{\partial e_s}\right|_{e_s=e_s^*(z)} - \left.\frac{\partial x_{ns}^*(e_s)}{\partial e_s}\right|_{e_s=e_s^*(z)} = \frac{-z(1-\psi)}{(n-z)(1-\psi)-\phi}$ while (A8) and (A9) yield $\left.\frac{\partial e_{ns}^*(e_s)}{\partial x_s}\right|_{x_s=x_s^*(z)} - \left.\frac{\partial x_{ns}^*(e_s)}{\partial x_s}\right|_{x_s=x_s^*(z)} = \frac{z(1-\psi)}{(n-z)(1-\psi)-\phi}$. Therefore, the first two terms in (A14) can be written as,

$$\begin{aligned} & \frac{de_s^*(z)}{dz}\left[B'(e_s^*(z)) - zD'(NE^*(z))\right] \\ & -\frac{dx_s^*(z)}{dz}\left[C'(e_s^*(z)) + zD'(NE^*(z))\right] \\ =\ & D'(E^*(z))(n-z)\frac{z\,(1-\psi)}{(n-z)\,(1-\psi)-\phi}\left(\frac{de_s^*(z)}{dz} - \frac{dx_s^*(z)}{dz}\right). \quad \text{(A15)} \end{aligned}$$

Differentiating nonsignatories' first order conditions at the equilibrium, $B'(e_{ns}^*(z)) \equiv D'[z(e_s^*(z) - e_s^*(z) + (n-z)e_{ns}^*(z)$ and $C'(x_{ns}^*(z)) \equiv D'(ze_s^*(z) + (n-z)e_{ns}^*(z))$,with respect to z yields,

$$\begin{aligned} \frac{de_{ns}^*(z)}{dz} &= \frac{\Delta NE^* + z\left(\frac{de_s^*(z)}{dz} - \frac{dx_s^*(z)}{dz}\right) - (n-z)\frac{dx_{ns}^*(z)}{dz}}{\phi - (n-z)}, \\ \frac{dx_{ns}^*(z)}{dz} &= \frac{\Delta NE^* + z\left(\frac{de_s^*(z)}{dz} - \frac{dx_s^*(z)}{dz}\right) + (n-z)\frac{de_{ns}^*(z)}{dz}}{\frac{\phi}{\psi} + (n-z)}. \end{aligned}$$

From the above we calculate the difference $\frac{de_{ns}^*(z)}{dz} - \frac{dx_{ns}^*(z)}{dz}$ and thus, the last

terms in (A14) can be written as,

$$
\begin{aligned}
& D'(E^*(z))(n-z)\left[\frac{de^*_{ns}(z)}{dz}-\frac{dx^*_{ns}(z)}{dz}\right] \\
= \; & D'(E^*(z))(n-z)\frac{(\psi-1)\left[\Delta NE^*+z\left(\frac{de^*_s(z)}{dz}-\frac{dx^*_s(z)}{dz}\right)\right]}{(n-z)(1-\psi)-\phi} \qquad \text{(A16)}
\end{aligned}
$$

Substituting (A15) and (A16) into (A14) yields,

$$\frac{dw_s(z)}{dz}=\frac{\phi\Delta NE^*}{(n-z)(1-\psi)-\phi}\ .$$

We know that $\frac{\phi}{(n-z)(1-\psi)-\phi}<0$ for all z, and thus, the sign of $\frac{d\omega_s(z)}{dz}$ depends solely on ΔNE^*. From (A7) and (A12) we know that $\Delta NE^*=0$ at $z^{\min}$. Therefore, given the uniqueness of $z^{\min}$, we can conclude that $\omega_s(s)$ is U-shaped and hence $\left.\frac{d\omega_s(z)}{dz}\right|_{z\lessgtr z^{\min}} \lessgtr 0$. The conversion to integer values of coalition size is trivial.

[Part 5] Recall that $w_s(z)=B(e^*_s(z))-D(NE^*(z))-C\left(x^*_s(z)\right)$ and $w_{ns}(s)=B(e^*_{ns}(z))-D(NE^*(z))-C\left(x^*_{ns}(z)\right)$. Thus, $w_s(z)\lessgtr w_{ns}(z)\Leftrightarrow B(e^*_s(z))-C\left(x^*_s(z)\right)\lessgtr B(e^*_{ns}(z))-C\left(x^*_{ns}(z)\right)$ and since $B'>0$ and $C'>0$ we have $B(e^*_s(z))-C\left(x^*_s(z)\right)\lessgtr B(e^*_{ns}(z))-C\left(x^*_{ns}(z)\right)\Leftrightarrow e^*_s(z)\lessgtr e^*_{ns}(z)$ and $x^*_s(z)\lessgtr x^*_{ns}(z)\Leftrightarrow s\lessgtr z^{\min}$.

8 Appendix II: Proof of Proposition 2

Direct comparison of $z^{\min}=\frac{n(1-\psi)-\phi}{1-\phi-\psi}$ to $z^{\min}_{x=0}=\frac{n-\phi}{1-\phi}$ yields that $z^{\min}>z^{\min}_{x=o}\Leftrightarrow n>1$. In order to make the analysis is terms of C'' we express $z^{\min}$ in terms of the second derivatives and not the ratios ϕ and ψ, that is, $z^{\min}=\frac{nD''(NE_{nc})C''(x_{nc})-B''(e_{nc})C''(x_{nc})}{D''(NE_{nc})C''(x_{nc})-B''(e_{nc})C''(x_{nc})}$. The difference $z^{\min}>z^{\min}_{x=o}$ is decreasing in $C''(x_{nc})$, because $z^{\min}$ is decreasing in $C''(x_{nc})$, $\frac{\partial z^{\min}}{\partial C''(x_{nc})}=-\frac{(n-1)D''(NE_{nc})[B''(e_{nc})]^2}{[D''(NE_{nc})(C''(x_{nc})-B''(e_{nc}))-B''(e_{nc})C''(x_{nc})]^2}<0$, since $n>1$ and $D''(NE_{nc})>0$.

(i) Using the specification of $z^{\min}$, we have that $\lim_{C''\to 0} z^{\min}\to\frac{-nD''(NE_{nc})B''(e_{nc})}{-D''(NE_{nc})B''(e_{nc})}=n$.

(ii) Using the specification of $z^{\min}$, for moderate values of $B''(e_{nc})$, $\lim_{C''\to\infty}(C''(x_{nc})-B''(e_{nc}))\to C''(x_{nc})$. Therefore, we have that $\lim_{C''\to\infty} z^{\min}\to\frac{nD''(NE_{nc})C''(x_{nc})-B''(e_{nc})C''(x_{nc})}{D''(NE_{nc})C''(x_{nc})-B''(e_{nc})C''(x_{nc})}=$

$\frac{nD''(NE_{nc})-B''(e_{nc})}{D''(NE_{nc})-B''(e_{nc})} = z_{x=0}^{\min}$.

9 Appendix III: Proof of Lemma 1

From (11) we have that $e_s > 0 \Leftrightarrow \Omega^2 + \delta s^2 > (n-s)s\gamma\delta^2 \Longrightarrow \delta^2 + \delta s^2 + 2\delta(n-s) + (n-s)\left[\left(1+\gamma^2\delta^2+2\gamma\delta\right)(n-s) - \gamma\delta^2(s-2)\right] > 0$. We derive the size of the coalition that minimizes this expression $A(s) = \delta^2 + \delta s^2 + 2\delta(n-s) + (n-s)\left[\left(1+\gamma^2\delta^2+2\gamma\delta\right)(n-s) - \gamma\delta^2(s-2)\right]$, which is $\underline{s} = \frac{2\delta(1+\gamma\delta)+n(2+\gamma\delta(4+\delta+2\gamma\delta))}{2(1+\gamma\delta)(1+\delta+\gamma\delta)}$. Substituting the value $\underline{s}$ into the $A(s)$, we get $A(\underline{s}) = \frac{a\left(4\delta^2(1+\gamma\delta)+4n\delta(1+\gamma\delta)(2+\gamma\delta)+n^2(4+\gamma\delta(8+\gamma(4-\delta)\delta))\right)}{4\delta^2(1+\gamma\delta)+8n\delta(1+\gamma\delta)^2+n^2(4+\gamma\delta(12+\gamma\delta(12+\delta+4\gamma\delta)))}$. Then $A(\underline{s}) > 0$, if $4\delta^2(1+\gamma\delta) + 4n\delta(1+\gamma\delta)(2+\gamma\delta)+n^2(4+\gamma\delta(8+\gamma(4-\delta)\delta)) > 0$, which is definitely true for $\delta < 4$. Clearly this is a sufficient but not necessary condition. Notice that if we divide the last expression with δ^3 and let $\delta \to \infty$, the expression reduces to the condition presented in Proposition 1, in Diamantoudi and Sartzetakis (2006). That is, if abatement is not available (extremely expensive) then the only option is to decrease the economic activity and thus emissions, which restricts the size of the coalition to maximum four countries as shown in Diamantoudi and Sartzetakis (2006). However, as the cost of abatement decreases substantially, a large part of the necessary emission reduction is achieved through abatement which eases the restriction on γ. That is, we can have a situation with substantial damages relative to benefits from emissions so that countries want to take strong action and in the same time we have available abatement technologies that can achieve the necessary emission reductions at a reasonable cost.

From (14) we have that $e_{ns} > 0 \Longleftrightarrow (\Omega - n\gamma\delta)\,\Omega + s^2\delta(1+\gamma\delta) > 0$. Dividing the expression by δ^2 and denoting by $\tau = \frac{1}{\delta}$, yields, $(n-s)^2(\phi+\gamma)^2 + [s^2 - (n-s)(\gamma n-2)]\,(\phi+\gamma) - (\gamma n - 1) > 0$. If abatement is not available, that is $\delta \to \infty$, which implies $\tau \to 0$, the expression reduces to the condition presented in Proposition 1, in Diamantoudi and Sartzetakis (2006). However, as the cost of abatement decreases emission of nonsignatories are positive for higher values of γ. Although the necessary condition can be derived using the above technique, it is very complicated and we omit it. It is important to notice that the condition constraints both the values of γ and δ.