European Economic Review 127 (2020) 103423

Contents lists available at ScienceDirect

European Economic Review

journal homepage: www.elsevier.com/locate/euroecorev

The strategic dimension of financing global public goods

Ulrike Kornek[a,*], Ottmar Edenhofer[a,b,c]

[a] Mercator Research Institute on Global Commons and Climate Change, Torgauer Str. 12-15, Berlin 10829, Germany
[b] Potsdam Institute for Climate Impact Research, PO Box 601203, Potsdam 14412, Germany
[c] Technische Universität Berlin, Strasse des 17. Juni 152, Berlin 10623, Germany

ARTICLE INFO

Article history:
Received 18 January 2018
Accepted 9 March 2020
Available online 20 March 2020

JEL classification:
C72
F33
H41
Q52
Q54

Keywords:
Transfers
Global public goods
International environmental agreements
Climate change

ABSTRACT

One challenge in addressing transboundary problems such as climate change is the incentive to free-ride. Transfers from multilateral compensation funds are often used to counteract such incentives, albeit with varying success. We examine how such funds can change the incentive to free-ride in a global public-goods game. In our game, self-interested countries choose their own preferred course, deciding their voluntary public good provision, whether to join a fund that offers compensation for providing the public good and the volume of compensatory payments. We show that (i) total public-good provision is higher when those contributing are given more compensation; and (ii) non-participation in the fund can be punished if the remaining members decrease their public-good provision sufficiently. We then examine three specific fund designs. In the first, the compensation paid to each country is equal to the percentage of above-average total costs for public-goods provision. This design is best able to deter free-riding and can establish the social optimum as the equilibrium. In the second, the compensation paid to each country is a function of the marginal cost of their public-good provision. Here there are significant incentives to free-ride. In the third case, the monetary resources provided by the fund are fixed, a design frequently encountered in international funds. This design is the one least able to deter free-riding.

1. Introduction

International compensation funds are an instrument commonly used to support cooperation on transboundary externalities. Examples are The Global Fund to Fight AIDS, Tuberculosis and Malaria, the Multilateral Fund for the Implementation of the Montreal Protocol or the Green Climate Fund. While some funds have been relatively successful (Biermann and Simonis, 1999), stepping up public-good provision through transfers remains a challenge for climate-change mitigation (Paulsson, 2009; Kumar, 2015). We introduce compensation funds into a global-public-goods game. Transfers from the fund change the equilibrium strategies adopted by countries. Voluntary public-good provision increases because each member of the fund receives a larger transfer if it contributes more to the public good. Our analysis shows how transfers should increase with each country's contribution so as to reduce their incentive to free-ride.

* Corresponding author.
E-mail addresses: kornek@mcc-berlin.net (U. Kornek), Ottmar.Edenhofer@pik-potsdam.de (O. Edenhofer).

https://doi.org/10.1016/j.euroecorev.2020.103423

We provide relevant insights into the design of international compensation funds because, in our game, countries can free-ride by contributing little to the public good and by deciding not to participate in the fund. Transfers change both incentives to free-ride in the following sense. The compensation fund increases members' public-good provision because those that contribute more are compensated more. The total number of members determines by how much each member increases its provision. When a new country joins the fund and all other members contribute more, global benefits from the public good increase. The joining country may have an incentive to remain a member in order to enjoy higher benefits. We show that joining the fund is in the self-interest of countries even if the transfer of the new member turns out to be zero.

The way in which transfers increase with individual public-good provision critically influences the decision to join the fund. Our model is based on three specific designs that have been described in the literature. The first is based on effort-sharing. Transfers balance differences in the total cost of public-good provision among members (Falkinger, 1996). Different total costs arise from different voluntary contributions. The second design is based on the price for public-good provision, which is a policy variable that is easier to measure. The fund balances differences in marginal (as opposed to total) costs (Cramton and Stoft, 2012; McKay et al., 2015). The third design in the model is typical of funds frequently employed in international negotiations. The monetary resources provided by the fund are fixed. The resources are refunded in proportion to each member's contribution to the public good (Gersbach and Winkler, 2012; Gersbach and Hummel, 2016).

Previous literature shows how the three designs enhance voluntary contributions to the public good among a given group of members. The incentive of individual countries to participate is often neglected. To enforce participation, some authors assume that the fund will be discontinued once a single member leaves. The novelty of our approach is that each country chooses how much a country joining the fund will increase the volume of transfers by, its participation, and its contribution to public-good provision based on self-interest. The group of members of the fund forms endogenously and may comprise any number of countries.

Importantly for climate-change negotiations, we show that funds with fixed resources – such as the Green Climate Fund – do not effectively counteract free-riding incentives. Free-riding incentives are lower if the fund is based on marginal costs. Our main results show that the fund that balances total costs performs best in terms of enhancing the contributions of self-interested countries.

To analyze the impact of transfers on public-good provision, we model a three-stage game. First, countries vote on how much a single country will increase the fund's volume of transfers by when it joins. The volume of transfers is the scale factor that determines the extent to which members with higher public-good provision are compensated by members that contribute less. Different member contributions to the public good determine who is a donor and who is a recipient. When a country joins, the anticipated flow of transfers increases for all members. Second, countries choose whether to participate in the fund. Lastly, countries choose their own voluntary public-good provision. Only members of the fund can expect transfers.

For the purpose of this study we assume symmetric countries. Though we concede the importance of the heterogeneity of countries in global public-good provision, this assumption has the advantage of enabling us to focus on a central feature of global public-good provision that retains its importance even when countries are heterogeneous: free-riding incentives. Under symmetry, free-riding and transfers arise because countries choose different contributions to the public good. The heterogeneous case adds other transfer determinants while retaining the design elements identified in this study. Section 5 reviews the implications of heterogeneity.

We solve the three-stage game by using backward induction. In the third stage, countries choose their voluntary contribution to the public good. When a country increases its contribution, it faces individual costs. The benefits of global public-good provision increase for every country, which leads to the well-known free-riding incentive. See the intersection of marginal costs and marginal benefits in Fig. 1, panel (a), denoted "No transfer". With the compensation fund members anticipate transfers. The level of a member's individual transfer increases with its individual contribution. Marginal transfers are positive, i.e. there is an additional transfer for each extra unit of the public good provided. Note that marginal transfers are positive irrespective of whether a member has a positive or negative transfer. When contributing more, a donor will pay less money into the fund and a recipient will receive more money. As a result, each member has an incentive to contribute more, in addition to the benefits of public-good provision. Fig. 1, panel (a) illustrates how positive marginal transfers increase members' contributions compared to the absence of transfers, denoted "With fund". We call this transfer design strategic because it counteracts the incentive to free-ride.

In the second stage, countries decide whether or not to participate in the fund. Our analysis shows that a member has an incentive to participate if remaining members sufficiently decrease their public-good provision, should the member in question leave. Remaining members will contribute less to the public good if marginal transfers decrease as a result of non-participation by a single country. Fig. 1, panel (b) illustrates this effect. The benefits forfeited as a result of reduced public-good provision are a punishment for the member that has left.

The extent of this punishment depends on the extent to which transfers increase in proportion to costs. Strategic transfers are an additional incentive to provide the public good against individual costs. If transfers increase in proportion to individual costs, public-good provision is particularly attractive. This is the case in Fig. 1. Remaining members have a lower incentive to provide the public good if the proportionality factor of transfers to costs decreases when one member leaves the fund. Fig. 1, panel (b) illustrates this effect.

In the first stage, countries determine how much a country joining the fund will increase the fund's volume of transfers by. Countries anticipate whether non-participation will be punished sufficiently in the second stage. To increase public-good

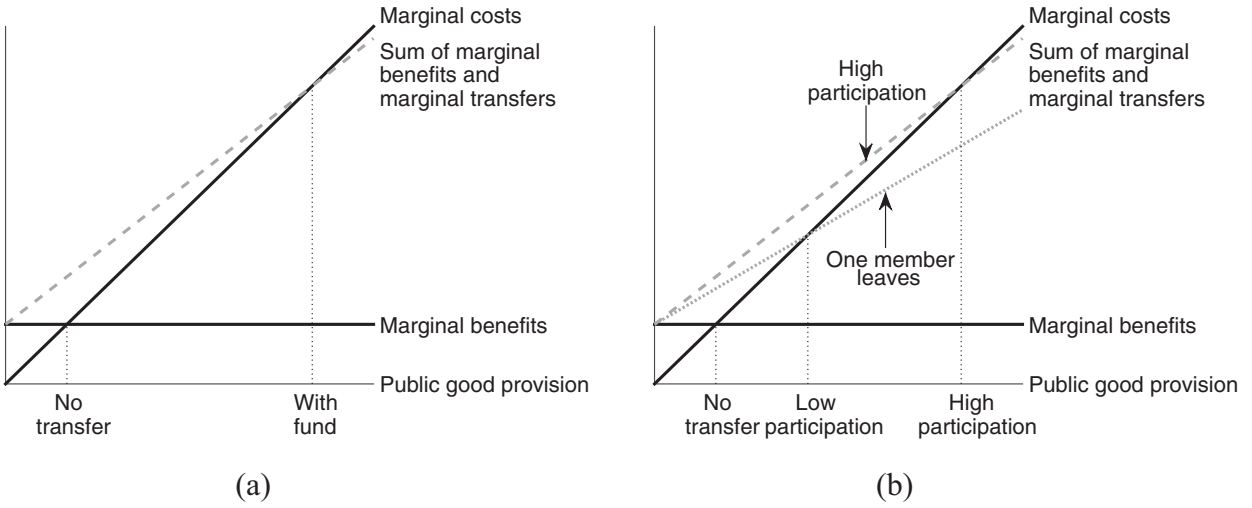

Fig. 1. Equilibrium public-good provision as strategic transfers are introduced. In the absence of the fund, countries equalize marginal costs of public-good provision to their individual marginal benefits. (a) Members of the fund equalize marginal costs to the sum of marginal benefits and marginal transfers of public-good provision, increasing equilibrium contributions. (b) When one member leaves the fund, abatement by remaining members decreases as an optimum response to a change in the incentives from strategic transfers.

provision in the last stage, the volume of transfers will only be chosen at an ambitious level if countries have an incentive to participate in the fund.

Based on these general insights, we analyze the equilibrium for three different transfer designs and specific payoff functions. It is necessary to be specific because complexity prohibits finding closed-form solutions. The volume of transfers takes on different meanings for total-cost, marginal-cost, and fixed funds. In the first design, the fund balances differences in total costs among members. The volume of transfers is the percentage of above-average total costs compensated for. In the second design, the fund balances differences in marginal costs. The volume of transfers determines the size of payment to members with an above-average price for the public good. In the third design, the monetary resources of the fund are fixed, the amount is equal to the volume of transfers. The resources are redistributed to members in proportion to their contribution to the public good.

Our main finding is that the fund that balances total costs performs best. It establishes the social optimum in equilibrium if benefits of public-good provision are linear and costs are quadratic. For this design, and irrespective of the shape of the payoff, members anticipate that they will only need to pay a small share of additional costs when they contribute more. The remaining share is compensated for by transfers from the other members. Transfers increase in proportion to costs. When a member leaves the fund, the fund's volume of transfers decreases, and remaining members have one country less to share their costs with. Remaining members anticipate that they will have to shoulder a higher share of their costs themselves. The proportionality factor of transfers to costs decreases with non-participation. Remaining members will find it significantly less attractive to provide the public good. If benefits are linear and costs are quadratic, the contributions of remaining members will be halved. The departing member is effectively punished. When countries vote on the fund's volume of transfers, they anticipate that all countries have an incentive to participate in the fund. Full compensation for above-average costs is the equilibrium volume of transfers, resulting in socially optimal public-good provision.

If payoffs are not linear/quadratic, the fund may not establish the social optimum. The fund will perform worse when marginal costs are strictly convex. In numerical experiments, we find that payoffs with the fund are equal or close to the social optimum with strictly convex marginal costs. We also consider decreasing marginal benefits. Payoffs in equilibrium with the fund are above the non-cooperative level if marginal benefits do not decrease too steeply. If the slope of marginal benefits is large in absolute terms, the fund will hardly improve on the non-cooperative equilibrium. In this case, however, there is also less need for cooperative public-good provision. Socially optimal payoffs are close to non-cooperative payoffs when marginal benefits decrease steeply. For climate change, marginal damages have been described as relatively flat in the short- to-medium term (Pizer, 2002), indicating that the case of linear benefits in which the total-cost fund significantly improves payoffs is also empirically relevant.

In the other two fund designs, free-riding prevails. Remaining members will not sufficiently decrease their contributions to punish a departing member. If the fund balances marginal rather than total costs, transfers will not increase in proportion to costs. The remaining members will decrease their provisions to a limited extent when non-participation occurs. For the fund with fixed resources, remaining members will share the resources of the fund with one country less when a member leaves. Due to this effect, the incentive to provide the public good actually increases for the remaining members. The result is limited punishment for non-participation.

Our study builds on two strands of literature. The first focuses on how funds facilitate voluntary public-good provision among a given group of members. Gerber and Wichardt (2009) introduce a compelling deposit-refunding mechanism. In their mechanism, each country pays a given deposit into a fund that is only refunded if it contributes a given amount to the public good. The level of the deposit is determined in such a way that paying the deposit and providing the public good is individually optimal. Individually optimal public-good provision constitutes a renegotiation-proof equilibrium. Individual participation in the fund is enforced by assuming that the entire fund will be discontinued if a single country does not pay the deposit.

Gersbach and Winkler (2012) and Gersbach and Hummel (2016) extend the mechanism proposed by Gerber and Wichardt (2009). In their models, each country's refund is a function of total deposits and of its individual public-good provision relative to all members. Deposits are determined so that socially optimal public-good provision is individually optimal. Participation by all countries is again ensured by assuming that the fund will be discontinued if a single country does not participate.[1] Bayer and Urpelainen (2013) study a model involving two donors and one recipient country that may form a fund. In their 2013 study, one of the two donors may still fund the public good in the recipient country if the other donor does not participate. No other studies in this strand of literature analyze the participation of single countries. Buchholz and Konrad (1995) and Ruebbelke (2006) focus on the individual incentives to provide payments and the resulting payoffs between two (groups of) countries. Cramton and Stoft (2012) and McKay et al. (2015) analyze how transfers between a given group of countries should be designed to increase the provision of a public good. We extend this literature by studying the individual incentive of arbitrarily many countries to participate in a fund without assuming that the entire fund will be discontinued if one country leaves.

We model the individual incentive to participate in accordance with the second strand of the literature. Equilibrium participation follows from the formation of a coalition (Barrett, 1994; Bayramoglu et al., 2018). In the standard model found in this literature, a coalition maximizes the sum of its members' welfare. Members comply with the resulting partially cooperative contributions. If a country leaves the coalition, the remaining members will decrease their contributions to the public good. A departing member is punished by the decrease in benefits from public-good provision. However, because members' contributions are partially cooperative, punishment is restricted, and participation in a stable coalition is low if the need for cooperation is large (Barrett, 1994; Karp and Simon, 2013).

The literature on coalition formation has identified ways and means of enhancing participation. Transfers have been shown to lead to larger coalitions, often assuming exogenously determined transfer levels (Carraro et al., 2006; Finus et al., 2006; Weikard, 2009; Lessmann et al., 2015; Kornek et al., 2017). In Barrett (2001), coalition members choose transfers endogenously to help ensure the participation of countries that are committed to non-participation, whereas we assume no asymmetry in the participation decision. Lessmann et al. (2009) and Nordhaus (2015) show that trade sanctions can punish a departing member and enhance participation, but the feasibility of implementing trade sanctions has been questioned (Bordoff, 2009).

We study the formation of a stable compensation fund by adopting the equilibrium concept found in the literature on coalition formation. We depart from the assumption that members will maximize their joint welfare and that transfers are given exogenously. Public-good provision and transfers result from the incentive for strategic transfers, as in the literature on deposit refunding. Members' contributions are an optimal response to the incentive from transfers and generally do not comply with partially cooperative levels. Incentives to participate differ from the literature on coalition formation.

The remainder of this article is structured as follows: Section 2 introduces the game and payoff functions. In Section 3 we indicate general properties of strategic transfers. The three specific transfer designs are discussed in Section 4. Section 5 concludes.

2. The model

We study the set-up of a compensation fund in a global public-goods game. The level of transfers within the fund is determined by a generic rule denoted by $\mathcal{T}$. Rule $\mathcal{T}$ depends on three decision variables chosen in the three stages of the following game:

1st stage: Countries decide unanimously how much a single country will increase the fund's volume of transfers by when it joins.
2nd stage: Countries individually choose whether or not to participate in the fund.
3rd stage: Countries individually choose their level of public-good provision. Members of the fund anticipate transfers.

In the first stage, countries decide how much a single country will increase the volume of transfers by when it joins the fund. The level of this increase is denoted by the variable t (the decision in the first stage). If a country joins in the second stage, it increases the volume of transfers by t. The fund's volume of transfers determines to what extent a member displaying above-average effort will be compensated by transfers from the fund. In the second stage, countries decide on their participation in the fund, knowing how much they will increase the volume of transfers by. Transfers also depend on

[1] Gersbach and Winkler consider a two-period model. Countries have an individual incentive to participate in the second but not the first period without cancelation of the fund. In Gersbach and Hummel, developing countries do not pay deposits giving them an incentive to participate.

the set of member countries. Finally, each country chooses its level of public-good provision, which in turn determines each member's level of transfers.

Our design approach motivates the transfer rule $\mathcal{T}$ based on these three decision variables. First, a member's incentive not to provide the public good is mitigated because transfers reward individual contributions. Second, a country's incentive not to participate is mitigated because the volume of transfers increases when a country joins the fund. Then, a higher incentive to provide the public good results for the other members, thus rewarding membership.

We solve the game using backward induction. In the last stage, countries choose to provide the public good by maximizing their individual payoff. Members of the fund anticipate transfers. This results in a Nash equilibrium in individual public-good provision. The equilibrium in the last stage shows how strategic transfers incentivize voluntary provision of the public good. The outcome can be compared with two benchmark cases: non-cooperative equilibrium and socially optimal provision levels.

The payoffs of the third stage form the basis for the decision in the second stage. We adopt the equilibrium concept from cartel theory and the literature on international environmental agreements (d'Aspremont and Gabszewicz, 1986; Hoel, 1992; Carraro and Siniscalco, 1993; Barrett, 1994). The concept says that a set of countries will form a stable fund if no member has an incentive to leave (internal stability) and no non-member has an incentive to join (external stability). The second stage reveals whether or not a (sub)set of self-interested countries will find it optimal to participate in the fund.

In the first stage, countries vote for their preferred level of t. They anticipate the funds that will be stable in the second stage and the resulting payoffs in the last stage. The level of t is unanimously agreed on by all countries, regardless of whether they participate in the fund. The volume of transfers between members is the product of t and the number of members.

2.1. The payoff model

We consider $N > 2$ symmetric countries. Each country i's individual provision of the global public good is denoted by q_i, which we will henceforth refer to as abatement. Countries incur individual abatement costs $C(q_i)$ with positive and increasing marginal costs ($C' > 0$, $C'' > 0$). Global abatement leads to benefits $B(Q = \sum_{i=1}^{N} q_i)$ that exhibit positive and decreasing marginal benefits ($B' > 0$, $B'' \leq 0$). The payoff in the absence of transfers is

$$\tilde{\pi}_i = B(Q) - C(q_i). \tag{1}$$

In the absence of a fund, each country will maximize (1) given abatement decisions by all other countries. The First-Order Condition (FOC)

$$B'(Nq^{\mathrm{NC}}) = C'(q^{\mathrm{NC}}) \tag{2}$$

defines the non-cooperative Nash equilibrium (NC). The Social Optimum (SO) is defined by Samuelson's rule of public-good provision:

$$NB'(Nq^{\mathrm{SO}}) = C'(q^{\mathrm{SO}}). \tag{3}$$

2.2. Transfers from a compensation fund

Transfer rule $\mathcal{T}$ is a function that depends on three decision variables: (i) the variable $t \geq 0$, which determines how much a new member will increase the fund's volume of transfers by; (ii) the number of members in the fund $k = |S|$ with the subset of members $S \subseteq \{1..N\}$; (iii) the abatement by members $q_i, i \in S$. Transfers are proportional to the volume of transfers. Each member's transfer is:

$$\forall\, i \in S: \quad \mathcal{T}_i = \mathcal{T}(t, k, q_i, q_{-i}) = t \cdot k \cdot \tau(q_i, q_{-i})$$

where q_{-i} denotes abatement by all members but i. A member can be a recipient or a donor depending on countries' abatement choices, i.e. the level of transfers can be positive or negative.

The fund's volume of transfers is $t \cdot k$. This specification is based on the literature on deposit-refunding, which assumes that each member contributes a deposit (in our case t) to the total amount of resources in the fund. Transfers refund the total amount of resources (in our case $t \cdot k$) as a function of individual abatement (Gersbach and Winkler, 2012; Gersbach and Hummel, 2016). We now extend this formulation. The transfer for each member (either received or paid) that arises as a result of different abatement levels is proportional to the volume of transfers. The transfer level is determined by τ. Function τ is the transfer design in abatement. It specifies how much more money is received from or paid less into the fund when a member increases abatement. We test three designs: τ is the difference between a member's (i) total abatement costs, (ii) marginal abatement costs, or (iii) share of abatement and the average among members.

Transfers are added to the payoff for each member $i \in S$:

$$\pi_i^m = B(Q) - C(q_i) + \mathcal{T}_i = B(Q) - C(q_i) + t \cdot k \cdot \tau(q_i, q_{-i}).$$

Transfers fulfill the following assumptions. Transfers add up to zero $\sum_{j \in S} \mathcal{T}(t, k, q_j, q_{-j}) = 0$ for every set of abatement choices by members. Hence, each member anticipates that when its level of transfer increases via larger individual abatement, the transfer each other member receives will decrease. Rule $\mathcal{T}$ is symmetric: two members with the same abatement

have the same transfer level. Since transfers reward abatement, marginal transfers are positive $\frac{\partial}{\partial q_i}\mathcal{T}(t,k,q_i,q_{-i}) \geq 0,\ \forall i \in S$. In turn, the transfer received by member i will decrease when another member increases its abatement: $\frac{\partial}{\partial q_{-i}}\mathcal{T}(t,k,q_i,q_{-i}) \leq 0$. Marginal costs net of transfers increase $\frac{\partial^2}{\partial q_i^2}(C(q_i) - \mathcal{T}(t,k,q_i,q_{-i})) > 0$ to ensure that the payoff with transfers is concave.

Positive marginal transfers hold whether transfers are positive or negative: money received from the fund will increase or money paid into the fund will decrease with individual abatement. Strategic transfers provide an additional incentive for members to abate, as we shall see in the next section.

3. The incentives for strategic transfers

This section analyzes how strategic transfers influence abatement (third stage), (ii) participation in the fund (second stage), and (iii) payoffs when choosing how much a country will increase the volume of transfers by (first stage). We hypothesize about what determines the effectiveness of transfers as free-riding deterrents before analyzing the three design specifications in Section 4.

3.1. The equilibrium in the third stage

In the last stage, countries decide how much to abate. The volume of transfers is known through t and members of the fund S from the first and second stages of the game. Countries will individually maximize their payoffs, given the abatement decisions of all other countries. As transfers add up to zero for every set of members' abatement choices, the maximization does not include this constraint. The FOCs of members $i \in S$ and non-members $j \notin S$ are:

$$\begin{aligned} 0 &= B'(Q) - C'(q_i) + \frac{\partial}{\partial q_i}\mathcal{T}(t,k,q_i,q_{-i}) \\ 0 &= B'(Q) - C'(q_j) \end{aligned} \tag{4}$$

To make the analysis more tractable, we assume that the FOCs define a unique interior Nash equilibrium in abatement, with the same level for all members.[2] The last stage is characterized by member $q^m(t,\ k)$ and non-member abatement $q^n(t,\ k)$ depending only on t and the number of members k.

The FOCs lead directly to the first result:

Proposition 1. *For given participation S and variable t, total and member abatement increase weakly with strategic transfers compared to the non-cooperative equilibrium.*

Proof. See Appendix A.1. □

Strategic transfers increase total abatement. Because transfers reward abatement, members will abate more than when the fund is non-existent. However, transfers by all members are zero in the equilibrium of the final stage. With symmetric member abatement $q^m(t,\ k)$ and symmetric transfer rule $\mathcal{T}$, the level of transfer is the same in equilibrium, and transfers are zero. But member abatement is still above the non-cooperate level. If a member were to abate at the non-cooperative level while the other members abate at $q^m(t,\ k)$, the compensation fund would require this member to make payments to the others to even out the asymmetry in abatement choices (while the characteristics of benefits, costs and transfers are symmetric). This member has an incentive to change its abatement to $q^m(t,\ k)$, thus increasing its transfer level to zero.

Marginal transfers determine the additional incentive to abate in the FOCs (4). The following break-down shows how marginal transfers depend on t and k:

$$\frac{\partial}{\partial q_i}\mathcal{T}(t,k,q_i,q_{-i}) = \underbrace{t \cdot k}_{\substack{\text{Volume} \\ \text{of transfers}}} \cdot \frac{\partial}{\partial q_i}\underbrace{\tau(q_i,q_{-i})}_{\substack{\text{Transfer design} \\ \text{in abatement}}}. \tag{5}$$

There are two components. First, higher t and participation k tend to increase member abatement through the volume of transfers. Second, the transfer design in abatement captures how transfers reward abatement, which generally depends on participation through the abatement by all other members, q_{-i}. Given full participation, the appropriate choice of t will lead to socially optimal abatement if marginal transfers increase sufficiently (the increase in marginal transfers may be bounded under the general functional form τ). When abstracting from the participation decision, it would be irrelevant to study different transfer designs as the previous literature has shown how t can be chosen to achieve the social optimum for deposit-refunding schemes (Gerber and Wichardt, 2009; Gersbach and Winkler, 2012; Gersbach and Hummel, 2016).

[2] Uniqueness can generally only be established under restrictive assumptions about the payoff function with transfers. Showing uniqueness is straightforward if marginal transfers by each member are independent of all other members' abatement $\frac{\partial^2\mathcal{T}(t,k,q_i,q_{-i})}{\partial q_i q_j} = 0\ \forall i \neq j$. This condition holds for the funds based on total and marginal costs in Sections. 4.1 and 4.2. The proof utilizes the fact that the payoff with transfers is concave and follows by contradiction. In Section 4.3, a proof based on contradiction also establishes uniqueness specifically for the fixed-size fund.

However, a large t may lead to an incentive not to participate in the fund. If a member leaves, remaining members may abate at an ambitious level as their additional incentive to abate also scales with t. The free-rider would benefit from the abatement by the remaining members while saving abatement costs. On the other hand, lower participation will change the additional incentive to abate for remaining members because (i) the volume of transfers will decrease and (ii) incentives from the transfer design in abatement will change. Free-riding in the form of non-participation could be punished if remaining members abated at a sufficiently lower level. In the next section, we discuss how strategic transfers counteract free-riding incentives in the second stage. The discussion indicates the characteristics of transfer design that avoid a trade-off between high abatement and high participation when choosing t.

3.2. The equilibrium in the second stage

In the second stage, countries choose to participate in the fund. From the first stage they know they will increase the volume of transfers by t. Payoffs from the last stage are anticipated. This section discusses the mechanism behind transfers that deters free-riding through non-participation. We show that the sum of abatement by all other countries needs to decrease when one member leaves. Remaining members will have a lower incentive to abate if marginal transfers decrease after non-participation. Abatement will decrease particularly if transfers are proportional to costs and the proportionality factor decreases after non-participation. Lastly, we also show how the shape of the payoff function influences the change in abatement.

Participation is determined by the stability of a fund among a set of members. Stability holds if internal and external stability are fulfilled (Barrett, 1994). For symmetric countries, both conditions for stability can be summarized by the stability function $\Delta\pi$ (Hoel, 1992). The stability function is the difference between payoff π^m as a member and payoff π^n as a non-member to the fund among the remaining members. Both payoffs depend only on t, which is given from stage one, and the number of members $k, k-1$ (due to symmetry):

$$\Delta\pi(t,k) = \pi^m(t,k) - \pi^n(t,k-1) \tag{6}$$

The fund with k members is internally stable if $\Delta\pi(t,k)$ is non-negative and externally stable if $\Delta\pi(t,k+1)$ is negative.

The stability function does not cover transfers because the level of transfers is zero in the equilibrium of the last stage. Transfers have no direct influence on the incentive to participate in the fund. The stability function is:

$$\begin{aligned}\Delta\pi(t,k) = {} & B\Big(kq^m(t,k) + (N-k)q^n(t,k)\Big) - B\Big((k-1)q^m(t,k-1) + (N-k+1)q^n(t,k-1)\Big) \\ & - [C(q^m(t,k)) - C(q^n(t,k-1))]\end{aligned} \tag{7}$$

Transfers influence stability indirectly through the abatement levels because the incentive to abate that comes from transfers changes with participation. Member abatement and total abatement change. The first difference in Eq. (7) is the change in benefits from global abatement when one member leaves. If the sum of abatement by all other countries decreases, the benefits forfeited are a positive contribution to the stability function, possibly leading to internal stability.

On the other hand, if member abatement $q^m(t,k)$ is above the outside level $q^n(t,k-1)$, a member can save abatement costs when leaving. If the second difference is high, the stability function may be negative, and a fund with k members may not be internally stable.

Hence, a fund that leads to ambitious member abatement compared to the outside level can only be stable if the abatement of all other countries decreases sufficiently to punish a departing member. In the next section we analyze the determinants of this decrease.

3.2.1. Determinants of abatement change after non-participation

In Fig. 2, panels (a) and (b) illustrate the mechanism behind strategic transfers leading to punishment of a departing member. The formal derivation of the mechanism can be found below. When participation decreases by one, marginal transfers will decrease. In Fig. 2, panels (a) and (b), the line plotting the sum of marginal benefits and marginal transfers shifts downwards as participation decreases from k to $k-1$. The remaining members will abate less because they equalize their marginal costs to the sum of marginal benefits and marginal transfers. The benefits forfeited punish the departing member. Fig. 2, panels (a) and (b), also show how different transfer designs lead to different levels of punishment. While marginal transfers decrease for both designs after non-participation, transfers are proportional to costs in panel (b). The proportionality factor decreases after non-participation. The abatement by the remaining members in panel (b) is lower than in panel (a).

In the following, we provide a formal analysis of the components of transfer design and payoff that contribute to a change in abatement by members and non-members when participation decreases by one. The abatement levels in the stability function (7) are interdependent through t. The FOCs in (4) determine $q^m(t,k)$ and $q^n(t,k)$ for k members. Abatement levels $q^m(t,k-1)$ and $q^n(t,k-1)$ follow from inserting $k-1$ into the FOCs, while holding t fixed.

For a given level of $\bar{t}$ from the first stage, member abatement $q^m(\bar{t},k)$ and $q^m(\bar{t},k-1)$ can be measured by comparing the derivatives of member abatement with t. The fundamental theorem of calculus gives us

$$q^m(\bar{t},k) - q^m(\bar{t},k-1) = \int_0^{\bar{t}} \left(\frac{dq^m(t,k)}{dt} - \frac{dq^m(t,k-1)}{dt} \right) dt. \tag{8}$$

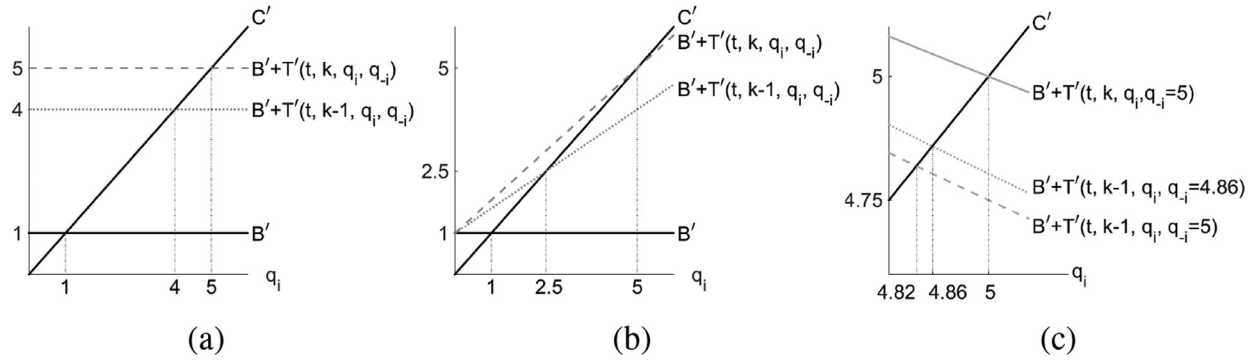

Fig. 2. Constant marginal benefits ($B = Q$), linear marginal abatement costs ($C_i = \frac{1}{2}q_i^2$) and the sum of marginal benefits and marginal transfers as a function of the individual abatement of a generic member. a) Fund based on marginal costs with $t = 1$. For $k = 5$ and $k = 4$, equilibrium abatement in the last stage is 5 and 4, respectively. b) Fund based on total costs with $t = 1/5$. For $k = 5$ and $k = 4$, equilibrium abatement in the last stage is 5 and 2.5, respectively. c) Fund of fixed size with $t = 25$. For $k = 5$, individually optimal abatement is 5 when the other members abate at 5. For $k = 4$, individually optimal abatement is 4.82 and 4.86 when fixing abatement by the other three members at 5 and 4.86, respectively.

The derivatives $\frac{dq^m(t,k)}{dt}$, $\frac{dq^m(t,k-1)}{dt}$ can be identified by applying the implicit-function theorem to the FOCs in Eq. (4). For k members this yields:

$$\frac{dq^m(t,k)}{dt} = \frac{k\cdot\tau' + (k-1)\frac{\partial}{\partial q_{-i}}\mathcal{T}'\cdot\frac{dq^m(t,k)}{dt} + B''\cdot\left((k-1)\frac{dq^m(t,k)}{dt} + (N-k)\frac{dq^n(t,k)}{dt}\right)}{C''(q^m(t,k)) - \mathcal{T}'' - B''}. \tag{9}$$

In Eq. (9) the functions are evaluated at the equilibrium abatement levels. To abbreviate (9) contains $\tau' = \frac{\partial}{\partial q_i}\tau(q_i, q_{-i})$, $\mathcal{T}' = \frac{\partial}{\partial q_i}\mathcal{T}(t,k,q_i,q_{-i})$, $\mathcal{T}'' = \frac{\partial^2}{\partial q_i^2}\mathcal{T}(t,k,q_i,q_{-i})$. The derivative $\frac{dq^m(t,k-1)}{dt}$ can be identified by inserting $k-1$ for k in Eq. (9).

If the derivative in Eq. (9) is larger for k than for $k-1$ at each t, the remaining members will abate less by Eq. (8) when a member leaves.

By Eq. (9) the change in abatement depends on the transfer design, i.e. on the specification of the function $\mathcal{T}$ (effects in τ', $\mathcal{T}''$ and $\frac{\partial\mathcal{T}'}{\partial q_{-i}}$) and the shape of the payoff function in the absence of transfers (effects in B'' and C'').

3.2.2. The influence of transfer design on abatement change

Transfer design determines how marginal transfers decrease after non-participation. The term $k\cdot\tau'$ in Eq. (9) encompasses two effects. First, there is the *volume of transfers* effect. After non-participation, abatement by remaining members will tend to decrease as the volume of transfers decreases. This can be considered a punishment for free-riders. If one member leaves the fund, marginal transfers will decrease with the volume of transfers. Abatement is less attractive for the remaining members. Benefits from public-good provision will tend to decrease.

The *transfer rate* effect captures the extent to which non-participation changes the way transfers increase linearly with abatement (τ'). The incentive to abate is lower if the transfer design in abatement leads to a further decrease in marginal transfers, i.e. if τ' decreases when going from k to $k-1$.

The extent to which marginal transfers increase with abatement ($\mathcal{T}''$) constitutes the *cost proportionality* effect. Strategic transfers provide an additional incentive to abate against the costs of abatement. Since marginal costs increase, abatement close to the social optimum is unattractive from an individual perspective. Transfers that increase proportionally to costs will mitigate this disincentive and make abatement attractive. Accordingly, abatement becomes less attractive if the proportionality factor of transfers to costs decreases when one member leaves. Fig. 2 illustrates the difference in member abatement between a fund in which transfers are not proportional to costs ($\mathcal{T}'' = 0$, panel a) and one in which they are ($\mathcal{T}'' \neq 0$, panel b). Compare the lines plotting the sum of marginal benefits and marginal transfers when participation is at k and when participation decreases by one to $k-1$. In both panel (a) and (b), marginal transfers decrease to the same extent at abatement level $q^m(t, k)$ when one member leaves. However, in panel (b) the decrease in marginal transfers results from a decrease in the proportionality factor of transfers to costs. In panel (a), the decrease in marginal transfers is a constant. As a result, abatement by the remaining members $q^m(t, k-1)$ is lower in panel (b) than in panel (a).

Last, there is a *member reaction* effect. Punishment for non-participation is less effective if $\frac{\partial}{\partial q_{-i}}\mathcal{T}' < 0$. In this case, abatement is more attractive for each member if the other members abate less. Hence, when non-participation induces the remaining members to abate less in the absence of this effect, the incentive to abate increases in its presence. There is less punishment for non-participation. Fig. 2, panel (c) illustrates this effect for a generic member. Abatement of the generic member is 5 when the number of members is $k = 5$. When another member leaves, the sum of marginal benefits and marginal transfers for the generic member is drawn assuming two abatement levels by the other three members. First, abatement by the three members is 5. Given $q_{-i} = 5$, the generic member would abate at 4.82, incurring a loss of 0.18 in global abatement to the departing member. Second, abatement by the three members is $q_{-i} = 4.86$. The sum of marginal

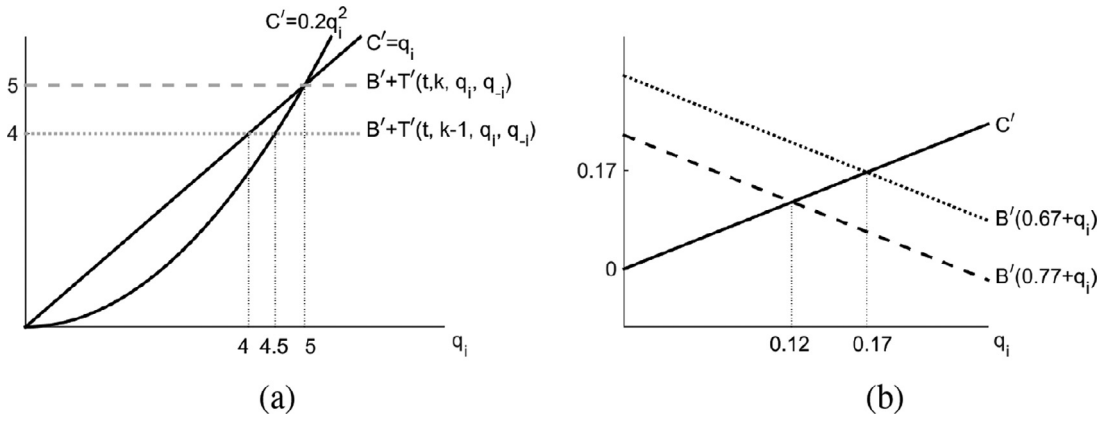

Fig. 3. Marginal abatement costs and the sum of marginal benefits and marginal transfers as a function of individual abatement. (a) Generic member. Marginal benefits and marginal transfers fixed at the same level as in Fig. 2a). Member abatement is 5 when participation is $k = 5$. For quadratic ($C_i = \frac{1}{2}q_i^2$) and cubic costs ($C_i = \frac{1}{15}q_i^3$), member abatement is 4 and 4.5 at participation $k = 4$, respectively. (b) Generic non-member. Linear marginal benefits ($B' = 1 - Q$) and linear marginal abatement costs ($C_i = \frac{1}{2}q_i^2$), $N = 5$. If the four other countries collectively abate at 0.77 and at 0.67, optimal abatement will be 0.12 and 0.17, respectively.

benefits and marginal transfers shifts upward, and the generic member abates more as a reaction to lower abatement by the other members. With an increase in abatement from 4.82 to 4.86 by the generic member, the departing member incurs less punishment from forfeited benefits compared to the absence of the member-reaction effect.

Section 4 discusses the consequences of the transfer-volume, transfer-rate, cost- proportionality, and member-reaction effects for the equilibrium of the game. The three transfer designs differ in the effects they exhibit. In all three designs, the remaining members will abate less due to the volume-of-transfers effect. In the first design (the fund based on total abatement costs), the transfer-rate and cost-proportionality effects are present. Here abatement by the remaining members decreases most. In the second design (the fund based on marginal costs), the transfer-rate effect is present, but transfers are not proportional to costs. Abatement decreases to a lesser extent than in the first design. In the third design (the fixed-size fund), the decrease in abatement is also limited. The transfer-rate effect is absent, transfers are not proportional to costs, and the member-reaction effect is present.

3.2.3. The influence of the payoff function on abatement change

The payoff function determines the extent to which members react to a change in marginal transfers, and hence determines the strength of the aforementioned effects. If costs increase sharply with abatement, element C'' in Eq. (9), the punishment for non-participation is less effective. A change in marginal transfers has less influence on abatement by the remaining members because changing abatement is very costly. The result is less punishment for non-participation. This effect is illustrated in Fig. 3, panel (a) for a generic member. Member abatement is the intersection of marginal costs with the sum of marginal benefits and marginal transfers. Marginal costs C' are either linear or quadratic. For both linear and quadratic marginal costs, member abatement is 5 with participation at $k = 5$. When participation goes down by one and marginal transfers decrease (the line with $k - 1$), member abatement is 4 for linear marginal costs (as in Fig. 2, panel a) and 4.5 for quadratic marginal costs. Member abatement decreases more under linear than strictly convex marginal costs.

Concave benefits ($B'' < 0$) also influence abatement by the remaining members in Eq. (9). In this case, each country reacts to a change in abatement by all other countries. This effect is similar to the member-reaction effect described above. Importantly, non-members will change their abatement as a reaction to the way members change their abatement. When the implicit-function theorem is applied to the FOC in Eq. (4), the derivative of non-member abatement with t is:

$$\frac{dq^n(t,k)}{dt} = \frac{B'' \cdot \left(k\frac{dq^m(t,k)}{dt} + (N-k-1)\frac{dq^n(t,k)}{dt}\right)}{C''(q^n(t,k)) - B''}. \tag{10}$$

With $B'' < 0$, non-members will increase abatement as a response to decreasing abatement by the remaining members. Abatement by all other countries is higher with the reaction of non-members. There is less punishment for non-participation. Fig. 3, panel (b) illustrates this effect for a generic non-member. Assuming that the sum of abatement by all other countries is 0.77, the generic non-member's optimal reaction is abatement at 0.12. If the sum of abatement by all other countries decreases to 0.67 when one member leaves, marginal benefits B' will shift upward for the generic non-member. As an optimal response, its abatement will increase to 0.17. The departing member will suffer fewer losses in benefits forfeited from global abatement with the increase in abatement by the generic non-member.

In conclusion, the remaining members will need to decrease their abatement sufficiently to punish free-riding by not participating in the fund. In the standard model of coalition formation, the remaining members will also abate less when one member leaves the stable coalition (Barrett, 1994; Finus, 2008). This literature finds that stable coalitions are small

Table 1
Three specific designs $\mathcal{T}(t,k,q_i,q_{-i}) = t\cdot k\cdot\tau(t,k,q_i,q_{-i})$ of strategic transfers. The derivatives entering Eq. (9) were calculated under the assumption of the linear-quadratic payoff.

$\mathcal{T}$	Fund balancing differences in total abatement costs $t\cdot k\cdot\{C_i-\frac{1}{k}\sum_{j\in S}C_j\}$[a]	Fund balancing differences in marginal abatement costs $t\cdot k\cdot\{C_i'-\frac{1}{k}\sum_{j\in S}C_j'\}$	Fund of fixed size $t\cdot k\frac{q_i}{\sum_{j\in S}q_j}-t$
Derivatives under linear-quadratic payoff			
τ'	$C'(q^m)\cdot\left(1-\frac{1}{k}\right)$	$C''(q^m)\left(1-\frac{1}{k}\right)$	$\frac{k-1}{k^2}\frac{1}{q^m}$
$\mathcal{T}''$	$t\cdot k\cdot C''\cdot\left(1-\frac{1}{k}\right)$	0	$-2\cdot t\frac{k-1}{k^2}\frac{1}{(q^m)^2}$
$\frac{\partial\mathcal{T}'}{\partial q_{-i}}$	0	0	$-t\frac{k-2}{k^2}\frac{1}{(q^m)^2}$

[a] if $t<\frac{1}{k-1}$ (to exclude convex payoffs), transfers are 0 otherwise.

when the gap between socially optimal and non-cooperative payoffs is large. The assumption this literature proceeds on is that the remaining members will maximize their aggregate payoff. Accordingly, when one member leaves, abatement by remaining members will decrease only to a limited extent to shield the coalition from a large decrease in benefits. The limited abatement decrease is insufficient to punish the leaving member if the gains from public-good provision are large. In our model, the decrease in abatement is an optimal response to the change in marginal transfers when one member leaves. It can be much larger than in the standard model of coalition formation. Section 4 demonstrates that ambitious funds can be stable.

3.3. The equilibrium in the first stage

When choosing t in the first stage, countries anticipate the number of members in the second stage and the resulting payoffs in the final stage.

The objective function in the first stage depends on stable participation in the second stage. From the perspective of the first stage, however, there is no unique subset S of countries forming a stable fund. First, if a fund with k members is stable in the second stage, it is unclear which k countries will form the fund due to symmetry. Second, stability may be fulfilled for different numbers of participating countries k given a specific t. Due to symmetry, we assume that countries expect each possible stable fund to form with equal probability. The function

$$\phi(t,k)=\begin{cases}1, & \Delta\pi(t,k)\geq 0 \text{ and } \Delta\pi(t,k+1)<0\\ 0, & \text{otherwise}\end{cases}$$

is equal to one if a fund among k countries is stable at t and is zero otherwise. Due to symmetry, the preferred t for each country and the subgame perfect equilibrium is found by maximizing the expected payoff $\mathrm{E}[\pi(t)]$ of a generic country:

$$\max_t \mathrm{E}[\pi(t)]=\frac{1}{\sum_{k=1}^{N}\binom{N}{k}\phi(t,k)}\sum_{k=1}^{N}\binom{N}{k}\left(\frac{k}{N}\pi^m(t,k)+\frac{N-k}{N}\pi^n(t,k)\right)\cdot\phi(t,k). \tag{11}$$

Here, $\binom{N}{k}$ different funds among k countries are possible in the second stage, to which a country expects to be a member/non-member with probabilities $\frac{k}{N}$ and $\frac{N-k}{N}$, respectively.

A larger t may increase abatement in the final stage. This will tend to increase payoffs in $\mathrm{E}[\pi(t)]$, but could in turn reduce the number of members, as incentives not to participate may also increase. This will tend to reduce the expected payoff $\mathrm{E}[\pi(t)]$. The optimization in Eq. (11) anticipates the possible trade-off between higher participation and higher abatement. The previous two sections indicated the elements of transfer design and payoff that are conducive to establishing stability in the second stage. These elements define the participation constraint on optimization in (11) posed by the stability function. The optimization, however, is difficult to solve for general functional forms. We therefore analyze the subgame perfect equilibrium for specific payoff functions and three transfer designs.

4. Three specific strategic-transfer designs

This section considers three design specifications for the fund. All three designs reward abatement so that members abate more in the final stage. They lead to different incentives to participate in the second stage. Different equilibrium outcomes result with each of the three designs.

The first design specification for the fund is based on total abatement costs and is the most effective in deterring free-riding incentives. $\mathcal{T}(t,k,q_i,q_{-i})$ is given in the first column of Table 1 (see Falkinger 1996 for a similar fund in a centralized setting). Transfers are proportional to the difference between individual costs and the average among members. The volume of transfers $t\cdot k$ determines the share of above-average abatement costs that is reimbursed by the fund.

For the other two designs, free-riding incentives prevail. In the second design, the total costs of abatement are replaced by marginal costs of abatement. Total abatement costs are not directly observable and may be difficult to measure. Marginal abatement costs are equal to the tax on emissions or the price for tradable allowances, which are directly observable. The

second design balances out differences in the marginal abatement costs of members. See the second column in Table 1. Cramton and Stoft (2012) analyze the same compensation fund but do not consider the incentive for single countries to participate. The volume of transfers $t \cdot k$ determines how much more transfer is received when a country increases its price for abatement above the average price.

For the designs that balance total and marginal costs, a country's transfer will increase unboundedly with abatement. In practice, compensation funds like the Global Environmental Facility or the Green Climate Fund operate with a fixed amount of resources. We therefore study a fund of fixed size. Each member contributes t to the total resources in the fund and receives a refund proportional to its abatement. $\mathcal{T}(t, k, q_i, q_{-i})$ is given in the third column of Table 1. It is similar to the literature on deposit refunding (Gersbach and Winkler, 2012; Gersbach and Hummel, 2016).

4.1. A fund based on differences in total abatement costs

We first analyze a compensation fund that balances out differences in total abatement costs. With transfers given in the first column of Table 1, the FOC for members in the final stage is:

$$0 = B' - C'(q_i) + t \cdot k \cdot C'(q_i)\left(1 - \frac{1}{k}\right). \tag{12}$$

if $t < \frac{1}{k-1}$ and otherwise $B' = C'$. The FOC shows that the fund effectively reduces abatement costs for members. When a member decides to abate more in the last stage, it anticipates that the share $t \cdot k\left(1 - \frac{1}{k}\right)$ of additional abatement costs will be covered by fewer payments to, or more payments from, the other members.

To understand the equilibrium in the second stage, we first analyze how strategic transfers trigger a decrease in abatement by the remaining members. Section 3.2.2 identifies four effects based on the design of transfers. First, there is the volume-of-transfers effect. When one member leaves, the remaining members will abate less because $t \cdot 100$ percent less of additional abatement costs will be reimbursed through the fund. Second, the transfer-rate effect contributes to a further decrease in the incentive to abate. There is one country less to share the abatement costs. The derivative τ' determines the size of this effect. For the total-cost fund, this term is given in the first column of Table 1. Third, transfers are proportional to costs. The second derivative of transfers $\mathcal{T}''$ is positive in Table 1. The proportionality factor of transfers to costs will decrease when one member leaves the fund. Fourth and last, the member-reaction effect is absent. The derivative $\frac{\partial \mathcal{T}'}{\partial q_{-i}}$ is zero for the total-cost fund in Table 1. In sum, the additional incentive to abate declines because the remaining members anticipate that they will have to sustain a higher share of their abatement costs themselves when one member leaves the fund.

4.1.1. The linear-quadratic payoff

In this section we show that the social optimum is the equilibrium of the game for the linear-quadratic payoff function.

Proposition 2. *Consider the case of linear benefits ($B(Q) = b \cdot Q$), quadratic costs of abatement ($C(q_i) = \frac{c}{2} \cdot q_i^2$), and a fund that balances differences in total abatement costs. In the equilibrium of the game, countries will vote for $t = \frac{1}{N}$, which establishes the social optimum.*

Proof. See Appendix A.2. □

In the final stage, the members of the fund (in which all countries participate) abate at socially optimal levels, $q^m(t = \frac{1}{N}, N) = N\frac{b}{c}$. Socially optimal abatement is individually optimal because countries anticipate that they will only pay a fraction of $\frac{1}{N}$ of additional costs when they increase their abatement.

Fig. 2, panel (b) illustrates how marginal transfers decrease when a member leaves in the second stage. Remaining members' abatement then halves: $q^m(t = \frac{1}{N}, N-1) = \frac{1}{2}N\frac{b}{c}$. The departing member incurs a major reduction in benefits, a reduction that outweighs the abatement costs saved. The fund, which is made up of all countries abating at the socially optimal level, is stable in the second stage.

All countries vote for $t = \frac{1}{N}$ in the first stage (unanimity of vote follows endogenously). With $t = \frac{1}{N}$ countries anticipate that every country will participate in the fund and abate at the social optimum. If a single member left the fund, the remaining members would sufficiently decrease their abatement to punish the free-rider. Every country chooses $t = \frac{1}{N}$ because the social optimum is the highest attainable payoff that can be expected in the first stage.

The social optimum is the equilibrium. With $t = \frac{1}{N}$, every country finds it attractive to join the fund and chooses socially optimal abatement. However, a commitment issue arises. A member that chooses to abate at the non-cooperative level can simply decide to withhold the payment required to compensate the other members. National sovereignty means that it cannot be forced to pay. The defecting member could free-ride on the higher abatement levels of the other members. In Appendix B we discuss a possible solution to this problem. In the spirit of Gerber and Wichardt (2009) and Gersbach and Winkler (2012), each member is required to make a payment to the fund before the abatement stage. In this additional "deposit" stage, each member pays an amount

$$d(t, k) = \frac{1}{2}\frac{b^2}{c}\left[\frac{1}{(1 - t(k-1))^2} - 1\right]t(k-1).$$

into the fund. The deposit is refunded net of transfers based on differences in total costs: $\mathcal{T}(t,k,q_i,q_{-i}) = d(t,k) + t \cdot k \cdot (C(q_i) - 1/k \sum_{j \in S} C(q_j))$. Transfers in the final stage are positive. Amount d is fully refunded if all members choose abatement $q^m(t,k)$. If one member chooses non-cooperative abatement while all others abate at $q^m(t,k)$, its transfer is zero. Its deposit d will be distributed to all the other members to reimburse them for their higher abatement efforts.

4.1.2. Other payoff functions

In this section we diverge from the linear-quadratic payoff function. For other payoff functions the introduction of the fund does not always lead to the social optimum. The first payoff function introduces decreasing marginal benefits (Barrett, 1994):

$$\tilde{\pi}_i = b \cdot Q - \frac{\gamma}{2} Q^2 - \frac{c}{2}(q_i)^2, \quad \gamma \geq 0. \tag{13}$$

The parameter γ is the slope of marginal benefits. Setting $\gamma = 0$ recovers the linear-quadratic payoff function. Decreasing marginal benefits are introduced with $\gamma > 0$.

The second payoff function introduces strictly convex marginal costs:

$$\tilde{\pi}_i = b \cdot Q - \frac{c}{e}(q_i)^e, \quad e \geq 2. \tag{14}$$

Parameter e is the exponent of the abatement cost function. The case $e = 2$ recovers the linear-quadratic payoff function.

The game is solved numerically (complexity prohibits analytical solutions, see Barrett 1994). The numerical algorithm is described in Appendix C. The algorithm finds the equilibrium with the fund by solving the maximization problem in (11). This delivers an optimal t^* that results in stable participation $k^* \in \{1..N\}$. The stable participation size k^* is unique in all numerical calculations. Member and non-member payoffs at the equilibrium are denoted by $\pi^m(t^*, k^*)$ and $\pi^n(t^*, k^*)$.

Two indicators assess how the fund improves upon the non-cooperative equilibrium. The first indicator is based on the global payoff

$$\sum_i^N \pi_i = k\pi^m(t,k) + (N-k)\pi^n(t,k)$$

The environmental effectiveness measures the extent to which the global payoff at the equilibrium with the fund closes the gap between the social optimum $N\pi^{SO}$ and the non-cooperative outcome $N\pi^{NC}$:

$$100 \cdot \frac{[k^*\pi^m(t^*,k^*) + (N-k^*)\pi^n(t^*,k^*)] - N\pi^{NC}}{N\pi^{SO} - N\pi^{NC}}. \tag{15}$$

If environmental effectiveness equals 100%, the fund leads to socially optimal abatement in equilibrium. It is straightforward to see that the social optimum will obtain if the equilibrium is $t^* = \frac{1}{N}$ and $k^* = N$.[3] Environmental effectiveness equals zero if the non-cooperative outcome is the equilibrium with the fund.

The second indicator, the cooperation gap, measures the extent to which the social optimum improves payoff for the countries:

$$(\pi^{SO} - \pi^{NC})/\pi^{NC}. \tag{16}$$

If the cooperation gap is close to zero, the social optimum will not deviate much from the non-cooperative equilibrium. Hence, there is less need for cooperation.

We now report the results of our numerical simulations. To analyze the outcome in terms of payoffs, Fig. 4 shows the environmental effectiveness on the left vertical axes. The equilibrium with the fund t^* and resulting participation k^* are shown in Fig. D.5 in the Appendix. Fig. 4, panel (a) adopts decreasing marginal benefits (the payoff in Eq. (13)). Parameter $\frac{\gamma}{c}$ and the number of countries N are varied.[4]

For the linear-quadratic payoff ($\gamma = 0$), the numerical algorithm reproduces the analytic finding of Proposition 2. Environmental effectiveness is 100%, the social optimum is the equilibrium of the game. With $\gamma/c > 0$, environmental effectiveness is below 100% and decreases as marginal benefits become steeper. The social optimum can no longer be sustained as the equilibrium because members and non-members react to changes in abatement by all other countries. Section 3.2.3 describes this effect for $B'' < 0$.

To understand why the equilibrium diverges from the social optimum, consider the non-equilibrium values $t = 1/N$ and $k = N$, for which all countries would abate at socially optimal levels. If a member leaves the fund, marginal transfers will decrease because the volume-of-transfers, the transfer-rate, and the cost-proportionality effects are present. Abatement by the remaining members will tend to decrease. The departing member will also decrease its abatement as there is no additional incentive to abate without participating in the fund. However, the remaining members will react and compensate for lower total abatement by increasing their abatement with $\gamma > 0$. The departing member comes in for less punishment

[3] The FOC with the fund is $0 = B'(Nq^m) - C'(q^m) + \frac{1}{N} \cdot N \cdot (1 - \frac{1}{N})C'(q^m)$, which equals the social optimum defined in Eq. (3).

[4] Environmental effectiveness only depends on the values of $\frac{\gamma}{c}$ and N. Any combinations of b, c, γ, and N which hold $\frac{\gamma}{c}$ and N fixed exhibit the same environmental effectiveness, see Appendix D.

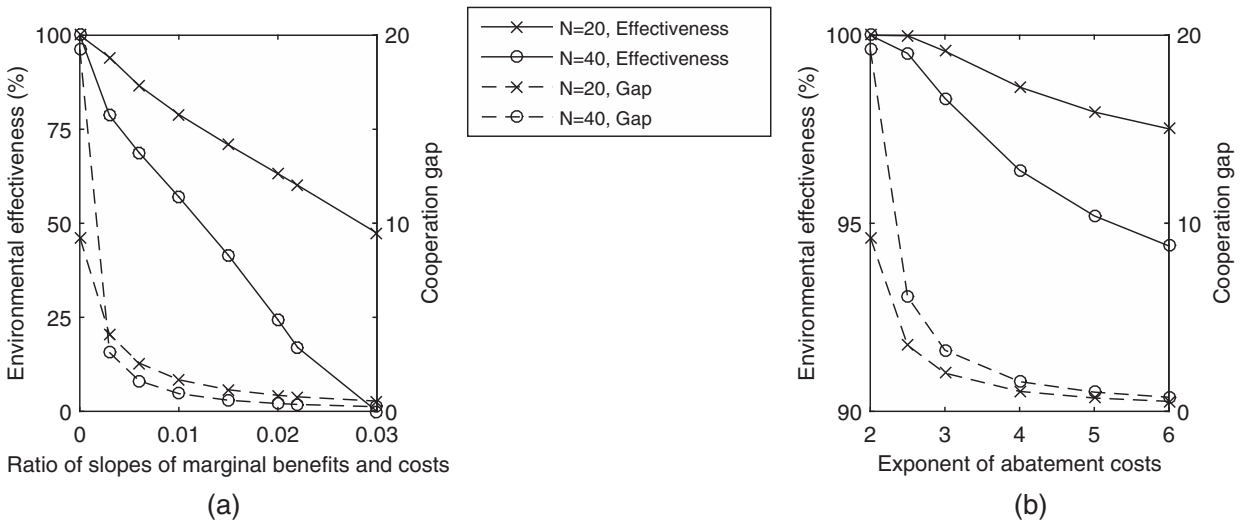

Fig. 4. Environmental effectiveness (Eq. (15)) and cooperation gap (Eq. (16)) as a function of (a) the ratio γ/c in payoff Eq. (13) and (b) the exponent e in payoff Eq. (14).

in the form of a decrease in benefits. In addition, with more ambitious abatement by remaining members, the abatement of the departing member will fall below the non-cooperative level. More abatement costs can be saved by the departing member than if non-member abatement remained at the non-cooperative level. Non-participation is individually optimal at $t = 1/N$ and $k = N$. The social optimum cannot be sustained as a stable fund.

To counteract this free-riding incentive, countries choose an equilibrium t^* that is below $\frac{1}{N}$. Member abatement is below the social optimum, and this reduces the benefits from free-riding when leaving the fund. The fund is stable with a $t^* < \frac{1}{N}$, and all countries participate in equilibrium ($k^* = N$) for small γ/c. Please see Fig. D.5, panel (a) for the numerical details. Payoffs are below the social optimum. A larger t would enhance payoffs, but would also destabilize the fund.

With increasing γ/c, not all countries participate in the fund in equilibrium. For $N = 40$ and $\gamma/c = 0.02,\ 0.022$, equilibrium participation is $k^* = 34, 32$, respectively. In this case, a lower level of t than the equilibrium would lead to higher participation than k^*. However, members would abate less. The expected payoff in the first stage is maximal with lower participation and higher member abatement.

Fig. 4, panel (a) shows that when marginal benefits decrease steeply, the fund does not greatly improve on the non-cooperative outcome. However, the cooperation gap also decreases, as shown in the right vertical axis of the same panel. This effect has been discussed by Barrett (1994). With decreasing marginal benefits, non-cooperative abatement is substantial, and there are fewer gains in cooperation. Fig. 4, panel a) illustrates this analytical finding numerically.[5] The gap approaches zero as the slope of marginal benefits increases.

Consider next the strictly convex marginal costs in Eq. (14). Fig. 4, panel (b) shows the environmental effectiveness. The exponent of abatement costs e and the number of countries N are varied.[6] Again reproducing Proposition 2, environmental effectiveness is 100% when the payoff is of the linear-quadratic variety ($e = 2$). The social optimum is the equilibrium of the game. With an increasing exponent e, environmental effectiveness is below 100%. The social optimum cannot generally be sustained as the equilibrium because members react less to a change in marginal transfers when marginal abatement costs are steep. Section 3.2.3 describes this effect considering the convexity of costs C''.

To understand why the social optimum is not the equilibrium, consider the fund that would implement the social optimum for $t = 1/N$ and $k = N$. Because marginal costs are steep, remaining members react less to a change in marginal transfers. After non-participation, abatement by remaining members changes only to a limited extend. There is little punishment for a departing member. Leaving the fund is (almost always) individually optimal for $e > 2$. As for concave benefits, countries choose a t^* that is below $1/N$ in equilibrium to counteract the incentive to free-ride by not participating. Member abatement is below the social optimum, so that leaving the fund is less attractive. For $t^* < 1/N$, all countries will find it optimal to participate in the fund, $k^* = N$ for all parameter values. See Fig. D.5, panel (b) for the numerical details.

Unlike with decreasing marginal benefits, environmental effectiveness is still close to the social optimum for strictly convex marginal costs. It is above 90% in Fig. 4, panel (b) for all parameter values.

[5] As in footnote 4, the cooperation gap only depends on γ/c and N.

[6] Environmental effectiveness only depends on the values of e and N. It is independent of b and c, see Appendix D.

In sum, we find that when the social optimum significantly improves payoffs in comparison with the non-cooperative equilibrium, the fund based on differences in total abatement costs will also significantly improve payoffs in equilibrium. When benefits are linear and costs are quadratic, the fund will always establish the social optimum.

4.2. A fund based on differences in marginal costs

In this section we consider a fund that bases transfers on marginal abatement costs. Marginal costs are generally more easily measurable than total costs as they equal the tax on emissions or the market price for tradable permits. Transfers balance differences in marginal costs, as specified in the second column of Table 1. This design performs worse than the fund based on total costs.

Proposition 3. *Assume that payoffs are of the form in Eqs. (13) and (14).*

(a) *Consider a fund based on differences in marginal costs that is stable in the second stage of the game for some t and k. There is a stable fund based on differences in total costs with weakly larger participation. The stable fund based on total costs also leads to a weakly larger global payoff than the stable fund based on marginal costs.*

(b) *Consider the parameter range in Section 4.1. In equilibrium, environmental effectiveness is weakly larger with the fund based on total costs than with the fund based on marginal costs.*

(c) *In equilibrium with the fund based on marginal costs, countries vote for $t^* = \frac{2b}{c(N-1)}$ in the case of linear benefits ($B(Q) = b \cdot Q$) and quadratic costs ($C(q_i) = \frac{c}{2} \cdot q_i^2$). All countries participate ($k^* = N$). Environmental effectiveness is $100 \cdot \frac{4(N-2)}{(N-1)^2}$, which is 100% if $N = 3$ and approaches zero as the number of countries increases.*

Proof. Parts (a) and (b) are shown in Appendix A.3. Part (c) is shown in Appendix A.4. □

The fund based on marginal costs is less effective in punishing non-participation than the fund based on total costs. Proposition 3(a) shows that for any stable fund based on marginal costs, there is a stable fund based on total costs that performs better. When a member leaves the fund based on marginal costs, remaining members will decrease their abatement less than with the fund based on total costs. Free-riding by not participating in the fund is more attractive for the marginal-cost than for the total-cost fund.

To see why, consider the four effects influencing member abatement identified in Section 3.2.2. When the volume-of-transfers and the transfer-rate effects are compared, marginal transfers decrease in a similar way for the marginal-cost fund as for the total cost fund. First, the volume of transfers decreases when one member leaves the fund. Second, the transfer-rate effect is present. In Table 1, the second column specifies element $\tau' = C''(1 - \frac{1}{k})$. In the case of non-participation, there is one country less to share marginal abatement costs with.

However, transfers are not proportional to costs. For the linear-quadratic payoff, the second derivative is zero in Table 1, $\mathcal{T}'' = 0$. Marginal abatement costs net of marginal transfers increase irrespective of participation in the fund. Even though marginal transfers decrease with participation, for participation k and $k-1$ convex costs will counteract the additional incentive to abate to the same extent. In comparison, abatement changes less with the fund based on marginal costs than with the fund based on total costs. Comparatively, there is less punishment for non-participation.

The first part of Proposition 3(a) is a direct consequence of the comparative lessening of punishment. Participation is larger when the fund balances total and not marginal costs. Consider a case where the stability function in Eq. (7) is positive for the fund based on marginal costs for some t and participation k. A certain degree of member abatement will result. The proof shows that there is some t for the fund based on total costs, such that the stability function has a larger value at the same k and member abatement in comparison to the fund based on marginal costs. Recall that stability in the second stage is achieved when the stability function switches signs. Since the stability function is larger for the total-cost fund, participation is weakly larger than for the marginal-cost fund. The proof shows that the global payoff for the stable fund based on total costs is also weakly larger than for the fund based on marginal costs because participation is larger. This establishes the second part of Proposition 3(a).

Proposition 3(b) is a direct consequence of Proposition 3(a). The proof utilizes the fact that if there is only one stable fund size k^*, the expected payoff in the first stage (defined in Eq. (11)) is $1/N$ times the global payoff. Uniqueness was established in Section 4.1 for the parameter range considered. Hence, in the case of equilibrium with both fund designs, the global payoff will achieve its highest value because the expected payoff is optimal. Proposition 3(a) establishes that the global payoff is higher for the stable fund based on total costs. Hence, the global payoff in equilibrium with that fund is at least as high as in equilibrium with the fund based on marginal costs. The fund based on total costs establishes higher environmental effectiveness since environmental effectiveness increases with global payoff (see Eq. (15)). The proof shows 3(a) and (b) without deriving an explicit solution to the game with the fund.

In terms of linear-quadratic payoff, Proposition 3(c) shows how much better the fund based on total costs performs than the fund based on marginal costs. While environmental effectiveness is 100% in the former, it is below 100% in the latter if $N > 3$. It will even approach zero if the number of countries is comparatively large. For example, if there are 100 countries, environmental effectiveness is 4% for the fund based on marginal costs. Environmental effectiveness can also be high for small numbers of countries (the fund establishes the social optimum for $N = 3$). If there are only a small number of countries, socially optimal abatement is not much more ambitious than non-cooperative abatement. A small decrease in abatement by remaining members is sufficient to punish non-participation.

4.3. A fund of fixed size

The last design we consider is a fund of fixed size. Members pay an amount t of resources into the fund. Each member receives a share of total resources proportional to its abatement. The third column of Table 1 specifies the design. This approach is most similar to funding mechanisms that have been implemented in practice. However, significant incentives to free-ride prevail, as we show in the next proposition.

Proposition 4. *Assume that payoffs are of the form in Eqs. (13) and (14).*

(a) *Consider a fund of fixed size that is stable in the second stage of the game for some t and k. There is a stable fund based on differences in marginal costs with weakly larger participation. The stable fund based on marginal costs also leads to a weakly larger global payoff than the stable fund of fixed size.*
(b) *Consider the parameter range in Section 4.1. In equilibrium, environmental effectiveness with the fund based on total costs is weakly larger than with the fund of fixed size.*
(c) *In equilibrium, environmental effectiveness with the fund of fixed size is no greater than in equilibrium with the fund that balances marginal abatement costs if benefits are linear ($B(Q) = b \cdot Q$) and costs are quadratic ($C(q_i) = \frac{c}{2} \cdot q_i^2$). Environmental effectiveness approaches zero as the number of countries increases.*

Proof. See Appendix A.5. □

The fund of fixed size is less effective in punishing non-participation than the fund that balances marginal costs, Proposition 4(a). By Proposition 3, it therefore also performs worse than the fund based on total costs, Proposition 4(b). When a member leaves the fund, the drop in the total amount of resources triggers little abatement change. Non-participation is not effectively punished.

Again, we can gain insights into the mechanism by analyzing the four effects identified in Section 3.2.2. Abatement by the remaining members will decrease when one member leaves because the volume of transfers decreases. However, there is one member less to share the (smaller) amount of resources with, which increases the incentive to abate. In contrast to the other fund designs, the transfer-rate effect does not contribute to the punishment of a departing member. Table 1 specifies the transfer-rate effect for the linear-quadratic payoff, element $\tau' = \frac{k-1}{k^2} \frac{1}{q^m}$. In addition, the member reaction effect is present with the derivative $\frac{\partial \tau'}{\partial q_{-i}}$ being negative in Table 1. When one member leaves the fund, all remaining members will abate less. The incentive to abate increases for all remaining members. Finally, transfers are not proportional to costs.

Of the three designs studied, the fixed-size fund entails the smallest change in abatement by remaining members. As a result, the statements about global payoff and environmental effectiveness in Proposition 4 are equivalent to Proposition 3. From a strategic perspective, the fixed-size fund performs worse than the other funds.

5. Conclusions

We demonstrate that international transfers via multilateral compensation funds can be designed to significantly increase the provision of a global public good. In our game, countries face free-riding incentives when deciding on (a) how much one country will increase the volume of transfers by in the fund, (b) their individual participation and (c) their voluntary level of public-good provision. Our results are important for the strategic design of transfers because all decisions taken by countries are based on self-interest. Public-good provision increases in the absence of a global authority with enforcement power.

We provide insights for climate change negotiations by showing that a design similar to the Green Climate Fund does not effectively counteract free-riding incentives. For this design, a fixed amount of resources is distributed to the members of the fund. When a single country leaves and withdraws its contribution from the resources of the fund, the provision of the public good by the remaining members is too ambitious to punish non-participation. A fund that balances differences in the marginal costs of public-good provision is better able to counteract free-riding incentives. In this design, the remaining members will lower their public-good provision to a larger extent when a member leaves than they will in a fund that provides a fixed amount of monetary resources. The design that balances differences in the total costs of public-good provision is best able to counteract free-riding incentives. Equilibrium with the fund is the social optimum among an arbitrary number of symmetric countries when the marginal benefits of public-good provision are linear and costs are quadratic.

We extend the linear-quadratic case to other payoff functions. When marginal costs are strictly convex and marginal benefits are close to the constant case, the fund based on differences in total costs will significantly improve the welfare of countries. Where marginal benefits decrease steeply, the fund does not greatly improve on the non-cooperative equilibrium. In this case, however, the welfare of countries in the social optimum is also relatively close to the non-cooperative equilibrium. Marginal damages of climate change have been described as relatively flat in the short-to-medium term (Pizer, 2002), indicating that the case of linear benefits in which the total-cost fund significantly improves payoffs is also empirically relevant.

Our three-stage model provides insights on how best to design transfers to finance global public goods. Further research would do well to investigate the assumptions underlying our simple model. For example, we assume that all countries

commit to the level by which a joining country increases the volume of transfers. The incentive to change this could be studied by means of renegotiation-proof equilibria.

The second assumption is that of symmetric countries. This assumption casts light on the mechanism behind the way in which transfers deter free-riding. Transfers arise because countries contribute to the public good at different levels. When countries are assumed to be heterogeneous, the benefit and cost functions in their public-good provision will differ. In the heterogeneous setting, constructing a fund based on differences in total costs is straightforward. Here, transfers arise due to different cost and benefit functions, as well as different contributions to the public good. The volume-of-transfer, transfer-rate and cost-proportionality effects that we have identified will still be present when countries are assumed to be heterogeneous. However there will also be donors and recipients in equilibrium. It is then readily convincing to suggest that a net donor country will only have an incentive to participate if its benefits from public-good provision are high enough to compensate for its negative payments. Hence, the effectiveness of the fund designed to enhance the provision of the public good will depend on the type of heterogeneity obtaining among countries.

Acknowledgments

The authors would like to thank Axel Ockenfels, Peter Cramton and Steven Stoft for valuable discussions when this article was taking shape. Kenneth Arrow and the other participants of the Stanford workshop "Promoting Cooperation in International Climate Negotiations" provided valuable comments on the idea behind the article. We are grateful to Christian Traeger, Ralph Winkler, Michael Finus, Christian Flachsland, Michael Pahle, and Kai Lessmann for looking at earlier versions of the paper and providing valuable comments and suggestions.

Appendix A. Proofs

A.1. Proof of Proposition 1

We first show that total abatement is weakly larger with the fund compared to the non-cooperative equilibrium. Denote $\mathcal{T}' = \frac{\partial}{\partial q_i}\mathcal{T}(t,k,q_i,q_{-i})$. Assume the opposite claim of the proposition: $k\cdot q^m(t,k)+(N-k)q^n(t,k)=Q(t,k)<Q^{NC}=Nq^{NC}$. Then $B'(Q(t,k))+\mathcal{T}'(t,k,q^m(t,k),q_{-i}=q^m(t,k))\geq B'(Q^{NC})\ \forall i\in S$ and $B'(Q(t,k))\geq B'(Q^{NC})$. Hence $C'(q^m(t,k))$, $C'(q^n(t,k))\geq C'(q^{NC})$, from which the contradiction $q^m(t,k)$, $q^n(t,k)\geq q^{NC}$ follows as marginal costs increase.

Second, $q^m(t,k)\geq q^n(t,k)$ follows from comparing the FOCs for members and non-members: $C'(q^m(t,k))=B'(Q(t,k))+\mathcal{T}'(t,k,q^m(t,k),q_{-i}=q^m(t,k))\geq B'(Q(t,k))=C'(q^n(t,k))$. Since total abatement is weakly larger than the non-cooperative level, it follows that $q^m(t,k)\geq q^{NC}$.

A.2. Proof of Proposition 2

In the last stage, member abatement for given t and k is $q^m(t,k)=\frac{b}{c}\frac{1}{1-t(k-1)}$, non-member abatement is at the non-cooperative level $q^n(t,k)=b/c$.

The stability function for a fund among k countries is:

$$\Delta\pi(t,k)=\frac{b^2}{c}\left(k\frac{1}{1-t(k-1)}-(k-1)\frac{1}{1-t(k-2)}-\frac{1}{2}\frac{1}{(1-t(k-1))^2}-\frac{1}{2}\right).$$

There is a unique number of members k such that the fund is stable given t (proof is available upon request). Inserting $t=1/N$ and $k=N$, the stability function has a positive value at $\frac{1}{2}(N-1)\geq 0$. Hence the fund among all countries with socially optimal abatement is internally stable. Since it is also externally stable by definition, $t=1/N$ establishes the social optimum as a stable fund in the second stage.

Since there is a unique number of members $k(t)$ such that the fund is stable given t, the expected payoff in Section 3.3 is $E[\pi(t)]=\frac{k(t)}{N}\pi^m(t,k(t))+\frac{N-k(t)}{N}\pi^n(t,k(t))$, which is the global payoff divided by the number of countries. By definition, the global payoff is highest at the social optimum, which can only be attained if $t=1/N$ and all countries participate. Hence, the expected payoff of each individual country in the first stage of the game is highest for $t=1/N$, for which all countries vote in the first stage.

A.3. Proof of Proposition 3(a) and (b)

Consider a stable fund of participation $k=\bar{k}$ for a $t=\bar{t}|_{MC}$, which leads to member abatement $\bar{q}\geq q^{NC}$ for the fund based on marginal costs. The proof considers the fund based on total costs that also leads to abatement $\bar{q}$ when $\bar{k}$ members participate.[7]

[7] Note that any abatement level of members $q^m(t,k)\geq q^{NC}$ can be attained by increasing t for the fund based on total costs. Solving the FOC for the payoff function in Eq. (14) leads to $q^m(t,k)=(b/c)^{1/e-1}(1-t(\bar{k}-1))^{1/1-e}$. At the upper bound on t, $t=1/(\bar{k}-1)$ (see Table 1), member abatement approaches infinity. For the payoff in Eq. (13), note that total abatement Q above b/γ violates positive marginal benefits $B'(Q)=b-\gamma\cdot Q\leq 0$. Hence, member

We proceed in four steps. The first step shows that the drop in abatement by members after non-participation is weakly larger for the fund based on total costs than for the fund based on marginal costs. The second step shows that the stability function is weakly larger for the fund based on total costs than for the fund based on marginal costs. Third, global payoff is also shown to be weakly larger, showing Proposition 3(a). Lastly, environmental effectiveness is weakly larger for the fund based on total costs, showing Proposition 3(b). All steps are first done for the payoff in Eq. (14) and then for the payoff in Eq. (13).

A.3.1. Payoff with strictly convex marginal costs in Eq. 14

We start by showing that, after non-participation, member abatement decreases more for the total cost fund than for the marginal cost fund. Because it is not possible to give an explicit expression for abatement levels given the more general payoff functions, we proceed with comparing first-order conditions.

Drop in abatement is weakly larger for the fund based on total costs than for the fund based on marginal costs

The $t = \bar{t}|_{\mathrm{TC}}$ that leads to abatement $\bar{q}$ for participation $\bar{k}$ in the fund based on total costs is defined by the FOC for members:

$$\begin{aligned} 0 &= b - C'(\bar{q}) + \bar{t}|_{\mathrm{TC}}(\bar{k}-1)C'(\bar{q}) \\ &\Rightarrow \bar{t}|_{\mathrm{TC}} = \frac{C'(\bar{q}) - b}{(\bar{k}-1)C'(\bar{q})}. \end{aligned} \tag{A.1}$$

With $\bar{t}|_{\mathrm{TC}}$ given, the FOC of members to the fund with participation $\bar{k}-1$ defines the extent to which members decrease their abatement after non-participation. To reduce notation, member abatement with participation $\bar{k}-1$ is denoted with $q_2|_{\mathrm{TC}}$. $q_2|_{\mathrm{TC}}$ is then an implicit function of $\bar{q}$:

$$\begin{aligned} 0 &= b - C'(q_2|_{\mathrm{TC}}) + \bar{t}|_{\mathrm{TC}}(\bar{k}-2)C'(q_2|_{\mathrm{TC}}) \\ &= b - C'(q_2|_{\mathrm{TC}}) + \left[C'(\bar{q}) - b\right]\frac{\bar{k}-2}{\bar{k}-1}\frac{C'(q_2|_{\mathrm{TC}})}{C'(\bar{q})}. \end{aligned} \tag{A.2}$$

For the fund based on marginal costs, the $t = \bar{t}|_{\mathrm{MC}}$ that leads to abatement $\bar{q}$ for participation $\bar{k}$ is again defined by the FOC for members:

$$\begin{aligned} 0 &= b - C'(\bar{q}) + \bar{t}|_{\mathrm{MC}}(\bar{k}-1)C''(\bar{q}) \\ &\Rightarrow \bar{t}|_{\mathrm{MC}} = \frac{C'(\bar{q}) - b}{(\bar{k}-1)C''(\bar{q})}. \end{aligned}$$

Let, as above, $q_2|_{\mathrm{MC}}$ denote member abatement after non-participation. Inserting $\bar{t}|_{\mathrm{MC}}$ to the FOC for members when participation is at $\bar{k}-1$ yields:

$$\begin{aligned} 0 &= b - C'(q_2|_{\mathrm{MC}}) + \bar{t}|_{\mathrm{MC}}(\bar{k}-2)C''(q_2|_{\mathrm{MC}}) \\ &= b - C'(q_2|_{\mathrm{MC}}) + \left[C'(\bar{q}) - b\right]\frac{\bar{k}-2}{\bar{k}-1}\frac{C''(q_2|_{\mathrm{MC}})}{C''(\bar{q})} \\ &= b - C'(q_2|_{\mathrm{MC}}) + \left[C'(\bar{q}) - b\right]\frac{\bar{k}-2}{\bar{k}-1}\frac{C'(q_2|_{\mathrm{MC}})}{C'(\bar{q})}\frac{\bar{q}}{q_2|_{\mathrm{MC}}}. \end{aligned} \tag{A.3}$$

The last equality follows from inserting $C''(q) = (e-q)C'(q)/q$ for this payoff function. Subtracting Eqs. (A.2) and (A.3) yields:

$$\begin{aligned} 0 &= C'(q_2|_{\mathrm{MC}}) - C'(q_2|_{\mathrm{TC}}) + \frac{C'(\bar{q}) - b}{C'(\bar{q})}\frac{\bar{k}-2}{\bar{k}-1}\left(C'(q_2|_{\mathrm{TC}}) - C'(q_2|_{\mathrm{MC}})\frac{\bar{q}}{q_2|_{\mathrm{MC}}}\right) \\ &= \left[C'(q_2|_{\mathrm{MC}}) - C'(q_2|_{\mathrm{TC}})\right]\underbrace{\left(1 - \frac{C'(\bar{q}) - b}{C'(\bar{q})}\frac{\bar{k}-2}{\bar{k}-1}\right)}_{(1)} + \underbrace{\frac{C'(\bar{q}) - b}{C'(\bar{q})}\frac{\bar{k}-2}{\bar{k}-1}C'(q_2|_{\mathrm{MC}})}_{(2)}\underbrace{\left(1 - \frac{\bar{q}}{q_2|_{\mathrm{MC}}}\right)}_{(3)}. \end{aligned}$$

abatement $\bar{q}$ for the fund based on marginal costs has to be below $b/(\gamma\bar{k})$, for which non-member abatement is zero. Solving the FOCs for the fund based on total costs, member abatement is $q^m(t,k) = \frac{b}{c}\left[\frac{\gamma}{c}N + 1 - t(k-1)\left(\frac{\gamma}{c}(N-k)+1\right)\right]^{-1}$. At the upper bound on t, member abatement approaches at $q^m(t = 1/(\bar{k}-1), k) = b/(\gamma\bar{k})$, which is the upper bound on $\bar{q}$.

The desired relationship $q_2|_{\text{MC}} \geq q_2|_{\text{TC}}$ follows. This is due to the fact that (1) $\left(1 - \frac{C'(\bar{q})-b}{C'(\bar{q})}\frac{\bar{k}-2}{\bar{k}-1}\right) \geq 0$ because $\bar{q} \geq q^{\text{NC}}$; (2) $\frac{C'(\bar{q})-b}{C'(\bar{q})}\frac{\bar{k}-2}{\bar{k}-1}C'(q_2|_{\text{MC}}) \geq 0$ because $\bar{q} \geq q^{\text{NC}}$; (3) $\left(1 - \frac{\bar{q}}{q_2|_{\text{MC}}}\right) \leq 0$ because $\bar{q} \geq q_2|_{\text{MC}}$, which follows from the FOC for member abatement.[8]

The stability function is weakly larger for the fund based on total costs than for the fund based on marginal costs

For the fund based on marginal costs we know that the stability function Eq. (7) is positive at $\bar{k}$ by assumption: $\Delta\pi(\bar{t}|_{\text{MC}}, \bar{k}) \geq 0$. The stability function is then higher for the fund based on total costs than for the fund based on marginal costs:

$$\begin{aligned}
\Delta\pi(\bar{t}|_{\text{TC}}, \bar{k}) &= b \cdot (\bar{k}\bar{q} + (N-\bar{k})q^{\text{NC}}) - C(\bar{q}) - b \cdot ((\bar{k}-1)q_2|_{\text{TC}} + (N-\bar{k}+1)q^{\text{NC}}) - C(q^{\text{NC}}) \\
\Delta\pi(\bar{t}|_{\text{MC}}, \bar{k}) &= b \cdot (\bar{k}\bar{q} + (N-\bar{k})q^{\text{NC}}) - C(\bar{q}) - b \cdot ((\bar{k}-1)\bar{q}_2|_{\text{MC}} + (N-\bar{k}+1)q^{\text{NC}}) - C(q^{\text{NC}}) \\
&\Rightarrow \Delta\pi(\bar{t}|_{\text{TC}}, \bar{k}) - \Delta\pi(\bar{t}|_{\text{MC}}, \bar{k}) = b(\bar{k}-1)(q_2|_{\text{MC}} - q_2|_{\text{TC}}) \geq 0
\end{aligned}$$

Hence, the fund that balances total costs is also internally stable. Recall that stability is fulfilled if the stability function switches sign at $\bar{k}$. If external stability does not hold for the total cost fund at $\bar{k}$, then there is a stable fund of larger participation $\bar{k} + \Delta k$, $\Delta k > 0$. This is the first part of claim a) for the payoff function with strictly convex marginal costs.

Global payoff is weakly larger for the stable fund that balances total costs than for the fund that balances marginal costs

At $\bar{k}$, both funds (based on marginal costs and on total costs) lead to the same abatement $\bar{q}$ by design of this proof. Hence, the global payoff is the same for both fund designs:

$$\bar{k}\pi^m(\bar{t}|_{\text{TC}}, \bar{k}) + (N-\bar{k})\pi^n(\bar{t}|_{\text{TC}}, \bar{k}) = \bar{k}\pi^m(\bar{t}|_{\text{MC}}, \bar{k}) + (N-\bar{k})\pi^n(\bar{t}|_{\text{MC}}, \bar{k}).$$

We now establish two other relationships. First, from Proposition 1 we know that member payoff is smaller than non-member payoff: $\pi^m(t,\ k) \leq \pi^n(t,\ k)$.[9] Second, with positive stability function until $\bar{k} + \Delta k$ we have $\pi^m(\bar{t}|_{\text{TC}}, \bar{k} + \ell) \geq \pi^n(\bar{t}|_{\text{TC}}, \bar{k} + \ell - 1),\ \forall\ \ell \in \{0, \ldots, \Delta k\}$. With these two inequalities we can establish that the global payoff for the stable fund based on total costs weakly decreases from $\bar{k} + \Delta k$ to $\bar{k} + \Delta k - 1$. At $\bar{k} + \Delta k$ the global payoff is

$$\begin{aligned}
&(\bar{k}+\Delta k)\pi^m(\bar{t}|_{\text{TC}}, \bar{k}+\Delta k) + (N-\bar{k}-\Delta k)\underbrace{\pi^n(\bar{t}|_{\text{TC}}, \bar{k}+\Delta k)}_{\geq \pi^m(\bar{t}|_{\text{TC}}, \bar{k}+\Delta k)} \\
&\geq N\pi^m(\bar{t}|_{\text{TC}}, \bar{k}+\Delta k) \geq N\pi^n(\bar{t}|_{\text{TC}}, \bar{k}+\Delta k - 1) \\
&= (\bar{k}+\Delta k - 1)\underbrace{\pi^n(\bar{t}|_{\text{TC}}, \bar{k}+\Delta k - 1)}_{\geq \pi^m(\bar{t}|_{\text{TC}}, \bar{k}+\Delta k - 1)} + (N-\bar{k}-\Delta k+1)\pi^n(\bar{t}|_{\text{TC}}, \bar{k}+\Delta k - 1) \\
&\geq (\bar{k}+\Delta k - 1)\pi^m(\bar{t}|_{\text{TC}}, \bar{k}+\Delta k - 1) + (N-\bar{k}-\Delta k+1)\pi^n(\bar{t}|_{\text{TC}}, \bar{k}+\Delta k - 1).
\end{aligned}$$

The same consecutive manipulations establish $(\bar{k}+\Delta k)\pi^m(\bar{t}|_{\text{TC}}, \bar{k}+\Delta k) + (N-\bar{k}-\Delta k)\pi^n(\bar{t}|_{\text{TC}}, \bar{k}+\Delta k) \geq \bar{k}\pi^m(\bar{t}|_{\text{TC}}, \bar{k}) + (N-\bar{k})\pi^n(\bar{t}|_{\text{TC}}, \bar{k})$, from which we have the desired relationship. Hence, the second part of claim (a) is shown for the payoff with strictly convex marginal costs.

Environmental effectiveness in equilibrium is weakly larger for the fund that balances total costs than for the fund that balances marginal costs

Consider the equilibrium with the fund based on marginal costs $t^*|_{\text{MC}}$. This establishes an equilibrium payoff in Eq. (11) that is a convex combination of expected payoffs $\frac{k}{N}\pi^m(t^*|_{\text{MC}}, k) + \frac{N-k}{N}\pi^n(t^*|_{\text{MC}}, k)$ for each stable k that follows from $t^*|_{\text{MC}}$. Take the stable $\bar{k}$ with the highest expected payoff $\bar{\pi}|_{\text{MC}} = \frac{\bar{k}}{N}\pi^m(t^*|_{\text{MC}}, \bar{k}) + \frac{N-\bar{k}}{N}\pi^n(t^*|_{\text{MC}}, \bar{k})$. Following the proof above, a stable fund based on total costs exists that exhibits a global payoff that is weakly larger than for the fund based on marginal costs. By noting that the expected payoff $\frac{k}{N}\pi^m + \frac{N-k}{N}\pi^n$ is equal to $1/N$ times the global payoff, we know that the stable fund based on total costs also establishes an expected payoff $\bar{\pi}|_{\text{TC}} = \frac{\bar{k}+\Delta k}{N}\pi^m(t|_{\text{TC}}, \bar{k}+\Delta k) + \frac{N-\bar{k}-\Delta k}{N}\pi^n(t|_{\text{TC}}, \bar{k}+\Delta k)$ that is weakly larger than for the fund based on marginal costs $\bar{\pi}|_{\text{MC}}$. The expected payoff $\bar{\pi}|_{\text{TC}}$ for the stable total cost fund is then also weakly higher than the equilibrium payoff in Eq. (11) for the fund based on marginal costs. This is because the equilibrium payoff is a convex combination of expected payoffs $\frac{k}{N}\pi^m(t^*|_{\text{MC}}, k) + \frac{N-k}{N}\pi^n(t^*|_{\text{MC}}, k)$ that are each weakly smaller than $\bar{\pi}|_{\text{TC}}$ (since we took the $\bar{k}$ with the largest $\bar{\pi}|_{\text{MC}}$).

Now consider the equilibrium expected payoff in Section 4.1.2 for the fund based on total costs $\frac{k^*}{N}\pi^m(t^*|_{\text{TC}}, k^*) + \frac{N-k^*}{N}\pi^n(t^*|_{\text{TC}}, k^*)$. From the numerical algorithm, we know that this equilibrium expected payoff is weakly larger than any

[8] With the implicit function theorem on the member FOC $0 = b - C'(q) + t(k-1)C''(q)$, $\frac{dq}{dk} = \frac{tC''}{\underbrace{C''}_{=(e-1)C'} \underbrace{-t(k-1)}_{=(C'-b)/C''} \underbrace{C'''}_{=(e-2)C''/q}} = \frac{tC''}{C'/q+b(e-2)/q} \geq 0.$

[9] Proposition 1 showed that total abatement increases above q^{NC} through larger member abatement. As marginal benefits decrease, non-members decrease their abatement below q^{NC}.

expected payoff attainable for a stable fund among k members. It is hence weakly larger than the expected payoff $\bar{\pi}|_{\text{TC}}$. We have just shown that this expected payoff is weakly larger than the maximized payoff for the equilibrium with the fund based on marginal costs. It follows that the equilibrium with the fund based on total costs establishes an equilibrium expected payoff that is weakly larger than in equilibrium with the fund based on marginal costs.

The environmental effectiveness of the fund based on total costs is equal to Eq. (15). For the fund based on marginal costs, we adjust the equation for the environmental effectiveness to allow for multiple equilibria in participation. We replace the global payoff for the equilibrium with the fund based on total costs $[k^*\pi^m(t^*|_{\text{TC}}, k^*) + (N-k^*)\pi^n(t^*|_{\text{TC}}, k^*)]$ in Eq. (15) with the convex combination

$$\frac{1}{\sum_{k=1}^{N}\binom{N}{k}\phi(t^*|_{\text{MC}}, k)}\sum_{k=1}^{N}\binom{N}{k}(k\pi^m(t^*|_{\text{MC}}, k) + (N-k)\pi^n(t^*|_{\text{MC}}, k))\cdot\phi(t^*|_{\text{MC}}, k) \tag{A.4}$$

of global payoffs for each stable k under $t^*|_{\text{MC}}$. This is a convex combination of global payoffs that are each no greater than the global payoff at equilibrium for the fund based on total costs (as shown above). It follows that the environmental effectiveness is weakly smaller for the fund based on marginal costs than for the fund based on total costs. This establishes part (b) of the proposition for the payoff with convex marginal costs.

A.3.2. Payoff with decreasing marginal benefits in Eq. 13

We now show the statements of Proposition 3 for payoffs with strictly decreasing marginal benefits following the exact steps as above.

Drop in abatement is weakly larger for the fund based on total costs than for the fund based on marginal costs

For the fund based on total costs, the FOCs of members and non-members in the last stage can be combined:

$$\begin{aligned}
0 &= b - \gamma(kq^m(t,k) + (N-k)q^n(t,k)) - cq^m(t,k) + t(k-1)cq^m(t,k)\\
0 &= b - \gamma(kq^m(t,k) + (N-k)q^n(t,k)) - cq^n(t,k)\\
&\Rightarrow q^n(t,k) = (1-t(k-1))q^m(t,k)\\
\Rightarrow 0 &= b - (N\gamma + c)q^m(t,k) + t(k-1)[(N-k)\gamma + c]q^m(t,k)
\end{aligned}$$

With the last equation we proceed as after Eq. (A.1) to find an expression relating member abatement when participation is at $\bar{k}-1$ (denoted $q_2|_{\text{TC}}$ as above) with abatement of members when participation is at $\bar{k}$ (denoted $\bar{q}$ as above). It follows:

$$0 = b - (N\gamma + c)q_2|_{\text{TC}} + [(N\gamma + c)\bar{q} - b]\frac{\bar{k}-2}{\bar{k}-1}\frac{(N-\bar{k}+1)\gamma + c}{(N-\bar{k})\gamma + c}\frac{q_2|_{\text{TC}}}{\bar{q}}. \tag{A.5}$$

For the fund based on marginal costs, the FOCs of members and non-members in the last stage can again be combined:

$$\begin{aligned}
0 &= b - \gamma(kq^m(t,k) + (N-k)q^n(t,k)) - cq^m(t,k) + t(k-1)c\\
0 &= b - \gamma(kq^m(t,k) + (N-k)q^n(t,k)) - cq^n(t,k)\\
&\Rightarrow q^n(t,k) = q^m(t,k) - t(k-1)\\
\Rightarrow 0 &= b - (N\gamma + c)q^m(t,k) + t(k-1)[(N-k)\gamma + c]
\end{aligned}$$

With the last equation we proceed again as after Eq. (A.1) to find an expression relating member abatement when participation is at $\bar{k}-1$ (denoted $q_2|_{\text{MC}}$ as above) with abatement of members when participation is at $\bar{k}$ (denoted $\bar{q}$ as above). It follows:

$$0 = b - (N\gamma + c)q_2|_{\text{MC}} + [(N\gamma + c)\bar{q} - b]\frac{\bar{k}-2}{\bar{k}-1}\frac{(N-\bar{k}+1)\gamma + c}{(N-\bar{k})\gamma + c}. \tag{A.6}$$

Subtracting Eqs. (A.5) and (A.6) yields:

$$0 = \underbrace{(N\gamma + c)}_{\geq 0}(q_2|_{\text{MC}} - q_2|_{\text{TC}}) + \underbrace{[(N\gamma + c)\bar{q} - b]}_{(1)}\underbrace{\frac{\bar{k}-2}{\bar{k}-1}\frac{(N-\bar{k}+1)\gamma + c}{(N-\bar{k})\gamma + c}}_{\geq 0}\left[\frac{q_2|_{\text{TC}}}{\bar{q}} - 1\right].$$

Based on the last equation, the desired ranking $q_2|_{\text{MC}} \geq q_2|_{\text{TC}}$ only follows if member abatement decreases after non-participation, $q_2|_{\text{TC}} \leq \bar{q}$, for the fund based on total costs. We now proceed with the parameter range in which $q_2|_{\text{TC}} \leq \bar{q} \Leftrightarrow \frac{\bar{k}-2}{\bar{k}-1}\frac{\gamma(N-k+1)+c}{\gamma(N-k)+c} \leq 1$.[10] We return to the other parameter range $q_2|_{\text{TC}} > \bar{q}$ below.

With $q_2|_{\text{TC}} \leq \bar{q}$, we immediately have $q_2|_{\text{MC}} \geq q_2|_{\text{TC}}$ because (1) $[(N\gamma + c)\bar{q} - b] \geq 0$ with $\bar{q} \geq q^{\text{NC}}$.

[10] This follows from the combined FOCs derived above:
$0 = b - (N\gamma + c)q + t(k-1)[(N-k)\gamma + c]q$
$\Rightarrow 0 = \frac{b}{\bar{q}} - (\gamma N - c) + t(\bar{k}-1)(\gamma(N-\bar{k}) + c), 0 = \frac{b}{q_2|_{\text{TC}}} - (\gamma N - c) + t(\bar{k}-2)(\gamma(N-\bar{k}+1) + c)$
$\Rightarrow 0 = \frac{b}{\bar{q}} - \frac{b}{q_2|_{\text{TC}}} + t((\bar{k}-1)(\gamma(N-\bar{k}) + c) - (\bar{k}-2)(\gamma(N-\bar{k}+1) + c))$
$\Rightarrow 0 = \frac{b}{\bar{q}} - \frac{b}{q_2|_{\text{TC}}} + t(\bar{k}-1)(\gamma(N-\bar{k}) + c)\left[1 - \frac{\bar{k}-2}{\bar{k}-1}\frac{\gamma(N-\bar{k}+1)+c}{\gamma(N-\bar{k})+c}\right]$.

The stability function is weakly larger for the fund based on total costs than for the fund based on marginal costs

For this payoff function, when one member leaves, abatement of remaining members decreases more for the total cost fund. As a result, total abatement (and abatement of all other countries but the leaving member) also decreases more for the total cost fund. This happens in conjunction with the reaction of non-members to the change in member abatement. These relationships are established separately in the following as they are needed to show how the stability functions relate.

Recall that $\bar{t}|_{TC}$ is the level of t that establishes member abatement $\bar{q}$ for $\bar{k}$ members of the fund based on total costs. Recall also that non-member abatement is $q^n(t,k)|_{TC} = (1 - t(k-1))q^m(t,k)|_{TC}$ for the total costs fund and $q^n(t,k)|_{MC} = q^m(t,k)|_{MC} - t(k-1)$ for the marginal costs fund. (See FOCs above for both funds). Consider the abatement of all others but the leaving member when participation is at $\bar{k}-1$, first for the fund based on total costs, then for the fund based on marginal costs:

$$\begin{aligned} Q_{-i}(\bar{t}|_{TC}, \bar{k}-1) &= (\bar{k}-1)q_2|_{TC} + (N-\bar{k})(1 - \bar{t}|_{TC}(\bar{k}-2))q_2|_{TC} \\ Q_{-i}(\bar{t}|_{MC}, \bar{k}-1) &= (\bar{k}-1)q_2|_{MC} + (N-\bar{k})(q_2|_{MC} - \bar{t}|_{MC}(\bar{k}-2)) \end{aligned}$$

Some manipulation yields:

$$\begin{aligned} Q_{-i}(\bar{t}|_{TC}, \bar{k}-1) &= (N-1)q_2|_{TC} - (N-\bar{k})\underbrace{\bar{t}|_{TC}(\bar{k}-2)}_{=\frac{\gamma N + c - b/q_2|_{TC}}{\gamma(N-\bar{k}+1)+c}} q_2|_{TC} \\ Q_{-i}(\bar{t}|_{MC}, \bar{k}-1) &= (N-1)q_2|_{MC} - (N-\bar{k})\underbrace{\bar{t}|_{MC}(\bar{k}-2)}_{=\frac{(\gamma N + c)q_2|_{MC} - b}{\gamma(N-\bar{k}+1)+c}} \end{aligned} \tag{A.7}$$

The under-braced equations follow from the combined FOCs above when participation is at $\bar{k}-1$. Further manipulation delivers:

$$\begin{aligned} Q_{-i}(\bar{t}|_{TC}, \bar{k}-1) &= \frac{q_2|_{TC}(\gamma + c)(\bar{k}-1) + (N-\bar{k})b}{\gamma(N-\bar{k}+1)+c} \\ Q_{-i}(\bar{t}|_{MC}, \bar{k}-1) &= \frac{q_2|_{MC}(\gamma + c)(\bar{k}-1) + (N-\bar{k})b}{\gamma(N-\bar{k}+1)+c} \end{aligned} \tag{A.8}$$

Hence the abatement of all other countries is lower for the fund based on total costs than the fund based on marginal costs when participation decreases by one: $Q_{-i}(\bar{t}|_{TC}, \bar{k}-1) \leq Q_{-i}(\bar{t}|_{MC}, \bar{k}-1)$ since $q_2|_{TC} \leq q_2|_{MC}$. From this it is straightforward that non-member abatement is higher for the fund based on total costs than for the marginal costs fund: $q^n(\bar{t}|_{TC}, \bar{k}-1) \geq q^n(\bar{t}|_{MC}, \bar{k}-1)$.

Next, we show that total abatement is lower for the fund that balances total costs than for the fund that balances marginal costs when participation is at $\bar{k}-1$: $Q(\bar{t}|_{TC}, \bar{k}-1) \leq Q(\bar{t}|_{MC}, \bar{k}-1)$. Take the FOCs for non-members, first for the fund based on total costs, second for the fund based on marginal costs:

$$\begin{aligned} &b - \gamma Q(\bar{t}|_{TC}, \bar{k}-1) = cq^n(\bar{t}|_{TC}, \bar{k}-1) \\ &b - \gamma Q(\bar{t}|_{MC}, \bar{k}-1) = cq^n(\bar{t}|_{MC}, \bar{k}-1) \\ &\Rightarrow -\gamma(Q(\bar{t}|_{TC}, \bar{k}-1) - Q(\bar{t}|_{MC}, \bar{k}-1)) = c(q^n(\bar{t}|_{TC}, \bar{k}-1) - q^n(\bar{t}|_{MC}, \bar{k}-1)) \end{aligned}$$

Since $q^n(\bar{t}|_{TC}, \bar{k}-1) \geq q^n(\bar{t}|_{MC}, \bar{k}-1)$, the claim on global abatement follows.

These relationships provide the basis for showing that the stability function is higher at $\bar{k}$ for the fund that balances total costs than for the fund that balances marginal costs. By noting that at $\bar{k}$ both funds establish the same abatement by members $\bar{q}$ and therefore the same member payoff, we have:

$$\begin{aligned} &\Delta\pi(\bar{t}|_{TC}, \bar{k}) = \pi^m(\bar{t}|_{TC}, \bar{k}) - \pi^n(\bar{t}|_{TC}, \bar{k}-1) \\ &\Delta\pi(\bar{t}|_{MC}, \bar{k}) = \pi^m(\bar{t}|_{MC}, \bar{k}) - \pi^n(\bar{t}|_{MC}, \bar{k}-1) \\ &\Rightarrow \Delta\pi(\bar{t}|_{TC}, \bar{k}) - \Delta\pi(\bar{t}|_{MC}, \bar{k}) = \pi^n(\bar{t}|_{MC}, \bar{k}-1) - \pi^n(\bar{t}|_{TC}, \bar{k}-1) \geq 0 \end{aligned}$$

The last inequality follows as non-member payoff after non-participation is weakly larger for the fund based on marginal costs: global abatement is higher and abatement of non-members is lower for the marginal cost fund compared to the total cost fund.

Global payoff and environmental effectiveness are weakly larger for the fund that balances total costs than for the fund based on marginal costs

Since the stability function is also higher for this payoff function, the claims for the global payoff and environmental effectiveness are derived as in the respective subsections of Appendix A.3.1. (The arguments there do not rely on the payoff function). Hence, for the payoff we also established claims (a) and (b) with decreasing marginal benefits.

Parameter range, in which member abatement does not drop after non-participation

We now check the parameter range in which member abatement does not drop after non-participation for the fund based on total costs: $q_2|_{\text{TC}} > \bar{q} \Leftrightarrow \frac{\bar{k}-2}{\bar{k}-1}\frac{\gamma(N-k+1)+c}{\gamma(N-\bar{k})+c} > 1$. In this parameter range, member abatement also does not drop after non-participation for the fund based on marginal costs $q_2|_{\text{MC}} > \bar{q}$.[11] If member abatement increases after non-participation, the stability function cannot be positive (as shown below). Therefore this case is irrelevant for the proposition as the necessary condition of a stable fund based on marginal costs cannot hold.

If member abatement increases after non-participation, $q_2|_{\text{MC}} > \bar{q}$, it follows that the abatement of all other countries but the leaving member also increases. Denote $\bar{Q}_{-i}$ as the abatement of all countries but one member with at $\bar{k}$. As in Eq. (A.7), $Q_{-i}(\bar{t}|_{\text{MC}}, \bar{k}-1)$ is the abatement of all other countries but the leaving member with participation at $\bar{k}-1$. We have:

$$\begin{aligned}
&\bar{Q}_{-i} - Q_{-i}(\bar{t}|_{\text{MC}}, \bar{k}-1)\\
&= (\bar{k}-1)\bar{q} + (N-\bar{k})(\bar{q} - \bar{t}|_{\text{MC}}(\bar{k}-1)) - Q_{-i}(\bar{t}|_{\text{MC}}, \bar{k}-1)\\
&= (N-1)\bar{q} - (N-\bar{k})\bar{t}|_{\text{MC}}(\bar{k}-1) - [(N-1)q_2|_{\text{MC}} - (N-\bar{k})\bar{t}|_{\text{MC}}(\bar{k}-2)]\\
&= (N-1)\underbrace{(\bar{q} - q_2|_{\text{MC}})}_{<0} - (N-\bar{k})\bar{t}|_{\text{MC}}\underbrace{(\bar{k}-1-\bar{k}+2)}_{=1} < 0
\end{aligned}$$

The stability function is:

$$\Delta\pi(\bar{t}|_{\text{MC}}, \bar{k}) = B(\bar{Q}_{-i} + \bar{q}) - C(\bar{q}) - [B(Q_{-i}(\bar{t}|_{\text{MC}}, \bar{k}-1) + q^n(\bar{t}|_{\text{MC}}, \bar{k}-1)) - C(q^n(\bar{t}|_{\text{MC}}, \bar{k}-1))]. \tag{A.9}$$

Define $q^* = \text{argmax}_q\, B(\bar{Q}_{-i} + q) - C(q)$. By definition $B(\bar{Q}_{-i} + \bar{q}) - C(\bar{q}) \leq B(\bar{Q}_{-i} + q^*) - C(q^*)$. The FOC for q^* is $B'(\bar{Q}_{-i} + q^*) = C'(q^*)$. Comparing this FOC to the FOC for non-members at participation $\bar{k}-1$, $B'(Q_{-i}(\bar{t}|_{\text{MC}}, \bar{k}-1) + q^n(\bar{t}|_{\text{MC}}, \bar{k}-1)) = C'(q^n(\bar{t}|_{\text{MC}}, \bar{k}-1))$, it is straightforward that $q^* > q^n(t, \bar{k}-1)$ since $\bar{Q}_{-i} < Q_{-i}(\bar{t}|_{\text{MC}}, \bar{k}-1)$. With this we have:

$$\begin{aligned}
\Delta\pi(\bar{t}|_{\text{MC}}, \bar{k}) &\leq B(\bar{Q}_{-i} + q^*) - C(q^*) - [B(Q_{-i}(\bar{t}|_{\text{MC}}, \bar{k}-1) + q^n(\bar{t}|_{\text{MC}}, \bar{k}-1)) - C(q^n(\bar{t}|_{\text{MC}}, \bar{k}-1))]\\
&= \underbrace{B(\bar{Q}_{-i} + q^*) - B(Q_{-i}(\bar{t}|_{\text{MC}}, \bar{k}-1) + q^n(\bar{t}|_{\text{MC}}, \bar{k}-1))}_{=\Delta B} + \underbrace{C(q^n(\bar{t}|_{\text{MC}}, \bar{k}-1)) - C(q^*)}_{<0 \text{ since } q^* > q^n(\bar{t}|_{\text{MC}}, \bar{k}-1)}
\end{aligned}$$

$\Delta B < 0$ follows if $\bar{Q}_{-i} + q^* < Q_{-i}(\bar{t}|_{\text{MC}}, \bar{k}-1) + q^n(\bar{t}|_{\text{MC}}, \bar{k}-1)$. Take the two FOCs for q^* and $q^n(\bar{t}|_{\text{MC}}, \bar{k}-1)$:

$$\begin{aligned}
&b - \gamma(\bar{Q}_{-i} + q^*) = cq^*\\
&b - \gamma(Q_{-i}(\bar{t}|_{\text{MC}}, \bar{k}-1) + q^n(\bar{t}|_{\text{MC}}, \bar{k}-1)) = cq^n(\bar{t}|_{\text{MC}}, \bar{k}-1)\\
&\Rightarrow -\gamma(\bar{Q}_{-i} + q^* - [Q_{-i}(\bar{t}|_{\text{MC}}, \bar{k}-1) + q^n(\bar{t}|_{\text{MC}}, \bar{k}-1)]) = c(q^* - q^n(\bar{t}|_{\text{MC}}, \bar{k}-1))
\end{aligned}$$

Since $q^* > q^n(\bar{t}|_{\text{MC}}, \bar{k}-1)$, we have $\bar{Q}_{-i} + q^* < Q_{-i}(\bar{t}|_{\text{MC}}, \bar{k}-1) + q^n(\bar{t}|_{\text{MC}}, \bar{k}-1)$ and $\Delta B < 0$. Hence, the stability function is negative.

A.4. Proof of Proposition 3c

Solving the last stage with transfers from Table 1 and the linear-quadratic payoff, members abate at $q^m(t,k) = \frac{b}{c} + t \cdot k\left(1 - \frac{1}{k}\right)$. Non-members abate at $q^n(t,k) = b/c$.

The stability function from Eq. (7) for k members is:

$$\begin{aligned}
\Delta\pi(t,k) =& b\left(N\frac{b}{c} + kt(k-1)\right) - \frac{c}{2}\left(\frac{b}{c} + t(k-1)\right)^2 - b\left(N\frac{b}{c} + (k-1)t(k-2)\right) + \frac{1}{2}\frac{b^2}{c}\\
=& t(k-1)(b - \frac{1}{2}ct(k-1)).
\end{aligned}$$

The only k for which the stability function switches sign from positive to negative when increased by one is the nearest lowest integer to $\frac{2b}{ct} + 1$. This is the unique stable participation in stage 2. In turn, a fund with participation k is internally stable if $t \leq \frac{2b}{c(k-1)}$.

Given k, the expected payoff of the first stage increases in t if $t \leq \frac{2b}{c(k-1)}$:

$$\text{E}[\pi(t,k)] = \frac{k}{N}\left[b\left(N\frac{b}{c} + kt(k-1)\right) - \frac{c}{2}\left(\frac{b}{c} + t(k-1)\right)^2\right] + \frac{N-k}{N}\left[b\left(N\frac{b}{c} + kt(k-1)\right) - \frac{b^2}{2c}\right]$$

[11] This follows from the combined FOCs derived above:
$0 = b - (N\gamma + c)q + t(k-1)(\gamma(N-k) + c)$
$\Leftrightarrow 0 = b - (\gamma N + c)\bar{q} + t(k-1)(\gamma(N-\bar{k}) + c), 0 = b - (\gamma N + c)q_2|_{\text{MC}} + t(\bar{k}-2)(\gamma(N-\bar{k}+1) + c)$
$\Leftrightarrow 0 = (\gamma N + c)(q_2|_{\text{MC}} - \bar{q}) + t\left[(\bar{k}-1)(\gamma(N-\bar{k}) + c) - (\bar{k}-2)(\gamma(N-\bar{k}+1) + c)\right]$
$\Leftrightarrow 0 = (\gamma N + c)(q_2|_{\text{MC}} - \bar{q}) + \frac{t}{(\bar{k}-1)(\gamma(N-\bar{k})+c)}\left[1 - \frac{(\bar{k}-2)(\gamma(N-\bar{k}+1)+c)}{(\bar{k}-1)(\gamma(N-\bar{k})+c)}\right]$.

Setting the derivative $\frac{dE[\pi(t,k)]}{dt}$ to zero derives the t that would maximize the expected payoff without the stability constraint.

$$\frac{dE[\pi(t,k)]}{dt} = \frac{k}{N}(k-1)(N-1)b - \frac{k}{N}ct(k-1)^2 = 0$$
$$\Rightarrow t = \frac{b(N-1)}{c(k-1)} \geq \frac{2b}{c(k-1)} \quad \text{since } N > 2.$$

Hence, inserting the maximum $t = \frac{2b}{c(k-1)}$ achieves the largest expected payoff while the fund among k members is stable:

$$\mathrm{E}[\pi(t = \frac{2b}{c(k-1)}, k)] = \frac{k}{N}\left[b\left(N\frac{b}{c} + k2\frac{b}{c}\right) - \frac{c}{2}\left(\frac{b}{c} + 2\frac{b}{c}\right)^2\right] + \frac{N-k}{N}\left[b\left(N\frac{b}{c} + k2\frac{b}{c}\right) - \frac{b^2}{2c}\right]$$
$$= \frac{b^2}{c}\left(N + 2k - 4\frac{k}{N} - \frac{1}{2}\right)$$

This increases in k. Hence, the equilibrium of the game is $t^* = \frac{2b}{c(N-1)}$, leading to full participation $k^* = N$. All countries have a payoff of $3N - 4.5$. Inserting this member payoff under full participation in Eq. (15) leads to the stated environmental effectiveness in the proposition.

A.5. Proof of Proposition 4

The proof proceeds completely analogous to Appendix A.3. Consider a stable fund with $k = \bar{k}$ members for a $t = \bar{t}|_{\mathrm{FF}}$ which leads to abatement $\bar{q} \geq q^{\mathrm{NC}}$ for the fund of fixed size. The proof considers the fund based on marginal costs that also leads to abatement $\bar{q}$ when $\bar{k}$ members participate.[12] We proceed in three steps to show Proposition 4(a). The first step shows that the drop in abatement after non-participation is weakly larger for the fund based on marginal costs than for the fund of fixed size. The second step shows that, as a result, the stability function is weakly larger for the fund based on marginal costs. The third step shows that global payoff is also weakly larger, showing Proposition 4(a).

Proposition 4(b) follows because we know from Proposition 4(a) that the fund of fixed size performs worse than the fund based on marginal costs, which performs worse than the fund based on total costs from Proposition 3.

Lastly, we show Proposition 4(c) based on Proposition 4(a).

A.5.1. Showing Proposition 4a

Drop in abatement is weakly larger for the fund based on marginal costs than for the fund of fixed size

Consider the payoff in Eq. (14). With the FOC for members of the fixed fund

$$0 = b - C'(q^m(t,k)) + t\frac{k-1}{k}\frac{1}{q^m(t,k)}$$

we apply the same procedure as after Eq. (A.1) to find an expression relating member abatement when participation is at $\bar{k}-1$ (denoted $q_2|_{\mathrm{FF}}$) with abatement by members when participation is at $\bar{k}$ (denoted $\bar{q}$ as above):

$$0 = b - C'(q_2|_{\mathrm{FF}}) + (C'(\bar{q}) - b)\frac{\bar{k}-2}{\bar{k}-1}\frac{\bar{k}}{\bar{k}-1}\frac{\bar{q}}{q_2|_{\mathrm{FF}}}. \tag{A.10}$$

Subtracting Eq. (A.3) and Eq. (A.10) yields:

$$0 = C'(q_2|_{\mathrm{FF}}) - C'(q_2|_{\mathrm{MC}}) - \underbrace{(C'(\bar{q}) - b)\frac{\bar{k}-2}{\bar{k}-1}}_{(1)}\left[\underbrace{\frac{C''(q_2|_{\mathrm{MC}})}{C''(\bar{q})}}_{(2)} - \underbrace{\frac{\bar{k}}{\bar{k}-1}\frac{\bar{q}}{q_2|_{\mathrm{FF}}}}_{(3)}\right].$$

The desired relationship $q_2|_{\mathrm{FF}} \geq q_2|_{\mathrm{MC}}$ follows. This is due to the fact that (1) $(C'(\bar{q}) - b)\frac{\bar{k}-2}{\bar{k}-1} \geq 0$ because $\bar{q} \geq q^{\mathrm{NC}}$; (2) $\frac{C''(q_2|_{\mathrm{MC}})}{C''(\bar{q})} \leq 1$ because $\bar{q} \geq q_2|_{\mathrm{MC}}$ (see Appendix A.3) ; (3) $\frac{\bar{k}}{\bar{k}-1}\frac{\bar{q}}{q_2|_{\mathrm{FF}}} > 1$ because $\bar{q} \geq q_2|_{\mathrm{FF}}$, which follows from the FOC for member abatement.[13]

---

[12] Note that any member abatement $q^m(t,k) \geq q^{\mathrm{NC}}$ is attainable by increasing t for the fund based on marginal costs. The FOCs for members for the payoff in Eq. (14) is $0 = b - C'(q) + t(k-1)C''(q)$. Inserting $C'(q) = cq^{e-1}, C''(q) = c(e-1)cq^{e-2}$, we have $q - t(k-1)(e-1) = bq^{2-e}$. The intersection of the linear function $q - t(k-1)(e-1)$ and bq^{2-e} approaches infinity as t approaches infinity.

For the payoff in Eq. (13), member abatement is $q^m(t,k) = \frac{\frac{b}{c} + t(k-1)((N-k)\frac{\gamma}{c}+1)}{\frac{\gamma}{c}N+1}$, which approaches infinity as t approaches infinity.

It is straightforward to show that the second order conditions for a maximum are fulfilled.

[13] With the implicit function theorem on the member FOC $0 = b - C'(q) + t\frac{k-1}{k}\frac{1}{q}$, $\frac{dq}{dk} = \frac{t/k^2/q}{C''+t\frac{k-1}{k}/q^2} \geq 0$.

For the payoff function in Eq. (13), we can again establish everything from the combined FOCs for the fixed fund:

$$\begin{aligned}
0 &= b - \gamma\left(kq^m(t,k) + (N-k)q^n(t,k)\right) - cq^m(t,k) + t\frac{k-1}{k}\frac{1}{q^m(t,k)} \\
0 &= b - \gamma\left(kq^m(t,k) + (N-k)q^n(t,k)\right) - cq^n(t,k) \\
&\Rightarrow q^n(t,k) = q^m(t,k) - \frac{t}{c}\frac{k-1}{k}\frac{1}{q^m(t,k)} \\
\Rightarrow 0 &= b - (N\gamma + c)q^m(t,k) + t\frac{k-1}{k}\frac{1}{q^m(t,k)}\left(1 + \frac{\gamma}{c}(N-k)\right).
\end{aligned}$$

With the last equation we proceed as after Eq. (A.1) to find an expression relating member abatement when participation is at $\bar{k}-1$ (denoted $q_2|_{\mathrm{FF}}$ as above) with abatement of members when participation is at $\bar{k}$ (denoted $\bar{q}$ as above). It follows:

$$0 = b - (N\gamma + c)q_2|_{\mathrm{FF}} + [(N\gamma + c)\bar{q} - b]\frac{\bar{k}-2}{\bar{k}-1}\frac{\bar{k}}{\bar{k}-1}\frac{(N-\bar{k}+1)\gamma + c}{(N-\bar{k})\gamma + c}\frac{\bar{q}}{q_2|_{\mathrm{FF}}} \tag{A.11}$$

Subtracting Eqs. (A.6) and (A.11) yields:

$$0 = \underbrace{(N\gamma + c)}_{\geq 0}(q_2|_{\mathrm{FF}} - q_2|_{\mathrm{MC}}) + \underbrace{[(N\gamma + c)\bar{q} - b]}_{(1)}\underbrace{\frac{\bar{k}-2}{\bar{k}-1}\frac{(N-\bar{k}+1)\gamma + c}{(N-\bar{k})\gamma + c}}_{\geq 0}\underbrace{\left(1 - \frac{\bar{k}}{\bar{k}-1}\frac{\bar{q}}{q_2|_{\mathrm{FF}}}\right)}_{(2)}$$

The desired ranking of member abatement follows if member abatement decreases after non-participation, $q_2|_{\mathrm{FF}} \leq \bar{q}$, for the fund of fixed size. This holds in the parameter range $\frac{\bar{k}-1}{\bar{k}} - \frac{\bar{k}-2}{\bar{k}-1}\frac{\gamma(N-\bar{k}+1)+c}{\gamma(N-\bar{k})+c} \geq 0$.[14] We consider the parameter range where $\bar{q} < q_2|_{\mathrm{FF}}$ below.

The desired relationship $q_2|_{\mathrm{FF}} \geq q_2|_{\mathrm{MC}}$ then follows because (1) $[(N\gamma + c)\bar{q} - b] \geq 0$ because $\bar{q} \geq q^{\mathrm{NC}}$, and (2) $\left(1 - \frac{\bar{k}}{\bar{k}-1}\frac{\bar{q}}{q_2|_{\mathrm{FF}}}\right) < 0$ with $q_2|_{\mathrm{FF}} \leq \bar{q}$.

The stability function is weakly larger for the fund based on marginal costs than for the fund of fixed size

With $q_2|_{\mathrm{FF}} \geq q_2|_{\mathrm{MC}}$, a higher stability function for the fund based on marginal costs than for the fund of fixed size follows exactly as in Sec. *The stability function is weakly larger for the fund based on total costs than for the fund based on marginal costs* in Appendix A.3 for the payoff with convex marginal costs.

For the payoff with decreasing marginal benefits, it is important to show that, after non-participation, the abatement of all other countries but the leaving member decreases more for the fund based on marginal costs than for the fund of fixed size. Recalling that non-member abatement is $q^n(t,k) = q^m(t,k) - \frac{t}{c}\frac{k-1}{k}\frac{1}{q^m(t,k)}$ (see FOCs above), we have

$$\begin{aligned}
Q_{-i}(\bar{t}|_{\mathrm{MC}}, \bar{k}-1) &= \frac{q_2|_{\mathrm{MC}}(\gamma + c)(\bar{k}-1) + (N-\bar{k})b}{\gamma(N-\bar{k}+1) + c} \quad \text{from Eq. A.8} \\
Q_{-i}(\bar{t}|_{\mathrm{FF}}, \bar{k}-1) &= (\bar{k}-1)q_2|_{\mathrm{FF}} + (N-\bar{k})(q_2|_{\mathrm{FF}} - \underbrace{\frac{\bar{t}|_{\mathrm{FF}}}{c}\frac{\bar{k}-2}{\bar{k}-1}\frac{1}{q_2|_{\mathrm{FF}}}}_{\frac{(\gamma N + c)q_2|_{\mathrm{FF}} - b}{\gamma(N-\bar{k}+1)+c}}) \\
&= \frac{q_2|_{\mathrm{FF}}(\gamma + c)(\bar{k}-1) + (N-\bar{k})b}{\gamma(N-\bar{k}+1) + c}
\end{aligned}$$

The under-braced equation follows from the combined FOCs for the fund of fixed size. With $q_2|_{\mathrm{FF}} \geq q_2|_{\mathrm{MC}}$, $Q_{-i}(\bar{t}|_{\mathrm{FF}} \geq Q_{-i}(\bar{t}|_{\mathrm{MC}}, \bar{k}-1)$ follows. Showing that the stability function is weakly larger for the marginal-cost than for the fixed-size fund ($\Delta\pi(\bar{t}|_{\mathrm{MC}}, \bar{k}) \geq \Delta\pi(\bar{t}|_{\mathrm{FF}}, \bar{k})$) follows exactly as after Eq. (A.8).

Global payoff is weakly larger for the stable fund based on marginal costs than for the fund of fixed size

Showing that the global payoff is weakly larger for the stable fund based on marginal costs than for the fixed fund follows exactly as in Subsection *Global payoff is weakly larger for the stable fund that balances total costs than for the fund*

[14] This follows from the combined FOCs derived above:
$0 = b - (N\gamma + c)q + t\frac{k-1}{k}\frac{1}{\bar{q}}(1 + \gamma/c(N-k))$
$\Rightarrow 0 = b\bar{q} - (\gamma N + c)(\bar{q})^2 + \bar{t}\frac{\bar{k}-1}{\bar{k}}(\frac{\gamma}{c}(N-\bar{k}) + 1), bq_2|_{\mathrm{FF}} - (\gamma N + c)(q_2|_{\mathrm{FF}})^2 + \bar{t}\frac{\bar{k}-2}{\bar{k}-1}(\frac{\gamma}{c}(N-\bar{k}+1) + 1)$
$\Rightarrow 0 = b(\bar{q} - q_2|_{\mathrm{FF}}) - (\gamma N + c)((\bar{q})^2 - (q_2|_{\mathrm{FF}})^2) + \bar{t}\left(\frac{\bar{k}-1}{\bar{k}}(\gamma(N-\bar{k}) + c) - \frac{\bar{k}-2}{\bar{k}-1}\gamma(N-\bar{k}+1) + c\right)$
$\Rightarrow 0 = \underbrace{(\bar{q} - q_2|_{\mathrm{FF}})}_{\geq 0}\underbrace{(b - (\gamma N + c)(\bar{q} + q_2|_{\mathrm{FF}}))}_{\leq 0 \text{ because } \bar{q}, q_2|_{\mathrm{FF}} \geq q^{\mathrm{NC}}} + \frac{\bar{t}}{\gamma(N-\bar{k})+c}\underbrace{\left(\frac{\bar{k}-1}{\bar{k}} - \frac{\bar{k}-2}{\bar{k}-1}\frac{\gamma(N-\bar{k}+1)+c}{\gamma(N-\bar{k})+c}\right)}_{\geq 0}.$

that balances marginal costs in Appendix A.3. Hence, Proposition 4(a) is shown for the payoff functions in Eq. (14) and in Eq. (13) for the parameter range $q_2 \leq \bar{q}$ for the fund of fixed size.

A.5.2. Showing Proposition 4b

We showed that $q_2|_{FF} \geq q_2|_{MC}$. For the payoff in Eq. (13) this only holds if $\frac{\bar{k}-2}{\bar{k}-1}\frac{\gamma(N-\bar{k}+1)+c}{\gamma(N-\bar{k})+c} \leq \frac{\bar{k}-1}{\bar{k}}$.

In Appendix A.3, we showed that $q_2|_{MC} \geq q_2|_{TC}$. For the payoff in Eq. (13) this only holds in the parameter range $\frac{\bar{k}-2}{\bar{k}-1}\frac{\gamma(N-\bar{k}+1)+c}{\gamma(N-\bar{k})+c} \leq 1$. With $\frac{\bar{k}-2}{\bar{k}-1}\frac{\gamma(N-\bar{k}+1)+c}{\gamma(N-\bar{k})+c} \leq \frac{\bar{k}-1}{\bar{k}} < 1$, we also have $q_2|_{MC} \geq q_2|_{TC}$ when $q_2|_{FF} \geq q_2|_{MC}$. Hence, we also have $q_2|_{FF} \geq q_2|_{TC}$. With this relationship, the proof of Proposition 4(b) follows exactly as the proof of Proposition 3(b) in Appendix A.3.

A.5.3. Parameter range, in which member abatement does not drop after non-participation

Analogous to Appendix A.3, the stability function cannot be positive in the parameter range where $\bar{q} < q_2|_{FF} \Leftrightarrow \frac{\bar{k}-2}{\bar{k}-1}\frac{\gamma(N-k+1)+c}{\gamma(N-k)+c} > \frac{\bar{k}-1}{\bar{k}}$ for the fund of fixed size. We can show that the abatement of all other countries but the leaving member increases after non-participation. Denote $\bar{Q}_{-i}$ as the abatement of all countries but one member with participation at $\bar{k}$ in the fund of fixed size. $Q_{-i}(\bar{t}|_{FF}, \bar{k}-1)$ is the abatement of all countries but the leaving member with participation at $\bar{k}-1$. We have:

$$
\begin{aligned}
&\bar{Q}_{-i} - Q_{-i}(\bar{t}|_{FF}, \bar{k}-1) \\
&= (\bar{k}-1)\bar{q} + (N-\bar{k})\left(\bar{q} - \frac{\bar{t}|_{FF}}{c}\frac{\bar{k}-1}{\bar{k}} \cdot 1/\bar{q}\right) \\
&\quad - (\bar{k}-1)q_2|_{FF} - (N-\bar{k})\left(q_2|_{FF} - \frac{\bar{t}|_{FF}}{c}\frac{\bar{k}-2}{\bar{k}-1} \cdot 1/q_2|_{FF}\right) \\
&= (\bar{N}-1)\underbrace{(\bar{q} - q_2|_{FF})}_{<0} - \frac{\bar{t}|_{FF}}{c}(N-\bar{k})\underbrace{\left(\frac{\bar{k}-1}{\bar{k}} \cdot 1/\bar{q} - \frac{\bar{k}-2}{\bar{k}-1} \cdot 1/q_2|_{FF}\right)}_{=\frac{(\bar{k}^2-2\bar{k})(q_2|_{FF}-\bar{q})+q_2|_{FF}}{\bar{k}(\bar{k}-1)\bar{q}q_2|_{FF}}>0} < 0
\end{aligned}
$$

Based on this, the stability function can be manipulated as in Eq. (A.9). After this manipulation, it is shown to be negative. Hence, the proposition is shown for all parameter ranges.

A.5.4. Showing Proposition 4c

Lastly we show Proposition 4(c) for the linear-quadratic payoff. Consider the equilibrium for the fixed-size fund $t^*|_{FF}$. This establishes a maximized expected payoff in Eq. (11) that is a convex combination of expected payoffs $\frac{k}{N}\pi^m(t^*|_{FF}, k) + \frac{N-k}{N}\pi^n(t^*|_{FF}, k)$ for each stable k that follows from $t^*|_{FF}$. Take the stable $\bar{k}$ with the highest expected payoff $\frac{\bar{k}}{N}\pi^m(t^*|_{FF}, \bar{k}) + \frac{N-\bar{k}}{N}\pi^n(t^*|_{FF}, \bar{k})$. We then know that there is a stable fund based on marginal costs with participation $\bar{k}+\Delta k$, $\Delta k \geq 0$. The stable fund based on marginal costs also leads to a global payoff $(\bar{k}+\Delta k)\pi^m(\bar{t}|_{MC}, \bar{k}+\Delta k) + (N-\bar{k}-\Delta k)\pi^n(\bar{t}|_{MC}, \bar{k}+\Delta k)$ that is weakly larger than the global payoff for the fund of fixed size under participation $\bar{k}$. The stable fund with participation $\bar{k}+\Delta k$ is unique for the fund based on marginal costs (shown in Appendix A.4).

The maximum expected payoff in Eq. (11) for the equilibrium with the marginal-cost fund is by definition weakly larger than $\frac{\bar{k}+\Delta k}{N}\pi^m(\bar{t}|_{MC}, \bar{k}+\Delta k) + \frac{N-\bar{k}-\Delta k}{N}\pi^n(\bar{t}|_{MC}, \bar{k}+\Delta k)$. Since the expected payoff is $1/N$ times the global payoff, the equilibrium with the fund based on marginal costs exhibits a global payoff that is weakly larger than $(\bar{k}+\Delta k)\pi^m(\bar{t}|_{MC}, \bar{k}+\Delta k) + (N-\bar{k}-\Delta k)\pi^n(\bar{t}|_{MC}, \bar{k}+\Delta k)$.

The environmental effectiveness of the fund based on marginal costs is equal to Eq. (A.4). For the fixed fund, the equation for the environmental effectiveness is adjusted to allow for multiple equilibria in participation. We replace the global payoff for the equilibrium with the fund based on marginal costs $[k^*\pi^m(t^*|_{MC}, k^*) + (N-k^*)\pi^n(t^*|_{MC}, k^*)]$ in Eq. (15) with the convex combination

$$
\frac{1}{\sum_{k=1}^{N}\binom{N}{k}\phi(t^*|_{FF}, k)}\sum_{k=1}^{N}\binom{N}{k}(k\pi^m(t^*|_{FF}, k) + (N-k)\pi^n(t^*|_{FF}, k)) \cdot \phi(t^*|_{FF}, k)
$$

of global payoffs for each stable k under $t^*|_{FF}$. Since this is a convex combination of global payoffs that are each weakly smaller than the global payoff at equilibrium for the fund based on marginal costs, the environmental effectiveness is weakly smaller for the fund of fixed size than for the fund based on marginal costs.

Appendix B. A mechanism for non-negative transfers in the last stage

This appendix analyzes a solution to the following commitment problem: if a country joins the fund but abates at the non-cooperative level rather than at $q^m(t, k)$, it would have to make payments to the fund. In other words, its transfer is negative, but it cannot be forced to make that payment due to its national sovereignty.

Following Gerber and Wichardt (2009), an additional "deposit" stage is added between the participation and abatement stages of our game. Here, all members are required to make a payment $d(t, k)$, the deposit, to the fund. The deposit is then refunded to members in the last stage of the game net of the transfer they have to make based on individual abatement choices. To solve the commitment problem, we hence require:

$$\mathcal{T}(t,k,q_i,q_{-i}) = d(t,k) + t\cdot k\cdot (C(q_i) - \frac{1}{k}\sum_{j\in S} C(q_j)) \geq 0$$

$$\Leftrightarrow \quad d(t,k) \geq t\cdot \left[\sum_{j\in S, j\neq i} C(q_j) - (k-1)C(q_i)\right].$$

The last equation shows that the lower bound on $d(t, k)$ depends on the abatement of member i and the abatement of all other members. It is decreasing in the former abatement. Hence, a lower bound on d is established by assuming the non-cooperative abatement level for member i. For the linear-quadratic payoff, we insert $q_i = q^{NC} = \frac{b}{c}$. The lower bound on d increases in the abatement of all other members. Here, we assume that the lower bound on d is fixed by the individually optimal member abatement $q^m(t,k) = \frac{b}{c}\frac{1}{1-t(k-1)}$. Higher levels of abatement would be feasible, but would be irrational for other members.

With these assumptions, the lower bound on the deposit becomes:

$$d(t,k) = \frac{1}{2}\frac{b^2}{c}\left[\frac{1}{(1-t(k-1))^2} - 1\right]t(k-1).$$

Refunding this deposit in the last stage leads to a transfer of $d(t, k)$ if all members choose $q^m(t, k)$. However, if a member chooses to abate at the non-cooperative level and all others remain at $q^m(t, k)$, that country's transfer is zero.

Adding the extra stage does not change the equilibrium of the game. Individually optimal abatement choices in the last stage are not affected by the deposit, as it is a constant. In addition, payoffs in the participation stage are not affected in the subgame perfect equilibrium: for each t and k given, countries pay $d(t, k)$ in the deposit stage and get it refunded in the abatement stage.

Appendix C. Algorithm used in Section 4.1.2

The algorithm in pseudocode is as follows:

1. Define all parameters b, c, e, γ, N.
2. For each $k \in 1, \ldots, N$, maximize the expected payoff of a country (see Eq. (11)) over $\frac{1}{k-1} > t \geq 0$ subject to internal stability of a fund among k members: $t^*(k) = \text{argmax}_t \frac{k}{N}\pi^m(t,k) + \frac{N-k}{N}\pi^n(t,k)$ s.t. $\Delta\pi(t,k) \geq 0$.
3. Check whether for $t^*(k)$, k is the unique equilibrium in the second stage of the game with internal and external stability fulfilled; if not abort.
4. Find the maximum expected payoff over all k: $k^* = \text{argmax}_k \frac{k}{N}\pi^m(t^*(k),k) + \frac{N-k}{N}\pi^n(t^*(k),k)$.

The equilibrium is $t^*(k^*)$. Step 4 finds the maximum expected payoff even if there are multiple stable fund participation numbers k for a different t. This is the case because we know by step 2 that each of these stable funds exhibits an expected payoff $\frac{k}{N}\pi^m(t,k) + \frac{N-k}{N}\pi^n(t,k)$ that is weakly smaller than the maximized expected payoff $\frac{k}{N}\pi^m(t^*(k^*),k^*) + \frac{N-k^*}{N}\pi^n(t^*(k^*),k^*)$. Since the expected payoff in Eq. (11) is a convex combination of these smaller expected payoffs, the expected payoff cannot be larger than the equilibrium found by the algorithm.

Matlab was used to execute the pseudocode.

Appendix D. The equilibrium t^* and k^* for Section 4.1.2

Fig. D.5 shows the equilibrium t^* and k^* that were used to calculate the environmental effectiveness in Fig. 4.

In Figs. 4 and D.5, panel (a), the variation of the parameters $\frac{\gamma}{c}$ and N are sufficient to calculate the environmental effectiveness for all variations in the parameters b, c, γ and N. This is formally derived now. For the payoff in Eq. (13), the fund leads to an abatement of members and non-members:

$$q^m(t,k) = \frac{b}{c}\left[\frac{\gamma}{c}N + 1 - t(k-1)\left(\frac{\gamma}{c}(N-k)+1\right)\right]^{-1} = \frac{b}{c}\hat{q}\left(\frac{\gamma}{c},N,t,k\right)$$

$$q^n(t,k) = \frac{b}{c}(1-t(k-1))\hat{q}\left(\frac{\gamma}{c},N,t,k\right) = \frac{b}{c}\tilde{q}\left(\frac{\gamma}{c},N,t,k\right).$$

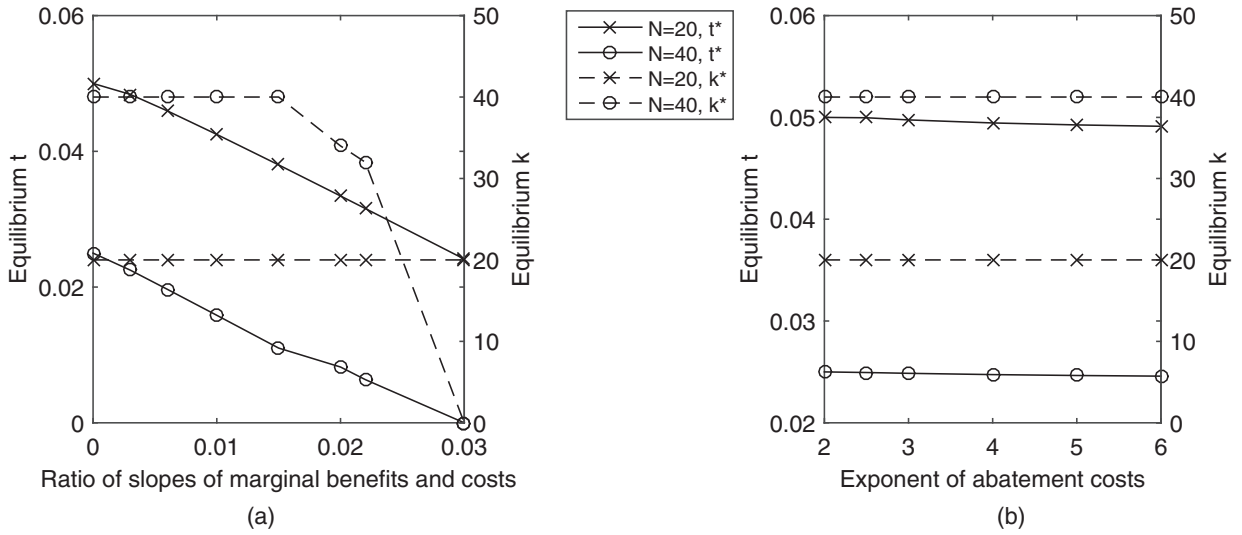

Fig. D.5. Equilibrium t and participation k as a function of (a) ratio γ/c in payoff Eq. (13) and (b) the exponent of the cost-function e in payoff Eq. (14). The range of parameter values is the same as in Fig. 4.

Member payoff is $\pi^m(t,k) = \frac{b^2}{c}\left\{\left[1-\frac{1}{2}\frac{\gamma}{c}(k\hat{q}+(N-k)\tilde{q})\right](k\hat{q}+(N-k)\tilde{q}) - \frac{1}{2}\hat{q}^2\right\}$, which scales with $\frac{b^2}{c}$ and is otherwise only dependent on $N, k, t, \frac{\gamma}{c}$. The same holds for the non-member payoff. Hence, the solution t^* of the optimization in Eq. (11) and resulting k^* only depends on $\frac{\gamma}{c}$ and N. As socially optimal and non-cooperative payoffs follow from $\pi^{SO} = \pi^m(t = 1/N, N)$, $\pi^{NC} = \pi^m(0, N)$ and the environmental effectiveness is the ratio of payoffs, the environmental effectiveness only depends on $\frac{\gamma}{c}$ and N.

In Fig. 4 and D.5, panel (b), the variation of the parameters e and N are sufficient to calculate the environmental effectiveness for all variations in the parameters b, c, e and N. For the payoff in Eq. (14), the fund leads to an abatement of members and non-members:

$$q^m(t,k) = (b/c)^{\frac{1}{e-1}} \frac{1}{(1-t(k-1))^{\frac{1}{e-1}}} = (b/c)^{\frac{1}{e-1}} \dot{q}(e,N,t,k)$$
$$q^n(t,k) = (b/c)^{\frac{1}{e-1}}.$$

The member payoff is $\pi^m(t,k) = b^{\frac{e}{e-1}}/c^{\frac{1}{e-1}}[k\dot{q} + (N-k) - \frac{1}{e}\dot{q}^e]$, which scales with $b^{\frac{e}{e-1}}/c^{\frac{1}{e-1}}$ and is only dependent on N, k, t, e. The same holds true for non-member payoff. As socially optimal and non-cooperative payoffs follow from $\pi^{SO} = \pi^m(t = 1/N, N)$, $\pi^{NC} = \pi^m(0, N)$, the environmental effectiveness and cooperation gap only depend on e and N.

Supplementary material

Supplementary material associated with this article can be found, in the online version, at doi:10.1016/j.euroecorev.2020.103423.

References

Barrett, S., 1994. Self-Enforcing international environmental agreements. Oxf. Econ. Pap. 46, 878–894.
Barrett, S., 2001. International cooperation for sale. Eur. Econ. Rev. 45, 1835–1850.
Bayer, P., Urpelainen, J., 2013. Funding global public goods: the dark side of multilateralism. Rev. Pol. Res. 30 (2), 160–189.
Bayramoglu, B., Finus, M., Jacques, J.-F., 2018. Climate agreements in a mitigation-adaptation game. J. Publ. Econ. 165, 101–113.
Biermann, F., Simonis, U.E., 1999. The multilateral ozone fund. Int. J. Soc. Econ. 26 (1/2/3), 239–273.
Bordoff, J.E., 2009. International Trade Law and the Economics of Climate Policy: Evaluating the Legality and Effectiveness of Proposals to Address Competitiveness and Leakage Concerns. Brookings Institution Press, pp. 35–68.
Buchholz, W., Konrad, K.A., 1995. Strategic transfers and private provision of public goods. J. Publ. Econ. 57, 489–505.
Carraro, C., Eyckmans, J., Finus, M., 2006. Optimal transfers and participation decisions in international environmental agreements. Rev. Int. Organ. 1 (4), 379–396.
Carraro, C., Siniscalco, D., 1993. Strategies for the international protection of the environment. J. Publ. Econ. 52 (3), 309–328.
Cramton, P., Stoft, S., 2012. Global climate games: how pricing and a green fund foster cooperation.. Econ. Energy Environ. Pol. 1 (2), 125–136. doi:10.5547/2160-5890.1.2.9.
d'Aspremont, C., Gabszewicz, J.J., 1986. New Developments in the Analysis of Market Structures. Macmillan, New York, pp. 243–264. Ch. On the stability of collusion
Falkinger, J., 1996. Efficient private provision of public goods by rewarding deviations from average. J. Publ. Econ. 62, 413–422.
Finus, M., 2008. Game theoretic research on the design of international environmental agreemets: insights, critical remarks, and future challenges. Int. Rev. Environ. Resour. Econ. 2, 29–67.

Finus, M., van Ierland, E.C., Dellink, R., 2006. Stability of climate coalitions in a cartel formation game. Econ. Govern. 7, 271–291. doi:10.1007/s10101-005-0009-1.
Gerber, A., Wichardt, P.C., 2009. Providing public goods in the absence of strong institutions. J. Publ. Econ. 93, 429–439.
Gersbach, H., Hummel, N., 2016. A development-compatible refunding scheme for a climate treaty. Resour. Energy Econ. 44, 139–168.
Gersbach, H., Winkler, R., 2012. Global refunding and climate change. J. Econ. Dyn. Control 36, 1775–1795.
Hoel, M., 1992. International environment conventions: the case of uniform reductions of emissions. Environ. Resour. Econ. 2 (2), 141–159.
Karp, L., Simon, L., 2013. Participation games and international environmental agreements: a non-parametric model. J. Environ. Econ. Manag. 65, 326–344.
Kornek, U., Steckel, J., Lessmann, K., Edenhofer, O., 2017. The climate rent curse: new challenges for burden sharing. Int. Environ. Agreem. Polit. Law Econ. 17 (6), 855–882. doi:10.1007/s10784-017-9352-2.
Kumar, S., 2015. Green climate fund faces slew of criticism. Nature 527, 419–420. doi:10.1038/nature.2015.18815.
Lessmann, K., Kornek, U., Bosetti, V., Dellink, R., Emmerling, J., Eyckmans, J., Nagashima, M., Weikard, H.-P., Yang, Z., 2015. The stability and effectivenss of climate coalitions. Environ. Resour. Econ. 62, 811–836. doi:10.1007/s10640-015-9886-0.
Lessmann, K., Marschinski, R., Edenhofer, O., 2009. The effects of tariffs on coalition formation in a dynamic global warming game. Econ. Model 26 (3), 641–649.
McKay, D., Cramton, P., Ockenfels, A., Stoft, S., 2015. Price carbon - I will if you will. Nature 526, 315–316.
Nordhaus, W., 2015. Climate clubs: overcoming free-riding in international climate policy. Am. Econ. Rev. 105 (4), 1339–1370. doi:10.1257/aer.15000001.
Paulsson, E., 2009. A review of the CDM literature: from fine-tuning to critical scrutiny? Int. Environ. Agreem. Polit. Law Econ. 9, 63–80.
Pizer, W.A., 2002. Combining price and quantity controls to mitigate global climate change. J. Publ. Econ. 85, 409–434.
Ruebbelke, D., 2006. Climate policy in developing countries and conditional transfers. Energy Policy 34, 1600–1610.
Weikard, H.-P., 2009. Cartel Stability Under an Optimal Sharing Rule, 77. The Manchester School, pp. 1463–6786. (5)